华大基因 BGI | Premier Scientific Partner

Erycina pusilla Genome Survey Report

Bioinformatics Center

Version 3.0

Jan 16, 2012

CONTENTS

1. CONVENTIONS

Table 1.1 Notational conventions used in this document

Notation	Description
Italic	A software or command of certain program. For example: “We mainly use corrected reads to get a complete assembly of the genome by *SOAP de novo* software”.
Bold-normal	Title. For example: “**1 CONVENTIONS**”.
Normal	Text. For example: “We extract sufficient DNA sample to construct different insert-size libraries”.

2. DESCRIPTION OF WORKFLOW

2.1 Pipeline of Experiment

The pipeline of the experiment is illustrated in Figure 2.1 as below:

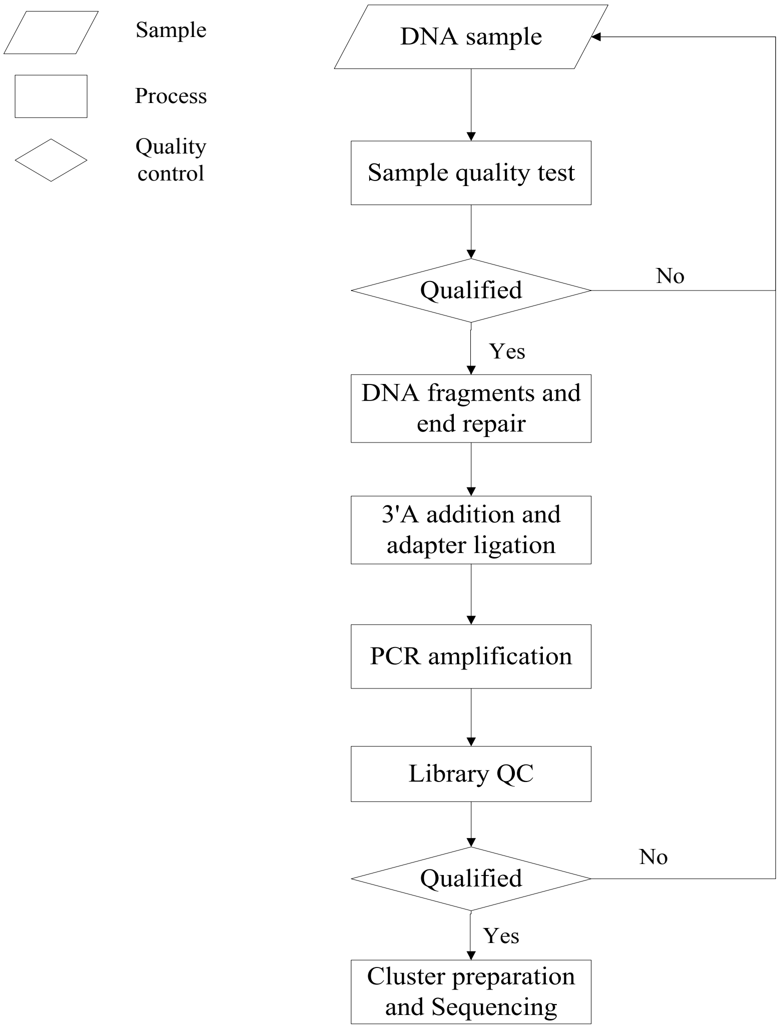

Figure 2.1 Pipeline of experiment.After the DNA sample(s) was(were) delivered, we did a sample quality test first. Then we used this(those) qualified DNA sample(s) to construct library, and we did a library quality test. At last, the qualified library would be used for sequencing.

2.2 Pipeline of Bioinformatics Analysis

The pipeline of the Bioinformatics Analysis is illustrated in Figure 2.2 as below:

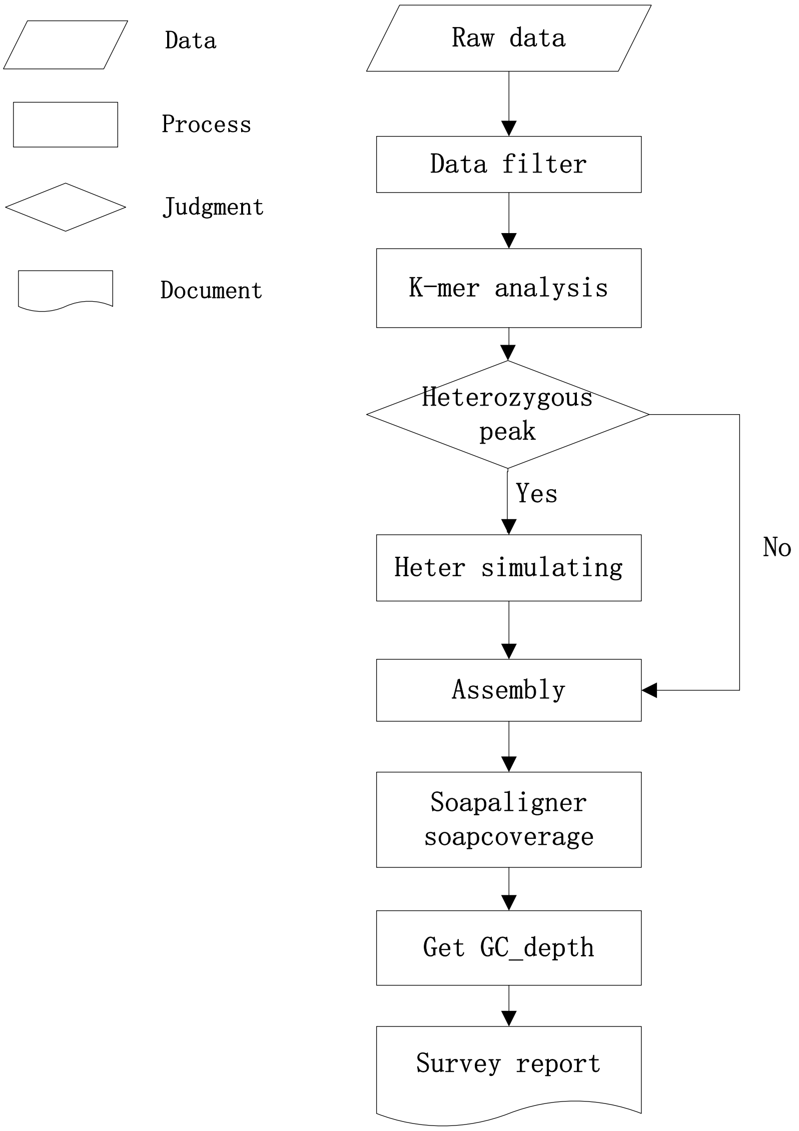

Figure 2.2 Pipeline of genome survey.When got the raw data, we filtered it first to get high quality reads. We used those clean data to do K-mer analysis and heterozygous simulation. Then we assembled them using *SOAPdenovo* software. We got the gc depth distribution by *SOAPaligner* result. After all, we knew the basic characteristics of the genome sequence and wrote the survey report.

3. BIOINFORMATICS RESULT

3.1 Background

a. Species name: *Erycina pusilla*

b. Sample name & Number:

(1)8 (0111103635): used for 170bp library

(2)10 (0111103637): used for 500bp library

c. Evaluate Genome size: 1.5Gb

d. Designated reference sequences: No

3.2 Data statistics

Table 3.1 Statistics of Clean Data

Lib ID	Insert Size(bp)	Read Length(bp)	Data(Mb)	Sequence Depth(X)
SZAXPI002377-12	170	100	39,431.64	26.29
SZAXPI002374-11	500	100	27,343.57	18.23
Total	-	-	66,775.20	44.52

This batch of sequencing produced 75.02Gb raw data. After low quality reads filtering, total 66.78Gb data was used for further analysis, if the genome size is estimated to be 1.5Gb in previous experiment, then the sequencing depth of filter data is expected to be 44.52X.

3.3 17-mer analysis 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.

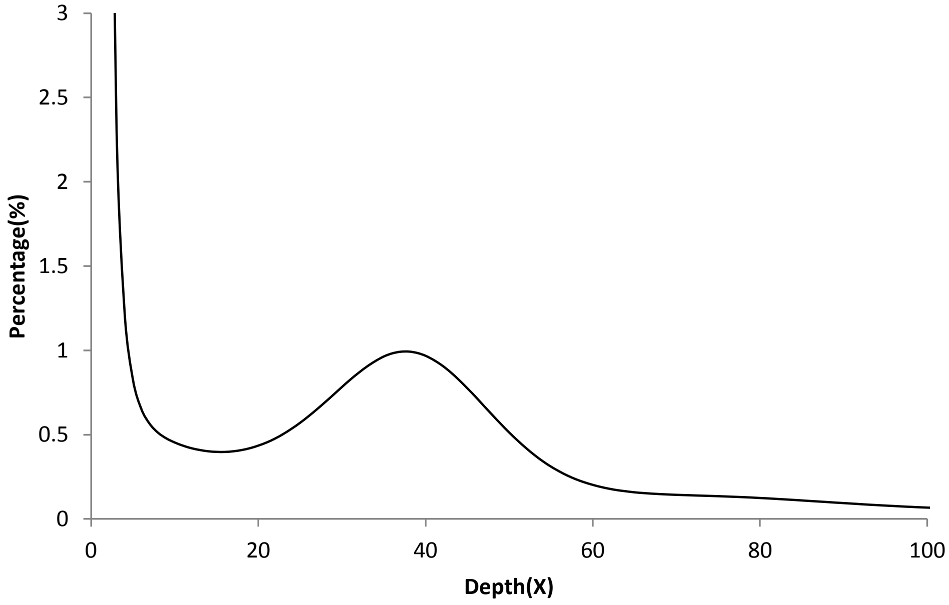

Figure 3.1 17-mer depth distribution

Table 3.2 17-mer Data statistics

K	K-mer_num	Peak_depth	Genome Size(bp)	Used Bases(bp)	Used Reads	X
17	56,059,846,752	38	1,475,259,125	66,737,912,800	667,379,128	45.24

Total 66.74Gb data was retained for 17-mer analysis. The 17-mer frequency distribution derived from the sequencing reads was plotted in Fig 3.1, the peak of the 17-mer distribution is about 38, and the total K-mer count is 56,059,846,752, then the genome size can be estimated (by formula: Genome Size=K-mer_num/Peak_depth) as 1.475Gb.

