Research use only

Raw Data

Report

March 2020

Humanizing Genomics
macrogen

Project Information

Client Name	Macrogen
Company / Institution	Macrogen
Order Number	HN12345678
Type of Read	Paired-end
Read Length	151
Number of Samples	22
Type of Sequencer	Illumina platform

Table of Contents

1. Data Download Information

1. 1. Raw Data and Analysis Results

Download link	File size	md5sum
Sample1_1.fastq.gz	21.6M	1bce36b41e9d3b485fc6c186ffd31660
Sample1_2.fastq.gz	23.2M	6782ed3061181c1f00a3b262e6d57547
Sample2_1.fastq.gz	24.7M	c84f191fa0443489fc8bfa1970bdaf17
Sample2_2.fastq.gz	26.5M	8be69abb07155054ff53dc4329d4a0ac
Sample3_1.fastq.gz	31.7M	46f632f96d53d1c2b4769c3de337daf9
Sample3_2.fastq.gz	33.9M	41217864402814e9510c96077843a6d4
Sample4_1.fastq.gz	27.7M	8540fa0a98f81575a3b463274773876f
Sample4_2.fastq.gz	29.7M	9e81f2db2d3824982fc3fadac84837d1
Sample5_1.fastq.gz	23.0M	a476fcf20214f92d03952c4bb3f74b99
Sample5_2.fastq.gz	24.6M	6e2ba3451696e31a4fca5464a5cd6c7c
Sample6_1.fastq.gz	28.6M	b80178292d443af29638ed6f1f4cc401
Sample6_2.fastq.gz	30.7M	91c1695ce85bffc83abb03952458c0b5
Sample7_1.fastq.gz	29.3M	dbb2ef7426b9af7f362ed7e9986dd13c
Sample7_2.fastq.gz	31.4M	fef8c85a4bd7c362489f8f3ade609585
Sample8_1.fastq.gz	26.3M	9ebcbe9504910d78e9e5989ea4a6a7cb
Sample8_2.fastq.gz	28.2M	fdb340cb13261372aab9350edddb38e0
Sample9_1.fastq.gz	22.1M	0f58be0fc036decf200f45c651aacdb2
Sample9_2.fastq.gz	23.7M	1fcfc59df41490d9376078d6c8c6a155
Sample10_1.fastq.gz	28.5M	c90588e883138f631ba05bed295ad58f
Sample10_2.fastq.gz	30.6M	33ceee5a32ee5b0ee8738eefb4c3fd1e
Sample11_1.fastq.gz	28.1M	dc62b10922b2d77716235a511897074e
Sample11_2.fastq.gz	30.1M	3335db8e01287f93688fcffd34d75caf
Sample12_1.fastq.gz	20.9M	2d8e2eca305cc444599a566404c2aff6
Sample12_2.fastq.gz	22.4M	2f7c6f9c8ebbdb011f80fb8c475378cf
Sample13_1.fastq.gz	31.5M	473d8a80948d484e694d35a2b326db3d
Sample13_2.fastq.gz	33.7M	c3030cf4f83e445f3e528d6490479709
Sample14_1.fastq.gz	30.9M	51ed8166534aeb4c357b9f14732d70fc
Sample14_2.fastq.gz	33.1M	9a8b3547bcda6fe5a2a7cb0093bed964
Sample15_1.fastq.gz	33.5M	5bed80fe97994a5ea238fa6142bec3d6
Sample15_2.fastq.gz	35.9M	a0534cd0eac46b263394fc323f022843
Sample16_1.fastq.gz	34.2M	701f2e1deeb4c3daaea1f9c724b1e5c5
Sample16_2.fastq.gz	36.7M	496215f1049ef5046a9abc713946a7f8
Sample17_1.fastq.gz	23.1M	abd310b82c4f8eeb432c2918aab30795

