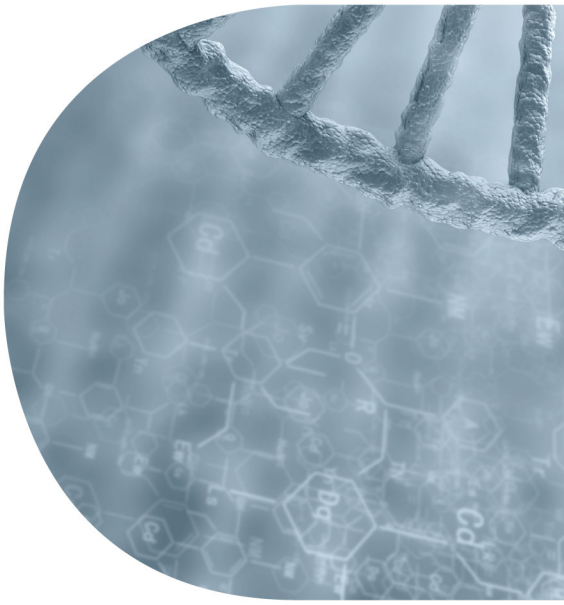
Research use only

Raw Data

Report

November 2020



Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen Oceania PL
Company / Institution	Macrogen Oceania PL
Order Number	HN00137457
Type of Read	Paired-end
Read Length	151
Number of Samples	12
Library Kit	TruSeq Stranded mRNA LT Sample Prep Kit
Library Protocol	TruSeq Stranded mRNA Sample Preparation Guide, Part # 15031047 Rev. E
Type of Sequencer	Illumina platform

Table of Contents

1. Data Download Information

1. 1. Raw Data and Analysis Results

Download link	File size	md5sum
D00GiR1FID001_1.fastq.gz	1.4G	7f0285263bcc8d9a3c0ec21133bd7445
D00GiR1FID001_2.fastq.gz	1.5G	f9be05f5e2821010eef98fb2bc2b111f
D00GiR2FID004_1.fastq.gz	1.6G	222f64c39b2e0201408b005b74b42889
D00GiR2FID004_2.fastq.gz	1.7G	8ad19b2bd131e528d0a1eca313b566b6
D00GiR3FID007_1.fastq.gz	1.8G	2fb8784ecdbe0b5ebb47e9161493e78f
D00GiR3FID007_2.fastq.gz	1.8G	40625c80caad4e8bfc7e86f5bb4a7c7b
D08GiR1FID064_1.fastq.gz	1.7G	62b5b37d98b718137cefa6abd4e449af
D08GiR1FID064_2.fastq.gz	1.7G	1ff13d2370cd7bc43b3a05b7bf36e8df
D08GiR2FID067_1.fastq.gz	1.5G	41b7b2c6b2e1a05e7b6c5cfd9e094d4b
D08GiR2FID067_2.fastq.gz	1.5G	cfd469d49fde2adb434e663d6683fb80
D08GiR3FID070_1.fastq.gz	1.6G	ceeb0ee0b3d919edf86648aaa2743848
D08GiR3FID070_2.fastq.gz	1.7G	ba3e0fe3cec37b4b4b520dc018158e5c
D12GiR1FID085_1.fastq.gz	1.7G	0f6530fe35ec68d67f1b0f5129323792
D12GiR1FID085_2.fastq.gz	1.8G	fb97b6736b898e25ad9c9fbf60a266c9
D12GiR2FID088_1.fastq.gz	1.7G	ad161cc58a981d39ece478dee68dee16
D12GiR2FID088_2.fastq.gz	1.7G	401204034212c05d53f8d2ce864a2cb6
D12GiR3FID091_1.fastq.gz	1.4G	cf6de22af124021999ab40e886327579
D12GiR3FID091_2.fastq.gz	1.5G	a1ef11394eef7db1402662e9be8f5bc2
D19GiR1FID163_1.fastq.gz	1.5G	8ecdf7b8ebb8e71d6be3706f0b1235b3
D19GiR1FID163_2.fastq.gz	1.6G	7608ab9b1a4a4f72a1f9da633b1954d5
D19GiR2FID166_1.fastq.gz	1.9G	0a8b4daa62daa5d1c18d4d2baa64d97a
D19GiR2FID166_2.fastq.gz	2.0G	5884f3986f2e2027fa2a2cc84a35e8a0
D19GiR3FID169_1.fastq.gz	1.5G	77ec65ead6d37f5190dc8b1bc9dbe625
D19GiR3FID169_2.fastq.gz	1.6G	6da01d2cfd864cedc9c2694314551321

- 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 12 samples. For example, in D00GiR1FID001, 40,945,514 reads are produced, and total read bases are 6.2G bp. The GC content (%) is 50.0% and Q30 is 94.84%.

