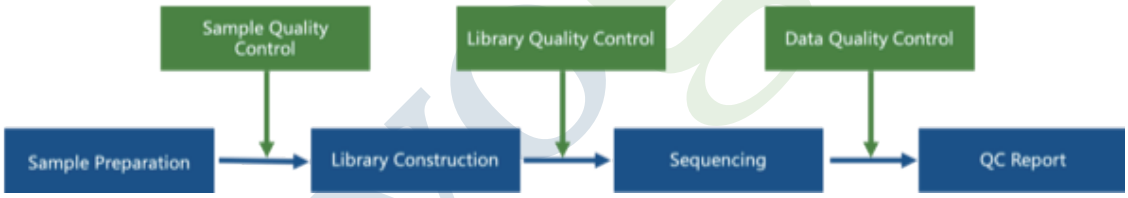
QC Analysis Report

Contract Information	Contract Content
Contract ID	H204SC19122554
Batch ID	X204SC19122554-Z01-F001
Report Time	2020-02-29

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

2 Library Construction and Quality Control

After passing the QC procedures, mRNA in eukaryotic total RNA samples is enriched using oligo (dT) beads. The enriched 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. After a series of end terminal repair, A tailing, sequencing adapter ligation, the double-stranded cDNA library is completed through size selection and PCR enrichment.

The quality control of library consists of three steps:

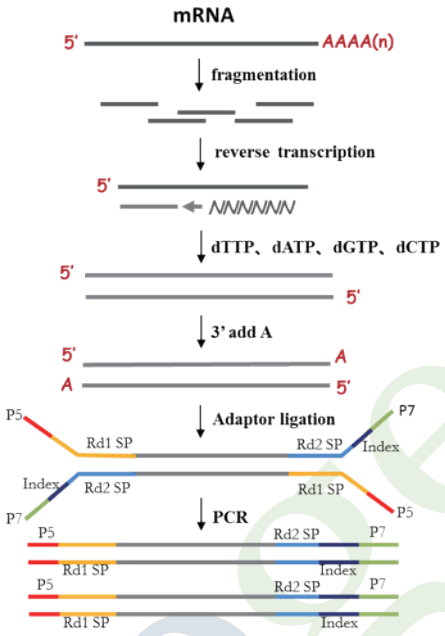
(1)Qubit 2.0: Tests the library concentration preliminarily.

(2)Agilent 2100: Tests the insert size.

(3)qPCR: Quantifies the library effective concentration precisely.

The workflow chart is as follows :

The genomic DNA of each sample will be randomly sheared into short fragments of about 350 bp.



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 Qphred represents the base quality value,Qphred=-10log10(e).

The relationship between sequencing error rate (e) and sequencing base quality value (Qphred) 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:

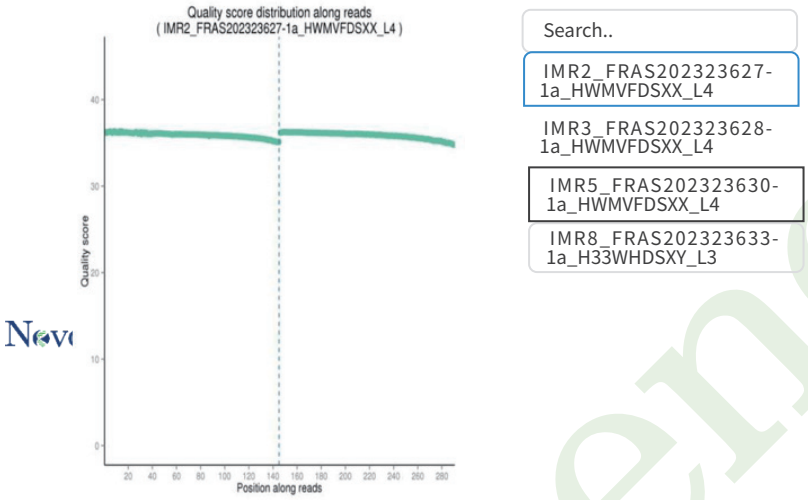


Fig. Distribution of cing Quality

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

1.2 Dstribution 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:

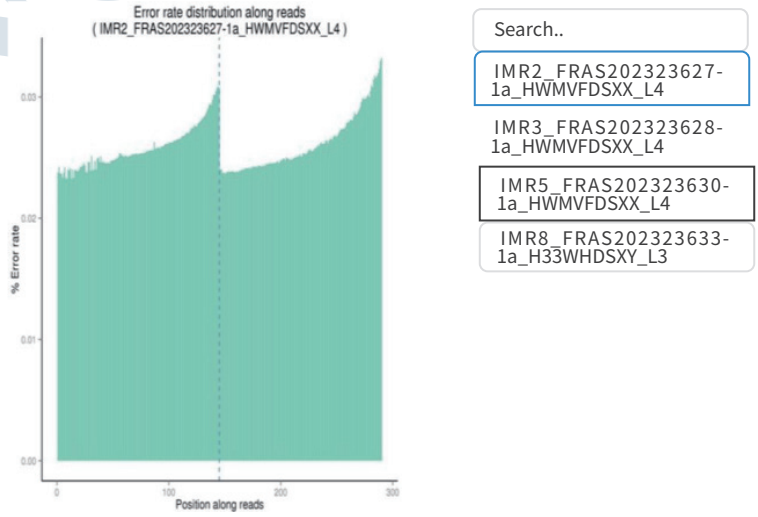


Fig.2 Error Rate Distribution

YThe 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

Base distribution is used to identify abnormal AT and GC fluctuations and separations. 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:

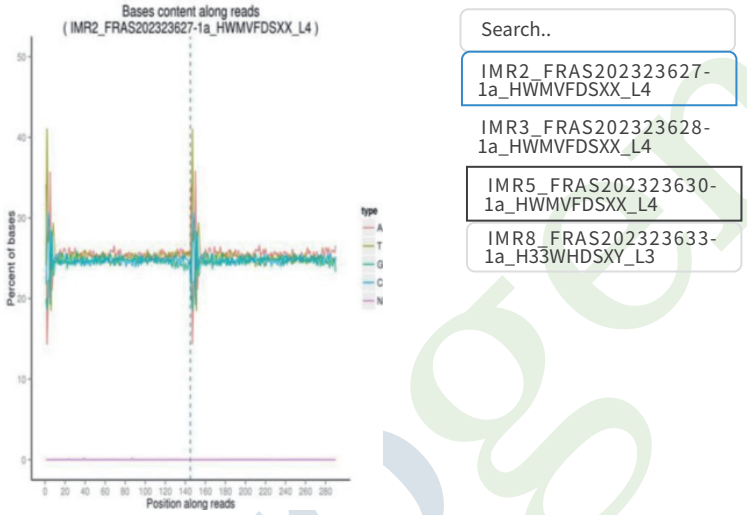


Fig.3 A/T/G/C Distribution

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

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. As a result, it is necessary to filter the raw reads and get the clean (filtered) reads. The filtering process is as follows:·

(1)Remove reads containing adapters.
(2)Remove reads containing N >10% (N represents an unresolvable base).
(3)Remove reads containing >50% of low quality (Qphred <= 5) bases.

Sequences of adaptor

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 :

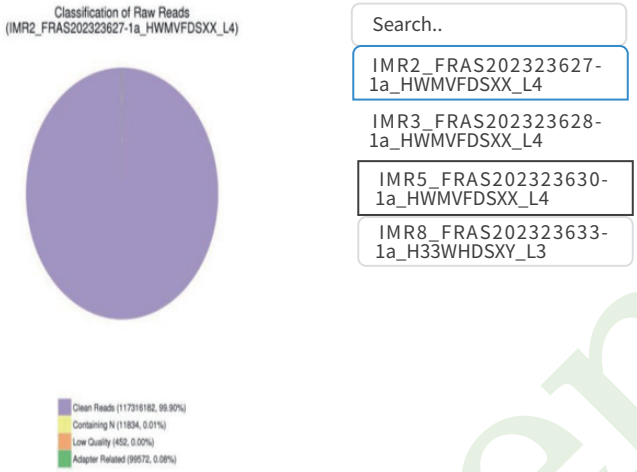


Fig.4 Composition of Raw Data

2 Summary of Sequencing Data Information

The total output of data on the sequencer: Raw data 69.4 G.

The detailed 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(%)
IMR2	FRAS202323627-1a_HWMVFDSXX_L4	117428040	17.6	99.90	0.03
IMR3	FRAS202323628-1a_HWMVFDSXX_L4	110474636	16.6	99.92	0.03
IMR5	FRAS202323630-1a_HWMVFDSXX_L4	124316786	18.6	99.92	0.03
IMR8	FRAS202323633-1a_H33WHDSXY_L3	110753644	16.6	99.92	0.03

Note:

(1)Sample: Sample name.

(2)Library_Flowcell_Lane: The naming format of the raw data files are as follows: Library ID_Flowcell ID_lane ID.

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

(4)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.

(5)Effective Rate (%): (Clean Reads / Raw Reads) * 100% Error Rate: Base Error RateQ20, Q30: (Base Count of Qphred > 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 NAAGAACACGTTCGGTCACCTCAGCACACTTGTGAATGTC ATGGGATCCAT
+ #55???BBBBB?BA@DEEFFCFFHHFFCFFHHHHHHHHFAE0ECFFD/AEHH

Note:
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	Lane Number – Positive integer
1101	Tile Number – 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	Read Number - 1 for single reads; 1 or 2 for paired ends
N	Filtered flag - 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 to allow for data integrity verification. 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 does not 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 can be divided by 4 to obtain the number of reads.
(3)The date size is the space occupied by the data in the hard disk. It is related to the format of disk and compression ratio and has no influence on the quantity of sequenced bases. Consequently, the size of the read1 file may not be equal to the size of the read2 file.
(4)When a large amount of data is needed, e.g. whole genome sequencing data, separate-lane sequencing strategy is often used to ensure the quality of data. As a result, it is 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)The following are details regarding the quality control standard. If clean data delivery is requested, the data will be filtered strictly according to a high standard to obtain high quality clean data which can be used for further research and paper writing. Paired reads will be discarded in the following situations: when either one read contains adapter contamination; when either one read contains uncertain nucleotides more than 10%; when either one read contains more than 50% low quality nucleotides (base quality less than 5). The data analysis results based on this standard can be approved by high level magazines (Yan L.Y. et al . 2013). For more information, please refer to the official website of Novogene (www.novogene.com).
(6)The following are details regarding 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 itis not necessary to dispose the beginning and end of reads in the downstream analysis (e.g. mapping).
(7)Ninety days after the data delivery, the data will be considered outdated and will be deleted, so please store your data carefully and promptly. If you have any questions or doubts, please contact us as soon as possible. Have a nice day!

3 Explanation of Sequencing Data Related

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. Nature Structural & Molecular Biology 20, 1131-1139.