Humanizing Genomics
macrogen

Raw Data

Report

August 2018

Research use only

COPYRIGHT © MACROGEN INC

Project Information

Client Name	James Wynne
Company / Institution	CSIRO
Order Number	1806KHP-0041
Type of Read	Paired-end
Read Length	101
Number of Samples	15
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

Download link	File size	md5sum
NEGcontrolR1_1.fastq.gz	2.1G	20e577ba4060508f8e124f4e9666c1ca
NEGcontrolR1_2.fastq.gz	2.1G	a72ff8b87cccfcd49f709e635fd45461
NEGcontrolR2_1.fastq.gz	2.2G	2a7178ed28ba8960331f9ce032597604
NEGcontrolR2_2.fastq.gz	2.2G	1d6b11e322a1860c1ae7bf99f4eeee15
NEGcontrolR3_1.fastq.gz	2.1G	c33746bf517ba0105487d7eac421e328
NEGcontrolR3_2.fastq.gz	2.1G	c7801c8c691c795d7d0be1494281fc18
POMV6HPIR1_1.fastq.gz	2.3G	57f24478372b15a8e706a24d60130187
POMV6HPIR1_2.fastq.gz	2.3G	3065cd5a7b1a273a35988b404969baae
POMV6HPIR2_1.fastq.gz	2.3G	79c88e649ec7226da56c8c4f5528c2a8
POMV6HPIR2_2.fastq.gz	2.4G	b0a9092dfeba769a886b386f50a5df08
POMV6HPIR3_1.fastq.gz	2.2G	4b44e7591936b71463292354ebccf405
POMV6HPIR3_2.fastq.gz	2.2G	05edeb8a082e7dbc21e037b9559c6d98
ISAV6HPIR1_1.fastq.gz	2.0G	0292e966c1c5bc0aec4a3edf5618df8f
ISAV6HPIR1_2.fastq.gz	2.0G	bc2fb283a40746a1e5bfe0e5cd41bc38
ISAV6HPIR2_1.fastq.gz	2.1G	1c9768b7551433e35015ac440f25861d
ISAV6HPIR2_2.fastq.gz	2.1G	da19b6851b3f29629447e9564069bb43
ISAV6HPIR3_1.fastq.gz	2.2G	fd04ea45de70c1b500313624732959f4
ISAV6HPIR3_2.fastq.gz	2.2G	90da7ac6d7de7bd49aaffe2c9e0215f1
POMV24HPIR1_1.fastq.gz	2.0G	0c1e9ef6d3882f61e25c2ae47e8d5da5
POMV24HPIR1_2.fastq.gz	1.9G	a6368879b61dd0a2119429284fa31bb5
POMV24HPIR2_1.fastq.gz	1.8G	4de8458181d77f9dc038d00c16cd192e
POMV24HPIR2_2.fastq.gz	1.8G	04c6d36b380c927c21002b26d32205ee
POMV24HPIR3_1.fastq.gz	1.8G	c5a45b3663effe2c5716e15355890bbb
POMV24HPIR3_2.fastq.gz	1.7G	950d7ef4b34c000ed67f3ba58ecba197
ISAV24HPIR1_1.fastq.gz	2.4G	671e3d4d9dce36e3940e53fdb95190a3
ISAV24HPIR1_2.fastq.gz	2.5G	b525ff86f53aec8ece6745d7459c664b
ISAV24HPIR2_1.fastq.gz	2.2G	4dd826ca664a2ded4bc5ace418531c40
ISAV24HPIR2_2.fastq.gz	2.1G	38e4d6a5d3711489ed7c941b4d85040e
ISAV24HPIR3_1.fastq.gz	1.8G	d397c1f0e9fdf4e264e161b1cadc18ea
ISAV24HPIR3_2.fastq.gz	1.8G	2a58bb80bac51d6131da984a9dc074ad

