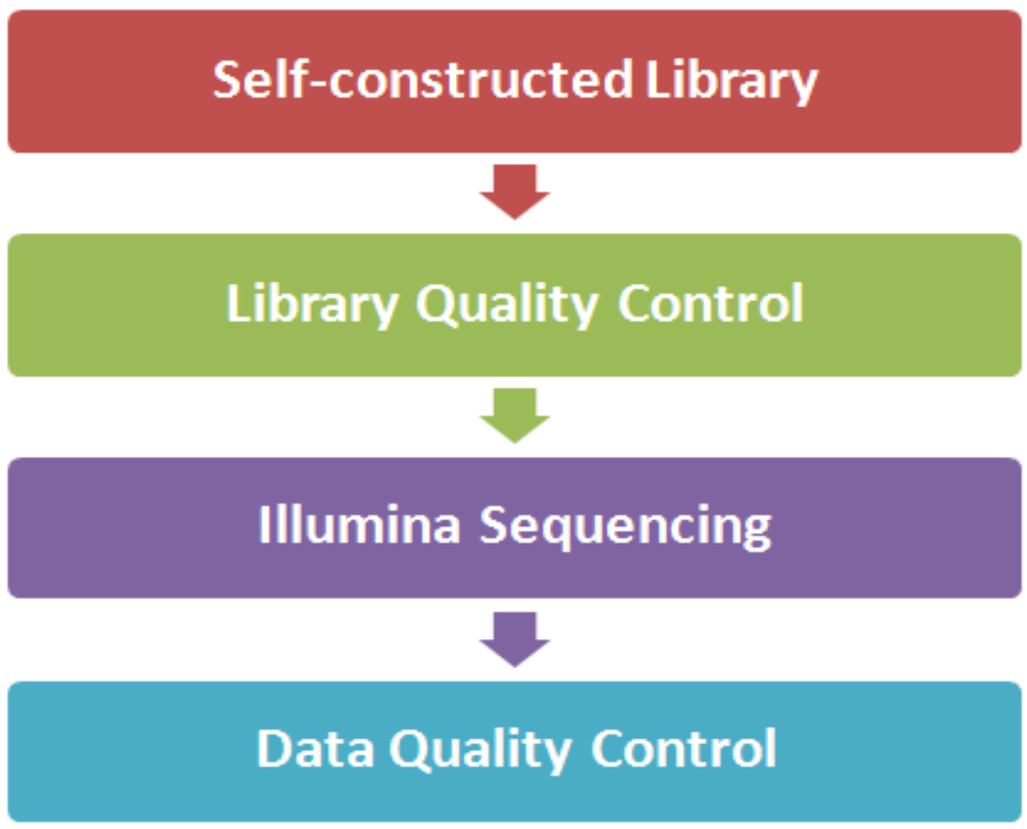
X201SC19100215-Z01-F001 QC Analysis Report

20-October-2019

A. Library Preparation and Sequencing

From the DNA samples to the final data, each step (including sample testing, library preparation and sequencing) will influence the data quality. The quality of data would have direct impacts on the analysis results. To guarantee the reliability of the data, Quality Control will be performed on each step of the procedure. For customer' self-constructed libraries, the quality detection will be performed before sequencing to ensure producing high quality of data.



1 Library Quality Control

There are mainly three methods in QC for library quality control:

(1) Qubit 2.0: tests the library concentration preliminarily.

(2) Agilent 2100: tests the insert size.

(3) Q-PCR: quantifies the library effective concentration precisely.

2 Sequencing

The qualified libraries are fed into Illumina sequencers after pooling according to its effective concentration and expected data volume.

B. Results and Instructions

1 Data Quality Control

1.1 Distribution of Sequencing Quality

The "e" represents the sequence error rate and Q_{phred} represents the base quality value,Q_{phred}=-10log$_{10}$(e). The relationship between sequencing error rate (e) and sequencing base quality value (Q_{phred}) is as below:

Phred score	error base	right base	Q-score
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

The distribution of quality score is shown in **Fig.1**:

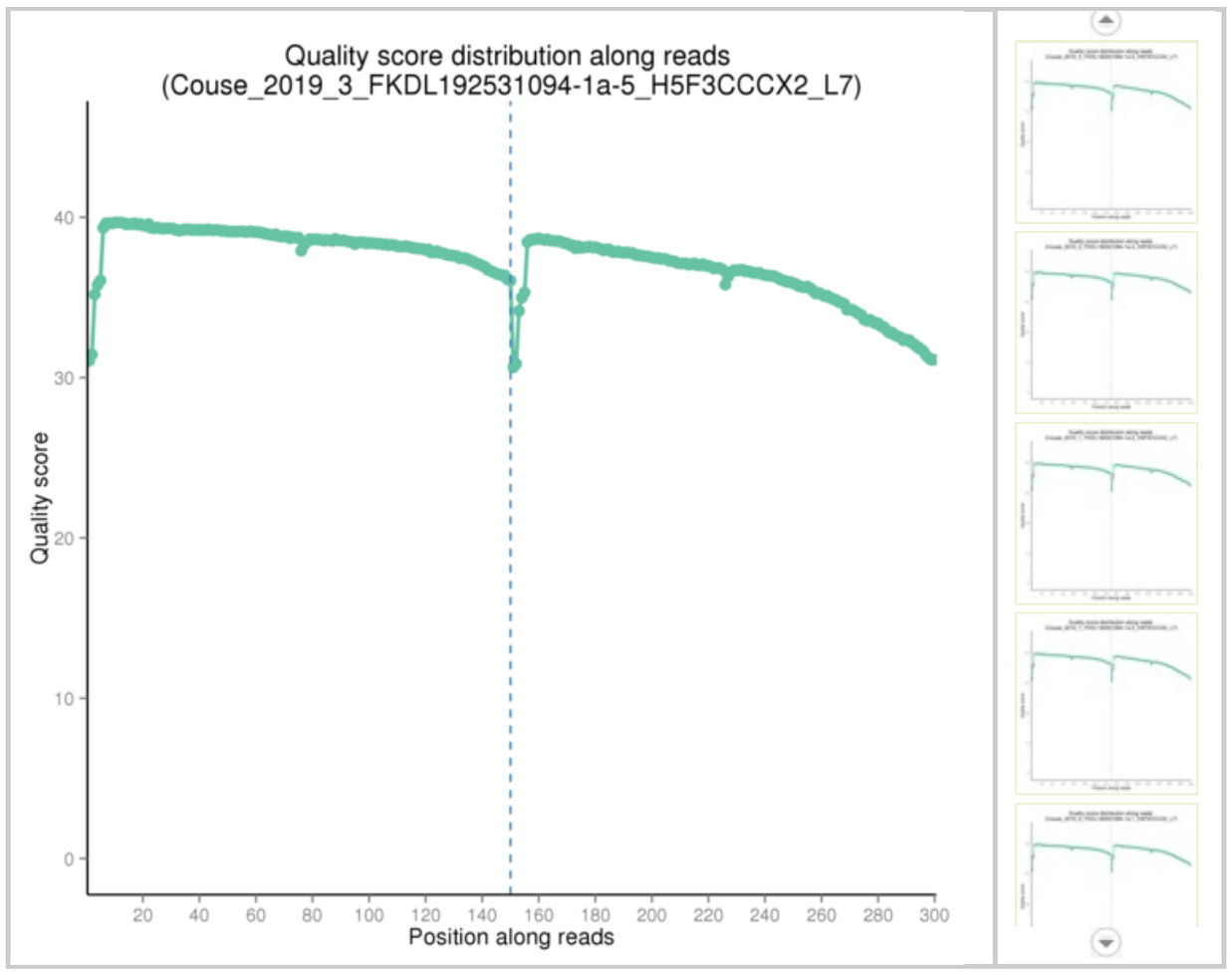


Fig.1 Distribution of Sequencing Quality

The base position is on the horizontal axis and the sequencing quality is on the vertical axis
The first half part of the distribution is for reads1 and the latter half part is for reads2

1.2 Distribution of Sequencing Error Rate

For Illumina SBS technology, the distribution of sequencing error rate has two features:

(1) Error rate grows with sequenced reads extension because of the consumption of sequencing reagent. The phenomenon is common in the Illumina high-throughput sequencing platform (Erlich Y. et al. 2008; Jiang et al. 2011).

(2) The first several bases have higher sequencing error rate than others. At the beginning of sequencing, the focusing of the sequencer's fluorescence image sensor sensing element is not sensitive enough, thus, the quality of acquired fluorescence image is low.

