



Re-sequencing Filter Report

2021/9/25



@2021 BGI All Rights Reserved



Table of Contents

Results	3
1 Data production	3
2 Quality control	3
Methods	3
1 Experimental procedure of re-sequencing	4
2 Bioinformatic analysis workflow	4
3 Parameters for data filtering	4
Help	4
1 FASTQ format	4
FAQs	5
References	5

Results

1 Data production

After sequencing, the **raw reads** were filtered. Data filtering includes removing adaptor sequences, contamination and low-quality reads from **raw reads**. Next, we get the statistics of data production. Table 1 shows statistical results after data treatment.

Table 1 Project information ([Download](#))

information	detail
Project_id	F21FTSEUHT1365
Subproject_id	MOUuftvR
sample_number	1

Table 2 Statistics of clean data ([Download](#))

Sample Name	Clean Reads	Clean Base	Read Length	Q20(%)	GC(%)
G	611,166,828	91,675,024,200	150;150	94.44	42.00

Note: The above table is the information of all samples, please click download to view. The annotation of header is as follows:

Sample Name: the name of sample

Clean Reads: total number of sequencing reads

Clean Bases: the product of sequencing reads and reads length, representing the total number of sequencing bases

Read Length: the length of reads

Q20_rate(%): the proportion of nucleotides with a quality score ≥ 20 in all nucleotides

GC_rate(%): the proportion of nucleotides G and C in all nucleotides

2 Quality control

The quality of data was examined after filtering.

2.1 The distribution of base percentage and qualities along reads after data filtering

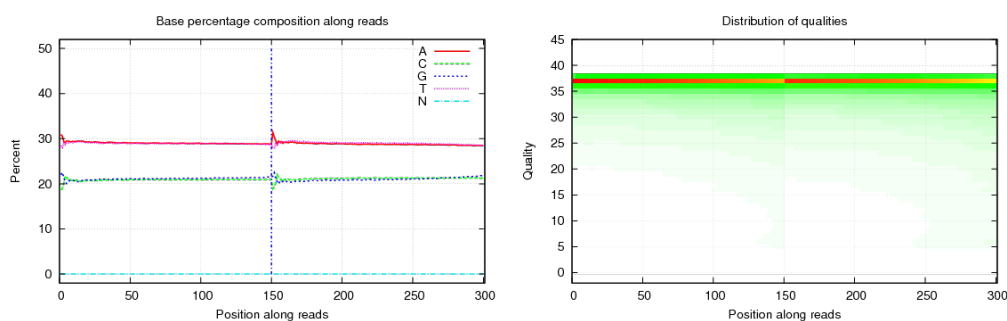


Figure 1 G base percentage composition (left) and base quality distribution (right).

Note: In the left figure, x-axis represents base position along reads, y-axis represents base percentage at the position; each color represents a type of nucleotide. Under normal conditions, the sample does not have AT/GC separation. It is normal to see fluctuations in the first several bp positions, which is caused by random primer and the instability of enzyme-substrate binding at the beginning of the sequencing reaction. In the right figure, x-axis represents base position along reads, y-axis represents base quality; each dot represents the base quality of the corresponding position along reads, color intensity reflects the number of nucleotides, a more intense color along a quality value indicates a higher proportion of this quality in the sequencing data.

