

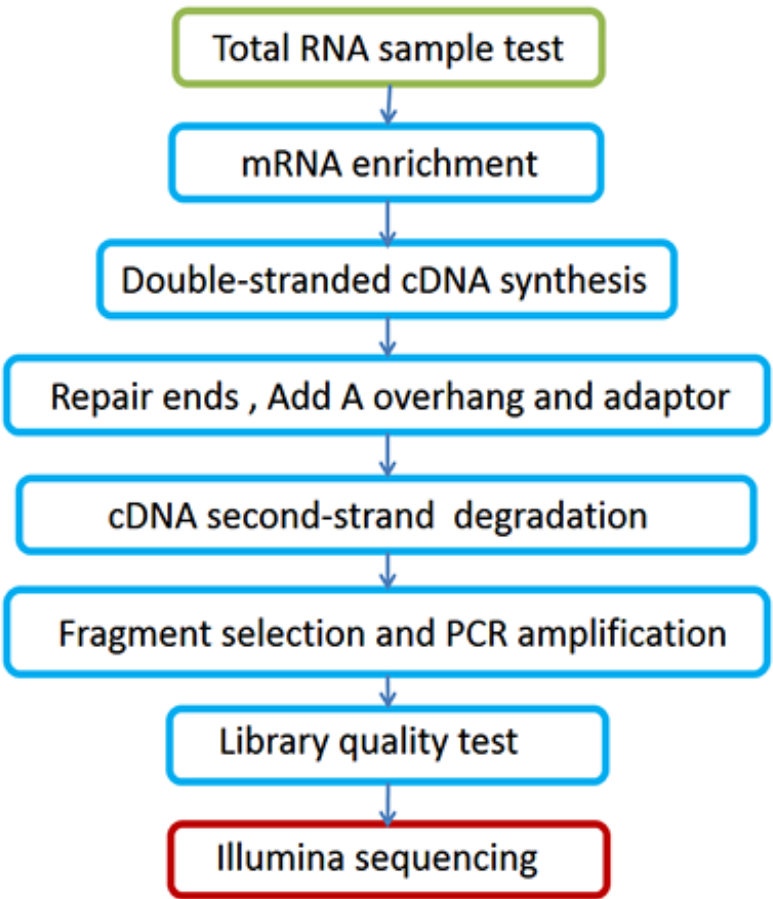
## C201SC18122416 QC Analysis Report

04-March-2019

- Library Preparation and Sequencing
  - Sample Quality Control
  - Library Construction and Quality Control
  - Sequencing
- Results and Instructions
  - Data Quality Control
    - Distribution of Sequencing Quality
    - Distribution of Sequencing Error Rate
    - Distribution of A/T/G/C Base
    - Results of Raw Data Filtering
  - Summary of Sequencing Data Information
- Appendix
  - Introduction of Sequenced Data Format
  - Explanation of Sequencing Data Related
  - References

A. Library Preparation and Sequencing

From the RNA samples to the final data, each step, including sample test, library preparation, and sequencing, influences the quality of the data, and data quality directly impacts the analysis results. To guarantee the reliability of the data, quality control (QC) is performed at each step of the procedure. The workflow is as follows:



1 Sample Quality Control

There are three main methods of QC for RNA samples:

- (1) Nanodrop: Preliminary quantitation
- (2) Agarose Gel Electrophoresis: tests RNA degradation and potential contamination
- (3) Agilent 2100: checks RNA integrity and quantitation

