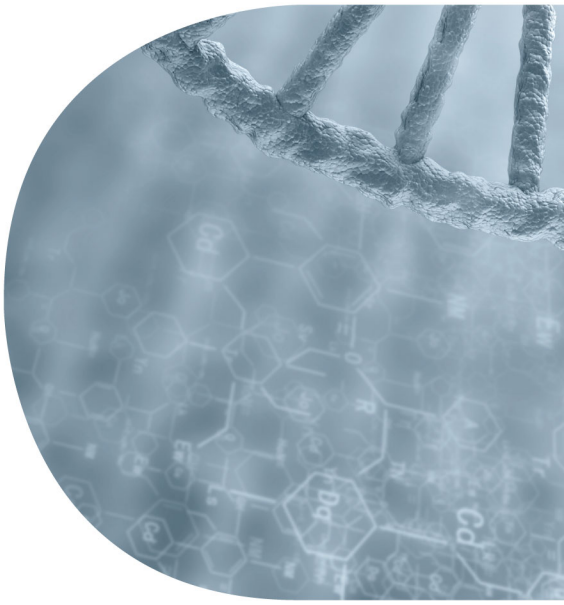
Research use only

Raw Data

Report

September 2020



Humanizing Phenomics 120
psomagen

Project Information

Client Name	Gregory Randolph
Company/Institution	University of Wyoming
Order Number	2007UNHS-0288
Type of Read	paired-end
Read Length	251
Number of Samples	1
Type of Sequencer	Illumina platform

Table of Contents

1. Experimental Methods and Workflow

1. 1. Experiment overview

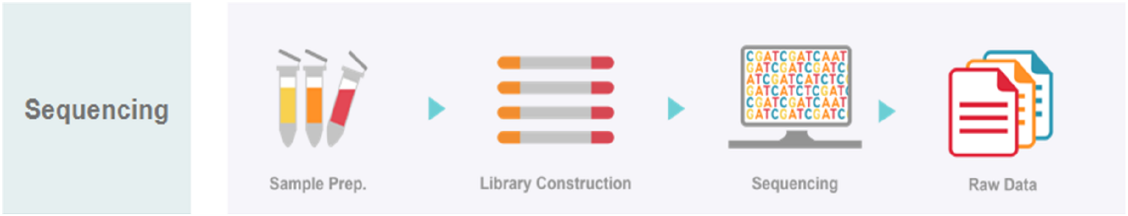


Fig1. Experiment overview

The Illumina NGS workflows include 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.

1. 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.

2. Summary of Data Production

2. 1. Raw data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for the 1 samples. For example, in NovaSeq2-Full, 1,763,377,328 reads are produced, and total read bases are 442.6G bp. The GC content (%) is 61.59% and Q30 is 88.29%.

LINK 2007UNHS-0288.xlsx : Download

Table 1. Raw data Stats

Sample ID	Total read bases (bp)	Total reads	GC(%)	AT(%)	Q20(%)	Q30(%)
NovaSeq2-Full	442,607,709,328	1,763,377,328	61.59	38.41	94.37	88.29

- 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 read1 and read2.
- GC(%) : GC content.
- AT(%) : AT content.
- Q20(%) : Ratio of bases that have phred quality score greater than or equal to 20.
- Q30(%) : Ratio of bases that have phred quality score greater than or equal to 30.

2. 2. Total Read Bases

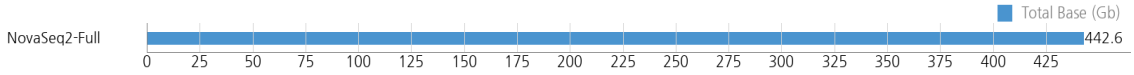


Figure 2.Throughput of Raw data

2. 3. Total Reads

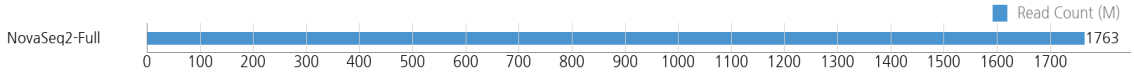


Figure 3. Total read count of Raw data

2. 4. GC/AT Content

Figure 4. GC/AT Content of Raw data

2. 5. Q20/Q30 (%)

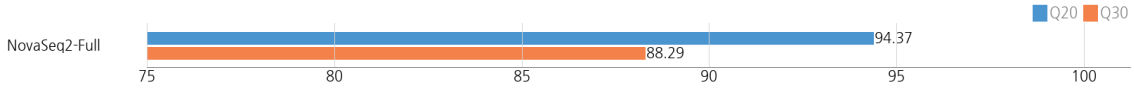


Figure 5. Q20/Q30 scores of Raw data

3. Data Download Information

3. 1. Raw Data and Analysis results

LINK 2007UNHS-0288.xlsx : Download

Fastq.gz	File size	md5sum
NovaSeq2-Full_R1.fastq.gz	77.8G	e9503783c28dd8d892b9196d650a07a4
NovaSeq2-Full_R2.fastq.gz	82.1G	086cd19c85a0638890d627554e2c8e8c

- 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.

The data retention period is three months. Please email (ngs@psomagen.com) or contact your sales representative for a longer retention period.

4. Appendix

4. 1. FAQ

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

A: Large volume zip file that is provided by our compnay is not user-friendly in Windows environment, so it is recommended to use linux environment for smooth 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

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 Phenomics 120
psomagen