Research use only

Raw Data

Report

August 2019

Humanizing Genomics
macrogen

Project Information

Client Name	Yale Passamaneck
Company/Institution	US Bureau of Reclamation
Order Number	1907UNHS-0108
Type of Read	paired-end
Read Length	151
Number of Samples	6
Library Kit	TruSeq stranded mRNA
Library Protocol	TruSeq Stranded mRNA LT Sample Prep Kit/TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E
Type of Sequencer	Illumina platform

Table of Contents

1. Experimental Methods and Workflow

1. 1. Experiment overview

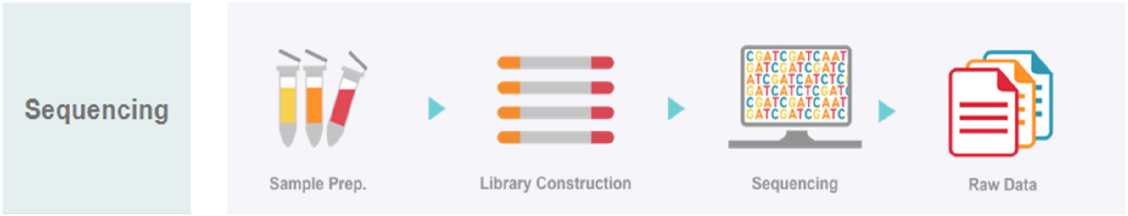

Fig1. Experiment overview

The Illumina NGS workflows include 4 basic steps :

1) Sample Preparation

For library construction, DNA/RNA is extracted from a sample. After performing quality control(QC), qualified samples proceed to library construction.

2) Library Construction

The sequencing library is prepared by random fragmentation of the DNA or cDNA sample, followed by 5' and 3' adapter ligation. Alternatively, "tagmentation" combines the fragmentation and ligation reactions into a single step that greatly increases the efficiency of the library preparation process. Adapter-ligated fragments are then PCR amplified and gel purified.

3) Sequencing

For cluster generation, the library is loaded into a flow cell where fragments are captured on a lawn of surface-bound oligos complementary to the library adapters. Each fragment is then amplified into distinct, clonal clusters through bridge amplification. When cluster generation is complete, the templates are ready for sequencing.

Illumina SBS technology utilizes a proprietary reversible terminator-based method that detects single bases as they are incorporated into DNA template strands. As all 4 reversible, terminator-bound dNTPs are persent during each sequencing cycle, natural competition minimizes incorporation bias and greatly reduces raw error rates compared to other technologies. The result is highly accurate base-by-base sequencing that virtually eliminates sequence-context-specific errors, even within repetitive sequence regions and homopolymers.

4) Raw Data

Sequencing data is converted into raw data for the analysis.

1. 2. Generation of Raw Data

The Illumina sequencer generates raw images utilizing sequencing control software for system control and base calling through an integrated primary analysis software called RTA (Real Time Analysis). The BCL (base calls) binary is converted into FASTQ utilizing illumina package bcl2fastq. Adapters are not trimmed away from the reads.

2. Summary of Data Production

2. 1. Raw data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for the 6 samples. For example, in Drb006-Foot, 255,723,784 reads are produced, and total read bases are 38.6G bp. The GC content (%) is 45.54% and Q30 is 89.95%.

LINK 1907UNHS-0108.xlsx : Download

Table 1. Raw data Stats

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
Drb006-Foot	38,614,291,384	255,723,784	45.54	54.46	95.77	89.95
Drb023-Testes	30,130,501,646	199,539,746	47.69	52.31	97.34	92.91
Drb029-Gill	29,988,592,450	198,599,950	45.79	54.21	96.8	91.75
Drb029-Ovary	32,371,041,760	214,377,760	47.83	52.17	96.77	91.76
Drb-Whole2	30,625,833,892	202,820,092	48.24	51.76	97.01	92.2
Drb-Whole3	30,239,464,152	200,261,352	48.46	51.54	97.15	92.53

- Sample ID : Sample name.
- Total read bases : Total number of bases sequenced.
- Total reads : Total number of reads. For Illumina paired-end sequencing, this value refers to the sum of read1 and read2.
- GC(%) : GC content.
- AT(%) : AT content.
- Q20(%) : Ratio of bases that have phred quality score greater than or equal to 20.
- Q30(%) : Ratio of bases that have phred quality score greater than or equal to 30.

