



# Statistical Report

## F26A430000218\_HOMqclaR

2026/5/19



@2026 BGI All Rights Reserved

## Table of Contents

Results	3
1 Sequencing production	3
Methods	4
1 Background Introduction	4
2 Whole Genome Sequencing Technolog	4
3 Overview of Bioinformatics Analysis	6
Help	6
1 How to visualize genomic variation	6
2 How to unzip files	7
3 List Of Delivery Data File	7
4 Cyclone platform data quality value	7
FAQs	8
References	8

## Results

### 1 Sequencing production

In this project, a total of 2 DNA samples were sequenced using CycloneSEQ platform. The data output and quality statistics are as follows: CYC sequencing produces an average of 2,359,869 reads per sample, and the total number of bases is 43.44 Gbp. The average N50 of sequencing read reaches 25,915 bp, the average read length reaches 18,421 bp, and the average sequencing quality reaches 17.55. See the table **Table1** for specific statistical results. The sequence read length distribution is shown in the figure (**Figure1**). The estimate of read length and sequencing mass nuclear density (kde) is shown in (**Figure2**).

**Table 1** Data Statistics [\(Download\)](#)

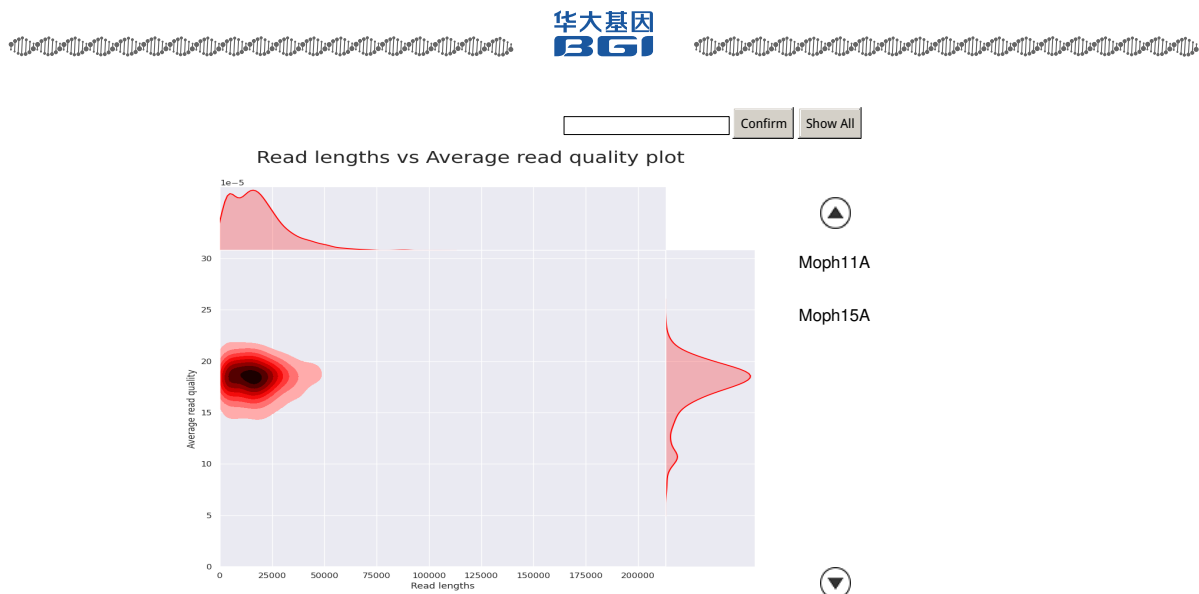
Sample	Total bases (Gb)	Read length N50 (bp)	Mean read length (bp)	Median read length (bp)	Mean read quality	Median read quality	Number of reads
Moph15A	42.62	26,727	18,805	15,736	17.60	18.20	2,266,456
Moph11A	44.25	25,103	18,037	15,471	17.50	18.20	2,453,283
Average	43.44	25,915	18,421	15,603	17.55	18.20	2,359,869

The annotation of table is as follows:  
 Samples: Sample ID  
 Total bases ( Gb ) : Total bases  
 Read length N50 ( bp ) : N50 length of reads  
 Mean read length ( bp ) : Mean length of reads  
 Median read length (bp): Median length of reads  
 Mean read quality: Mean quality of reads  
 Median read quality: Median quality of reads  
 Number of reads: Total reads



**Figure 1** Read length distribution .

The X-axis represents the length of sequencing reads (after log conversion), and the Y-axis represents the number of reads with the corresponding length. Under normal circumstances, the average length of DNA we can obtain is more than 10,000 bp (different according to different special requirements), reflecting the main peak of the measured read length distribution in the figure to the right of 10,000 bp.