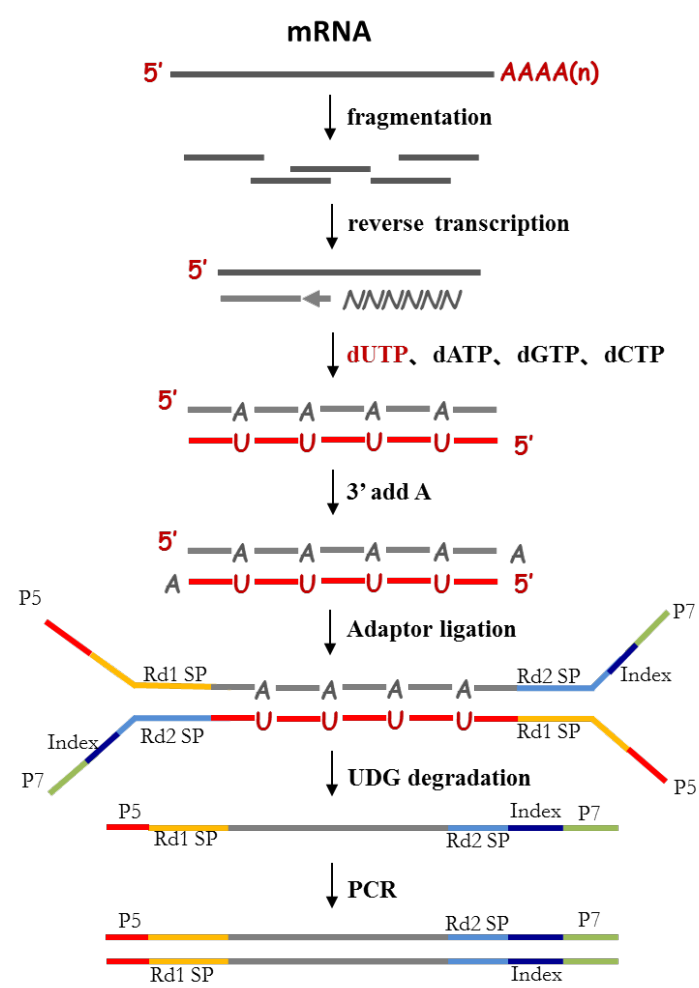
2 Library Construction and Quality Control

After the QC procedures, mRNA from organisms is enriched using oligo(dT) beads. For prokaryotic samples, rRNA is removed using the Ribo-Zero kit that leaves the mRNA. First, the mRNA is fragmented randomly by adding fragmentation buffer, then the cDNA is synthesized by using mRNA template and random hexamers primer, after which a custom second-strand synthesis buffer (Illumina) , dNTPs, RNase H and DNA polymerase I are added to initiate the second-strand synthesis. Second, after a series of terminal repair, A ligation and sequencing adaptor ligation, the double-stranded cDNA library is completed through size selection and PCR enrichment.

The quality control of library consists of three steps:

- (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.

The workflow chart is as follows:



### 3 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,  $Q_{phred} = -10\log_{10}(e)$ . 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**:

