Analysis Report
(De novo assembly)

Genome Sequencer FLX

Institution	:	국립원예특작과학원
Name	:	성기호
Order number	:	**1108KFT-0042**
Sample name	:	**2312ps3**

Index

1. Description of Workflow

1.1 Data processing ---- 3

1.1.1 Image processing ---- 3

1.1.2 Signal processing ---- 4

1.2 Data analysis ---- 5

1.2.1 De novo assembly ---- 5

1.2.2 Blast ---- 6

2. Results of Data processing

2.1 Raw data ---- 7

3. Results of Analysis

3.1 Results of assembly ---- 12

4. Understanding Report

4.1 Raw data ---- 14

4.1.1 SFF file ---- 14

4.1.2 Reads sequence ---- 14

4.1.3 Reads quality ---- 14

4.2 Results of analysis ---- 15

4.2.1 Large contigs(>500bp) ---- 15

4.2.2 All contigs ---- 15

4.2.3 Scaffolds ---- 15

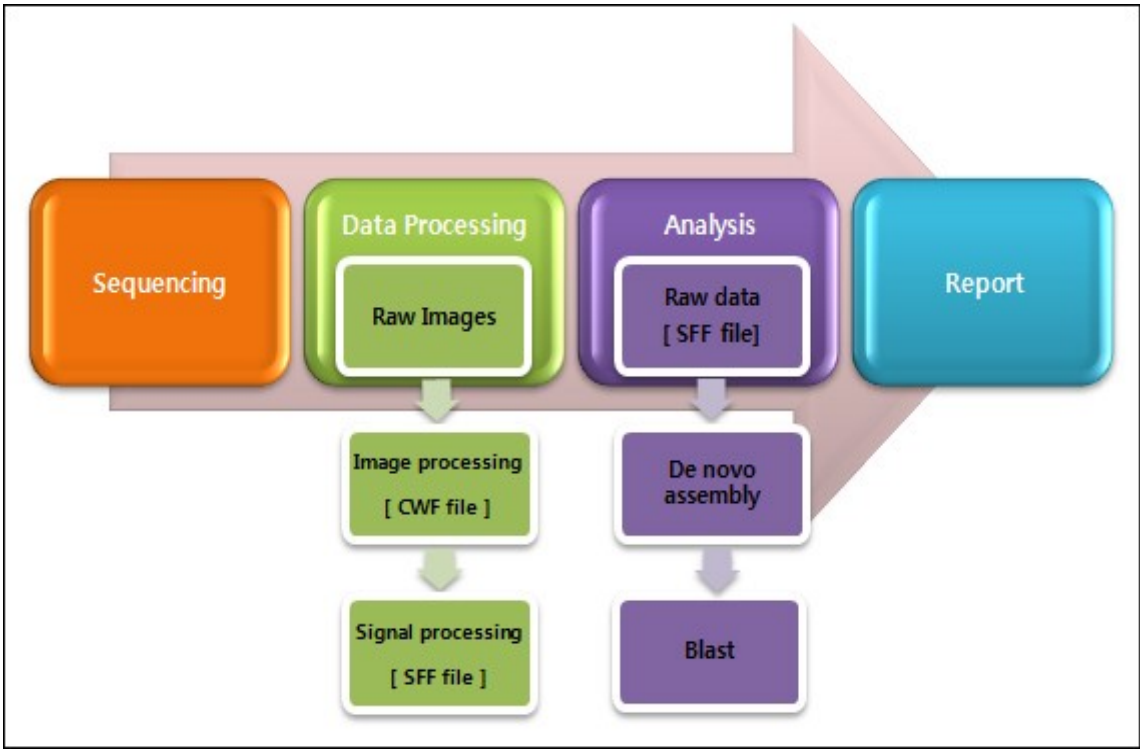
4.2.4 Result of Blast ---- 15

5. Data Download

5.1 Raw data ---- 16

4.2 Results of analysis ---- 16

1. Description of Workflow



1.1. Data processing

GS FLX data processing was performed using the Roche GS FLX software (v 2.6).

