Research use only

Raw Data

Report

September 2021

Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen Europe
Company / Institution	Macrogen Europe
Order Number	HN00155645
Type of Read	Paired-end
Read Length	151
Number of Samples	3
Type of Sequencer	Illumina platform

Table of Contents

1. Data Download Information

1. 1. Raw Data and Analysis Results

Download link	File size	md5sum
DN1_KO_rep1_1.fastq.gz	1.0G	c2095e9bc993755383be9247d830e2e3
DN1_KO_rep1_2.fastq.gz	991.1M	8d654303f1407119a14ccf671cd4d375
DN1_WT_rep1_1.fastq.gz	347.0M	31260d4e330390984afb1aabf1f39e60
DN1_WT_rep1_2.fastq.gz	322.8M	ff40f5e78e477fc6632206f1cc6ece8f
DN1_WT_rep2_1.fastq.gz	295.6M	339faea0788651328629e0dd3722b12f
DN1_WT_rep2_2.fastq.gz	283.9M	ee417d9b45ea135496b10b6edebf49ce

- 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.

Your data will be retained in our server for 3 months. Should you wish to extend the retention period, please contact us.

2. Experimental Methods and Workflow

2. 1. Experiment Overview

Fig1. Experiment overview

The Illumina NGS workflow includes 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.

2. 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.

3. Summary of Produced Data

3. 1. Raw Data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for the 3 samples. For example, in DN1_KO_rep1, 33,323,730 reads are produced, and total read bases are 5.0G bp. The GC content (%) is 51.04% and Q30 is 89.32%.

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
DN1_KO_rep1	5,031,883,230	33,323,730	51.04	48.96	94.66	89.32
DN1_WT_rep1	1,474,346,182	9,763,882	51.46	48.54	94.92	89.57
DN1_WT_rep2	1,013,035,444	6,708,844	51.28	48.72	89.67	77.98

- 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 read 1 and read 2.
- GC(%) : GC content.
- AT(%) : AT content.
- Q20(%) : Ratio of bases that have phred quality score of over 20.
- Q30(%) : Ratio of bases that have phred quality score of over 30.

3. 2. Total Read Bases

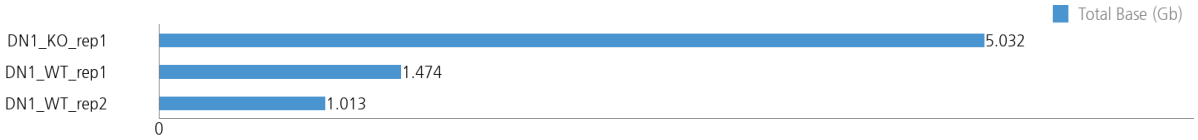

Figure 2.Throughput of Raw data

3. 3. Total Reads

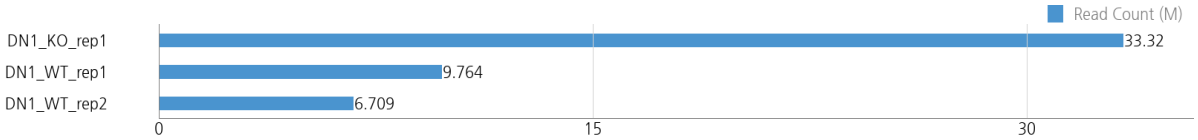

Figure 3. Total read count of Raw data

3. 4. GC/AT Content

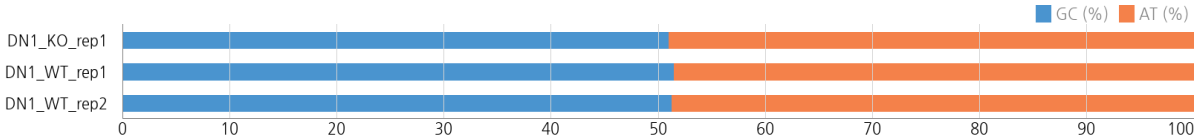

Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

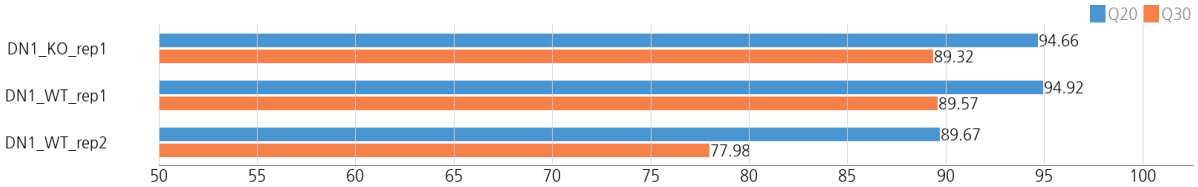

Figure 5. Q20/Q30 scores of Raw data

4. Appendix

4. 1. FAQ

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

A: As the large size zip files provided by our company are hard to process in the Windows environment, we highly recommend using Linux environment for a smoother 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)

Research use only

HEADQUARTER

Macrogen, Inc.

Laboratory, IT and Business Headquarter & Support Center

[08511] 1001, 10F, 254, Beotkkot-ro, Geumcheon-gu, Seoul, Republic of Korea (Gasan-dong, World Meridian 1)
Tel: +82-2-2180-7000
Email1: ngs@macrogen.com(Overseas)
Email2: ngskr@macrogen.com (Republic of Korea)
Web: www.macrogen.com
LIMS: dna.macrogen.com

SUBSIDIARY

Macrogen Europe

Laboratory, Business & Support Center

Meibergdreef 57, 1105 BA, Amsterdam, the Netherlands
Tel: +31-20-333-7563
Email: ngs@macrogen.eu

Macrogen Singapore

Laboratory, Business & Support Center

3 Biopolis Drive #05-18, Synapse, Singapore 138623
Tel: +65-6339-0927
Email: info-sg@macrogen.com

Psomagen (Macrogen USA)

Laboratory, Business & Support Center

1330 Piccard Drive, Suite 103, Rockville, MD 20850, United States
Tel: +1-301-251-1007
Email: inquiry@psomagen.com

Macrogen Japan

Laboratory, Business & Support Center

16F Time24 Building, 2-4-32 Aomi, Koto-ku, Tokyo 135-0064 JAPAN
Tel: +81-3-5962-1124
Email: ngs@macrogen-japan.co.jp

BRANCH

Macrogen Spain

Laboratory, Business & Support Center

Av. Sur del Aeropuerto de Barajas, 28. Office B-2, 28042 Madrid, Spain
Tel: +34-911-138-378
Email: info-spain@macrogen.com

Humanizing Genomics
macrogen