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
D00GiR1FID001	6,182,772,614	40,945,514	50.0	50.0	98.18	94.84
D00GiR2FID004	7,235,868,358	47,919,658	50.22	49.78	98.09	94.75
D00GiR3FID007	7,799,270,196	51,650,796	50.71	49.29	98.32	95.22
D08GiR1FID064	7,298,122,336	48,331,936	50.1	49.9	98.25	95.08
D08GiR2FID067	6,401,095,662	42,391,362	49.81	50.19	98.2	94.94
D08GiR3FID070	7,195,104,398	47,649,698	50.47	49.53	98.18	94.96
D12GiR1FID085	7,644,444,460	50,625,460	49.22	50.78	98.22	94.97
D12GiR2FID088	7,432,333,854	49,220,754	50.75	49.25	98.27	95.16
D12GiR3FID091	6,322,608,278	41,871,578	49.22	50.78	98.27	95.07
D19GiR1FID163	6,629,932,840	43,906,840	49.3	50.7	98.22	94.99
D19GiR2FID166	8,430,189,268	55,829,068	50.21	49.79	98.16	94.92
D19GiR3FID169	6,791,657,464	44,977,864	49.05	50.95	98.15	94.78

- 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

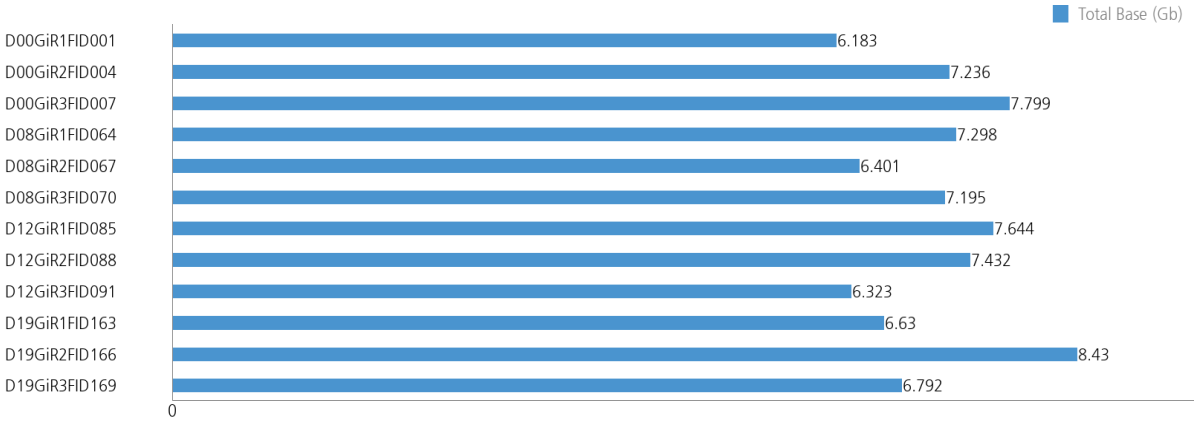


Figure 2.Throughput of Raw data

3. 3. Total Reads

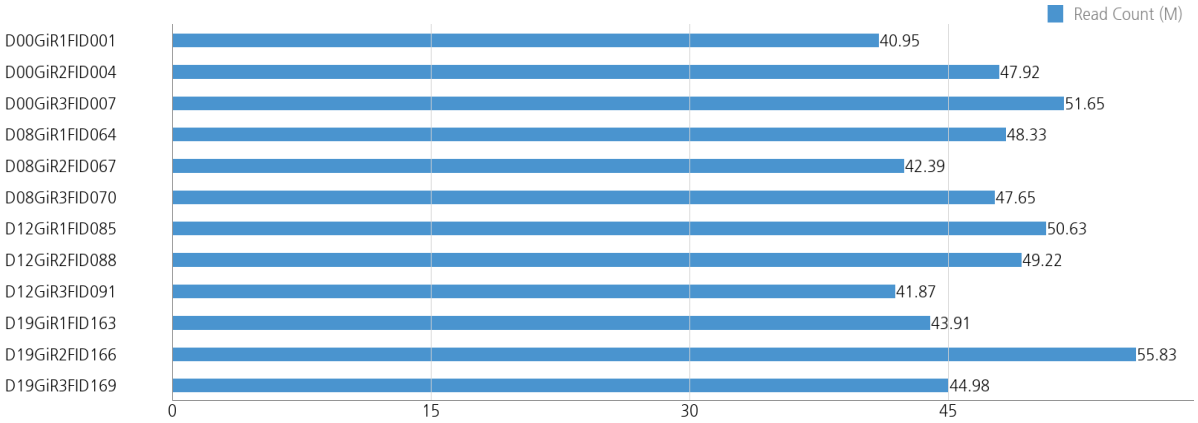


Figure 3. Total read count of Raw data

3. 4. GC/AT Content

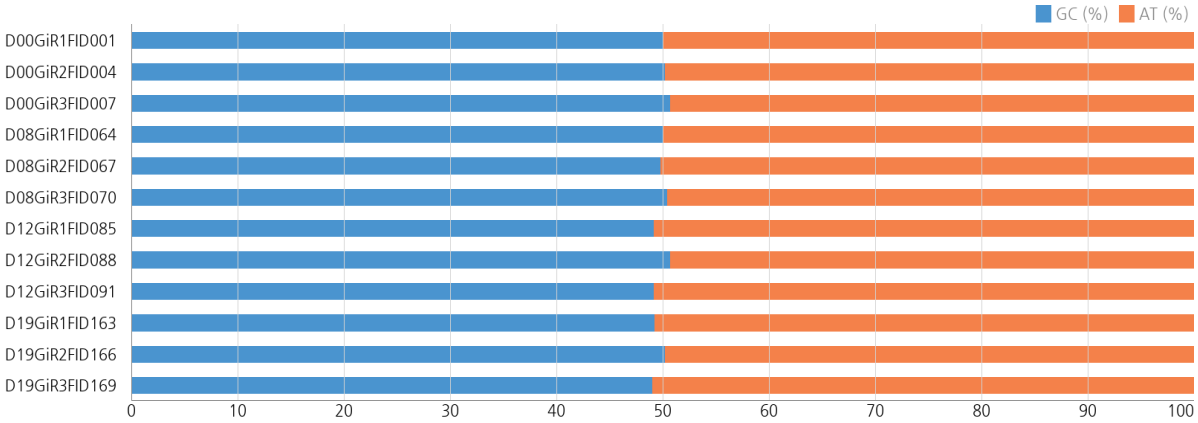


Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

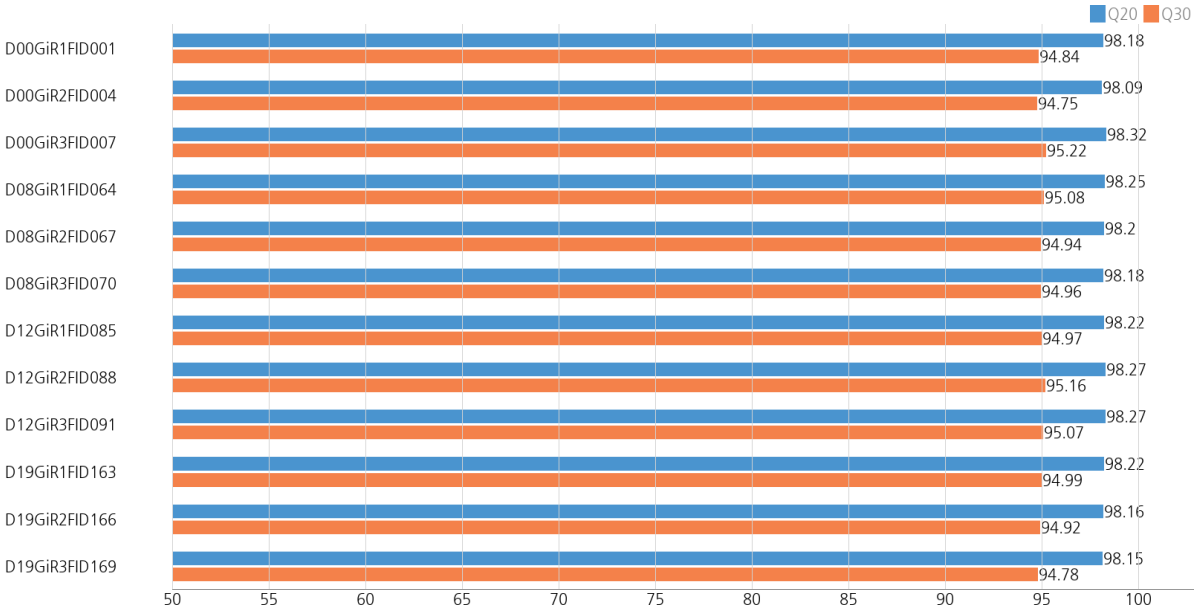


Figure 5. Q20/Q30 scores of Raw data

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 31, 1105 AZ, 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