RNA Library Preparation and Sequencing

Quality Control Demo Report

May 1, 2016

Contents

1 Sample Information

Table 1. Sample information

Patient ID	Sample ID	Library ID
XX1270	XX1270	XXX17562
XX1311	XX1311	XXX17612
XX1309	XX1309	XXX17610

2 Experimental Procedure

2.1 RNA Quantification and Qualification

(1) Agarose Gel Electrophoresis: tests RNA degradation and potential contamination.
(2) Nanodrop: tests RNA purity (OD260/OD280).
(3) Agilent 2100: quantifies the RNA and checks RNA integrity.

2.2 Library Preparation for Sequencing

After QC, mRNA of Eukaryote organisms is enriched from total RNA by Oligo (dT) beads and the Ribo-Zero kit is used to remove rRNA from the Prokayote organisms. 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:

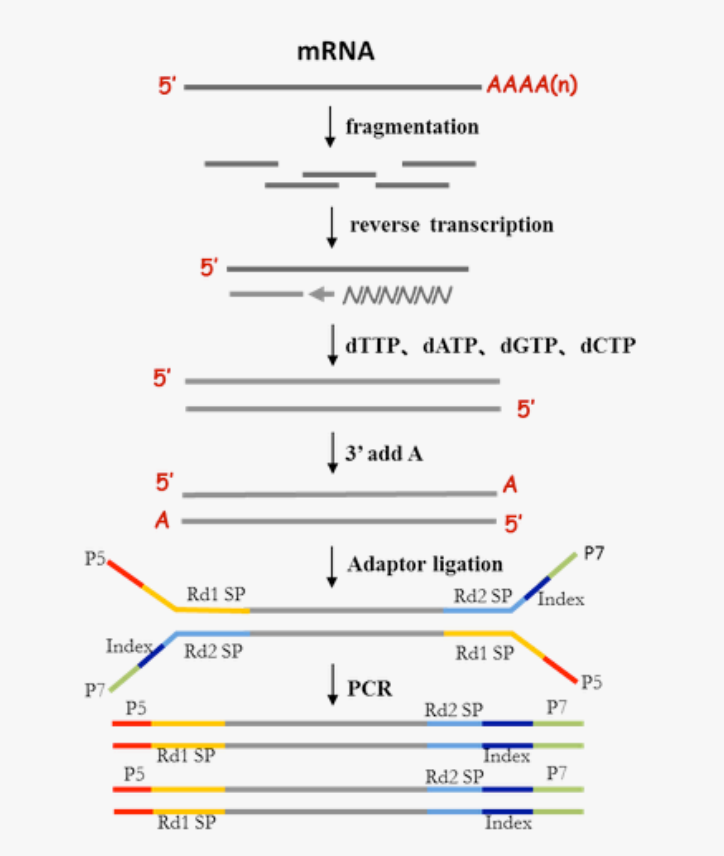

Figure 2.1 Library construction workflow

2.3 Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using TruSeq PE Cluster Kit v4-cBot-HS (Illumia, San Diego, USA) according to the manufacturer's instructions. After cluster generation, the libraries were sequenced on an Illumina sequencing platform.

3 Data Quality Control

3.1 Raw Data

The original raw image data obtained from high throughput sequencing platforms (e.g. Illumina platform) is transformed to sequenced reads by base calling. The sequenced reads are regarded as raw data or raw reads, which is recorded in FASTQ file (fq) containing sequence information (reads) and corresponding sequencing quality information.