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

2. Experimental Methods and Workflow

2. 1. Experiment Overview

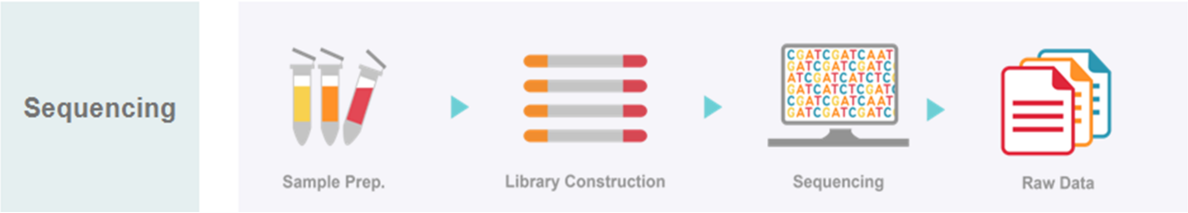


Fig1. Experiment overview

The Illumina NGS workflow includes 4 basic steps :

1) Sample Prep.(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 15 samples. For example, in NEGcontrolR1, 60,785,352 reads are produced, and total read bases are 6.1G bp. The GC content (%) is 48.8% and Q30 is 90.84%.

Table 1. Raw data Stats (maximum 20 samples)

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
NEGcontrolR1	6,139,320,552	60,785,352	48.8	51.2	95.6	90.84
NEGcontrolR2	6,769,899,710	67,028,710	49.17	50.83	95.78	91.16
NEGcontrolR3	6,382,134,248	63,189,448	49.03	50.97	96.02	91.73
POMV6HPIR1	6,980,068,388	69,109,588	48.76	51.24	95.56	90.86
POMV6HPIR2	6,976,214,026	69,071,426	48.69	51.31	95.55	90.93
POMV6HPIR3	6,742,240,658	66,754,858	49.07	50.93	96.01	91.64
ISAV6HPIR1	5,888,307,676	58,300,076	48.5	51.5	95.34	90.33
ISAV6HPIR2	6,203,947,018	61,425,218	48.96	51.04	95.33	90.39
ISAV6HPIR3	6,579,659,342	65,145,142	49.18	50.82	95.68	91.0
POMV24HPIR1	6,637,234,392	65,715,192	49.08	50.92	96.12	91.78
POMV24HPIR2	6,399,113,358	63,357,558	49.69	50.31	96.43	92.41
POMV24HPIR3	6,011,140,846	59,516,246	49.25	50.75	96.16	91.96
ISAV24HPIR1	7,545,567,994	74,708,594	49.01	50.99	95.46	90.77
ISAV24HPIR2	6,614,399,908	65,489,108	48.45	51.55	95.89	91.43
ISAV24HPIR3	5,406,071,056	53,525,456	48.18	51.82	95.54	90.76