1.1.1. Image Processing

a. Subtract background and normalize the images

b. Find the active wells on the PicoTiterPlate device

c. Extract the raw signals for each flow, in each active well

d. Write the resulting flow signals into “composite wells format”(CWF) files

1.1.2. Signal Processing

a. Screen out ghost wells (Amplicon pipelines only)

b. Balance the signal strengths of the different nucleotides

c. Correct for interwell cross-talk between neighboring wells

d. Correct for anomalous signal spikes or interruptions due to a reagent flow valve event

e. Correct for known out-of-phase errors

f. Rebalance the signal strengths of the different nucleotides

g. Correct for signal droop

h. Subtract residual background signal

i. Screen out ghost wells

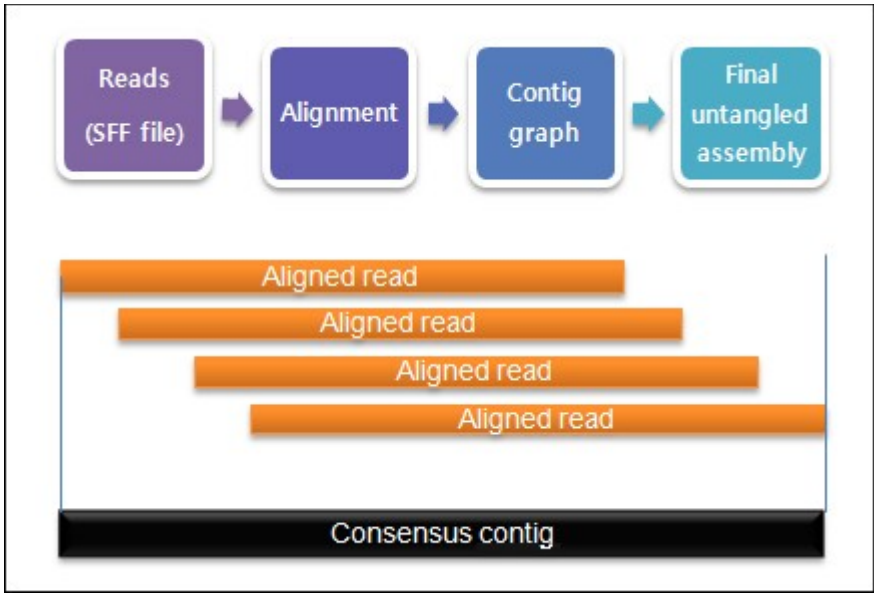
j. Filter the processed reads based on signal quality

k. Trim read ends for low quality and primer sequence

l. Generate flowgrams and base called sequences with corresponding quality scores for all individual, high quality reads and output to CWF and SFF files, one per PTP regio processed.

1.2. Data analysis

1.2.1. De novo assembly



Software : GS De Novo Assembler (v 2.6)

During the assembly process, the software:

a. Identifies pairwise overlaps between reads

b. Constructs multiple alignments of overlapping reads and divides or introduces breaks into the multiple alignments in regions where consistent differences are found between different sets of reads.

c. Attempts to resolve branching structures between contigs

d. Generates consensus basecalls of the contigs by using quality and flow signal information for each nucleotide flow included in the contigs multiple alignments

e. Outputs the contig consensus sequences and corresponding quality scores, along with an ACE file of the multiple alignments and assembly metrics files.

When Paired End data is available, the assembler performs these extra steps:

f. Organizes the contigs into scaffolds using Paired End information to order and orient the contigs and to approximate the distance between contigs

g. Outputs scaffolded consensus sequences and corresponding quality scores, along with an AGP file of the scaffolds and specific metrics tables.

Parameter :

a. Seed step

The number of bases between seed generation locations used in the exact k-mer matching part of the overlap detection.
Default Value : 12

b. Seed length

The number of bases used for each seed in the exact k-mer matching part of the overlap detection.
Default Value : 16

c. Seed count

The number of seeds required in a window before an extension is made.
Default Value : 1

d. Minimum overlap length

The minimum length of overlaps used by the assembler for the pairwise alignment step.
Default Value : 40

e. Minimum overlap identity

The minimum percent identity of overlaps used by the assembler for the pairwise alignment step.
Default Value : 90

f. Alignment identity score

When multiple overlaps are found, the per-overlap-column identity score used to sort the overlaps for use in the progressive alignment.
Default Value : 2

g. Alignment difference score

When multiple overlaps are found, the per-overlap-column difference score used to sort the overlaps for use in the progressive multi-alignment.
Default Value : -3

1.2.2. Blast (Basic Local Alignment Search Tool)

Software : Blast

E-value : 1.0E-3

Databases : NT (ftp://ftp.ncbi.nih.gov/blast/db/)

Blast finds regions of local similarity between sequences.