Generally, single base error rate should be lower than 1%. The error rate of this project is shown in **Fig.2**:

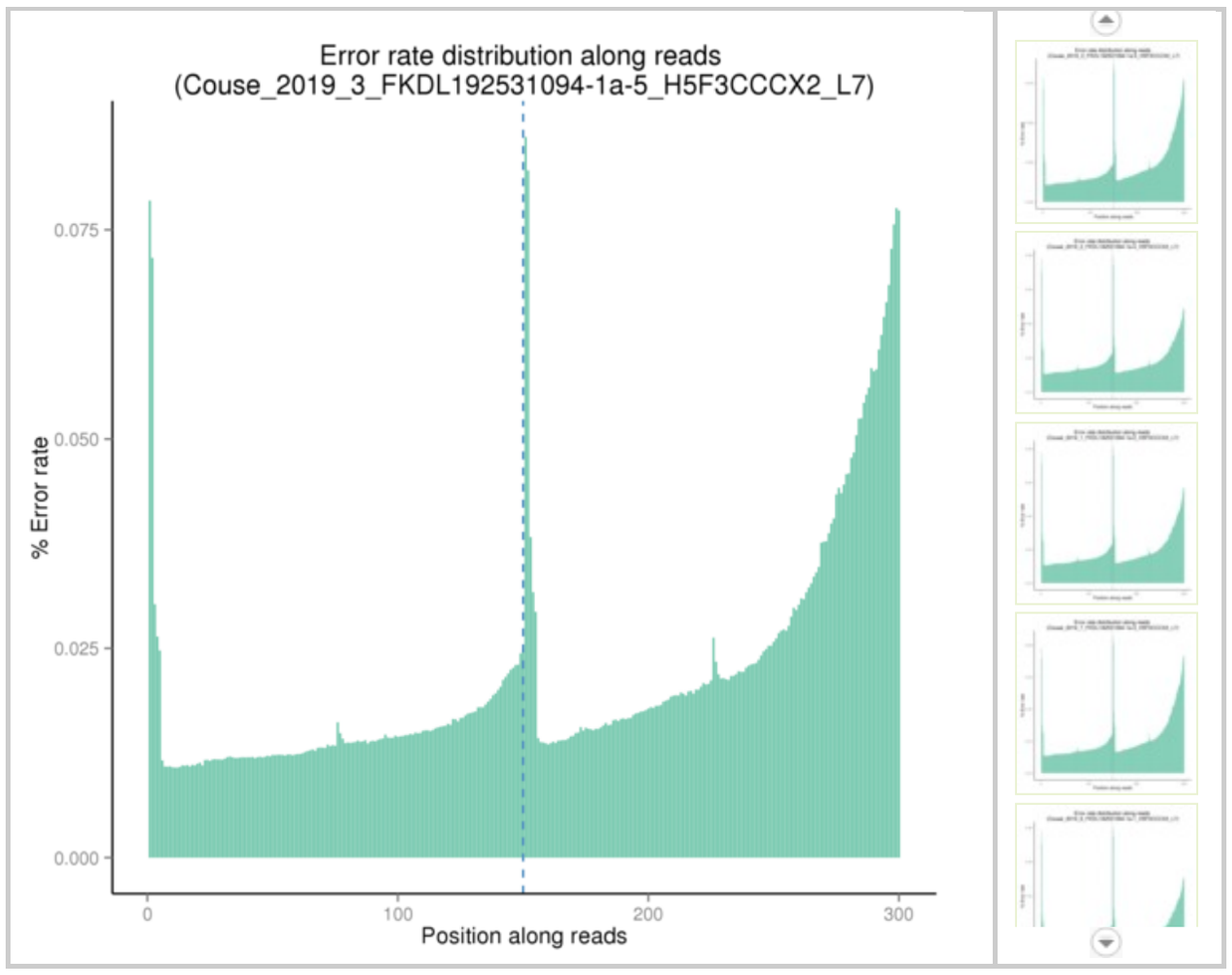


Fig.2 Error Rate Distribution

The base position is on the horizontal axis and the single base error rate is on the vertical axis

The first half part of the distribution is for reads1 and the latter half part is for reads2

1.3 Distribution of A/T/G/C Base

It is used to identify the separation situation of AT and GC by checking the distribution of GC content. According to the principle of complementary bases, the content of AT and GC should be equal at each sequencing cycle and be constant and stable in the whole sequencing procedure. But in practical measurement, due to the primer amplification bias and some other reasons, the first 6 to 7 nucleotides will fluctuate which is normal and reasonable.

