SUGENOMICS

Sequencing Data Quality Control Report

9 Bacteria Whole Genome Sequencing Samples

Report Date: 2021-09-06

1. Data Production

After sequencing, the raw reads were filtered. Data filtering includes removing adapter sequences, contamination and low-quality reads from raw reads. Table 1-1 shows statistical results of data production.

Table 1-1 Reads statistics results

Sample	Library	Raw Reads	Clean Reads	Raw Base(G)	Clean Base(G)	Effective Rate(	Error Rate(%)	Q20 (%)	Q30 (%)	GC Content (%)
13	FDSW210261158-1r	4,921,502	4,896,657	1.48	1.47	99.50	0.03	97.94	94.02	60.11
15	FDSW210261160-1r	5,228,125	5,216,705	1.57	1.57	99.78	0.03	97.89	93.98	60.30
16	FDSW210261161-1r	6,131,842	6,092,034	1.84	1.83	99.35	0.03	97.91	94.19	61.25
17	FDSW210261162-2r	5,167,511	5,161,460	1.55	1.55	99.88	0.03	97.15	92.11	57.75
18	FDSW210261163-1r	5,322,857	5,311,215	1.60	1.59	99.78	0.03	97.55	93.16	58.64

19	FDSW210261164-1r	6,282,151	6,262,110	1.88	1.88	99.68	0.03	97.69	93.74	61.38
20	FDSW210261165-1r	4,684,275	4,676,729	1.41	1.40	99.84	0.03	97.87	93.88	54.91
21	FDSW210261166-1r	4,104,465	4,084,541	1.23	1.23	99.51	0.03	97.81	93.95	60.50