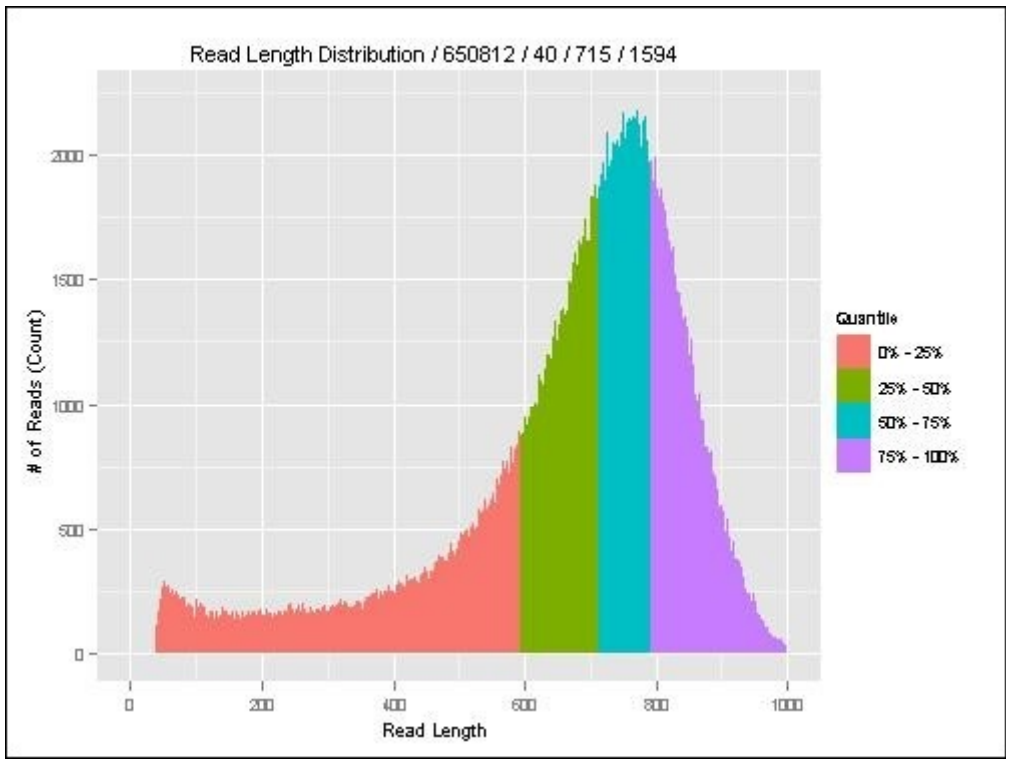
2. Results of Data processing

2.1. Raw data

a. General Library (2012.01.26)

