Research use only

# Sample Genome Survey

# Report

November 2018



Humanizing Genomics
macrogen

# Project Information

| | |
|---|---|
| Client Name | Macrogen |
| Company/Institute | Macrogen Inc. |
| Order Number | 1807BHP-1111 |
| Library Type | TruSeq DNA PCR-Free kit |
| Read Length | 101 |
| Sample | Sample |
| Type of Analysis | Genome Survey |
| Type of Sequencer | Illumina platform |

# Summary of Project Result

In this study, genome survey was performed in order to estimate the genome size of Sample.

Analysis was successfully performed on Sample sample. Figure 1 shows the throughput of raw data. Figure 2 shows the Q20 and Q30 percentage (% of bases with quality over phred score 20,30) of the sample's raw data.


Figure 1. Throughput of Raw Data

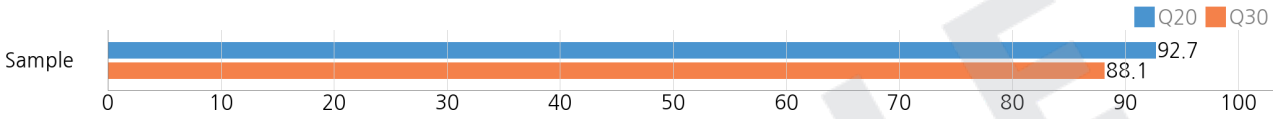


Figure 2. Q20/Q30 scores of Raw Data

After filtering, Jellyfish was used for k-mer analysis. A k-mer graph was drawn and the estimated genome size of Sample sample was calculated.

Table 1. K-mer Analysis Summary

| | K-mer Coverage | Genome Length |
|---|---|---|
| 17mer | 179.5 | 3,468,019 |
| 21mer | 170.5 | 3,466,176 |
| 25mer | 161.5 | 3,464,248 |

# Table of Contents

# 1. Data Download Information

## 1. 1. Raw Data and Analysis results

| File Name | File Size | md5sum |
|---|---|---|
| Sample_1.fastq.gz | 698M | 371111664d4dada2482877fdc84f9af9 |
| Sample_2.fastq.gz | 765M | c3d205422900f95ca750df8856681a55 |
| Sample_1.filtered.fastq.gz | 512M | 543ebc6799f92349112889ca6a470295 |
| Sample_2.filtered.fastq.gz | 564M | d6534f8d20a269c49f2645f0ec6fcf60 |

- fastq.gz - Compressed file of raw data.
- filtered.fastq.gz - Compressed file of adapter trimmed 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.

## 1. 2. Details of File Extensions

**Raw Data**

| File Extensions | Details |
|---|---|
| *.fastq | This file format is typically used in NGS technology, and includes ID, sequence and quality value. |

**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


Figure 3. Workflow Overview

**Sequencing**

1) Sample Prep.(Sample Preparation)
For library construction, DNA/RNA 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 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.

**Preprocessing**

1) Quality Control
After sequencing, quality control is performed on the sequenced raw reads. Overall reads' quality, total bases, total reads, GC (%) and basic statistics are calculated.

2) Preprocessing
In order to reduce biases in analysis, adapter trimming and quality filtering are performed. The filtered reads' quality, total bases, total reads, GC (%) and basic statistics are calculated again.

**Analysis**

1) K-mer Analysis
This process provides information of k-mer coverage, heterozygosity and estimated genome size.

# 3. Summary of Produced Data

## 3. 1. Raw Data Statistics

The total number of bases, reads, GC (%), Q20 (%), Q30 (%) were calculated for the Sample sample. For example, in Sample, 19,302,322 reads were produced, and total read bases are 1,949,534,522. The GC content is 46.37 % and Q30 is  88.11 %.

Table 2. Raw Data Stats

| Library Name | Total Read Bases (bp) | Total Reads | GC (%) | Q20 (%) | Q30 (%) |
|---|---|---|---|---|---|
| Sample | 1,949,534,522 | 19,302,322 | 46.37 | 92.65 | 88.11 |

- Library Name : Sample name.
- Total Read Bases (bp) : Total number of bases sequenced.
- Total Reads : Total number of reads. In Illumina paired-end sequencing, this value refers to the sum of read1 and read2.
- GC (%) : GC 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. Filtered Data Statistics

Trimmomatic was used to remove adapter sequences and low quality reads in order to reduce biases in analysis. The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) were calculated for the Sample sample after filtering.