The distribution of GC content is shown in **Fig.3**:

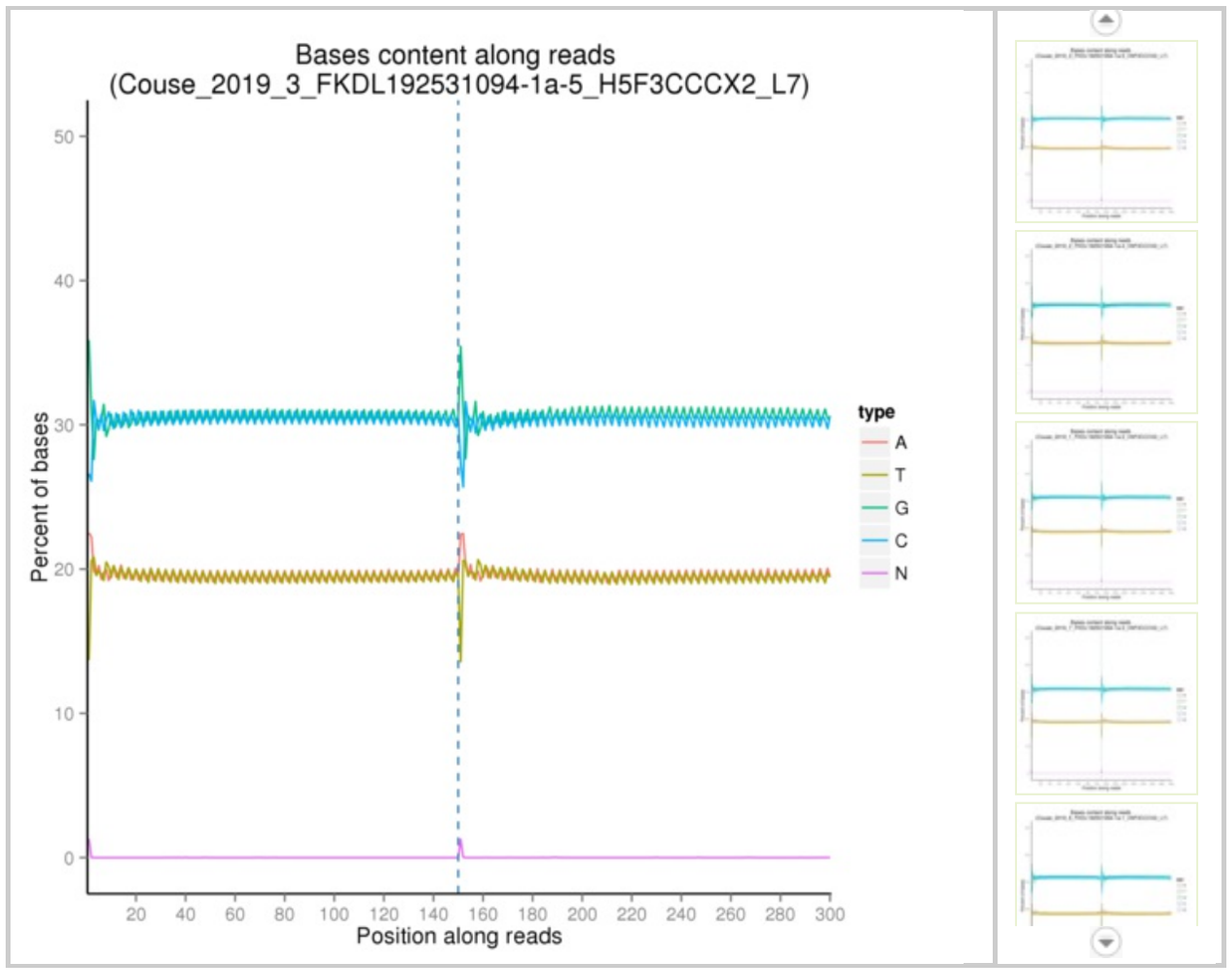


Fig.3 A/T/G/C Distribution

The base position is on the horizontal axis and the single base percentage is on the vertical axis

The first half part of the distribution is for reads1 and the latter half part is for reads2

1.4 Results of Raw Data Filtering

The sequenced reads (raw reads) often contain low quality reads and adapters, which will affect the analysis quality. So it's necessary to filter the raw reads and get the clean reads. The filtering process is as follows:

(1) Remove reads containing adapters.