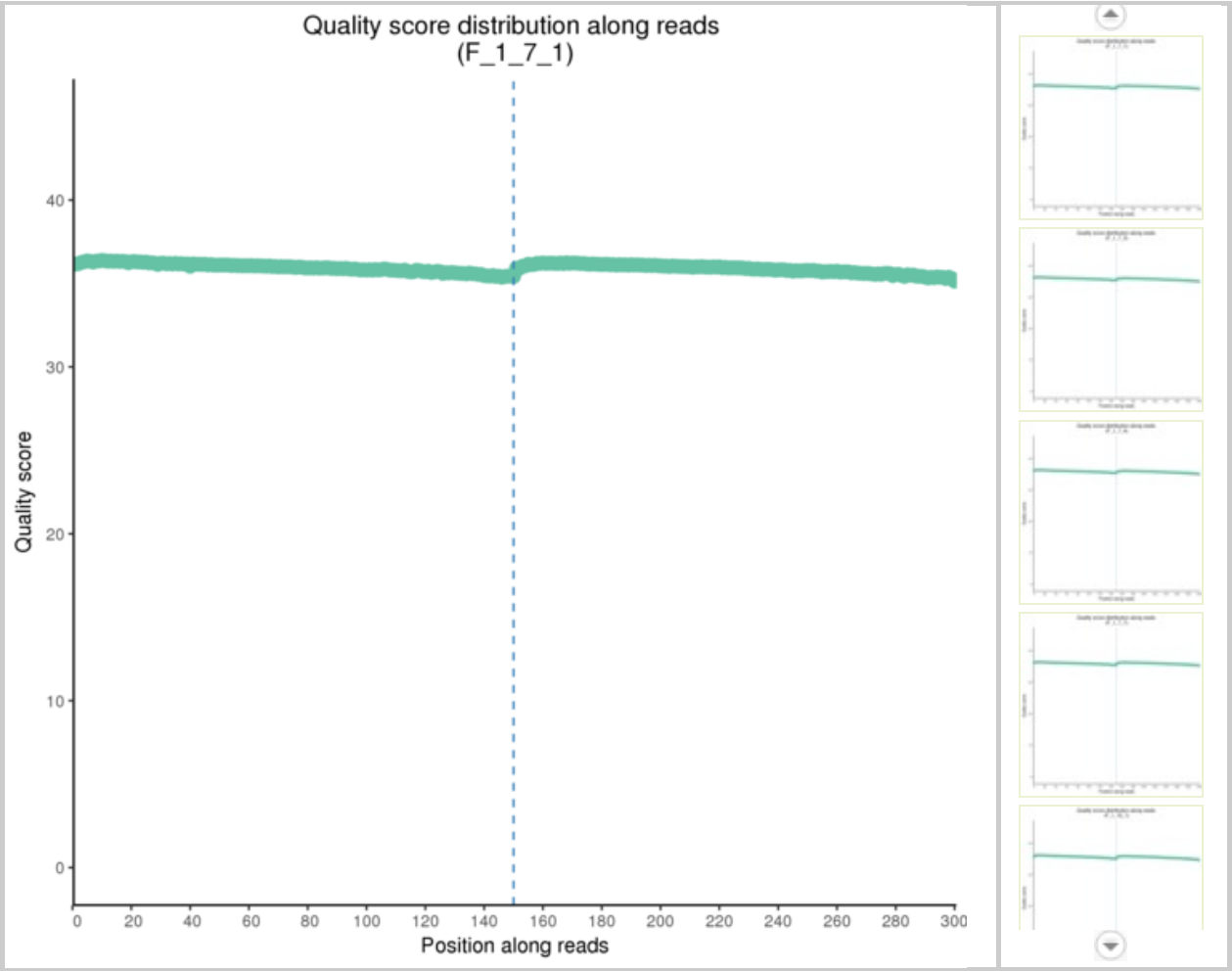


Fig.1 Distribution of Sequencing Quality

The base position is on the horizontal axis and the sequencing quality is on the vertical axis

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 reason for the high error rate of the first six bases is that the random hex-primers and RNA template bind incompletely in the process of cDNA synthesis (Jiang et al.2011).

The error rate of this project is shown in **Fig.2**:

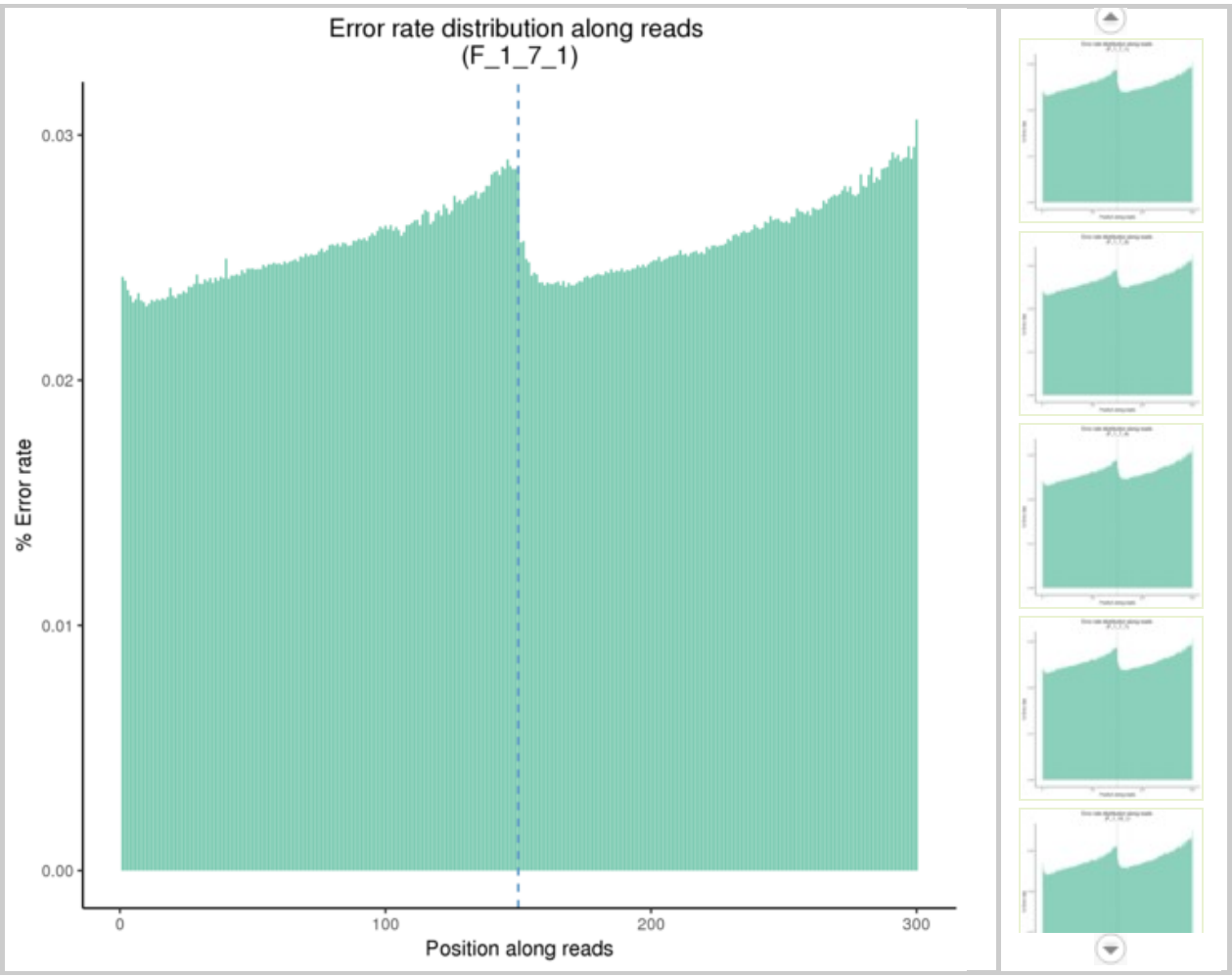


Fig.2 Error Rate Distribution

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

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. For the stranded-specific library (dUTP library), which remains only single strand information, the distribution of GC contents fluctuates obviously. So it is normal of occurring GC separation.