Table 3. Filtered Data Stats

| Library Name | Total Read Bases (bp) | Total Reads | GC (%) | Q20 (%) | Q30 (%) |
|---|---|---|---|---|---|
| Sample | 1,508,024,678 | 15,213,442 | 46.38 | 98.45 | 97.0 |

- Library Name : Sample name.
- Total Read Bases (bp) : Total number of bases sequenced.
- Total Reads : Total number of reads. In Illumina paired-end sequencing, this value refers to the sum of read1 and read2.
- GC (%) : GC 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. 3. Average Base Quality at Each Cycle after Filtering

The 'per base sequence quality' plot generated by FastQC was used to check the overall quality of the produced data. This plot shows the average quality at each cycle.

The x-axis and y-axis are respectively the number of cycles, and phred quality score. Phred quality score of 20 means 99% accuracy and reads with quality score over 20 are generally accepted as good quality reads.

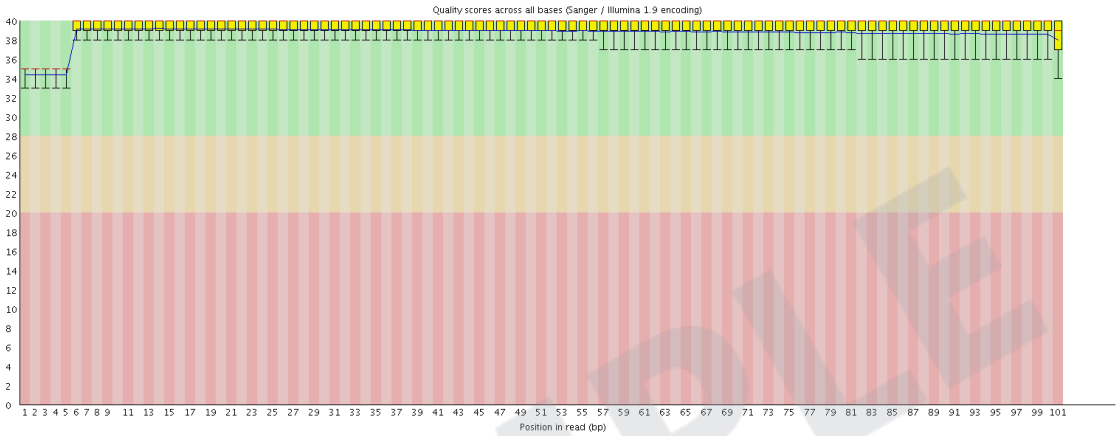


Figure 4. Read 1 Quality at each cycle of Sample

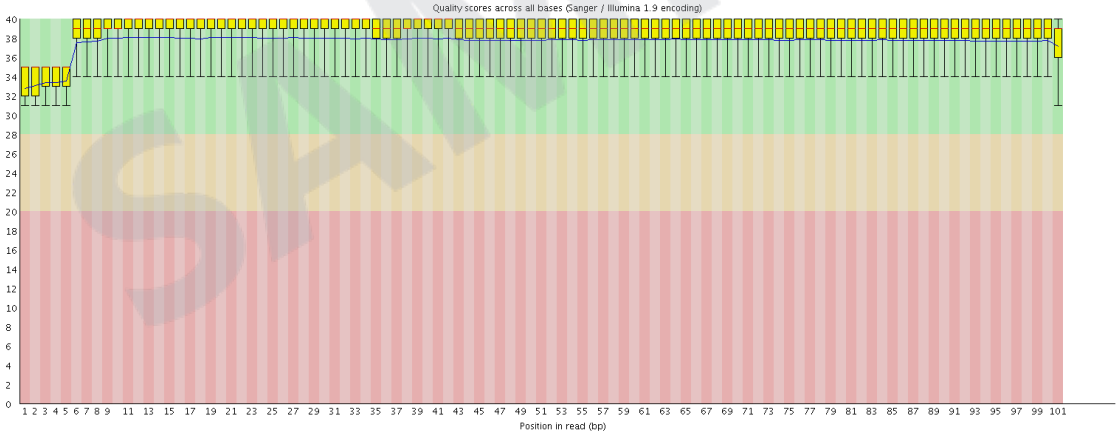


Figure 5. Read 2 Quality at each cycle of Sample

- Yellow box : Interquartile range (25-75%) of phred score at each cycle.
- Red line : Median phred score at each cycle.
- Blue line : Average phred score at each cycle.
- Upper & Lower whiskers : Point of 10% and 90%.
- Green background : Good quality.
- Orange background : Acceptable quality.
- Red background : Bad quality.

# 4. Analysis Results

## 4. 1. K-mer analysis

The graph is plotted with the coverage and frequency of k-mers. The sharp left-side peak represents random sequencing error while the right represents appropriate data. The genome size can be estimated using total k-mer number and volume peak. The top of the peak will not intersect the kmer-peaks line because of the over dispersion in real data.