(2) Remove reads containing N > 10% (N represents the base cannot be determined).

(3) Remove reads containing low quality (Qscore<= 5) base which is over 50% of the total base.

Adapter sequences :

5' Adapter **:**

5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

3' Adapter(The underlined 6bp bases is Index)**:**

5'-GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTTG-3'

The Sequencing data filtration of this project can be seen in **Fig.4** :

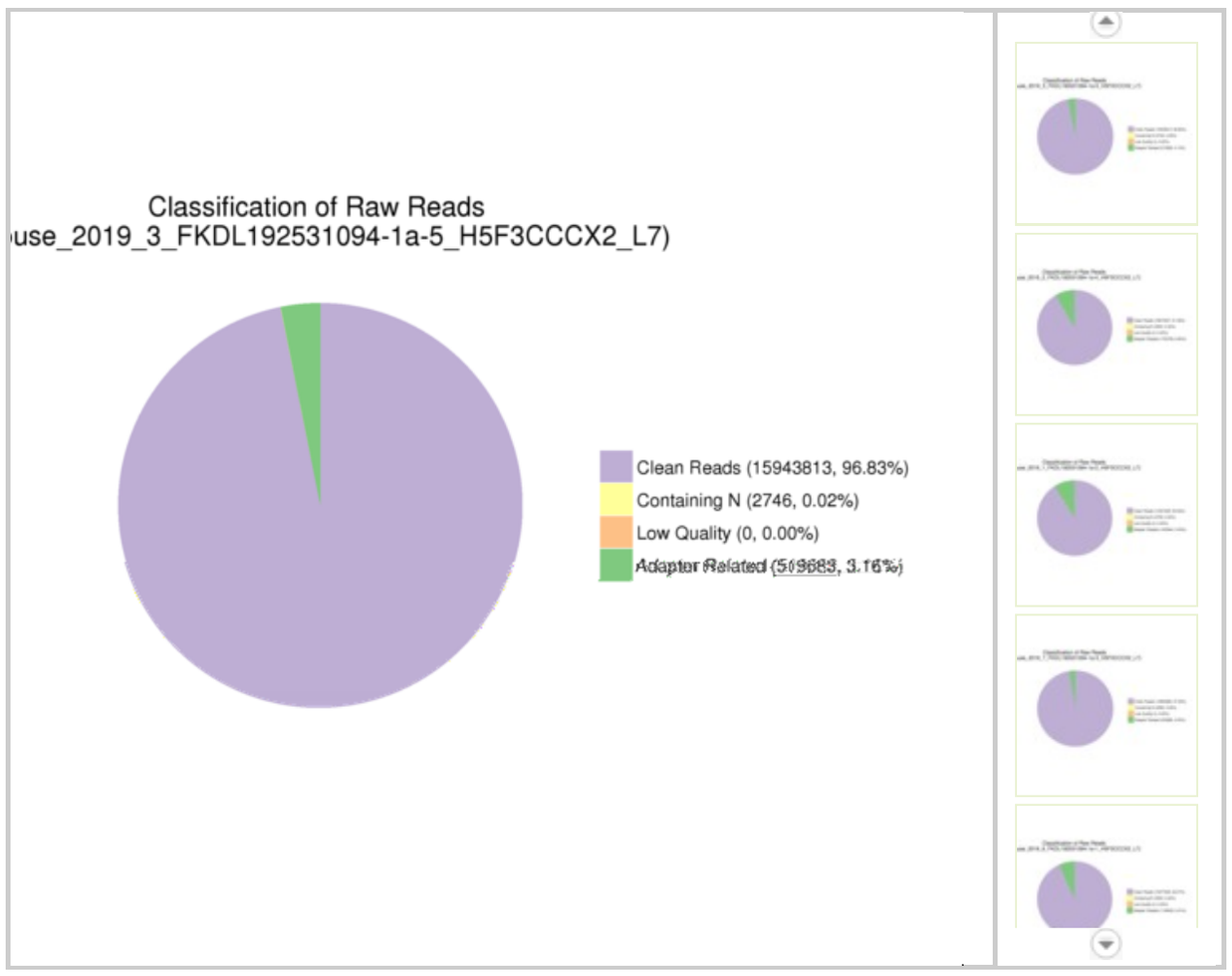


Fig.4 Composition of Raw Data

2 Summary of Sequencing Data Information

The total output of data on the sequencer: Raw data 120.2 G.