**Figure 2 Read length and sequencing quality nuclear density estimation.**

The X-axis represents the length of reads (the upper curve of the figure is the estimated distribution of read length and nuclear density), and the Y-axis represents the read sequencing quality (the curve on the right of the figure is the estimated distribution of reads sequencing quality > quantity nuclear density). The picture is a two-dimensional (length, quality) nuclear density estimation map. In theory, the higher the core, the better the sequencing quality of most reads, and the narrower the vertical distribution (especially the core area). It shows that the quality of sequencing is more stable.

## ● Methods

### 1 Background Introduction

With the development of high-throughput sequencing technology, the limitations of short-read sequencing are in complex regions, such as high repeats and high GC. While there are some problems of short-read sequencing, except for short read lengths, as well as it can not span high repeats and low complexity regions. There also are certain limitations in the detection of large structural variants (SV). The rapid development of long-read sequencing in recent years has solved these problems at this stage. Long-read sequencing uses modern optics, polymers, nanotechnology and other means to distinguish the difference of base signals, so as to directly read sequence information.

### 2 Whole Genome Sequencing Technolog

CycloneSEQ™ whole-genome resequencing products, using CycloneSEQ™ for long-read sequencing, which can effectively detect large structural variations such as insertions, deletions, inversions, and duplications.

#### 2.1 CycloneSEQ™ Sequencing

**Figure1** shows the CycloneSEQ™ library construction and sequencing process. The library construction kits provided with the CycloneSEQ™ platform can be used for library construction. The specific steps are as follows:

1 Size-Selection;

2 Library construction:



- 1) DNA Damage repair , End Repair, A-Tailing and clean-up;
- 2) Adapter Ligation;
- 3) Magnetic Beads Purication and Qubit Quantification
- 3 Priming and loading the flow cell;
- 4 On-board sequencing of the library.