22	FDSW210261167-1r	5,030,541	5,016,411	1.51	1.50	99.72	0.03	97.78	93.82	60.07

2. Data Format

The original image data is transferred into sequence data via base calling, which is defined as data or reads and saved as FASTQ file. Those FASTQ files are the original data provided for users, and they include the detailed read sequences and the read quality information. In each FASTQ file, every read is described by four lines, listed as follows:

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458     1:N:0:CGATGT
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHHFAE0ECFFD/AEHH
```

- Line 1 begins with a '@' character and is followed by a sequence identifier.
- Line 2 is the sequence letters.
- Line 3 is quality score identifier line, consisting only of a ‘+’ character.
- Line 4 encodes the quality values for the sequence in Line 2, and must contain the same number of symbols as letters in the sequence.

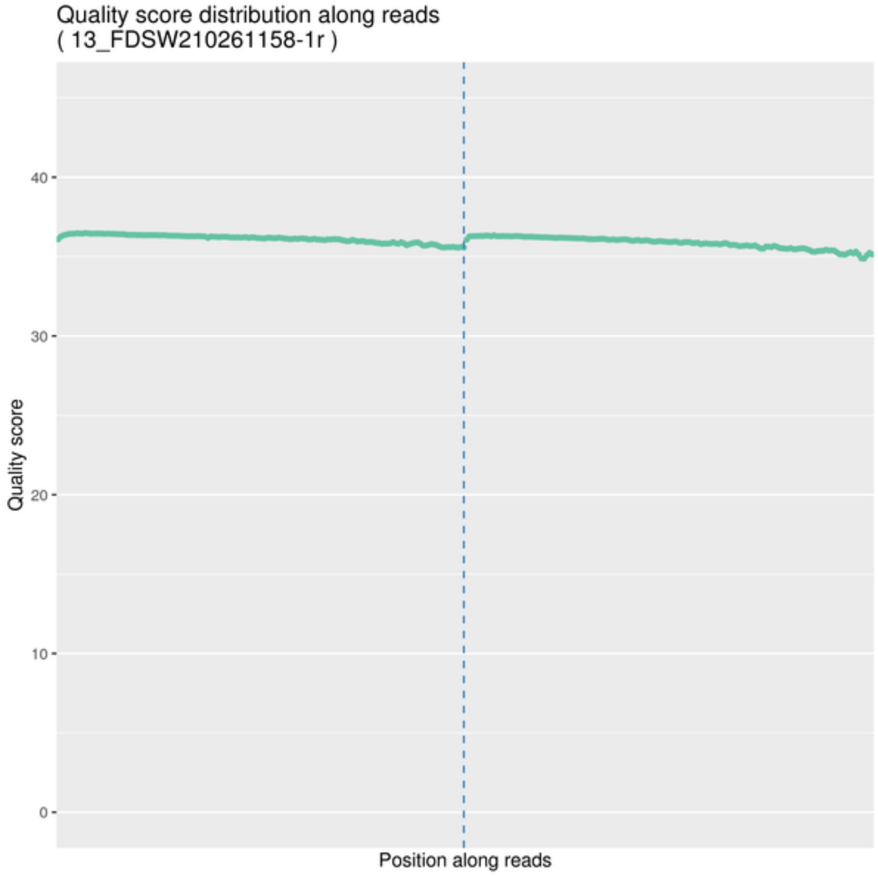
Illumina Sequence identifier details：

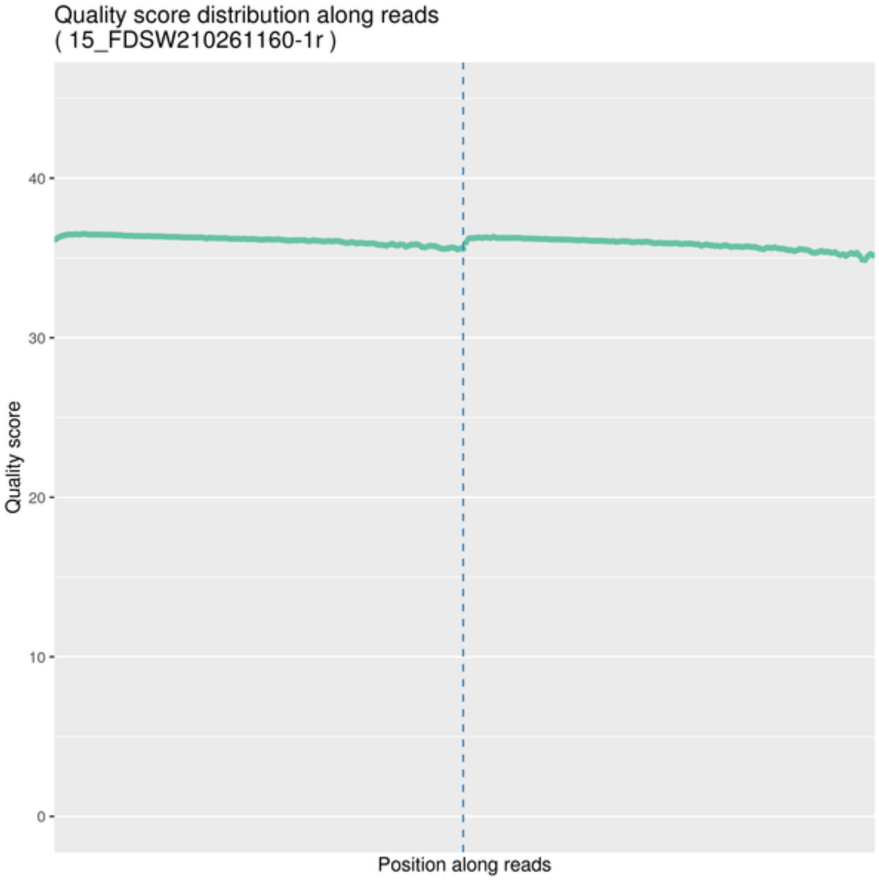
Identifier	Meaning
HWI-ST1276	Instrument – unique identifier of the sequencer