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

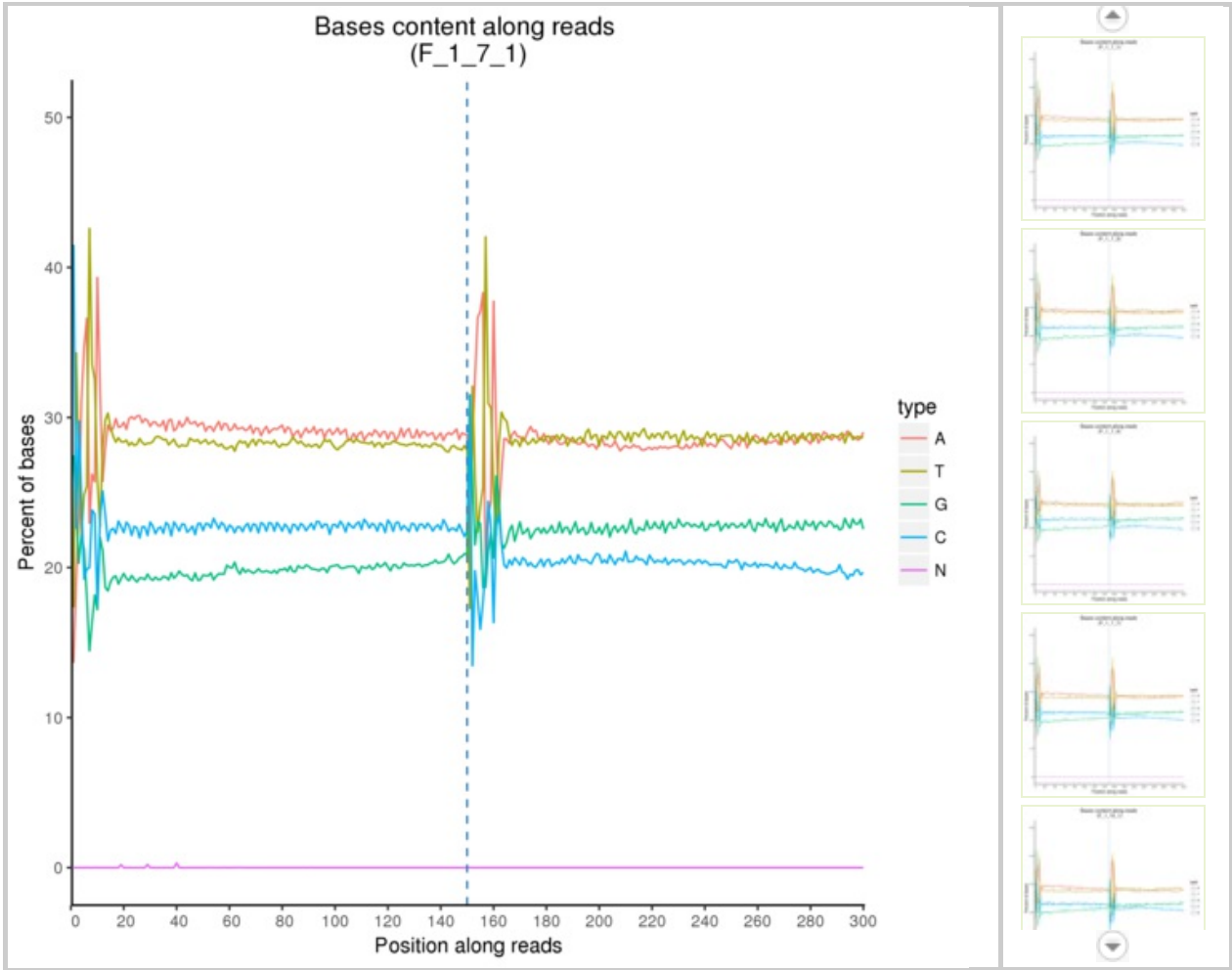


Fig.3 A/T/G/C Distribution

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

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 :

