DNA 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
XXX1	XXX1	XXX05644

2 Experimental Procedure

2.1 DNA Quantification and Qualification

(1) DNA degradation and contamination were monitored on 1% agarose gels.
(2) DNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA).
(3) DNA concentration was measured using Qubit® DNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA).
(4) Fragment distribution of DNA library was measured using the DNA Nano 6000 Assay Kit of Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA).

2.2 Library Preparation for Sequencing

A total amount of 1.0μg DNA per sample was used as input material for the DNA sample preparations. Sequencing libraries were generated using Truseq Nano DNA HT Sample Preparation Kit (Illumina USA) following manufacturer's recommendations and index codes were added to each sample. Briefly, the DNA sample was fragmented by sonication to a size of 350bp, and then DNA fragments were endpolished, A-tailed, and ligated with the full-length adapter for Illumina sequencing with further PCR amplification. At last, PCR products were purified (AMPure XP system) and libraries were analyzed for size distribution by Agilent 2100 Bioanalyzer and quantified using real-time PCR.

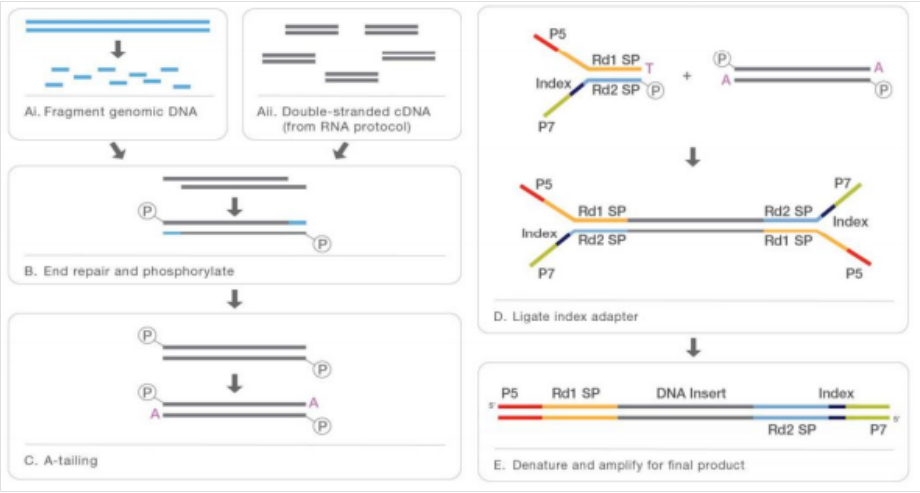

Figure 2.1 Library construction workflow

2.3 Clustering and Sequencing

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using Hiseq X HD PE Cluster Kit (Illumina) according to the manufacturer's instructions. After cluster generation, the libraries were sequenced on an Illumina sequencing platform and paired-end reads were generated.

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.

DNA-Seq Adapter (Adapter, Oligonucleotide sequences for TruSeq™ DNA Sample Prep Kits) information:
5' Adapter:
5′-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3′
3' Adapter:
5′-GATCGGAAGAGCACACGTCTGAACTCCAGTCAC(6-indexs)ATCTCGTATGCCGTCTTCTGCTTG-3′