Read count	Total bases	Average read length
650,812	433,267,606	665.734

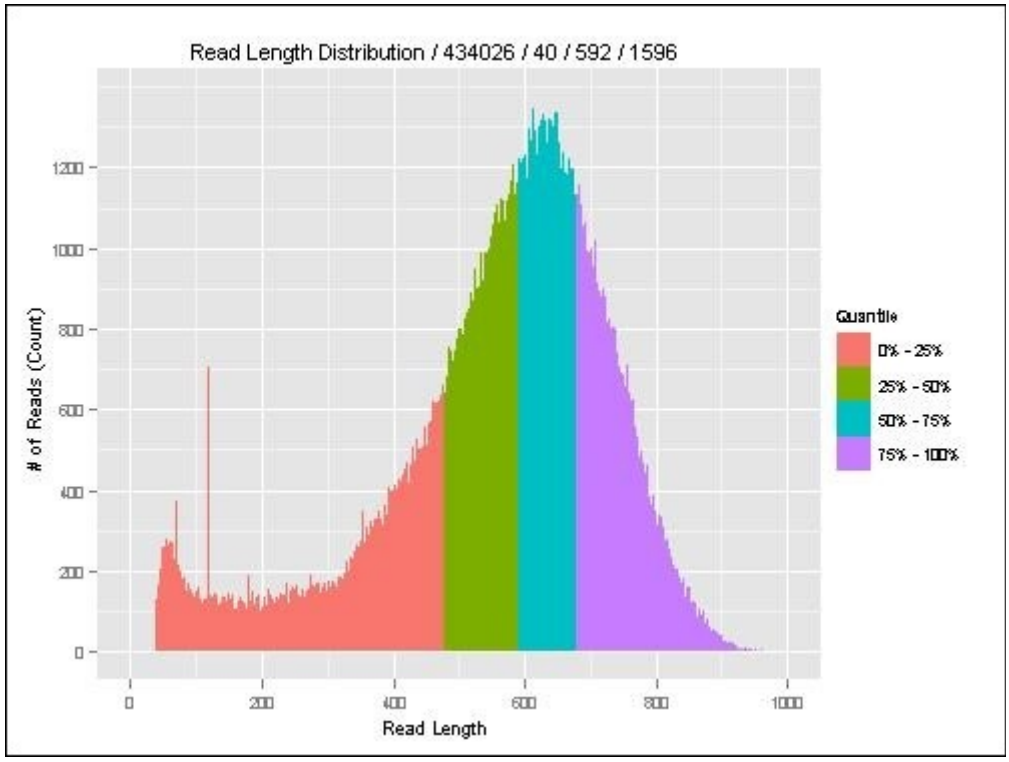
Read length distribution



b. General Library (2011.10.21-1)

Read count	Total bases	Average read length
434,026	242,359,043	558.398

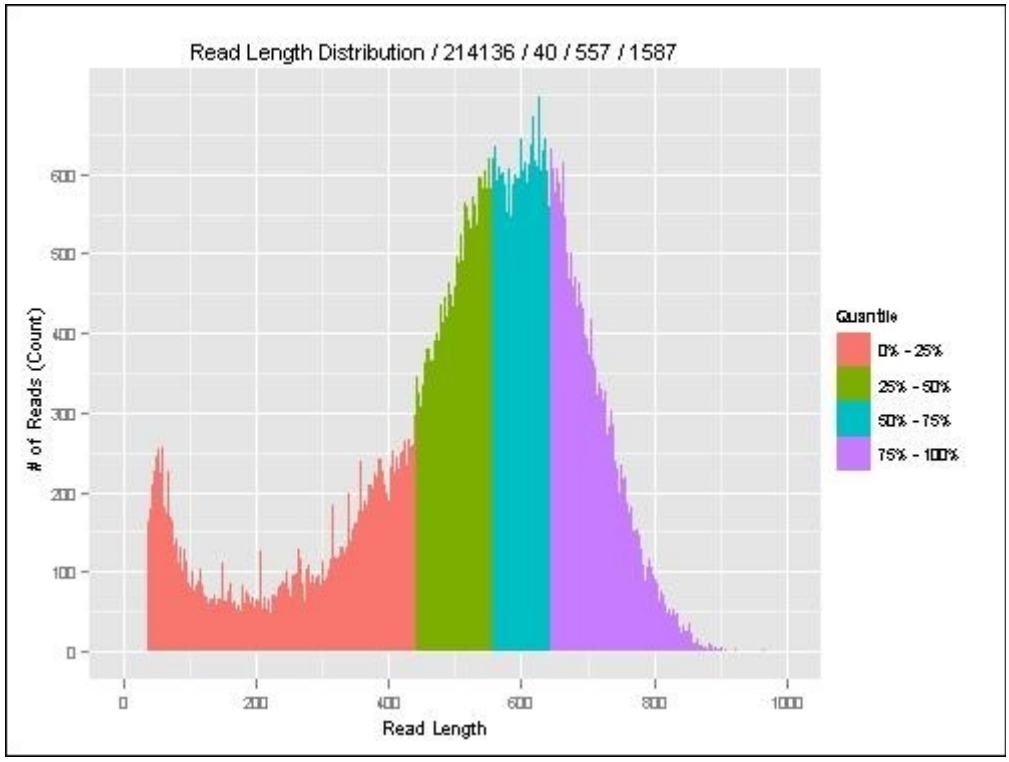
Read length distribution



c. General Library (2011.10.21-2)