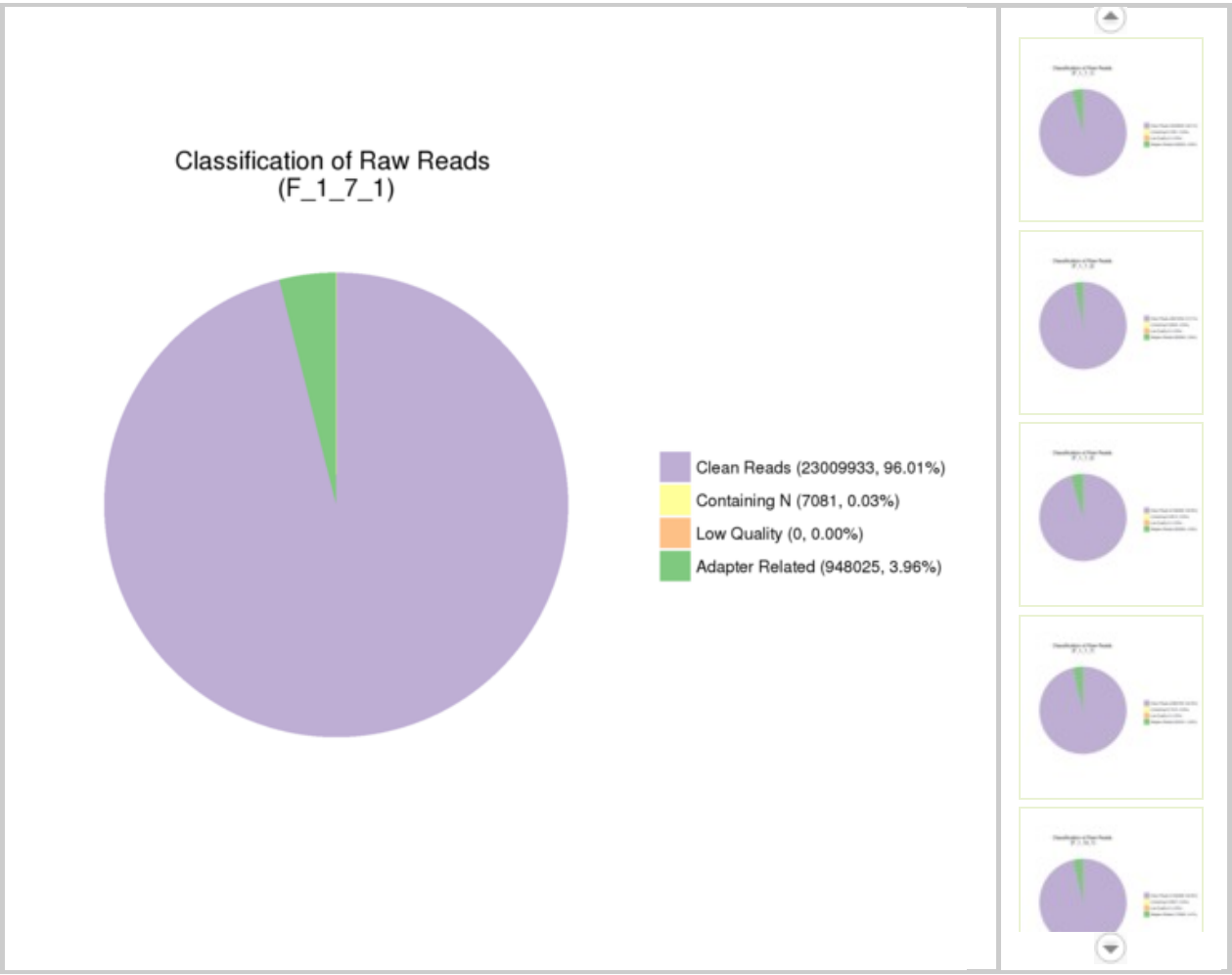


Fig.4 Composition of Raw Data

Different color for different components:

- (1)Adapter related: (reads containing adapter) / (total raw reads)
- (2)Containing N: (reads with more than 10% N) / (total raw reads)
- (3)Low quality: (reads of low quality) / (total raw reads)
- (4)Clean reads: (clean reads) / (total raw reads)

2 Summary of Sequencing Data Information

The total output of data on the sequencer: Raw data 143.5 G, and the data filtered from raw data: Clean data 137.5 G.