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 indicates more than expect proportion of the genome have a high sequencing depth which may due to sequence similarly.

Conclusion: Genome size is 1.475Gb. The distribution display a small tail which indicates several repeat sequences in *Erycina pusilla*. In addition, there is no obvious peak at 1/2 of peak depth, but the percentage has a slightly higher value, which indicates the genome heterozygosity of *Erycina pusilla* is slightly high.

3.4 Result of Assembly

Table 3.3 The result of assembly (using the data of 66.78Gb)

	Contig		**Scaffold**	
	Size(bp)	**Number**	**Size(bp)**	**Number**
N90	116	1,796,386	172	428,286
N80	138	1,255,155	591	203,752
N70	177	814,686	1,271	115,615
N60	260	491,191	2,337	71,454
N50	433	284,598	3,766	45,966
Longest	22,540	----	62,894	----
Total Size	681,771,944	----	756,111,956	----
Total Number(>=100bp)		2,431,836	----	1,047,905
Total Number(>=2kb)		38,521	----	80,990

Conclusion: This is an initial version of assembly without gap filling. The slightly high heterozygous rate and repeat sequences might be the reason that the contig N50 is shorter than expected.

3.5 GC-content and Sequencing depth analysis

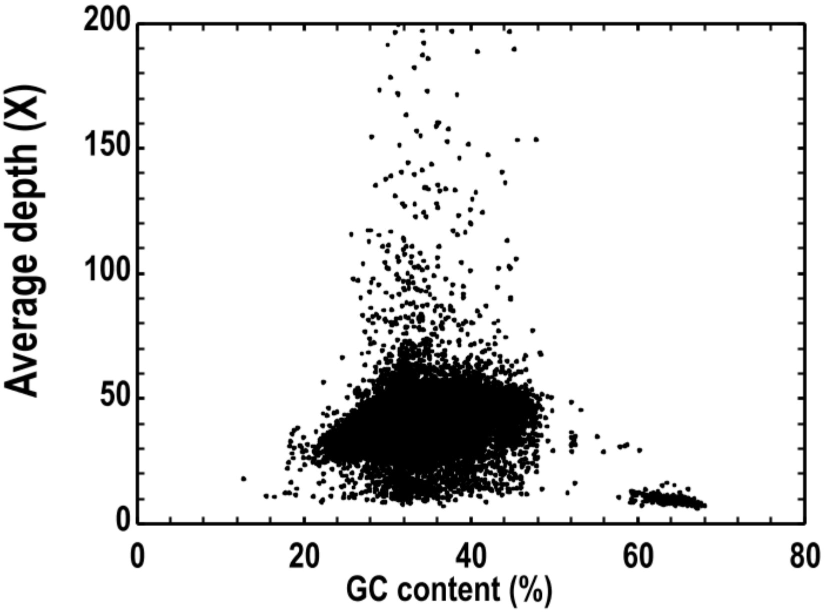

Figure 3.2 Distribution of GC depth.The x-axis represents as GC content. The y-axis represents the average depth. We used 5 kb non-overlapping sliding windows to calculate the GC content and average depth among the windows.

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.

Conclusion: According to the Distribution of GC depth, we align the sequences with GC content higher than 56% and the lower depth (less than 24) sequences against NCBI nt database. The lower depth sequences matched with *Oncidium gower ramsey* and *Cymbidium sinense*, suggests that these sequences are likely heterozygosis of this genome. Result of the sequences with GC higher than 56% shows that the samples were contaminated by *Mycobacterium*.

3.6 Conclusions and recommendations

1. The genome size of *Erycina pusilla* can be estimated as 1.475Gb.
2. Considering the distribution of 17-mer depth and the GC_depth, we infer the *Erycina pusilla* contains slightly high heterozygous rate.
3. According to the current data and assembly results, we think it might be a little risky to use the whole genome shotgun assembly strategy to complete the genome assembly.

4. DATA DOWNLOADING

Download

Host: *http://cdts.genomics.org.cn/*

ID: XXX

Password: XXX

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
|       |       |-- Raw Data
|       |       |  |-- lib_id/*_1.fq.gz
|       |       |  |-- lib_id/*_2.fq.gz
|       |       |  |-- InsertSize.txt
|       |       |-- Clean Data
|       |       |  |-- lib_id/*_1.fq.gz.clean.dup.clean.gz
|       |       |  |-- lib_id/*_2.fq.gz.clean.dup.clean.gz
|       |       |  |-- InsertSize.txt
|       |       |-- Assembly Result
|       |       |  |-- [Species name].scafSeq.gz
|          |          |  |-- [Species name].scaftig.gz
|          |          |-- Assembly Evaluation
|          |          |  |-- [Species name].GC_content_vs_depth.png
```

Figure 4.1 FTP directory structure

5. 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)