2. 2. Total Read Bases

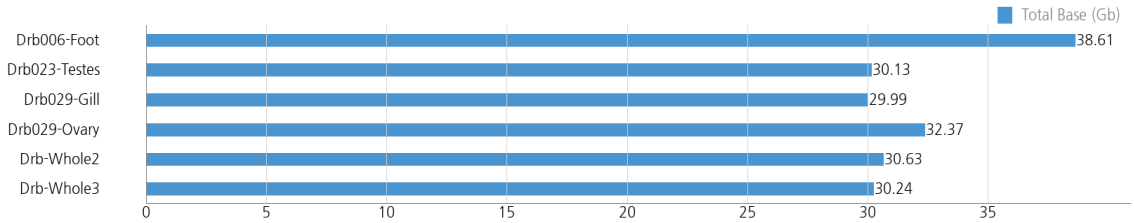

Figure 2.Throughput of Raw data

2. 3. Total Reads

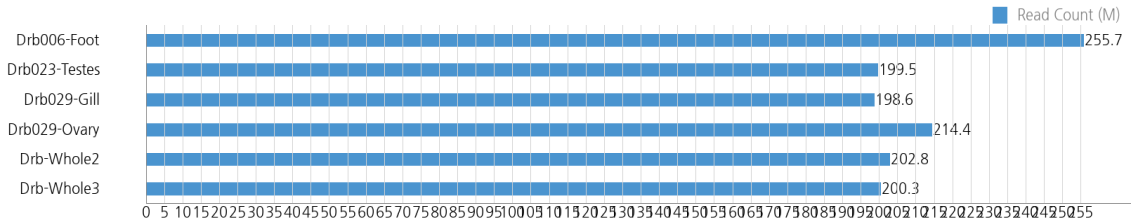

Figure 3. Total read count of Raw data

2. 4. GC/AT Content

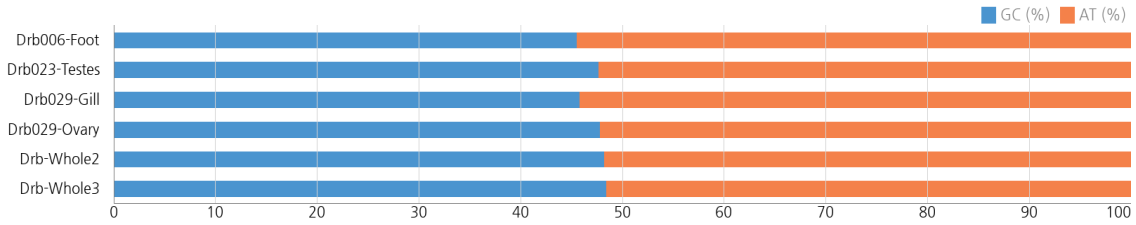

Figure 4. GC/AT Content of Raw data

2. 5. Q20/Q30 (%)

