华大基因 BGI | Premier Scientific Partner

Genome Survey Report

Bioinformatics Center

May 6, 2016

CONTENTS

1 DESCRIPTION OF WORKFLOW

1.1 Pipeline of Experiment

The pipeline of the experiment is illustrated in Figure 1.1 bellow:

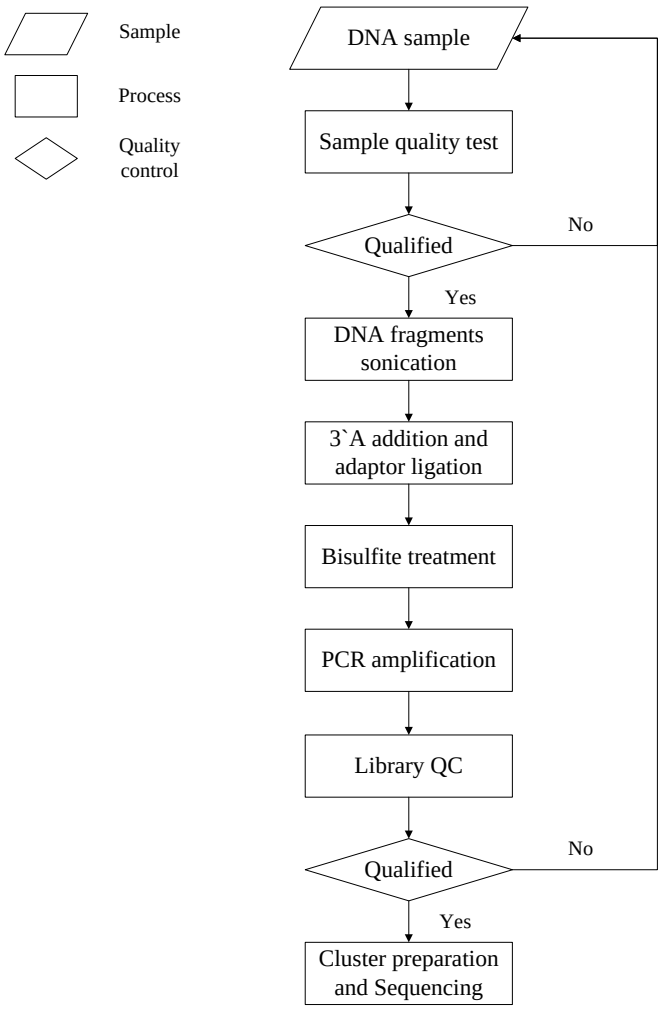

Figure 1.1 Pipeline of experiment.After the DNA sample(s) was(were) delivered, we did a sample quality test first. Then we used this(those) qualified DNA sample(s) to construct BS library, and we did a library quality test. At last, the qualified BS library would be used for sequencing

1.2 Pipeline of Bioinformatics Analysis

The pipeline of the Bioinformatics Analysis is illustrated in Figure 1.2 bellow:

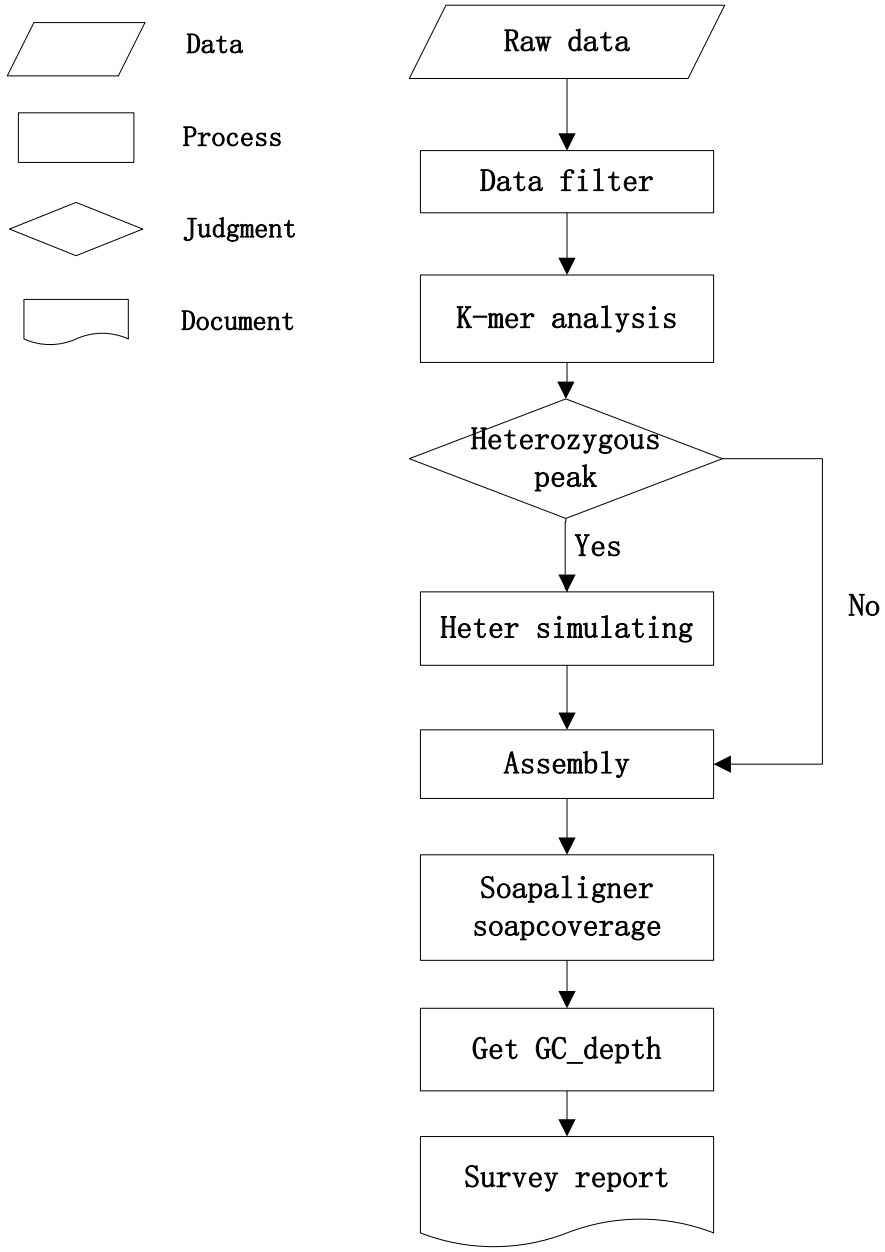

Figure 1.2 Pipeline of genome survey.When got the raw data, we filtered it first to get high quality reads. We used those clean data to do K-mer analysis and heterozygous simulation. Then we assembled them using *SOAP de novo* software. We got the gc depth distribution by *SOAPaligner* result. After all, we known the basic characteristics of the genome sequence and write the survey report.

2 BIOINFORMATICS RESULT

2.1 Background

a. Species name: Panopea generosa

b. Evaluate Genome size: 500 Mb

c. Designated reference sequences: No

2.2 Data statistics

Table 2.2.1 Statistics of Raw Data

Lib ID	Insert Size(bp)	Read Length(bp)	Data(Mb)	Sequence Depth(X)
wHAMPI023988-81	800	150	21781.6776	43.5634
wHAIPI023989-79	500	150	28301.1141	56.6022
wHAXPI023990-97	300	150	23277.0978	46.5542
Total	-	150	73359.8895	146.7198

This batch of sequencing produced 83.29GB raw data. After low quality reads filtering, total 73.36 Gb data was used for further analysis, if the genome size is estimated to be 500 Mb in previous experiment, then the sequencing depth of filter data is expected to be 146.72-fold.

2.3 17-mer analyses and genome size evaluation

A K-mer refers to an artificial sequence division of K nucleotides. A raw sequencing read with L bp contains (L-K+1) K-mers if the length of each K-mer is K bp. The frequency of each K-mer can be calculated from the raw genome sequencing reads. The K-mer frequencies along the sequencing depth gradient follow a Poisson distribution in a given data set. During deduction, the genome size G=K_num/Peak_depth, where the K_num is the total number of K-mer, and Peak_depth is the expected value of K-mer depth. Typically, K = 17.