Every read in FASTQ format is stored in four lines as follows:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18: ATCACG
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT
```

+

@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF

Line 1 beginning with a '@' character is followed by a sequence identifier and an optional description (like a FASTA title line). Line 2 is the raw sequence reads. 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 sequence in Line 2, and must contain the same number of characters as bases in the sequence.

Table 3.1.1 Illumina sequence identifier details

EAS139	The unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
2104	Tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence

The ASCII value for every character at the fourth line minus 33 will be the corresponding sequencing base quality value at the second line. If the sequencing error rate is recorded by "e" and the base quality for Illumina platform is expressed as Q_{phred}, the equation No.1 as below will be obtained:

Equation 1: $Q_{phred} = -10\log_{10}(e)$

The relationship between sequencing error rate (e) and sequencing base quality value (Q_{phred}) is listed as below (Table 4.2):

Table 3.1.2 Sequencing error rate and corresponding base quality value

Sequencing error rate	Sequencing quality value	Corresponding character
5%	13	.
1%	20	5
0.1%	30	?
0.01%	40	I

3.2 Quality Control

3.2.1 Sequencing Data Filtration

Raw sequencing data may contain adapter contaminated and low-quality reads. These sequence artifacts may increase the complexity of downstream analyses, which means that quality control is an essential step. All the downstream analyses will be based on clean reads that pass quality control.

We performed quality control according to the following procedure:

(1) Discard a read pair if either one read contains adapter contamination;
(2) Discard a read pair if more than 10% of bases are uncertain in either one read;
(3) Discard a read pair if the proportion of low quality bases is over 50% in either one read.

RNA-seq Adapter sequences (Oligonucleotide sequences of adapters from TruSeq™ RNA and DNA Sample Prep Kits):
5' Adapter:
5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
3' Adapter:
5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6-indexs)ATCTCGTATGCCGTCTTCTGCTTG-3′

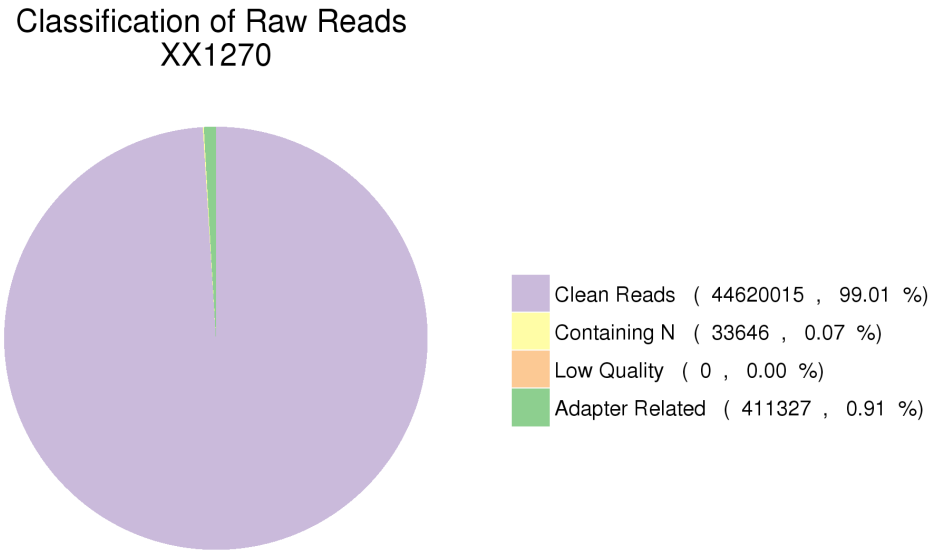

Figure 3.2.1 Raw data filtration result

Note: Reads were discarded in pairs.

(1) Containing N: the number of read pairs with either one read containing uncertain nucleotides more than 10%, and the proportion in raw data.