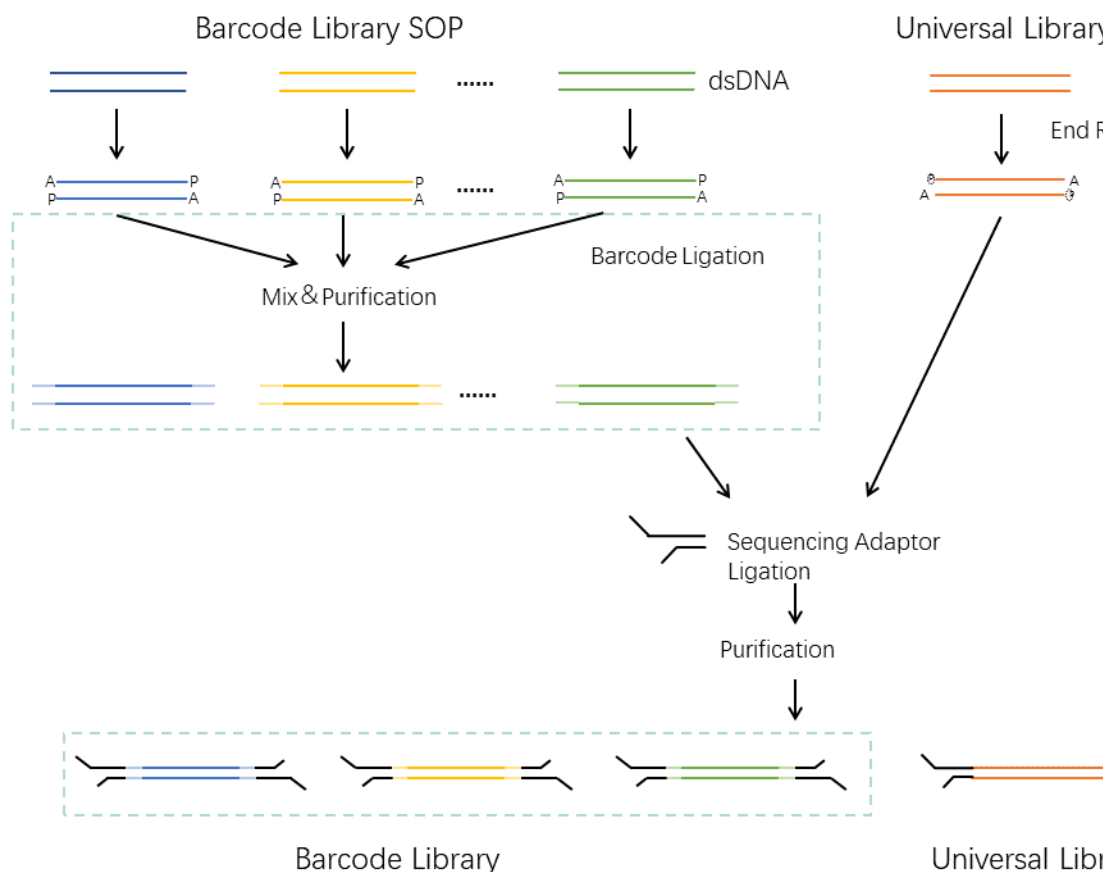


Figure1 CycloneSEQ™ Library Construction.

The sequencing process on CycloneSEQ™ is shown as **Figure2**:

CycloneSEQ™ is a nanopore sequencing technology based on the principle of nanopore sensors. This technology utilizes pore proteins spanning a nanoscale biomimetic membrane as the core sensor. During the sequencing process, voltage is applied on both sides of the biomimetic membrane. The double-stranded DNA/RNA bound to the motor protein is captured by the nanopore protein and unwound under the action of the motor protein. Driven by the electric field force, nucleic acids pass through the nanopore as continuous single-stranded molecules at a speed controlled by the motor protein. When nucleic acids pass through the pore, the current changes, and the current changes caused by different base sequences are different. A base calling algorithm is established through a machine learning model, and the sequence of the nucleic acid molecules passing through the hole is inferred from the current

changes. Thus, real-time and accurate sequencing of single-stranded nucleic acid molecules can be achieved.

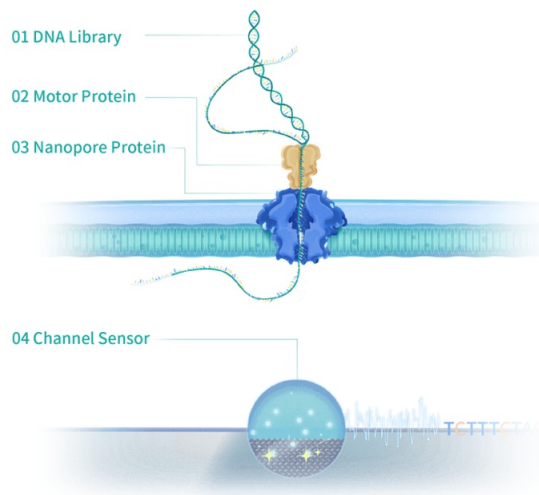


Figure2 CycloneSEQ™ Sequencing Process.

### 3 Overview of Bioinformatics Analysis

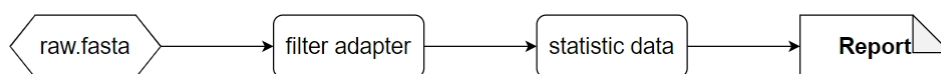


Figure3 Bioinformatics Analysis Pipeline.

**Figure3** shows the pipeline of Cyclone sequencing data filter bioinformatics analysis. For the download data of Cyclone long-read sequencing, first porechop<sup>[1]</sup> was used to filter adapter sequences, then seqtk<sup>[2]</sup> software was used to filter sequences below 100bp and NanoPlot<sup>[3]</sup> was used to do statistics with the filtered data. The software and parameters involved in each step are described as follows:

#### 3.1 Data Filtering

On the basis of passing the filter condition (fastq\_pass), porechop<sup>[1]</sup> software was used to remove the ligation, and seqtk<sup>[2]</sup> was used to filter the sequences less than 100 bp, and finally NanoPlot<sup>[3]</sup> software was used to stat the result. The following are the command line parameters:

```

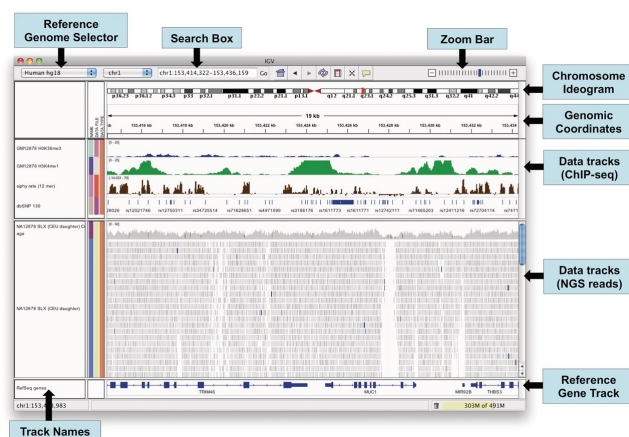
porechop -i fastq_pass.raw.FASTQ.gz -t 1 -o fastq_pass.FASTQ.gz
seqtk seq -L 100 fastq_pass.FASTQ.gz | gzip -c -> input.FASTQ.gz
NanoPlot --color red --N50 --FASTQ input.FASTQ.gz --outdir ./
  
```

## ● Help

### 1 How to visualize genomic variation



IGV (Integrative Genomics Viewer) is a high-performance genomic data visualization tool that can help users to synthesize and analyze different types of genomic data at the same time, and can flexibly amplify a specific area of the genome. IGV software can be downloaded for free from: <http://www.broadinstitute.org/igv>. IGV can view SAM/BAM comparison files and VCF mutation detection files. The figure below shows the IGV visualization window.



## 2 How to unzip files

All data is compressed into \*.tar.gz file format with "tar -czvf" under linux system, please decompress according to the following method:

Unix/Linux user: tar -zxvf \*.tar.gz

Windows user: suggest winRAR software

Mac user: shell command: tar -zxvf \*.tar.gz, suggest: 'stuffit expander'

## 3 List Of Delivery Data File

The following list shows the structure of delivery data file list, you can focus on the structure for the directories which are included in the directory.

```
cell1:
|-TB2000C007-202405241447070_read.fq.gz      ---fastq file
|-HistogramReadlength.png                    ---histogram of read lengths
|-LengthvsQualityScatterPlot_kde.png          ---read lengths vs average read quality plot
|-LogTransformed_HistogramReadlength.png      ---histogram of read length after log transformation
|-NanoStats.tsv                              ---read data statistics
\md5check                                    ---md5sum check file
```

## 4 Cyclone platform data quality value

The base quality of Cyclone **FASTQ** data is calculated using Phred. The Phred value is calculated during the base calling process using a model that predicts the probability of base errors. If the sequencing error rate is represented by E and the base quality value is represented by SQ, the following relationship exists:

$$SQ = -10\log_{10}E$$

The concise correspondence between the quality value and Phred score of the base is shown in the following table:

Phred Score	Incorrect Base Identification Rate	Correct Base Identification Rate
7	1/5	80.00%
10	1/10	90.00%
20	1/100	99.00%

## FAQs

### How to open files in BAM format in Microsoft Windows?

You can use the Picard tool to create an index file \*.bai for the BAM file, and then use the visualization software (IGV)

### If I extract it myself, how should long DNA fragments be stored?

Extracted long fragment gDNA samples can be stored at 4°C for half a year and at -20°C for one year. However, it is recommended to freeze and thaw only once for storage at -20°C. Repeated freezing and thawing will cause the sample to degrade and the sample DNA cannot be stored. Shake or mix vigorously to avoid fragmentation of long DNA fragments. We do not recommend you to send DNA samples, because the length of DNA directly affects the length of phasing. DNA samples can be damaged in transit.

## References

- [1] Porechop
- [2] seqtk
- [3] Wouter De Coster, Sven D'Hert, Darrin T Schultz, Marc Cruts, Christine Van Broeckhoven, NanoPack: visualizing and processing long-read sequencing data, Bioinformatics, Volume 34, Issue 15, 01 August 2018, Pages 2666–2669.