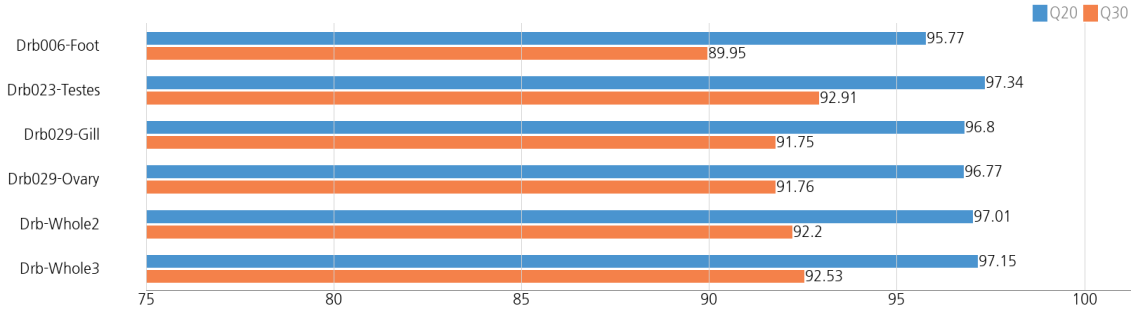

Figure 5. Q20/Q30 scores of Raw data

3. Data Download Information

3. 1. Raw Data and Analysis results

LINK 1907UNHS-0108.xlsx : Download

Download link	File size	md5sum
Drb006-Foot_1.fastq.gz	9.2G	eb1d7ed302e40356a8299ba0b67e9bf6
Drb006-Foot_2.fastq.gz	9.4G	84da188a44608b3f74155e2ddf4c73cd
Drb023-Testes_1.fastq.gz	7.5G	c46511ca022fa988c12370aae40bd66f
Drb023-Testes_2.fastq.gz	7.7G	94b09606a855aab1134a136c68b3aa25
Drb029-Gill_1.fastq.gz	7.6G	8bece4432191b874b0fcda4a7a514268
Drb029-Gill_2.fastq.gz	7.9G	7633bb11da08c259a33139eaf5365515
Drb029-Ovary_1.fastq.gz	8.1G	15d258e45eb2f252b0f5cc01a90ce2a5
Drb029-Ovary_2.fastq.gz	8.3G	41aa0402c461c6dd5637859efe6da4c8
Drb-Whole2_1.fastq.gz	7.7G	e832d504addb5fe299c1241e6aae5724
Drb-Whole2_2.fastq.gz	8.0G	e3b00c8a07222d0e2bdaaf9432941b40
Drb-Whole3_1.fastq.gz	7.6G	0771ff72ef3f7c1f5e64d4d59e9e038a
Drb-Whole3_2.fastq.gz	7.7G	1c3452cc51aa60006e901f08bef04b98

- fastq.gz : This is a zip file of raw data used in analysis.
- md5sum : In order to verify the integrity of files, md5sum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

The data retention period is three months. Please email (ngssales@macrogenlab.com) or contact our sales representative for longer retention period.

4. Appendix

4. 1. FAQ

Q: I want to see the produced data. How can I open those files?

A: Large volume zip file that is provided by our compnay is not user-friendly in Windows environment, so it is recommended to use linux environment for smooth operation.

4. 2. FASTQ File

Example of FASTQ

```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNNTNNNNNNNNNNNN
+
@@@BDDDDHHHHFHIIIIIIII#3AC##############################
```

FASTQ file is composed of four lines.
Line 1 : ID line includes information such as flow cell lane information.
Line 2 : Sequences line.
Line 3 : Separator line (+ mark).
Line 4 : Quality values line about sequences.

4. 3. Phred Quality Score Chart

Phred quality score numerically expresses the accuracy of each nucleotide. Higher Q number signifies higher accuracy. For example, if Phred assigns a quality score of 30 to a base, the chances of having base call error are 1 in 1000.

Phred Quality Score Q is calculated with $-10\log_{10}P$, where P is probability of erroneous base call.

Quality of phred score	Probability of incorrect base call	Base call accuracy	Characters
10	1 in 10	90%	!"#$%&'()*+
20	1 in 100	99%	,-./012345
30	1 in 1000	99.9%	6789:;h=i?
40	1 in 10000	99.99%	@ABCDEFGHIJ

- Encoding : Sanger Quality (ASCII Character Code=Phred Quality Value + 33)

Macrogen Corp. | 1330 Piccard Drive Suite 103 Rockville, MD 20850

Contact us
Tel: +1-301-251-1007
www.macrogenlab.com | ngssales@macrogenlab.com