Methods



1 Experimental procedure of re-sequencing

2 Bioinformatic analysis workflow

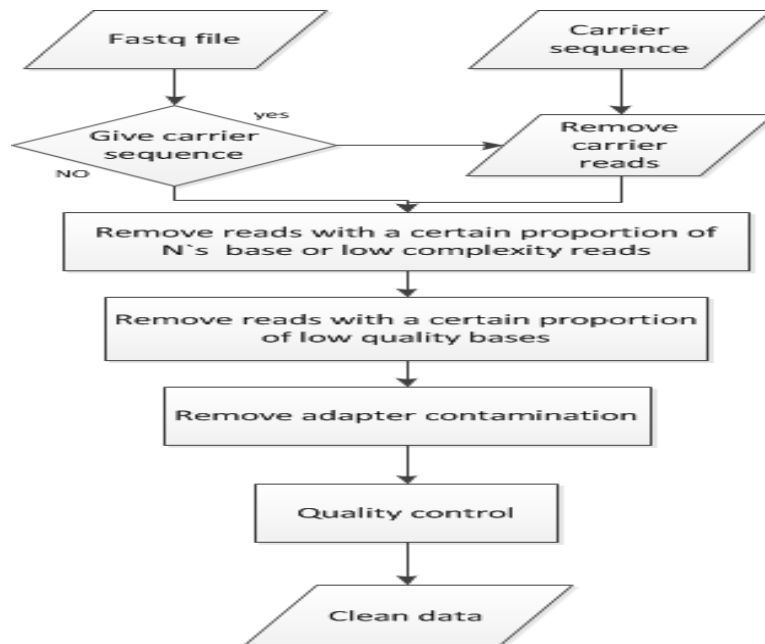


Figure 1 Bioinformatic analysis workflow

3 Parameters for data filtering

raw data with adapter sequences or low-quality sequences was filtered. We first went through a series of data processing to remove contamination and obtain valid data. This step was completed by SOAPnuke software developed by BGI.

SOAPnuke software filter parameters : "-l 20 -q 0.5 -n 0.03 -A 0.28", steps of filtering:

- 1) Filter adapter: Delete the entire read if more than 28% match the adapter sequence;
- 2) Filter low-quality data: Delete the entire read if there are more than 50% bases having a quality value lower than 20;
- 3) Remove N: Delete the entire read if there are more than 3% N in the read;
- 4) Obtain **clean reads**.

Help

1 FASTQ format

Images generated by sequencers are converted by **base calling** into nucleotide sequences, which are called **raw data** or **raw reads** and are stored in **FASTQ** format. **FASTQ** files are text files that store both reads sequences and their corresponding quality scores. Each read is described in four lines as follows:

```
@FCB068CABXX:6:1101:1403:2159#TAGGTTAT/1
GTAGAAGACTTATAGATTAAAATTCTCCAACATATAGATGTCCTTACACCGTTTTCTTTGCTCAGCAGGCTCCGT
GTTTGCTTGTCCTT
+
c`bcc_c^ccde_df\c_aeff`ffcffdfedadca^`b_eed`fe\fed\babdba^Yeebeccfdeae_eec^dbXbda`]bcbebc
```



The line 1 and 3 are sequence name generated by the sequencer; line 2 is sequence; line 4 is **sequencing quality** scores, in which every letter corresponds to a base in line 2; the base's **sequencing quality** is the **ASCII** value that the letter in line 4 refers to minus 64 (Specification). For example, the **ASCII** value of c is 99, so the corresponding **sequencing quality** value is 35. Solexa quality value of sequencing bases ranged from 2 to 35. Table 1 shows the brief correlation between **sequencing error** rate and **sequencing quality** value. Denote E as **sequencing error** rate and sQ as **sequencing quality** value, then we have:

$$SQ = -10 \times (\log \frac{E}{1-E}) / (\log 10)$$

$$E = \frac{Y}{1+Y}$$

$$Y = \frac{SQ}{e^{-10 \times \log 10}}$$

Table 1 Brief relationship between sequencing error rate and sequencing quality ([Download](#))

Sequencing error rate	Sequencing quality value	Character(Phred64)	Character(Phred33)
5%	13	M	.
1%	20	T	5
0.1%	30	^	?

FAQs

References

Request for information or Quotation

Contact your BGI account representative for the most affordable rates in the industry and to discuss how we can meet your specific project requirements or for expert advice on experiment design, from sample to bioinformatics.

info@bgi.com

www.bgi.com

BGI Americas
One Broadway, Cambridge, MA
02142, USA
Tel: +1 617 500 2741

BGI Europe
Ole Maaløes Vej 3, DK-2200
Copenhagen N, Denmark
Tel: +45 7026 0860

BGI Asia-Pacific
16 Dai Fu Street, Tai Po Industrial
Estate, New Territories, Hong
Kong
Tel: +852 36103510

Copyright ©2019 BGI. The BGI logo is a trademark of BGI. All rights reserved.
All brand and product names are trademarks or registered trademarks of their respective holders.
Information, descriptions and specifications in this publication are subject to change without notice.
DNBseq is a trademark of MGI Co., Ltd.
Published July 2019.

We Sequence, You Discover.
BGI