Research use only

Raw Data

Report

March 2021

Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen Europe
Company / Institution	Macrogen Europe
Order Number	HN00141902
Type of Read	Paired-end
Read Length	101
Number of Samples	9
Type of Sequencer	Illumina platform

Table of Contents

1. Data Download Information

1. 1. Raw Data and Analysis Results

Download link	File size	md5sum
F1_1.fastq.gz	791.8M	80343408e46268f5d430cb3cc9a84f8b
F1_2.fastq.gz	823.0M	87f22be6946804ec93953197375e04cb
F2_1.fastq.gz	869.7M	0c6bae6cbe80ca20434218b559c83a37
F2_2.fastq.gz	881.0M	be27a964d5e81273c3648958e1008881
F3_1.fastq.gz	954.0M	39ad5485881670ed0fb2a3f7d707c407
F3_2.fastq.gz	973.2M	2dc5adf7daf6682cec0ad8fc402be027
P1_1.fastq.gz	728.6M	430be2bc5feaae995a0d8f742014ea9b
P1_2.fastq.gz	751.2M	536a5be78121659419a9d947663388b5
P2_1.fastq.gz	757.8M	39cf2e1ff56ef87e2b99dd9e0dc3caf3
P2_2.fastq.gz	790.0M	7b60dd3cbd5247b5956e35b327b8e4f0
P3_1.fastq.gz	755.7M	3ef090459d5ec4e59eb3c9179eb35899
P3_2.fastq.gz	770.8M	67301223e0093a01cc6fed32b17e0064
Py1_1.fastq.gz	828.7M	15900cfe5e2bfa852fea2f4175e7eea3
Py1_2.fastq.gz	850.8M	de7f4c92399652fd7cf3acc8b950ec0c
Py2_1.fastq.gz	1009.2M	d70b47e51b3347ccd0b31775a059df93
Py2_2.fastq.gz	1.0G	a8de50f3a59f96e05bdbf53557b41fce
Py3_1.fastq.gz	901.9M	cfa962982661d07845634a45543d6cf7
Py3_2.fastq.gz	921.8M	8ec336afb95ca47a5370b7717fc00a6a

- 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 9 samples. For example, in F1, 39,596,138 reads are produced, and total read bases are 4.0G bp. The GC content (%) is 56.77% and Q30 is 96.53%.

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
F1	3,999,209,938	39,596,138	56.77	43.23	98.91	96.53
F2	4,216,483,158	41,747,358	56.95	43.05	98.97	96.7
F3	4,285,567,360	42,431,360	57.72	42.28	98.94	96.65
P1	4,014,333,678	39,745,878	59.46	40.54	98.92	96.61
P2	3,886,453,336	38,479,736	58.93	41.07	98.9	96.55
P3	4,126,939,386	40,860,786	58.38	41.62	98.95	96.68
Py1	4,001,917,344	39,622,944	57.48	42.52	99.04	96.83
Py2	4,458,612,882	44,144,682	59.52	40.48	98.93	96.62
Py3	4,271,503,514	42,292,114	57.65	42.35	98.98	96.73