- 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

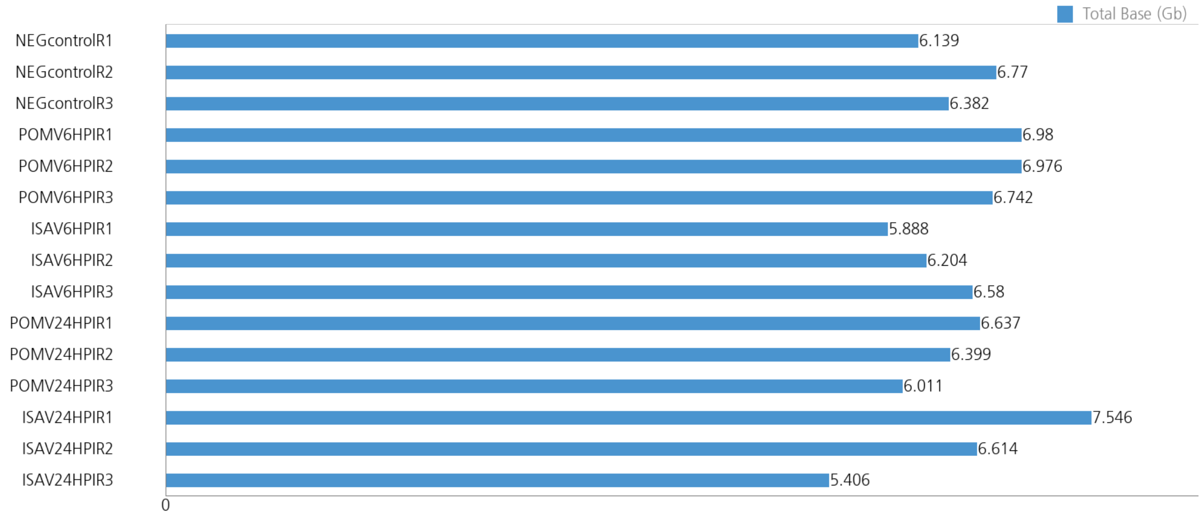


Figure 2.Throughput of Raw data

3. 3. Total Reads

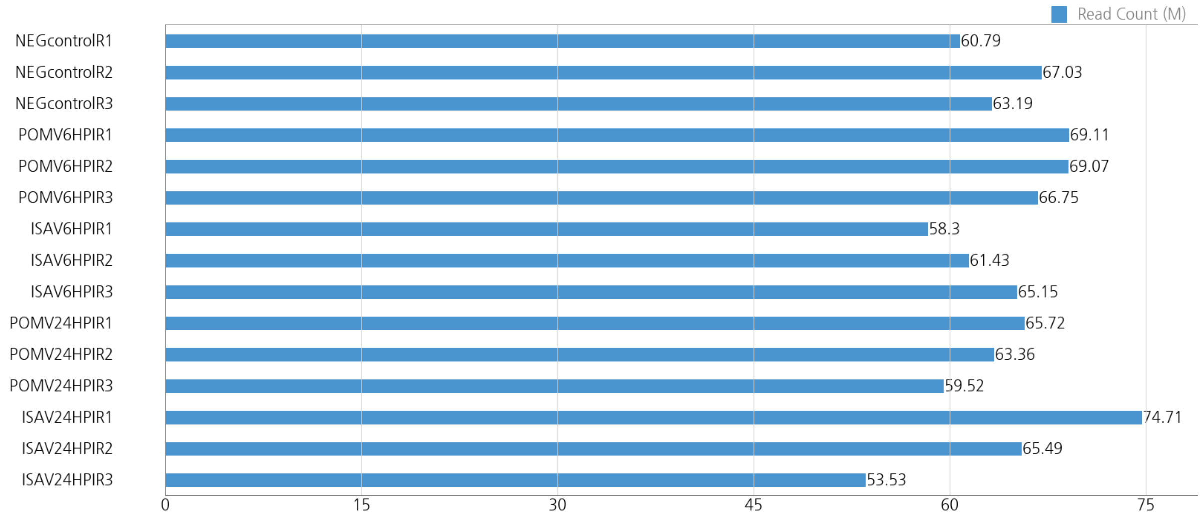


Figure 3. Total read count of Raw data

3. 4. GC/AT Content

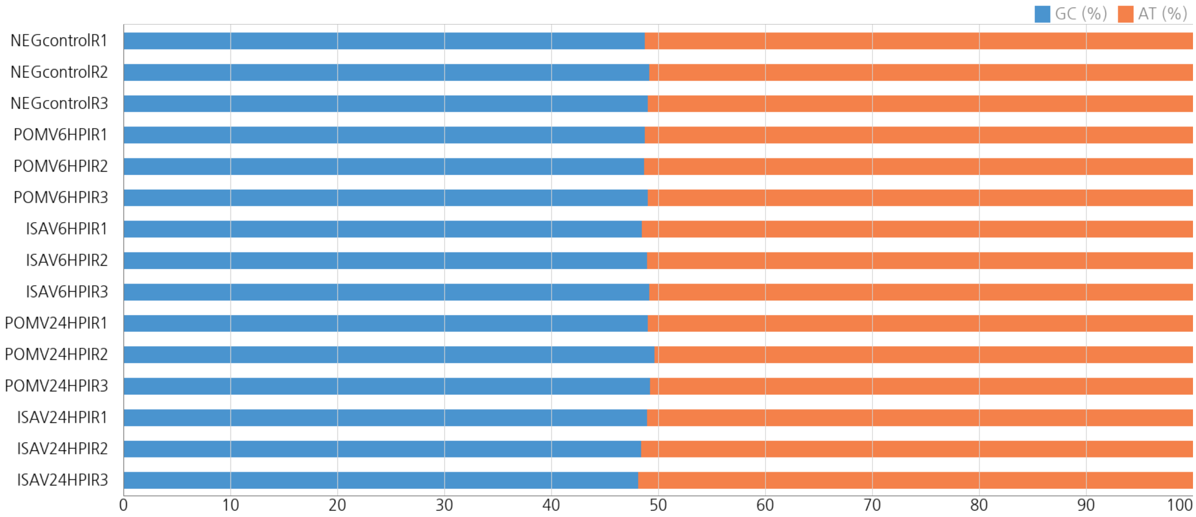


Figure 4. GC/AT Content of Raw data

3. 5. Q20/Q30 (%)