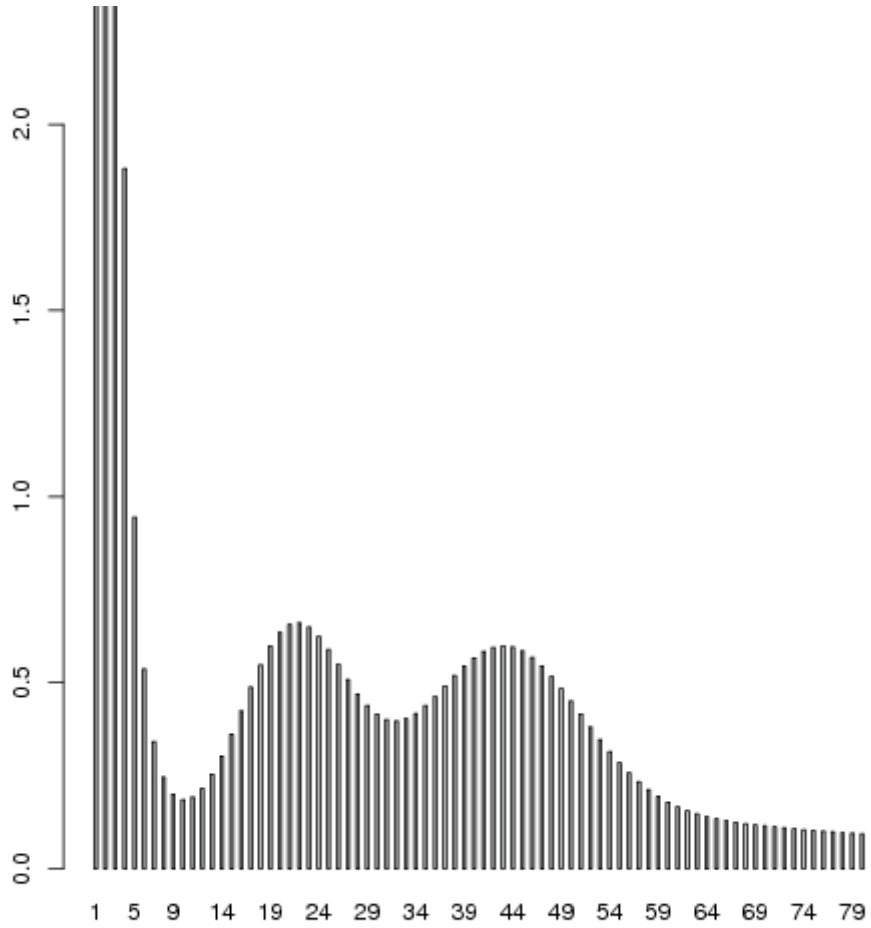

Figure 2.3.1 17-mer depth distribution

Table 2.3.1 17-mer Data statistics

K	K-mer_num	Peak_depth	Genome Size	Used Bases	Used Reads	X
17	**65404641962**	**22**	**2972938271**	**7321415145**	**488094343**	**24.6**

Total 83.29 Gb data was retained for 17-mer analysis .the 17-mer frequency distribution derived from the sequencing reads was plotted in Fig1, the peak of the 17-mer distribution is about 22, and the total K-mer count is 65404641962, then the genome size can be estimated (by formula: Genome Size=K-mer_num/Peak_depth) as 2972.38Mb.

If the heterozygous rate is higher, then a small peak will be presented at 1/2 of Peak_depth. So this K-mer analysis can be used to roughly determine the

heterozygous rate of a given genome.

Also, this distribution can be used to determine the repeat content of the genome. If this genome contains high proportion of repeat, the distribution will display a fat tail which indicate more than expect proportion of the genome have a high sequencing depth which may due to sequence similarly.

Conclusion: Genome size is 2972.38Mb.

2.4 Result of Assembly

Table 2.4.1 The result of assembly (using the data of 73.36G)

	Scaffold		Contig	
	Size(bp)	**Number**	**Size(bp)**	**Number**
max_len	154899	----	44528	----
N10	17578	5178	5374	14117
N20	11549	14492	3512	39569
N30	8268	27919	2506	76409
N40	6058	46411	1842	127098
N50	4432	71595	1357	195842
N60	3153	106496	978	290051
N70	2103	157052	672	423724
N80	1209	238185	402	632299
N90	443	410053	163	1052754
Total Size	1302267261		1083430729	
Total NumberNumber(>=100bp)	1296135		1956322	
Total Number(>=2kb)	163983		111859	
GC_rate	0.286		0.337	

Conclusion: This is a initial version of assembly without gap filling, due to the length of contig N50 is short than expected. Then WGS may not be suitable to assembly this genome.