Read count	Total bases	Average read length
214,136	111,691,598	521.592

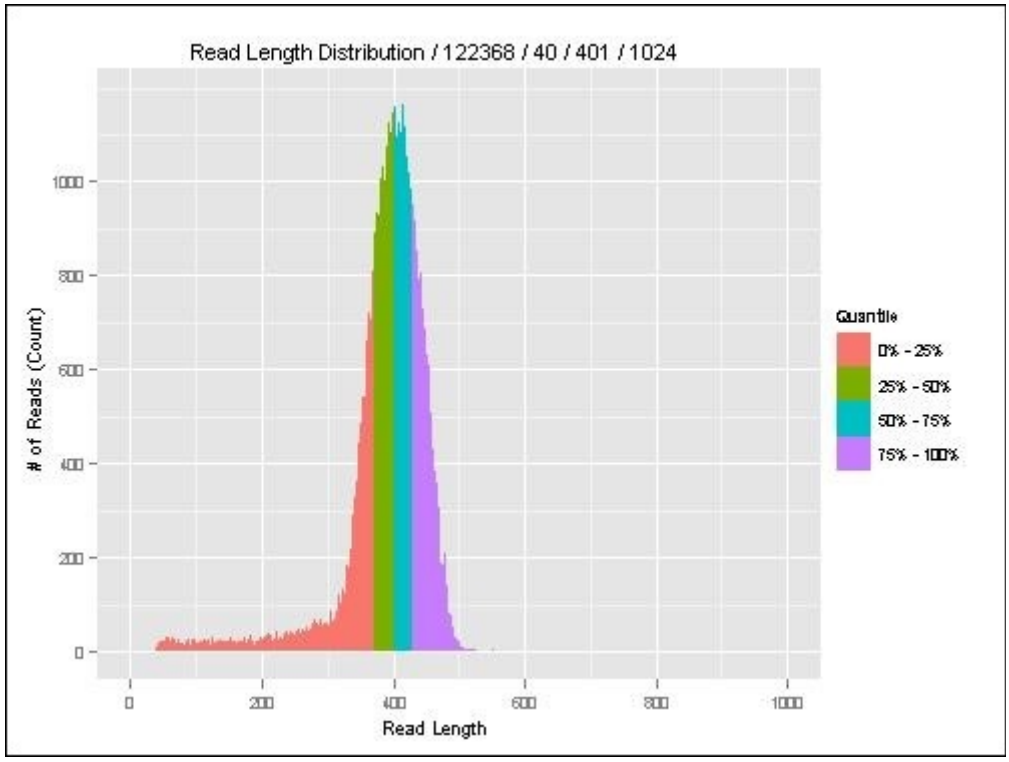
Read length distribution



d. Mate Paired Library (2012.01.25)