(2) Low Quality: the number of read pairs with either one read containing low quality (below 5) nucleotides more than 50 percent, and the proportion in raw data.

(3) Adapter related: the number of read pairs filtered out with adapter contamination, and the proportion of filtered read pairs in raw data.

(4) Clean reads: the number of read pairs passed quality control and the proportion in raw data.

3.2.2 Sequencing Error Rate Examination

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).

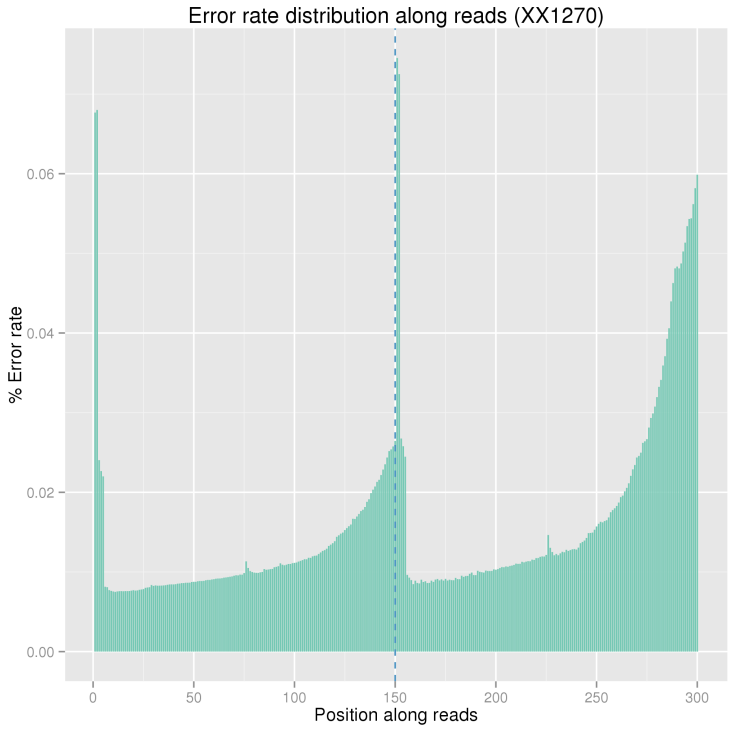

Figure 3.2.2 Sequencing error rate distribution

The x-axis represents position in reads, and the y-axis represents the average error rate of bases of all reads at a position.

3.2.3 Distribution of A/T/G/C Base

GC content distribution is evaluated to detect potential AT/GC separation. 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 phred-scaled quality scores of most bases should be greater than 20, which is required by downstream analyses. It is common to see that base quality decreases along reads, which is an inherent characteristic of next generation sequencing.

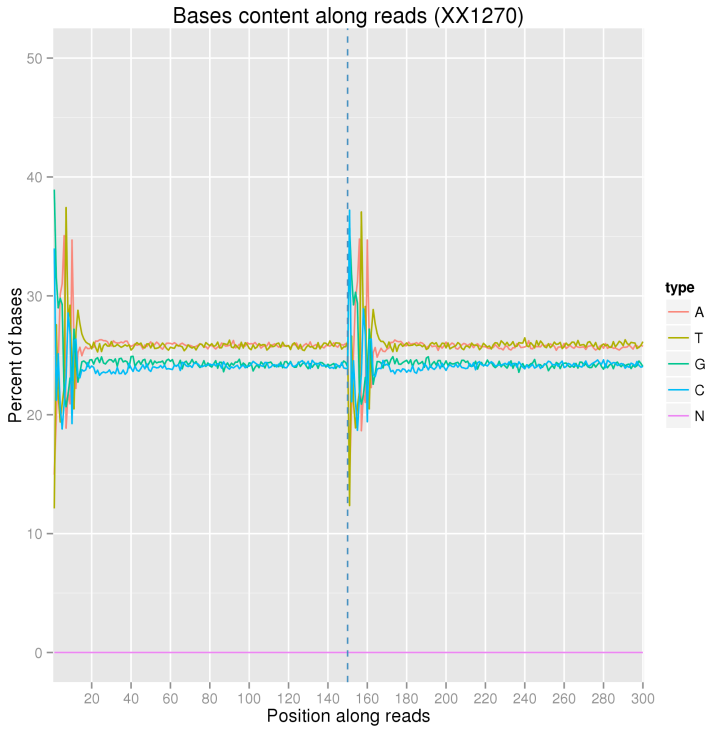

Figure 3.2.3 A/T/G/C Distribution

The x-axis is position in reads, and the y-axis is the single base percentage of all reads at a position.

3.2.4 Statistics of Sequencing Quality

According to the sequencing feature of Illumina platforms, for paired-end sequencing data we require that Q30 (the percent of bases with phred-scaled quality scores greater than 30) should be above 80%.

Table 3.2.4.1 Overview of data production quality

Sample name	Library	Raw Reads	Clean Reads	Raw Base(G)	Clean Base(G)	Effective (%)	Error(%)	Q20(%)	Q30(%)	GC(%)
XX1270	XXX17565	45064988	44620015	13.52	13.39	99.01	0.03	95.5	90.22	48.86
XX1271	XXX17564	38781464	38247266	11.63	11.47	98.62	0.03	96.25	91.55	49.49
XX1275	XXX17569	39416511	39008154	11.82	11.7	98.96	0.03	95.91	90.79	48.54

Note:

(1) Sample name: Sample name.

(2) Library: library name.

(3) Raw Reads: The number of sequencing reads pairs.

(4) Clean Reads: The number of sequencing reads pairs after data filtration.

(5) Raw Base (G): The original sequence data volume.

(6) Clean Base (G): The clean sequence data volume.

(7) Effective (%): The percentage of clean reads in all raw reads.

(8) Error (%): The average error rate of all bases.

(9) Q20: The percentage of bases with Phred score ≥20.

(10) Q30: The percentage of bases with Phred score ≥30.

(11) GC: The percentage of G and C in the total bases.

4 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.