Humanizing Genomics
macrogen

# HiSeq X Whole Genome Sequence Raw Data

# Report

May 2018

# Basic Information

| | |
|---|---|
| Client Name | Elza Duijm |
| Company/Institute | Naturalis Biodiversity Center |
| Order Number | 1804KHX-0077 |
| Library Kit | TruSeq DNA PCR Free (350) |
| 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 is extracted from a sample. After performing quality control(QC), passed sample is proceeded with the library construction.

### 2) Library Construction

The sequencing library is prepared by random fragmentation of the DNA 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 HiSeq X Ten generates raw images and base calling through an integrated primary analysis software called RTA 2(Real Time Analysis 2). The BCL (base calls) binary is converted into FASTQ using illumina package bcl2fastq2-v2.20.0. The demultiplexing option (--barcode-mismatches) was set to default (value : 1). 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 5 samples. For example, in group-1-EF, 399,363,124 reads are produced, and total read bases are 60.3G bp. The GC content (%) is 36.93% and Q30 is 89.25%.

Table 1. Raw data Stats

| Sample ID | Total read bases (bp) | Total reads | GC(%) | AT(%) | Q20(%) | Q30(%) |
|---|---|---|---|---|---|---|
| group-1-EF | 60,303,831,724 | 399,363,124 | 36.93 | 63.07 | 95.02 | 89.25 |
| group-2-IF | 56,804,209,216 | 376,186,816 | 36.92 | 63.08 | 94.84 | 88.94 |
| group-3-LF | 54,587,863,530 | 361,509,030 | 37.14 | 62.86 | 96.0 | 91.34 |
| group-4 | 82,470,135,538 | 546,159,838 | 37.05 | 62.95 | 96.08 | 91.51 |
| group-5-NF | 54,296,890,456 | 359,582,056 | 37.13 | 62.87 | 96.84 | 92.81 |

- Sample ID : Sample name.
- Total read bases : Total number of bases sequenced.
- Total reads : Total number of reads. In illumina paired-end sequencing, read1 and read2 are added.
- 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

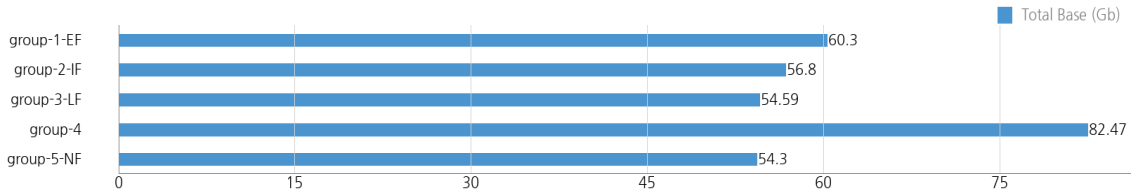


Figure 2.Throughput of Raw data

## 2. 3. GC/AT Content

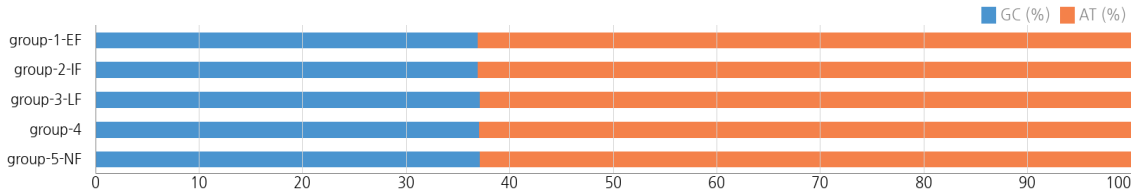


Figure 4. GC/AT Content of Raw data

## 2. 4. Q20/Q30 (%)

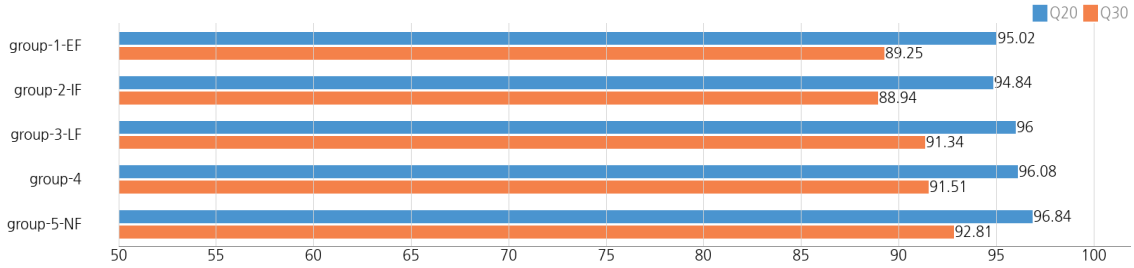


Figure 5. Q20/Q30 scores of Raw data

# 3. Data Deliverables

## 3. 1. FASTQ

FASTQ files are saved compressed in the GNU zip format, an open source data compression program. Each sample file is named as [Sample Name]_R1.fastq.gz (Read1) and [Sample Name]_R2.fastq.gz (Read2).

## 3. 2. md5sum

MD5 is a string of 32 hexadecimal values, which represents a 'fingerprint' of a file. By comparing the supplied MD5 value to the actual value computed by the MD5sums utility, you can make sure that the file that you downloaded off of the internet has not been tampered with or modified from the original file stored in our server.

# 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 company 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**

```
@ST-E00104:157:H03N0ALXX:1:1101:2837:1309 1:N:0:3
AAAACAACTCCCCTGGTATGATATATATTTGAGCAGAATTTATAATTTCAC
+
AAFFFJJJJJJFJJJJJJJFJJJJJJJJJJFJAJJJJJJJJJJJJJJJJJ
```

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 |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1000 | 99.9% |
| 40 | 1 in 10000 | 99.99% |

- Encoding : ASCII Character Code=Phred Quality Value + 33

### Quality Score Bins for Optimized 8-Level Mapping

Q score of HiSeq X Ten system : Q scores have been calibrated specifically to the HiSeq X Ten system and its consumables. It does use Q score binning. This is necessary for HiSeq X Ten runs due to the quantity of data being generated and since it cannot be turned off. Please refer to this table below, Q Scores for HiSeq X Ten are binned using the following criteria.

| Q-Score Bins | Example of Empirically Mapped Q-Scores |
|---|---|
| **N (no call)** | N (no call) |
| 2-9 | 7 |
| 10-19 | 11 |
| 20-24 | 22 |
| 25-29 | 27 |
| 30-34 | 32 |
| 35-39 | 37 |
| 40-45 | 42 |

- The quality score table above is typically updated when significant characteristics of the sequencing platform change, such as new hardware, software, or chemistry versions.

More information can be found here:

LINK http://support.illumina.com/help/SequencingAnalysisWorkflow/Content/Vault/Informatics/Sequencing_Analysis/CASAVA/swSEQ_mCA_FASTQFiles.htm


Research use only

**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