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

Table 1 Data Quality Summary									
Sample	Raw Reads	Clean Reads	Raw Base(G)	Clean Base(G)	Effective Rate(%)	Error Rate(%)	Q20(%)	Q30(%)	GC Content(%)
F_1_7_1	23965039	23009933	7.2	6.9	96.01	0.03	97.73	93.54	42.82
F_1_7_3	29146123	28274630	8.7	8.5	97.01	0.03	97.43	92.84	42.88
F_1_7_4	22040082	21080620	6.6	6.3	95.65	0.03	97.55	93.10	43.30
F_1_7_7	24805236	23897333	7.4	7.2	96.34	0.03	97.77	93.59	43.22
F_1_10_1	22345231	21562999	6.7	6.5	96.50	0.03	97.02	91.68	43.91
F_1_10_2	22956626	22107993	6.9	6.6	96.30	0.03	97.43	92.90	43.88
F_1_10_3	21598106	20915710	6.5	6.3	96.84	0.03	96.75	91.18	43.80
F_1_10_4	24641451	22588280	7.4	6.8	91.67	0.03	97.59	93.26	43.78
F_2_4_3	28380852	26450804	8.5	7.9	93.20	0.03	97.29	92.60	42.86
F_2_4_4	20414226	19656998	6.1	5.9	96.29	0.03	97.78	93.63	42.91
F_2_4_6	23491316	22619486	7.0	6.8	96.29	0.03	97.35	92.31	42.68
F_2_4_7	27082744	25862366	8.1	7.8	95.49	0.03	97.66	93.39	43.43
F_2_F_4	23823479	22819749	7.1	6.8	95.79	0.03	97.62	93.30	43.68
F_2_F_F	26315475	25421239	7.9	7.6	96.60	0.03	97.38	92.73	43.64
F_2_F_6	21551806	20731364	6.5	6.2	96.19	0.03	97.65	93.34	43.85
F_2_F_8	24122472	23170799	7.2	7.0	96.05	0.03	97.77	93.59	43.64
WT1	21397005	20196968	6.4	6.1	94.39	0.03	97.14	92.42	43.03
WT2	22954251	22111759	6.9	6.6	96.33	0.03	97.48	92.98	43.94
WT3	23256472	22568369	7.0	6.8	97.04	0.03	97.37	92.73	43.88
WT5	24141584	23334653	7.2	7.0	96.66	0.03	97.41	92.82	43.68

Sample: sample name

Raw reads: total amount of reads of raw data, each four lines taken as one unit. For paired-end sequencing, it equals the amount of read1 and read2, otherwise it equals the amount of read1 for single-end sequencing.

Clean reads: total amount of reads of clean data, each four lines taken as one unit. For paired-end sequencing, it means the amount of read1 and read2, otherwise it equals the amount of read1 for single-end sequencing.

Raw bases: (Raw reads) \* (sequence length), calculating in G. For paired-end sequencing like PE150, sequencing length equals 150, otherwise it equals 50 for sequencing like SE50.

Clean bases: (Clean reads) \* (sequence length), calculating in G. For paired-end sequencing like PE150, sequencing length equals 150, otherwise it equals 50 for sequencing like SE50.

Effective Rate(%): (Clean reads/Raw reads)\*100%

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
NAAGAACACGTTCCGGTCACCTCAGCACACTTGTGAATGTCATGGGATCCAT
+
#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:

Ident ifier	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 delivered 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 files (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 data 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 quality control standard. If we promise to deliver the clean data, we will filter the data strictly according to the standard to obtain high quality clean data which can be used for further research and paper writing. We will discard the paired reads in the following situation: when either one read contains adapter contamination; when either one read contains uncertain nucleotides more than 10 percent; when either one read contains low quality nucleotides (base quality less than 5) more than 50 percent, discard the paired reads. The data analysis results based on this standard can be approved by high level magazines (Yan L.Y. et al . 2013). If you want to get more information, please refer to the official website of Novogene ([www.novogene.com](http://www.novogene.com)).

(6) 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).

(7) 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.