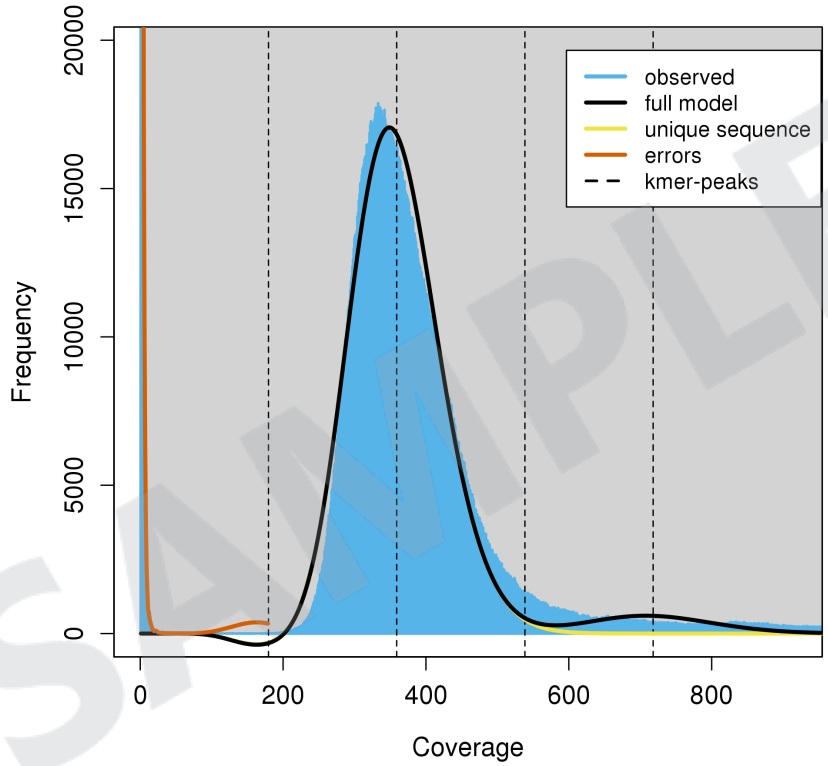


Figure 6. K-mer Graph (17mer)

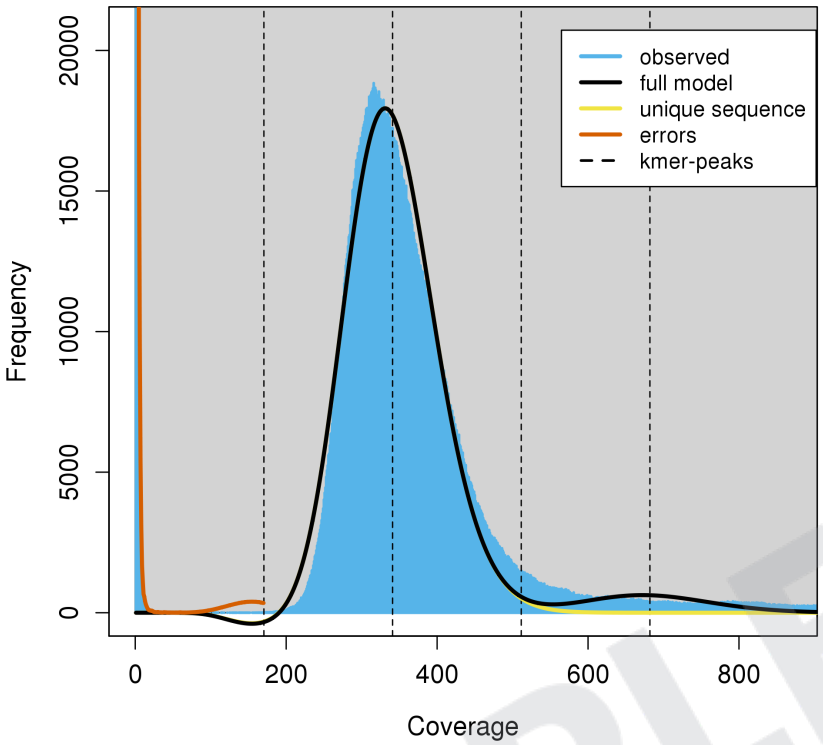


Figure 7. K-mer Graph (21mer)