71	run number – Run number on instrument
C1162ACXX	FlowCell ID – ID of flowcell
1	LaneNumber – positive integer
110	TileNumber – positive integer

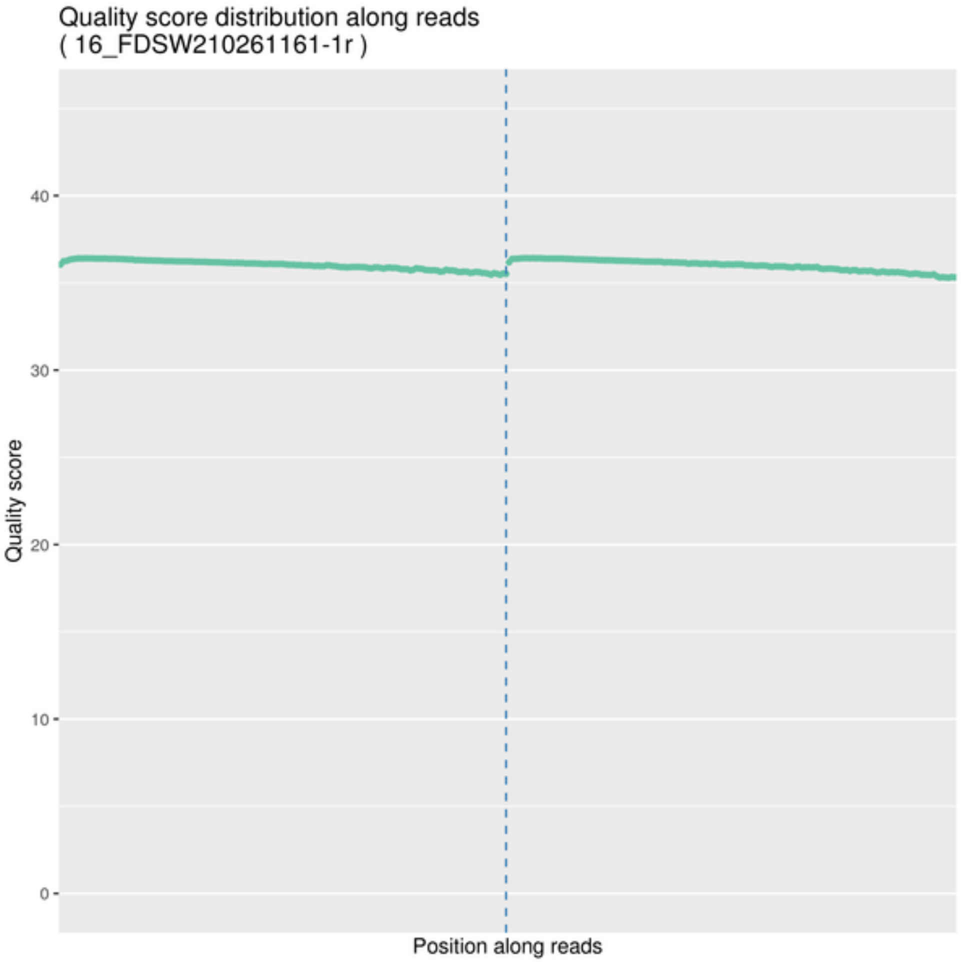
120	X – x coordinate of the spot. Integer which can be negative
245	Y – y coordinate of the spot. Integer which can be negative
1	ReadNumber - 1 for single reads; 1 or 2 for paired ends
N	whether it is filtered - NB：Y if the read is filtered out, not in the delivered fastq file, N otherwise
0	control number - 0 when none of the control bits are on, otherwise it is an even number