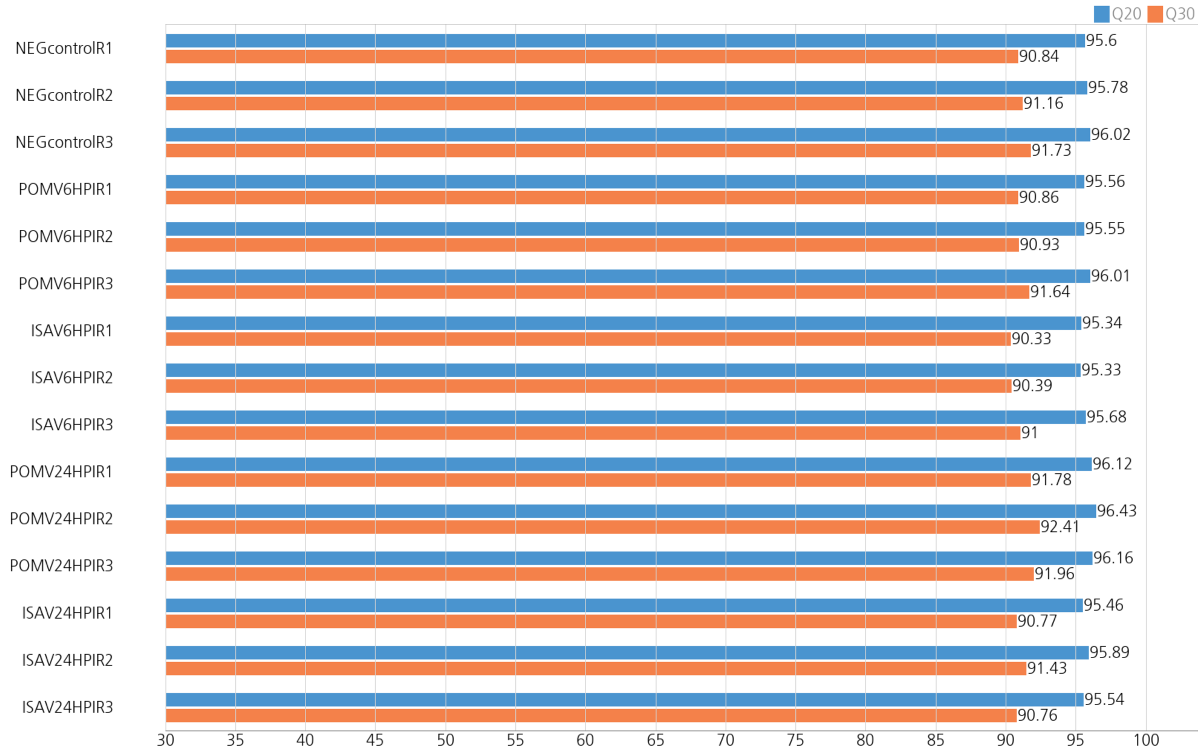


Figure 5. Q20/Q30 scores of Raw data

Research use only

4. Appendix

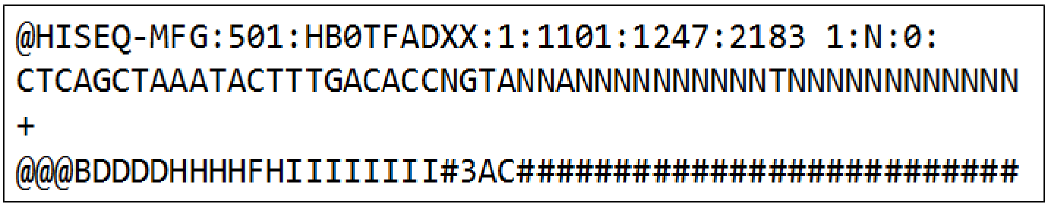
4. 1. FAQ

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)

Research use only

Humanizing Genomics
macrogen

Macrogen Korea

10F, 254 Beotkkot-ro,
Geumcheon-gu, Seoul
Rep. of Korea
Phone : +82-2113-7000

Contact

Web : www.macrogen.com
Lims : http://dna.macrogen.com

Research use only