- 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

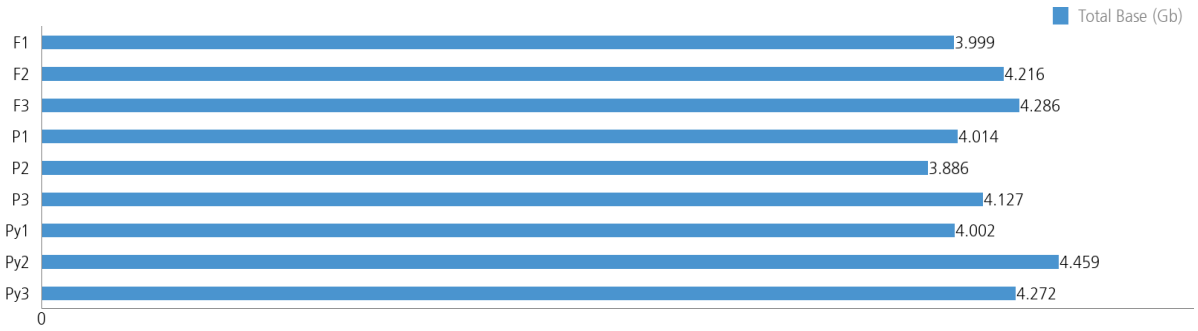

Figure 2.Throughput of Raw data

3. 3. Total Reads

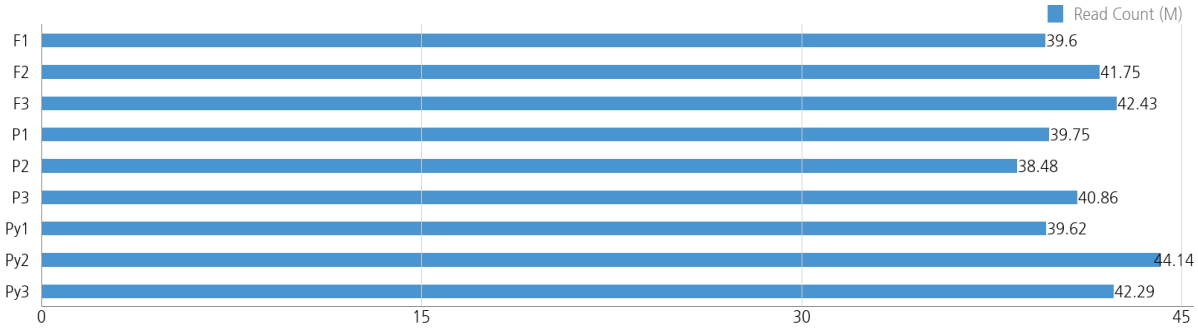

Figure 3. Total read count of Raw data

3. 4. GC/AT Content

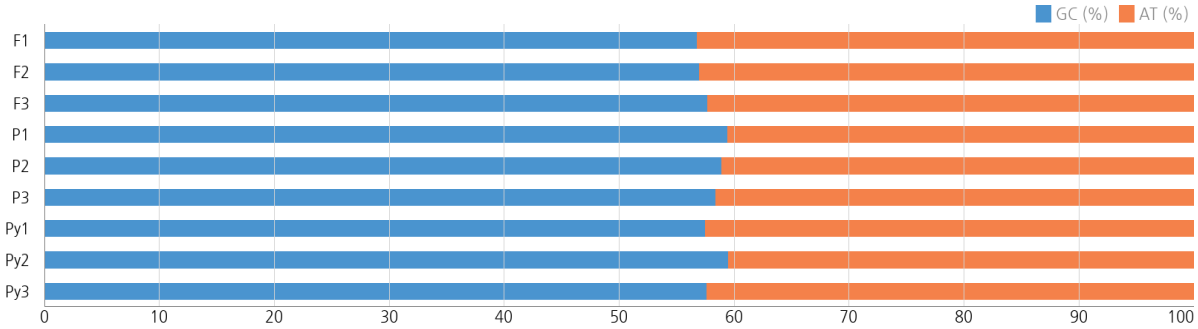

Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

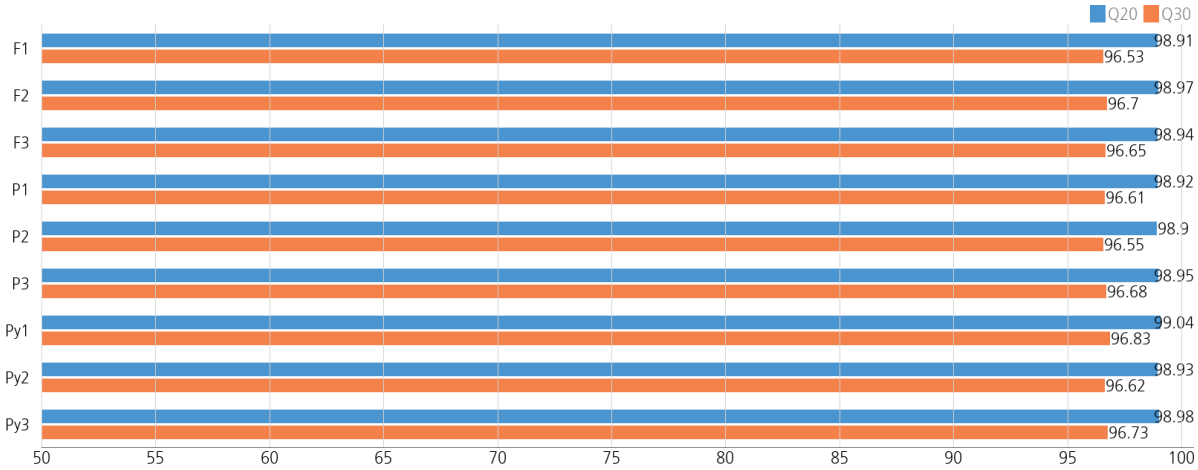

Figure 5. Q20/Q30 scores of Raw data

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