CGATGT	Illumina index sequences

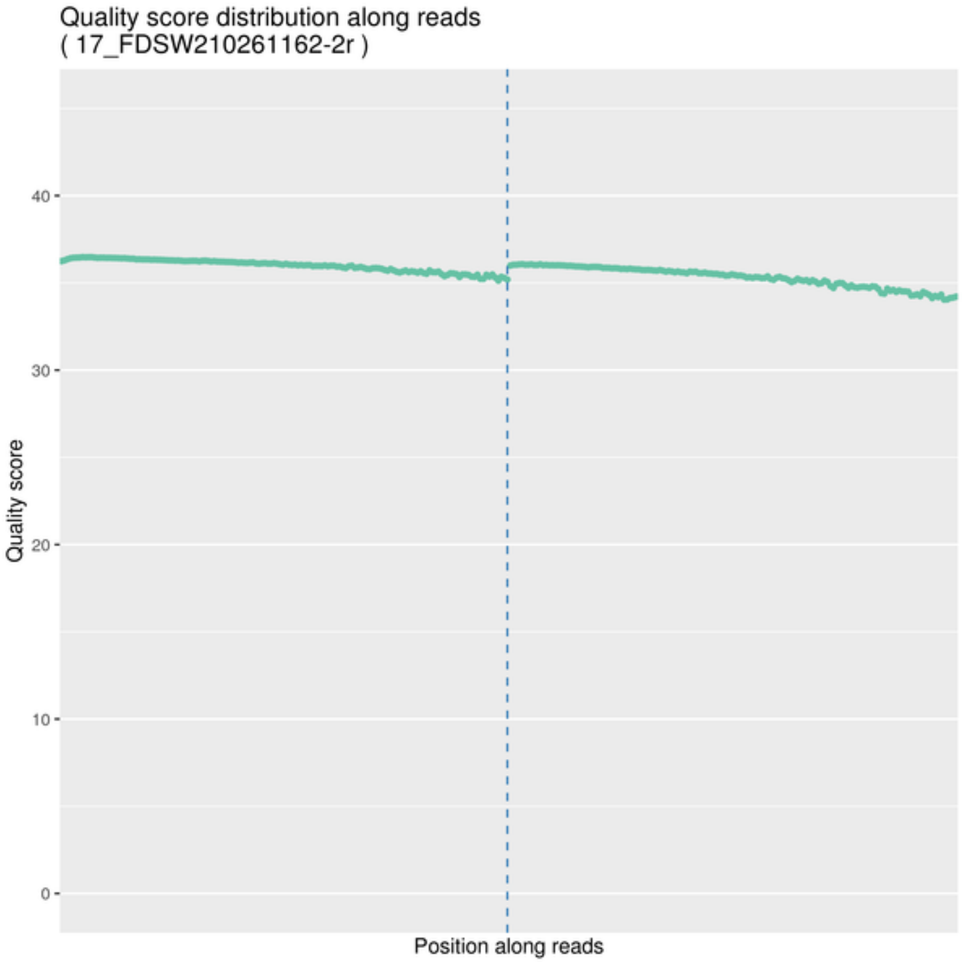
3. Data Quality Control

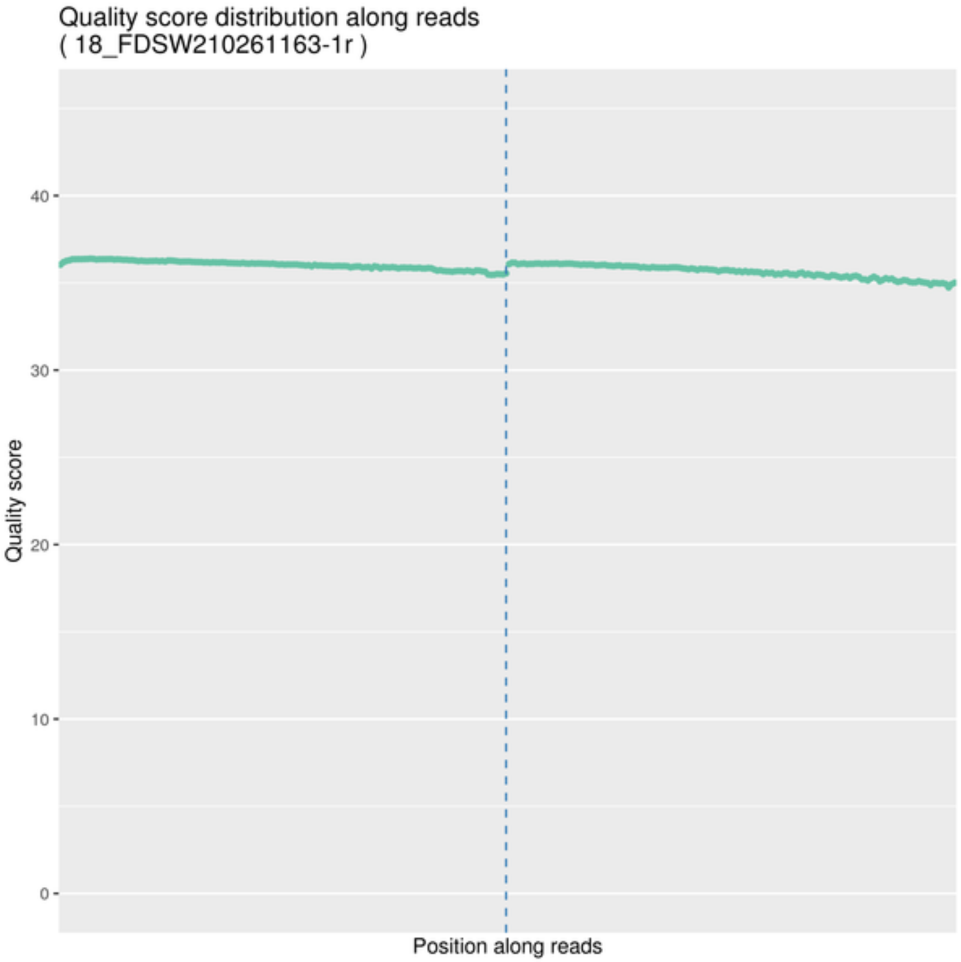
The distributions of quality score along reads in data filtering are shown in figures below:

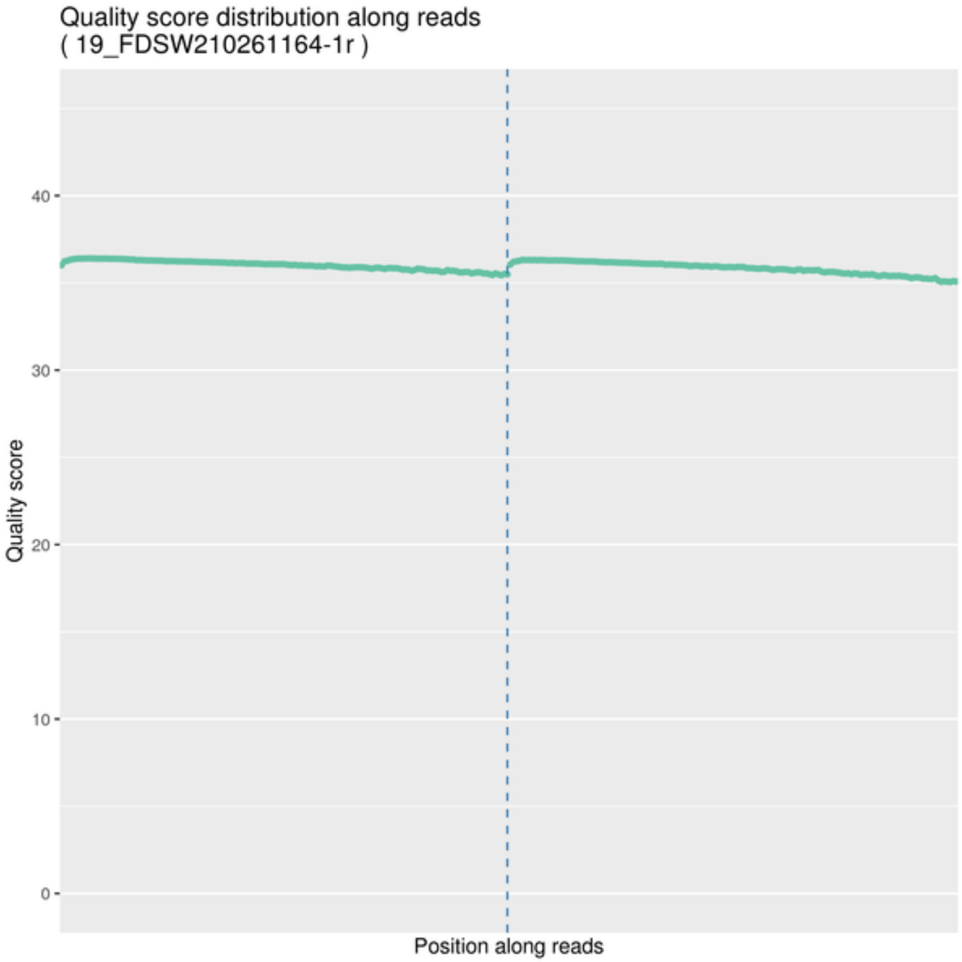


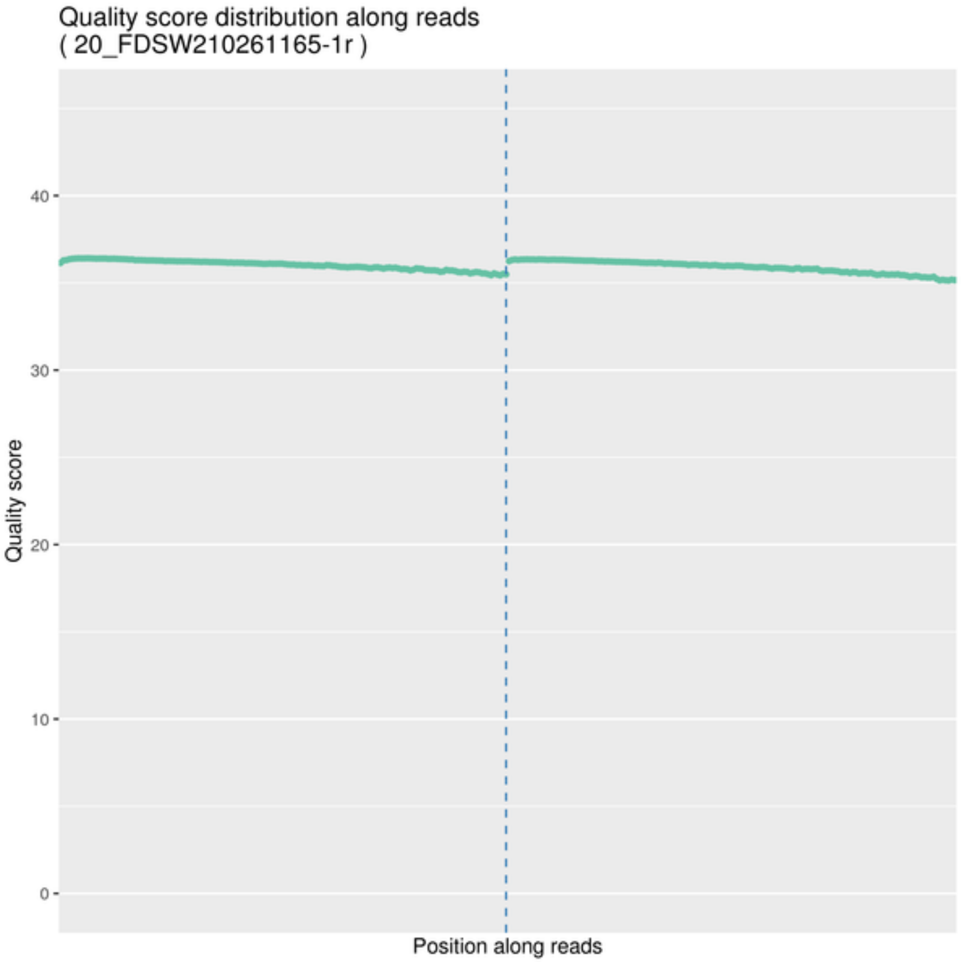


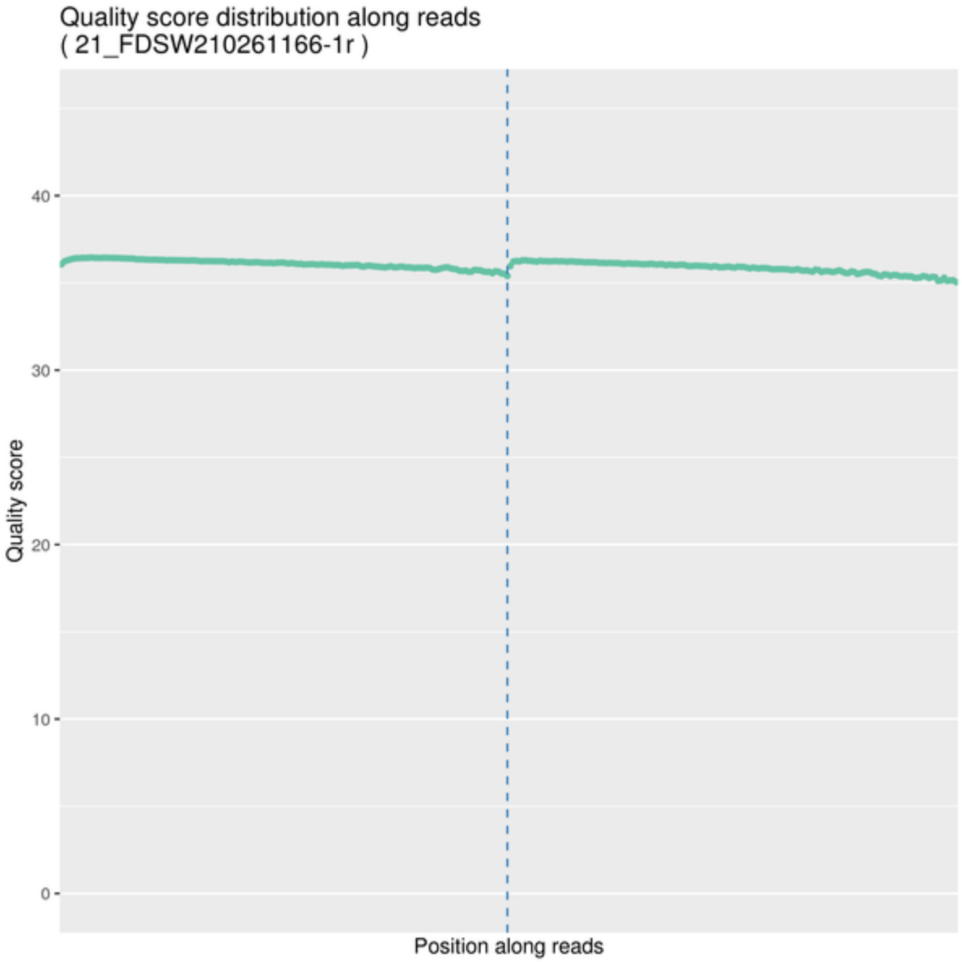

Quality score distribution along reads
(16_FDSW210261161-1r)
40
30
20
10
0
Quality score
Position along reads

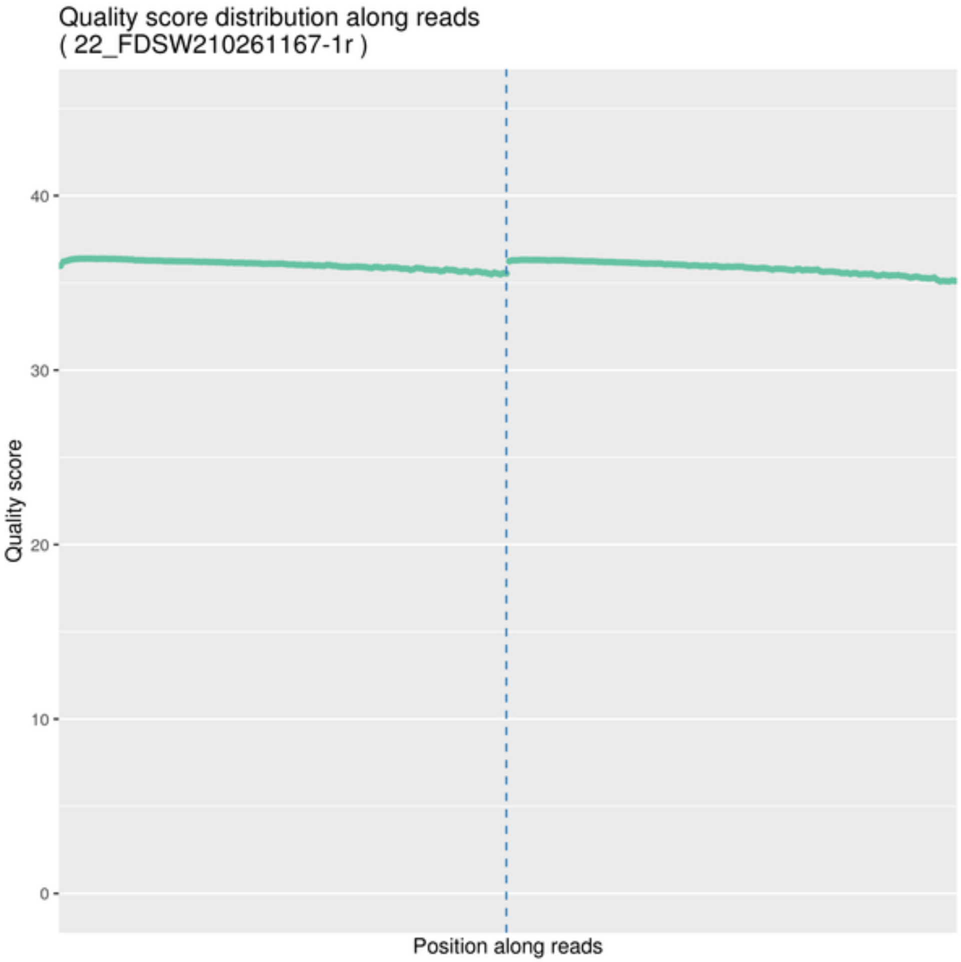

Quality score distribution along reads
(17_FDSW210261162-2r)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(18_FDSW210261163-1r)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(19_FDSW210261164-1r)
40
30
20
10
0
Quality score
Position along reads


Quality score distribution along reads
(20_FDSW210261165-1r)
Quality score
40
30
20
10
0
Position along reads


Quality score distribution along reads
(21_FDSW210261166-1r)
Quality score
40
30
20
10
0
Position along reads


Quality score distribution along reads
(22_FDSW210261167-1r)
Quality score
40
30
20
10
0
Position along reads