GAP
Genome Analysis Platform

CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
EXCELENCIA SEVERO OCHOA

MG-04_Total_RNA-Seq

Library_ Protocol	llumina-stranded-total-rna
Type of Sequencer	NovaSeq 6000
Type of Read	Paired-end
Read Length	151
Number of Samples	6

Sequencing Experimental Methods

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 present 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.

Raw data: Sequencing data is converted into raw data for the analysis (*.FASTQ).

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 Inc.'s package bcl2fastq. Adapters are not trimmed away from the reads.

FASTQ file name	File size	Checksum
S_01_1.fastq.gz	4,678,817,514	6b40f4a79eacca940ed1229237a86aec
S_01_2.fastq.gz	4,734,643,155	02f8f303b663b370089fd73fb4089ab9
S_02_1.fastq.gz	4,584,591,784	b1cabcb7279c9260b18f0127900ebfc3
S_02_2.fastq.gz	4,641,705,317	6accc5acb11650e838ba843ea59c927e
S_03_1.fastq.gz	4,656,924,421	da52806ade488e7d034492ab1b0f96ad
S_03_2.fastq.gz	4,881,174,700	d2d86f98a9ed7ac9804efe4ab432b154
S_04_1.fastq.gz	4,578,039,884	74f14377f5a51ebfb9ecb51a333728bf
S_04_2.fastq.gz	4,645,498,406	1a8eb367abab454c32173ae1f68439d9
S_05_1.fastq.gz	4,593,270,613	a5cf947bbbc6a333305551d9af768399
S_05_2.fastq.gz	4,662,041,216	29150cdc8b979a2b579cf9c4717e5be2
S_06_1.fastq.gz	4,630,410,502	e4a2596b7044a4f63abc0c251cc61732
S_06_2.fastq.gz	4,652,480,096	6195b2e4215c1a2ce78175832ad908d4

- FASTQ file name: This is a zip file of raw data used in analysis.
- Checksum: In order to verify the integrity of files, Checksum is used. If the values of md5sum are the same, there is no forgery, modification or omission.

Description of FASTQ Files

Example of FASTQ

```
@HISEQ-MFG:501:HB0TFADXX:1:1101:1247:21831:N:0:
CTCAGCTAAATACTTTGACACCNGTANNANNNN NNNNNN TNNNNN NNNNNN N
+
   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.

GAP
Genome Analysis Platform

CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
EXCELENCIA SEVERO OCHOA

Summary of Data Production: Raw Data Statistics

The total number of bases, reads, GC (%), Q20 (%), and Q30 (%) are calculated for all the samples.

Table 1. Raw data Stats

Sample name	Total bases	Read count	Library size	GC (%)	AT (%)	Q20 (%)	Q30 (%)
S_01	20,041,406,748	132,724,548	438	49.95	50.05	96.96	91.84
S_02	19,811,371,234	131,201,134	438	50.63	49.37	96.99	91.94
S_03	19,906,492,476	131,831,076	438	49.77	50.23	96.63	91.17
S_04	19,981,166,506	132,325,606	438	49.99	50.01	96.97	92.05
S_05	19,879,242,412	131,650,612	438	49.68	50.32	96.93	91.88
S_06	20,037,493,432	132,698,632	438	50.27	49.73	96.85	91.71

- Sample: Sample name.
- Total bases: Total number of bases sequenced.
- Read count: Total number of reads. In Illumina paired-end sequencing, read1 and read2 are added
- Library size: Size of the library in base pairs (adapters and insert size is taken into account)
- GC(%): GC content.
- AT(%): AT content.
- Q20(%): Ratio of reads that have phred quality score of over 20.
- Q30(%): Ratio of reads that have phred quality score of over 30.

GAP
Genome Analysis Platform

CIC bioGUNE
MEMBER OF BASQUE RESEARCH & TECHNOLOGY ALLIANCE
EXCELENCIA SEVERO OCHOA

Folder "FASTQC"

The FASTQC folder contains the quality control for each original FASTQ file per sample and per read if a paired-end run was performed.

For each sample, there is a "*.zip" archive and a "*_fastqc.html" file. Please, open the "**_fastqc.html**" file ONLY to check the quality of each FASTQ.

To interpret reports, each test is accompanied by a symbol:

***A green tick means the test comfortably passed the quality standard.**
***An orange exclamation mark means the test is barely passed.**
***A red cross means the test failed the quality control.**

You will find the following test reports for each QC file:

i. Basic Statistics: Statistics outline of the file.
ii. Per base sequence quality: Quality scores across all bases. Optimum values are included in a green area, acceptable in an orange area and bad results in a red one.
iii. Per sequence quality scores: Quality score distribution over all sequences.
iv. Per base sequence content: Percentage of each single base for each position.
v. Per base GC content: Percentage of GC content across all bases.
vi. Per sequence GC content: GC distribution over all sequences.
vii. Per base N content: Percentage of N content for each read position.
viii. Sequence Length Distribution: Distribution of sequence length over all sequences.
ix. Sequence Duplication Levels: Level of sequence duplication.
x. Overrepresented sequences: Sequences overrepresented on the file.
xi. Kmer Content: Relative enrichment of K-mer over read length.
For further information see: http://www.youtube.com/watch?v=bz93ReOv87Y