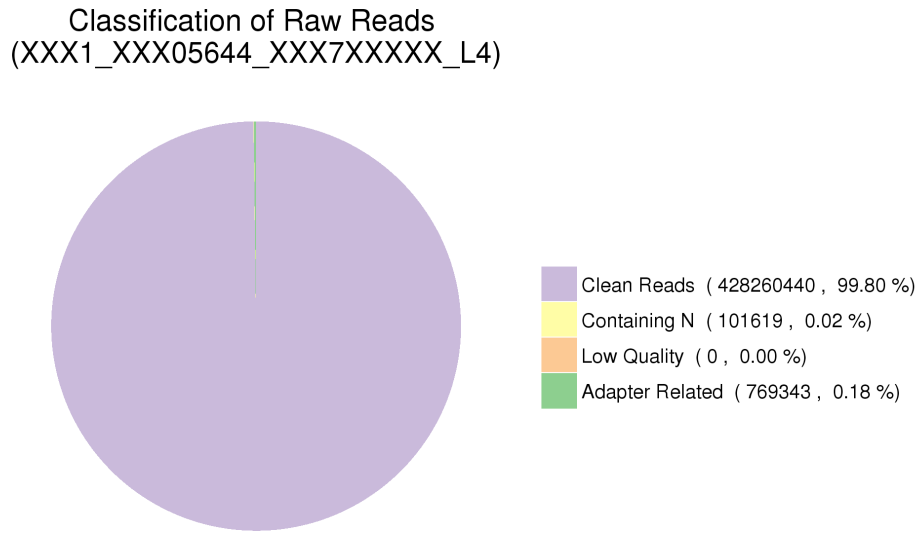

Figure 3.2.1 Raw data filtration result

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

Sequencing error rate and base quality can be affected by various factors such as sequencing platform, chemical reagent and sample quality. Due to the consumption of chemical reagents, error rate is increasing with read extension, which is a common feature of Illumina high throughput sequencing platforms.

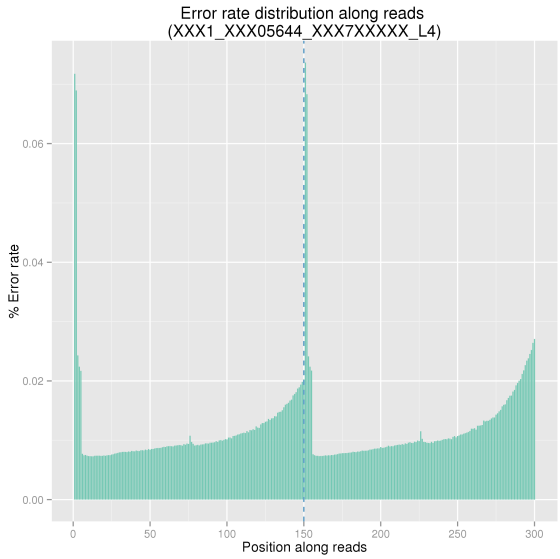

Figure 3.2.2 Sequencing error rate distribution

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 Sequencing Quality Distribution

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.

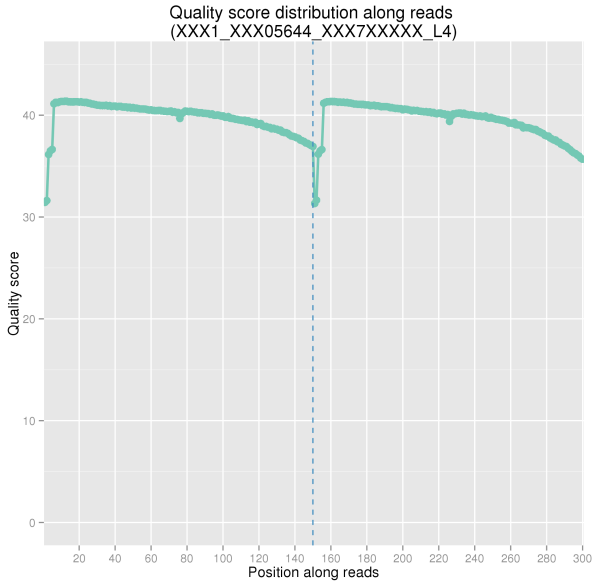

Figure 3.2.3 Sequencing quality distribution

The x-axis is position in reads, and the y-axis is the average quality score of bases 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	Flowcell/ Lane	Raw reads	Raw data(G)	Effective(%)	Error(%)	Q20(%)	Q30(%)	GC(%)
XXX1	XXX05644	XXX7XXXXX_L8	97757508	192.39	99.82	0.01	96.77	92.95	41.03
XXX1	XXX05644	XXX7XXXXX_L6	114403836		99.87	0.01	97.73	94.92	40.85
XXX1	XXX05644	XXX7XXXXX_L4	429131402		99.80	0.01	97.52	94.66	41.20

Note:

(1) Sample name: Sample name.

(2) Library: library name.

(3) Flowcell/Lane: the flowcell ID and lane number.

(4) Raw reads: The number of sequencing reads pairs; four lines will be considered as one unit according to FASTQ format.

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

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

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

(8) Q20: The percentage of bases with Phred score ⩾20.

(9) Q30: The percentage of bases with Phred score ⩾30.

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