Read count	Total bases	Average read length
122,368	47,889,870	391.36

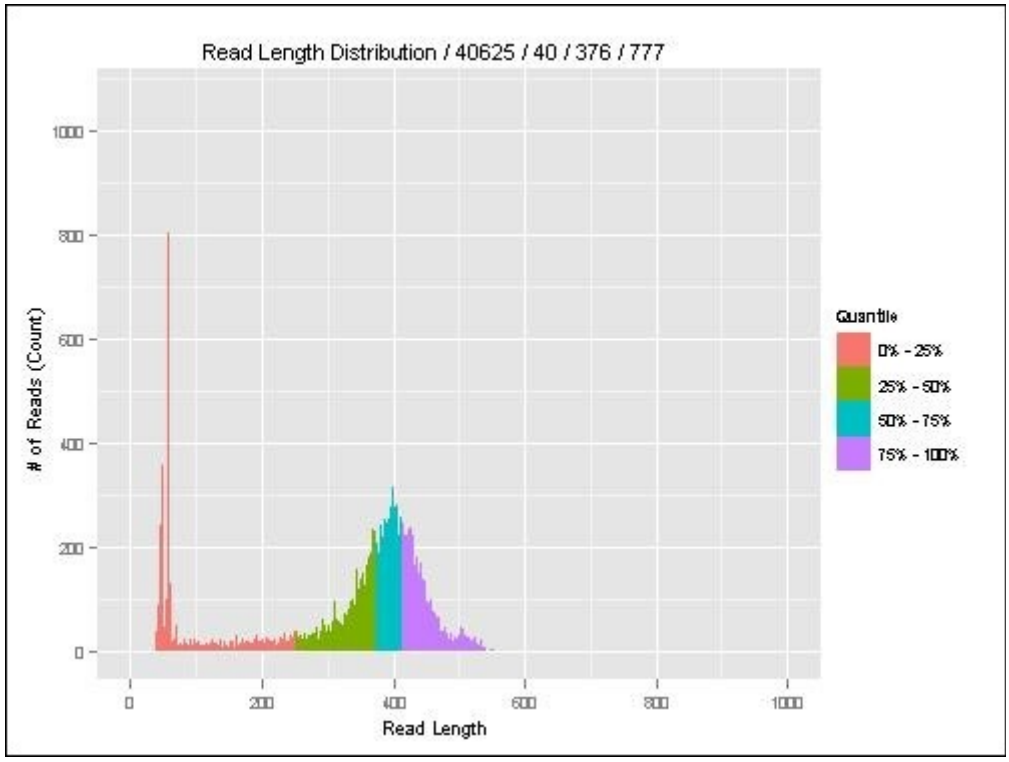
Read length distribution



e. Mate Paired Library (2011.10.23)

Read count	Total bases	Average read length
40,625	12,976,280	319.417

Read length distribution



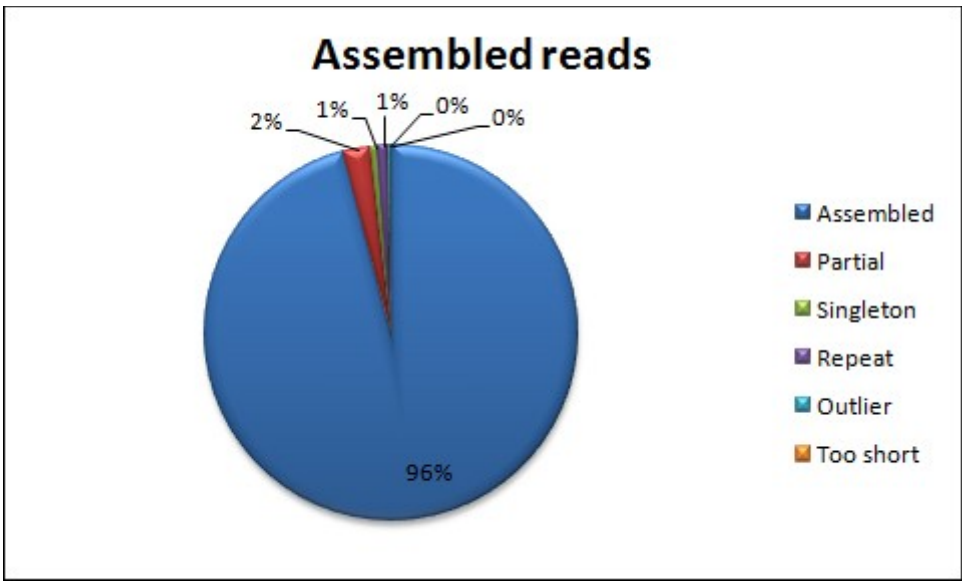
3. Results of Analysis

3.1. Results of assembly

3.1.1. Read status

Number of reads	Number of bases	Assembled	Partial	Singleton	Repeat	Outlier	Too short
1,547,124	838,079,703	1,483,958	36,089	9,470	13,513	4,094	0

- Number of reads : the read used in the assembly computation.
- Number of bases : the read's bases used in the assembly computation.
- Assembled : the read is fully incorporated into the assembly.
- Partial : only part of the read was included in the assembly.
- Singleton : the read did not overlap with any other reads in the input.
- Repeat : the read deemed to be from repeat regions.
- Outlier : the read was identified by the GS De Novo Assembler as problematic.
- Too short : the read was too short to be used in the computation.



3.1.2. Paired read status

Both mapped	One unmapped	Multiply mapped	Both unmapped	Distance Avg (2012.01.25)	Distance Dev (2012.01.25)	Distance Avg (2011.10.23)	Distance Dev (2011.10.23)
77,833	1,041	6,489	368	2782.9	754.8	2801.0	758.2

- Both mapped : both halves of the pair were aligned.
- One unmapped : one of the reads in the pair was unmapped.
- Multiply mapped : one or both of the reads in the pair were marked as Repeat.
- Both unmapped : both halves of the pair were unmapped.
- Distance Avg : the average distance between both halves, when both halves align on the same contig.
- Distance Dev : the standard deviation of the Distance Avg distribution.