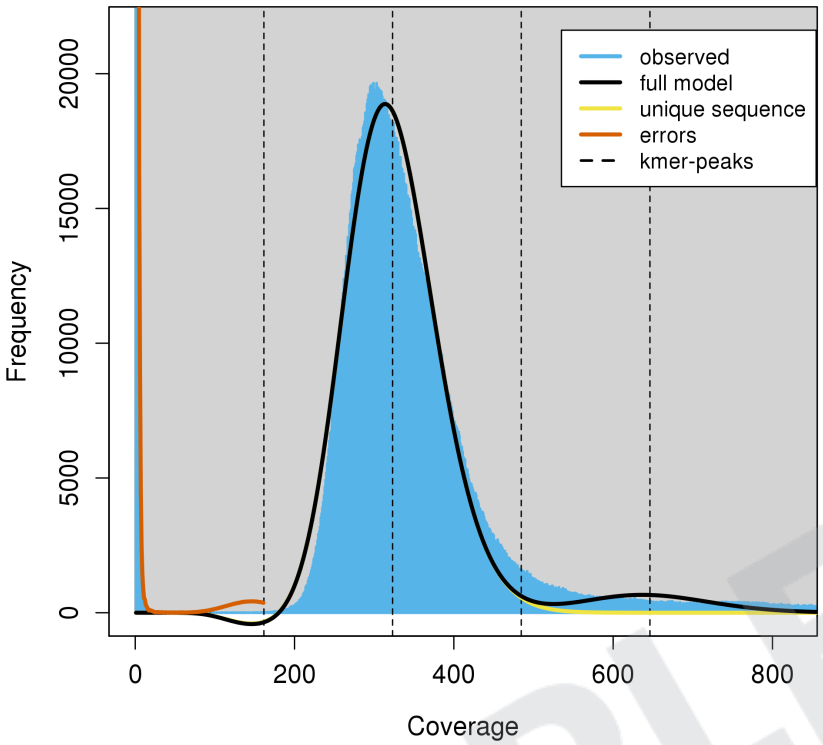


Figure 8. K-mer Graph (25mer)

Table 4. K-mer Analysis Result

| | K-mer Coverage | Heterozygosity | Genome Length | Genome Repeat Length | Error Rate |
|---|---|---|---|---|---|
| 17mer | 179.5 | 0.038 | 3,468,019 | 857,929 | 0.095 |
| 21mer | 170.5 | 0.03 | 3,466,176 | 844,862 | 0.09 |
| 25mer | 161.5 | 0.026 | 3,464,248 | 836,455 | 0.087 |

- K-mer Coverage : The mean k-mer coverage for heterozygous bases.
- Heterozygosity : The overall rate of heterozygosity.
- Genome Length : The inferred genome length.
- Genome Repeat Length : The length of the genome that is repetitive.
- Error Rate : The error rate of the reads.

# 5. Appendix

## 5. 1. FAQ

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

## 5. 2. FASTQ File

### 5. 2. 1. FASTQ Format

**Example of FASTQ**

```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:2183 1:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNNNNNNNTNNNNNNNNNNN
+
@@@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.

## 5. 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)

# 5. 4. Programs used in Analysis

## 5. 4. 1. FastQC

LINK http://www.bioinformatics.babraham.ac.uk/projects/fastqc

FastQC (V0.11.5) is a quality checking tool for high throughput sequencing data. The condition of the data can be checked through the various modules provided by the tool. Among the modules, 'Per base sequence quality' and 'Per tile sequence quality' modules are commonly used to validate whether the data can be used for analysis.

## 5. 4. 2. Trimmomatic

LINK http://www.usadellab.org/cms/?page=trimmomatic

Incomplete removal of adapter sequences from NGS data can ultimately affect the accuracy of analysis. To avoid this, Trimmomatic (v0.36) is used to remove adapter sequences. Depending on the library type used for data production, appropriate adapter sequence is used to remove the said sequences from the data. In addition to removing adapter sequences, Trimmomatic trims out bases of low quality.

## 5. 4. 3. Jellyfish

LINK http://www.genome.umd.edu/jellyfish.html

Jellyfish (v2.2.10) is a program that counts k-mers in DNA. It provides information that can be used in many analyses of DNA sequences such as genome size prediction, genome coverage confirmation and repeat sequence ratio calculation by counting the abundance of a particular k-mer in the sequence.

## 5. 4. 4. GenomeScope

LINK http://qb.cshl.edu/genomescope/

GenomeScope can infer the global properties of a genome from unassembled sequenced data. GenomeScope uses the k-mer count distribution and within seconds produces a report and several informative plots describing the genome properties.




**Macrogen Inc.** | 08511 10F, 254 Beotkkot-ro, Geumcheon-gu, Seoul, Rep of Korea

**Contact us**
Tel:+82-2-2180-7016
Site: www.macrogen.com | http://dna.macrogen.com