2.5 GC-content and Sequencing depth analysis

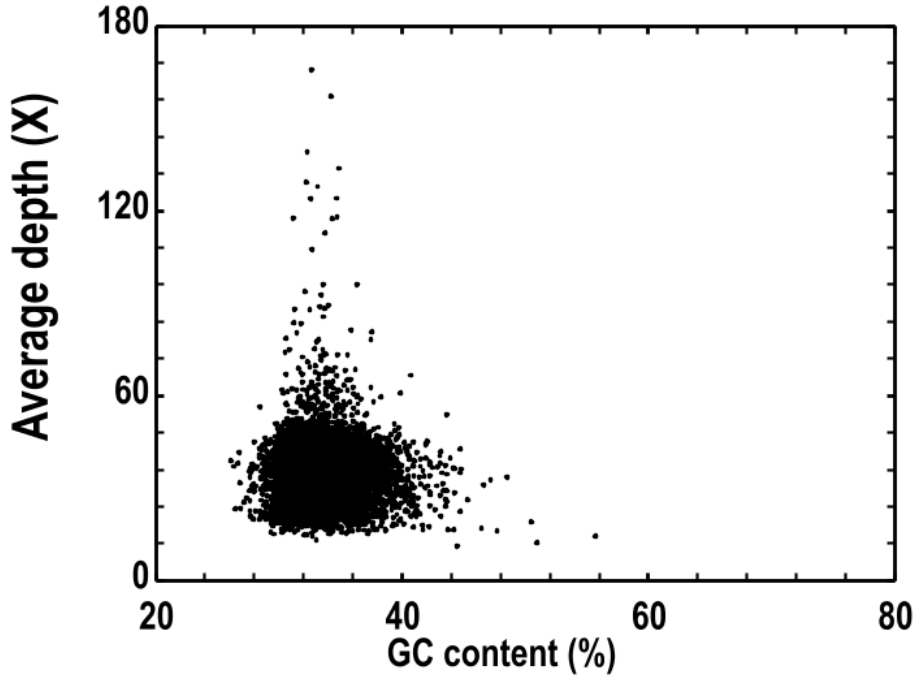

Figure 2.5.1 Distribution of GC depth.The x-axis represents as GC content;the y-axis represents the average depth. We used 10 kb non-overlapping sliding windows to calculate the GC content and average depth among the windows.

The distribution of GC content versus sequencing depth will provide an eye about the sequencing bias or contamination. Usually, the genomic region with high or low GC content will possess a low sequencing depth compare to median GC content region, if the distribution of a given genome project is different from the expected pattern, it may indicate sequencing bias of contamination. If predicted to be contaminated, then we can eliminate the polluted reads by aligned the reads against bacteria, virus and fungous database.

3 DATA DOWNLOADING

Decompress the files

Some of the documents have been compressed under Linux environment as *.gz, which can be decompressed by the following methods:

Unix/Linux user: gzip -d *.gz

Windows user: winRAR

Mac user: Shell：gzip -d *.gz

Some of the directories have been packed under Linux environment as *.tar, which can be unpacked by the following methods:

Unix/Linux user: tar -xvf *.tar

Windows user: winRAR

Mac user: Shell: tar -xvf *.tar

FTP directory structure

|-- Assembly
| |-- README.txt
| |-- Clean Data
| | |-- *lib_id_1.fq.gz.clean.dup.clean.gz*
| | |-- *lib_id_2.fq.gz.clean.dup.clean.gz*
| |-- Assembly Result
| | |-- Panopea generosa.*scafSeq*
| |-- Assembly Evaluation
| | |-- Panopea generosa.*GC_content_vs_depth.png*

Figure 5.1 FTP directory structure

4 CONTACT US

Service Hotline: 400-706-6615

Customer Service: customer@genomics.com.cn

Technical Support: tech@genomics.com.cn

Complaint Hotline: 010-80481175(Beijing) 0755-25273291(Shenzhen)