3.1.3. Scaffolds

Number of scaffolds	Number of bases	Avg. size	N50 size	Largest size
1,506	43,991,976	29,211	80,519	606,634

- Number of scaffolds : the number of scaffolds identified.
- Number of bases : the total number of bases in the scaffolds.
- Avg. size : the average scaffold size.
- N50 size : the N50 scaffold size.
- Largest size : the size of the largest scaffold.

3.1.4. Scaffold contigs

Number of contigs	Number of bases	Avg. size	N50 size	Largest size
2,454	43,479,573	17,717	49,208	568,760

- Number of contigs : the number of contigs identified in scaffold.
- Number of bases : the total number of bases in the scaffold contigs.
- Avg.size : the average scaffold contig size.
- N50 size : the N50 scaffold contig size.
- Largest size : the size of the largest scaffold contig.

3.1.5. Large contigs (Length >= 500bp)

Num of contigs	Num of bases	Avg.size	N50 size	Largest size	Q40Plus bases	%Q40
3,928	44,960,973	11,446	46,627	568,760	44,701,122	99.42%

- Num of contigs : the number of large contigs identified.
- Num of bases : the total number of bases in the large contigs.
- Avg. size : the average contig size.
- N50 size : An N50 means that half of all bases reside in contigs of this size or longer.
- Largest size : the size of the largest contig.
- Q40Plus bases : the number of bases called that have a quality score of 40 or above.
- %Q40 : the percentage of bases called that have a quality score of 40 or above.

3.1.6. All contigs (Length >= 100bp)

Number of contigs	Number of bases
5,428	45,364,602

- Number of contigs : the number of all contigs identified.
- Number of bases : the total number of bases in the all contigs.

4. Understanding Report

4.1. Raw data

4.1.1. SFF file

File name : ~_SFF.sff

Standard Flowgram Format file - output data from the 454 Genome Sequencer system.

The Standard Flowgram File is used to store the information on one or many 454 Sequencing reads and their trace data.

4.1.2 Reads sequence

File name : ~_reads.fasta

This file contains the nucleotide sequences for the filtered reads.

4.1.3. Reads quality

File name : ~_reads.fasta.qual

This file contains the nucleotide quality scores (Phred-equivalent) for the high quality reads.

$Q = -10 \log_{10}$ (error rate)

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%
50	1 in 100000	99.999%

4.2. Results of Analysis

4.2.1. Large contigs (>500bp)

File name : 454LargeContigs.fna, 454LargeContigs.qual

Fasta and quality file of all the 'large' consensus basecalled contigs contained in 454AllContigs.fna

4.2.2. All contigs

File name : 454Allcontigs.fna, 454AllContigs.qual

Fasta and quality file of all the consensus basecalled contigs longer than 100 bases.

4.2.3. Scaffolds

File name : 454Scaffolds.fna, 454Scaffolds.qual, 454Scaffolds.xlsx

454Scaffolds.xlsx :

An AGP file (NCBI's format for describing scaffolds of contigs) containing the scaffold layout.

(http://www.ncbi.nlm.nih.gov/projects/genome/assembly/agp/AGP_Specification.shtml)

4.2.4. Results of Blast

File name : ~_NT.xlsx

The excel file contains two sheets:

a. _Summary : Distribution of hits per accession number

b. _Alignment : Results of alignment

5. Data Download

5.1. Raw data

Download link : General Library(2011.01.26) - SFF file

Download link : General Library(2011.01.26) - Reads sequence/quality

Download link : General Library(2011.10.21-1) - SFF file

Download link : General Library(2011.10.21-1) - Reads sequence/quality

Download link : General Library(2011.10.21-2) - SFF file

Download link : General Library(2011.10.21-2) - Reads sequence/quality

Download link : Mate Paired Library(2012.01.25) - SFF file, reads sequence/quality

Download link : Mate Paired Library(2011.10.23) - SFF file, reads sequence/quality

5.2. Results of Analysis

Download link : sequence/quality, summary of analysis results

Thank you!