Sample17_2.fastq.gz	24.8M	030aa31784f005f84f51731e93b6ea21
Sample18_1.fastq.gz	24.6M	6e8c7da642a26694b1c21e52b4345778
Sample18_2.fastq.gz	26.4M	9e2806a4c7af43d7bd021642cb26b522
Sample19_1.fastq.gz	27.9M	46e8d08bf7a94cf2216b6c0ab9815e3d
Sample19_2.fastq.gz	29.9M	fa5f655157ccf454ba859480cac51e11
Sample20_1.fastq.gz	26.3M	43435fe51891445a274450e73a169138
Sample20_2.fastq.gz	28.2M	062bb50278764c8b6d90a4b3d7d7ab6e
Sample21_1.fastq.gz	22.0M	16aa4251135e837d6bd104b35d54f300
Sample21_2.fastq.gz	23.6M	525dbc44d21df07705964620ec1539f3
Sample22_1.fastq.gz	27.9M	045c21ef8134b3e8e22c47e5b558809f
Sample22_2.fastq.gz	29.9M	e4e6095e430979f515b658960bf24ffa

- 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 email (ngskr@macrogen.com) or contact our sales team.

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 20 samples. For example, in Sample1, 429,499 reads are produced, and total read bases are 64.9M bp. The GC content (%) is 51.26% and Q30 is 87.73%.

✔ **The following table only shows maximum of 20 samples. If your samples are more than 20, please refer to the attached excel file. View full table** : HN12345678_RawData_Stat.xlsx

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
Sample1	64,854,349	429,499	51.26	48.74	93.10	87.73
Sample2	74,011,140	490,140	48.40	51.6	96.05	85.12
Sample3	94,852,764	628,164	53.84	46.16	95.46	89.01
Sample4	83,091,374	550,274	46.67	53.33	95.82	89.04
Sample5	68,805,566	455,666	54.93	45.07	91.76	86.68
Sample6	85,718,019	567,669	46.35	53.65	91.70	86.82
Sample7	87,796,081	581,431	49.52	50.48	92.19	89.71
Sample8	78,744,235	521,485	46.29	53.71	93.02	86.93
Sample9	66,242,039	438,689	50.25	49.75	94.76	85.70
Sample10	85,478,231	566,081	54.40	45.6	92.57	87.17
Sample11	84,269,778	558,078	46.12	53.88	94.79	86.85
Sample12	62,705,468	415,268	52.97	47.03	91.56	85.99
Sample13	94,297,235	624,485	51.81	48.19	91.22	84.93
Sample14	92,475,269	612,419	51.59	48.41	92.72	85.72
Sample15	100,317,756	664,356	52.94	47.06	95.92	85.39
Sample16	102,566,901	679,251	47.62	52.38	97.31	88.60
Sample17	69,333,915	459,165	47.63	52.37	95.04	85.04
Sample18	73,705,818	488,118	52.46	47.54	92.11	89.82
Sample19	83,576,688	553,488	52.00	48.0	97.43	84.14
Sample20	78,758,429	521,579	48.15	51.85	94.51	86.45