The detail statistics for the quality of sequencing data are shown in **Table 1**.

Table 1 Data Quality Summary

Sample	Library	Flowcell/Lane	Raw reads	Raw data(G)	Effective(%)	Error(%)	Q20(%)	Q30(%)	GC(%)
Couse_2019_6	FKDL192531094-1a-1	H5F3CCCX2_L7	17274501	5.2	93.07	0.01	96.05	90.53	63.32
Couse_2019_10	FKDL192531094-1a-10	H5F3CCCX2_L7	18652209	5.6	97.85	0.02	95.52	89.50	65.92
Couse_2019_11	FKDL192531094-1a-11	H5F3CCCX2_L7	18133902	5.4	97.04	0.01	95.95	90.35	61.72
Couse_2019_12	FKDL192531094-1a-13	H5F3CCCX2_L7	14608582	4.4	92.64	0.03	94.63	88.03	62.92
Couse_2019_13	FKDL192531094-1a-14	H5F3CCCX2_L7	13997000	4.2	95.30	0.03	93.27	85.69	64.13
Couse_2019_14	FKDL192531094-1a-15	H5F3CCCX2_L7	1782579	0.5	84.77	0.03	93.71	86.49	63.55
Couse_2019_15	FKDL192531094-1a-16	H5F3CCCX2_L7	11034203	3.3	97.28	0.03	93.12	85.40	54.42
Couse_2019_16	FKDL192531094-1a-17	H5F3CCCX2_L7	10432294	3.1	96.07	0.03	94.08	87.03	65.18
Couse_2019_17	FKDL192531094-1a-18	H5F3CCCX2_L7	10309098	3.1	95.26	0.03	92.90	85.00	64.43
Couse_2019_18	FKDL192531094-1a-19	H5F3CCCX2_L7	14250962	4.3	92.14	0.01	96.06	90.56	64.58
Couse_2019_1	FKDL192531094-1a-2	H5F3CCCX2_L7	15760396	4.7	90.93	0.02	95.68	89.90	62.62
Couse_2019_19	FKDL192531094-1a-20	H5F3CCCX2_L7	18489236	5.5	83.16	0.01	96.09	90.72	60.19
Couse_2019_20	FKDL192531094-1a-21	H5F3CCCX2_L7	18681355	5.6	95.54	0.01	96.17	90.80	63.97
Couse_2019_21	FKDL192531094-1a-22	H5F3CCCX2_L7	13422972	4.0	96.50	0.02	95.87	90.19	63.81
Couse_2019_22	FKDL192531094-1a-24	H5F3CCCX2_L7	12870240	3.9	93.70	0.02	95.75	90.00	57.62
Couse_2019_7	FKDL192531094-1a-3	H5F3CCCX2_L7	15259450	4.6	97.36	0.03	95.00	88.52	62.12
Couse_2019_23	FKDL192531094-1a-32	H5F3CCCX2_L7	11603581	3.5	93.63	0.03	93.81	86.56	64.88
Couse_2019_24	FKDL192531094-1a-33	H5F3CCCX2_L7	13851851	4.2	94.84	0.03	93.41	85.89	64.61
Couse_2019_25	FKDL192531094-1a-34	H5F3CCCX2_L7	11932196	3.6	95.20	0.03	93.78	86.54	57.94
Couse_2019_26	FKDL192531094-1a-35	H5F3CCCX2_L7	16648454	5.0	89.40	0.02	95.36	89.37	63.12
Couse_2019_27	FKDL192531094-1a-36	H5F3CCCX2_L7	11220224	3.4	89.19	0.02	95.22	89.09	63.33
Couse_2019_2	FKDL192531094-1a-4	H5F3CCCX2_L7	20374727	6.1	91.36	0.01	96.11	90.69	64.03
Couse_2019_28	FKDL192531094-1a-44	H5F3CCCX2_L7	1035151	0.3	81.29	0.01	96.07	90.69	66.57
Couse_2019_29	FKDL192531094-1a-45	H5F3CCCX2_L7	2217902	0.7	78.52	0.01	96.33	91.19	65.78
Couse_2019_30	FKDL192531094-1a-46	H5F3CCCX2_L7	860440	0.3	90.82	0.03	95.12	89.11	61.38
Couse_2019_31	FKDL192531094-1a-47	H5F3CCCX2_L7	807068	0.2	92.77	0.03	94.88	88.53	59.58
Couse_2019_32	FKDL192531094-1a-48	H5F3CCCX2_L7	6540741	2.0	94.44	0.03	93.99	86.82	64.15
Couse_2019_3	FKDL192531094-1a-5	H5F3CCCX2_L7	16466242	4.9	96.83	0.03	94.73	88.12	60.95
Couse_2019_4	FKDL192531094-1a-6	H5F3CCCX2_L7	22406571	6.7	97.54	0.01	95.74	89.95	63.49
Couse_2019_5	FKDL192531094-1a-7	H5F3CCCX2_L7	11981183	3.6	98.48	0.03	93.25	85.53	62.89
Couse_2019_8	FKDL192531094-1a-8	H5F3CCCX2_L7	12572084	3.8	98.83	0.03	94.52	87.71	62.71
Couse_2019_9	FKDL192531094-1a-9	H5F3CCCX2_L7	14976480	4.5	90.69	0.01	95.82	90.19	61.19