- 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

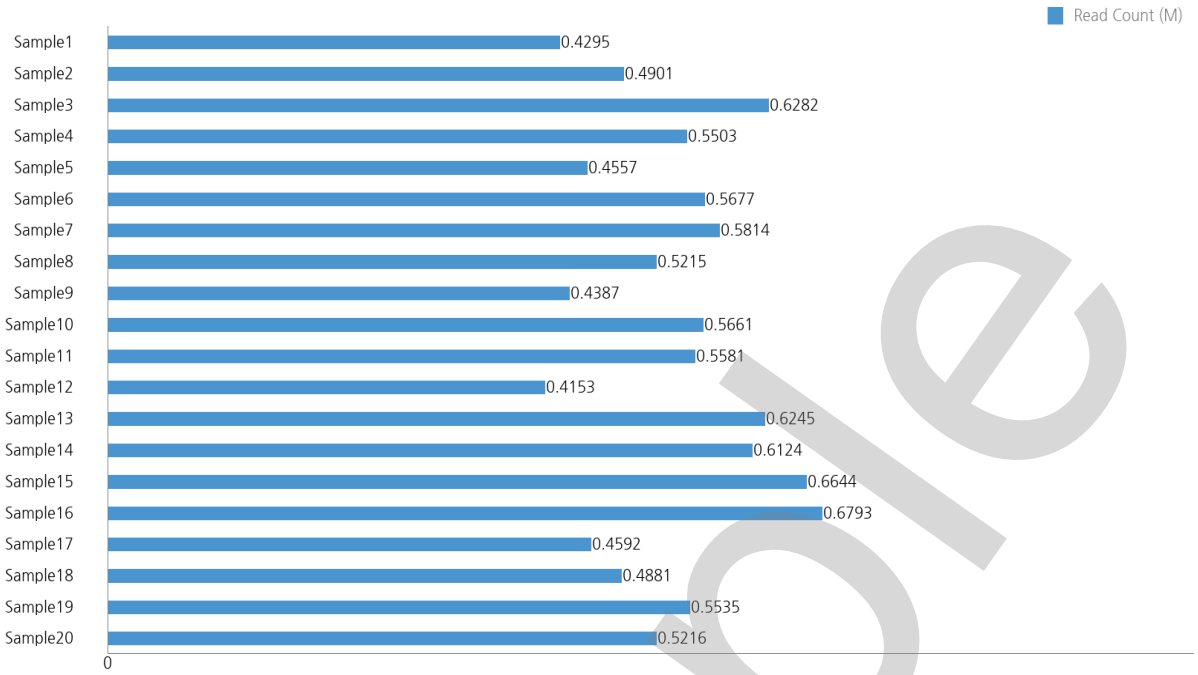

Figure 3. Total read count of Raw data

3. 4. GC/AT Content

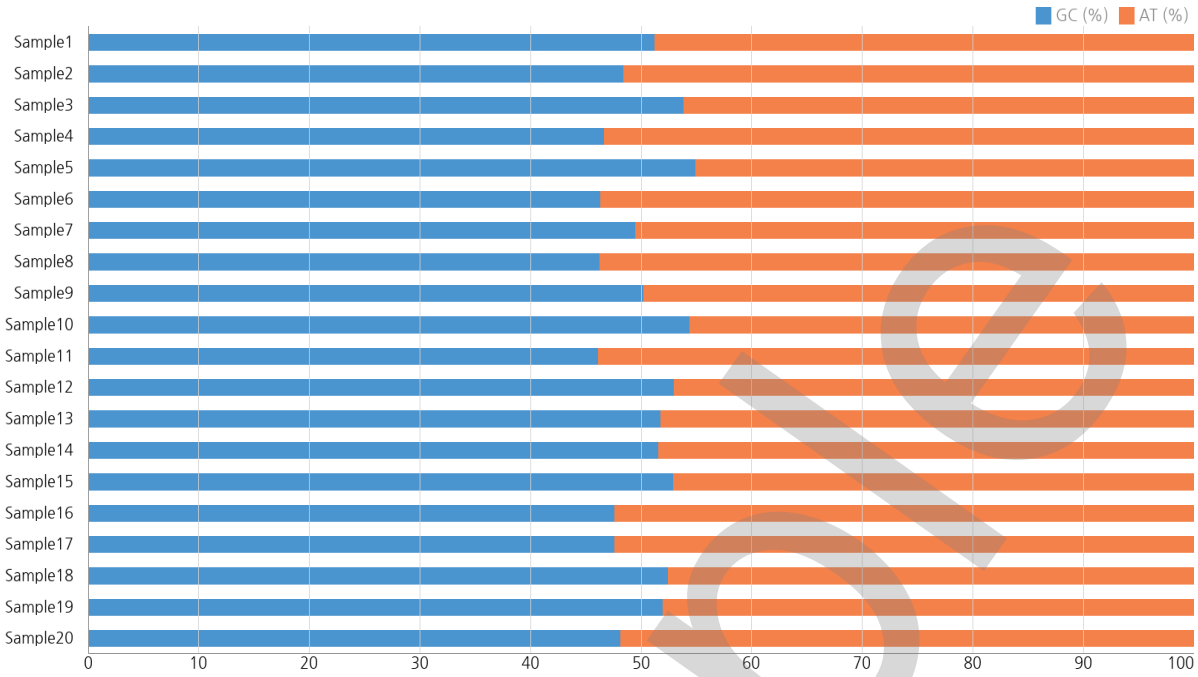

Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

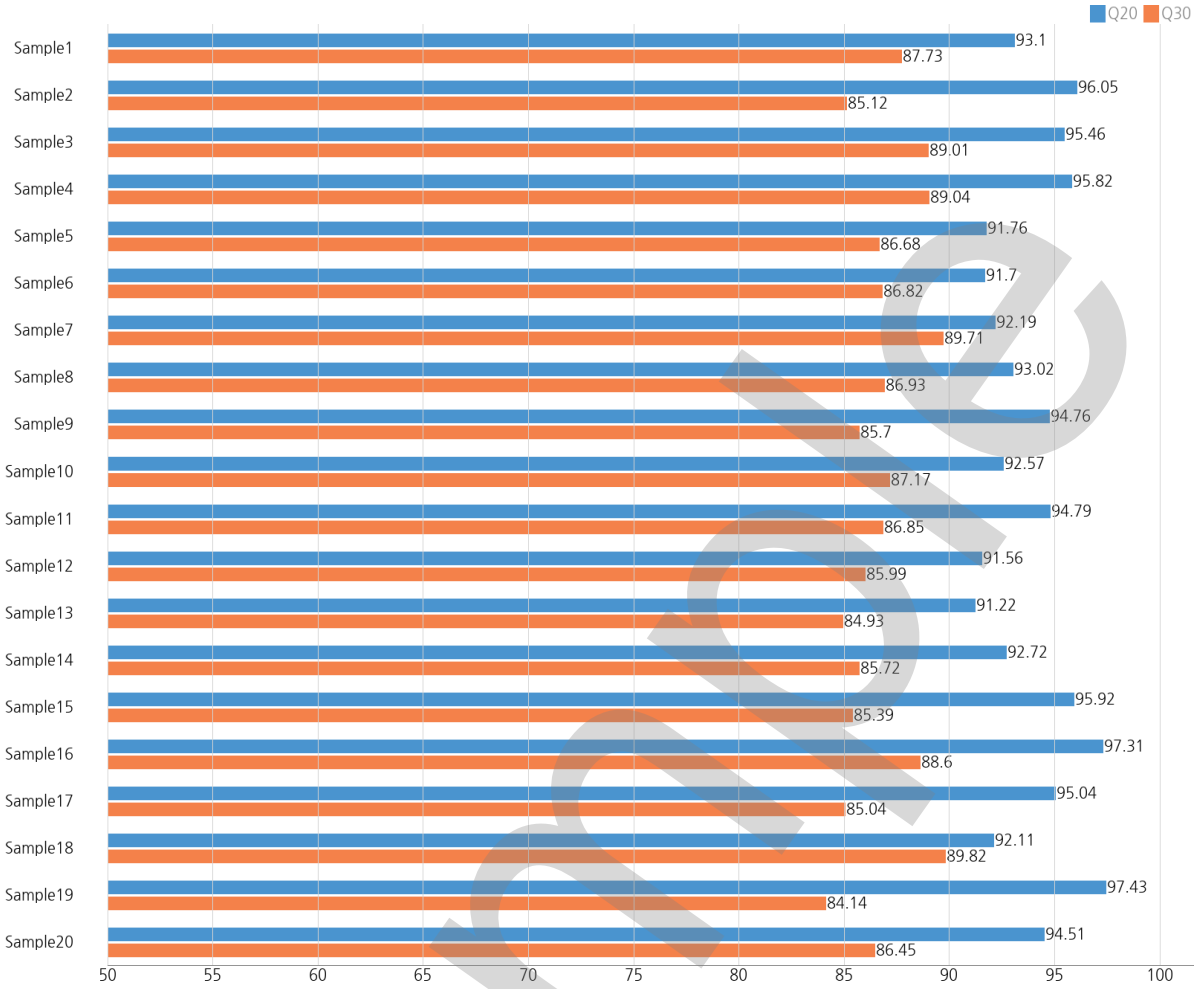

Figure 5. Q20/Q30 scores of Raw data

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:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNTNNNNNNNNNNNN
+
@@@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)

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
Email: ngs@macrogen.com
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

3F Kyoto University International Science Innovation Bldg.
36-1 Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501 JAPAN
Tel: +81-75-746-2773
Email: customer@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