Sample: sample name
Raw reads: four rows are taken as a unit to calculate the total amount of read1 and read2 in raw data files
Raw bases: (total raw reads) * (sequence length), calculating in G
Error rate: base error rate
Q20, Q30: (Base count of Phred value > 20 or 30) / (Total base count)
GC content: (G & C base count) / (Total base count)

C. Appendix

1. Introduction of Sequencing Data Format

The original data obtained from the high throughput sequencing platforms are transformed to sequenced reads by base calling. Raw data are recorded in a FASTQ file which contains sequenced reads and corresponding sequencing quality information. Every read in FASTQ format is stored in four lines as follows (Cock P.J.A. et al. 2010):

```
@HWI-ST1276:71:C1162ACXX:1:1101:1208:2458 1:N:0:CGATGT
NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHFAE0ECFFD/AEHH
```

Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description (such as a FASTA title line).

Line 2 is the sequence of the read.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the quality values for the bases in Line 2.

The details of Illumina sequence identifier are as follows:

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
1101	TileNumber – positive integer
1208	X – x coordinate of the spot. Integer which can be negative
2458	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

2. Explanation of Sequencing Data Related

(1) The data deliverd is a compressed file in format of '.fq.gz'. Before data delivery, we will calculate the md5 value of each compressed file and please check it when you get the data. There are two ways to check the md5 value. In Linux environment, you can use 'md5sum -c <*md5.txt>' command under the data directory. In Windows environment, you can use a calibration tool e.g. hashmyfiles. If the md5 value of compressed file doesn't match with the one we provide in md5 file in data directory, the file may have been damaged during the transmitting procedure.

(2) For paired-end (PE) sequencing, every sample should have 2 data flies (read1 file and read2 file). These 2 files have the same line number, you could use 'wc -l' command to check the line number in Linux environment. The line number divide by 4 is the number of reads.

(3) The date size is the space occupied by the data in the hard disk. It's related to the format of disk and compression ratio. And it has no influence on the quantity of sequenced bases. So the size of read1 file may be unequal to the size of read2 file.

(4) When customer’s samples need large amount of data e.g. whole genome sequencing data, we would use separate-lane sequencing strategy to make sure the quality of data. So it's possible that one sample has several parts sequencing data. For example, if sample 1 has two read1 files, sample1_L1_1.fq.gz and sample1_L2_1.fq.gz, that means this sample was sequenced on different lanes.

(5) About the sequenced reads. The Index is normally in the middle of the adapter during the process of experimenting and sequencing except the special library. We can get the Read1 sequence and Read2 sequence by Index read. They are all the sequence of samples so that it's no necessary to dispose the beginning and end of reads in the downstream analysis(e.g. mapping).

(6) Ninety days after the data delivery, we will delete outdated data. So please keep your data properly. If you have any question or doubt, please contact us as soon as possible. Have a nice day!

3 References

Cock P.J.A. et al (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic acids research 38, 1767-1771.

Hansen K.D. et al (2010). Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic acids research 38, e131-e131.

Erlich Y.et al (2008). Alta-Cyclic: a self-optimizing base caller for next-generation sequencing.Nature Methods,5,679-682.

Jiang L.C. et al (2011). Synthetic spike-in standards for RNA-seq experiments. Genome research 21, 1543-1551.

Yan L.Y. et al (2013). Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. Nat Struct Mol Biol.