Statistical Report

2023/12/7

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 Revio 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 PacBio(Pacific Bioscience) Revio platform. The data output and quality statistics are as follows:

We use the PBHiFi Revio sequence platform, produces 4,693,444 hifi ccs reads of the SM8T sample, and the total number of bases is 72.34 Gbp. The N50 of sequencing read reaches 15,419 bp, the mean read length reaches 15,413 bp, and the mean read quality reaches 30.80.

We use the PBHiFi Revio sequence platform, produces 2,876,869 hifi ccs reads of the SM8L sample, and the total number of bases is 39.47 Gbp. The N50 of sequencing read reaches 13,717 bp, the mean read length reaches 13,721 bp, and the mean read quality reaches 32.00.

See the table **Table1** for specific statistical results. The sequence read length distribution is shown in the **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
SM8T	72.34	15,419	15,413	14,829	30.80	30.10	4,693,444
SM8L	39.47	13,717	13,721	13,660	32.00	32.00	2,876,869

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

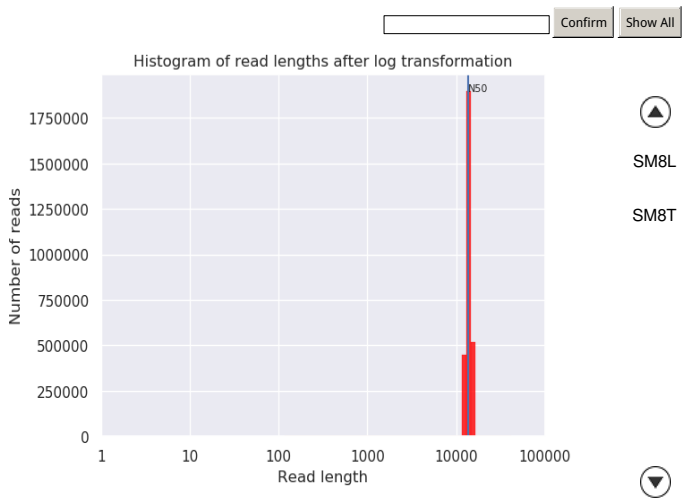


Figure 1 Read length distribution .

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.Long-read technology can effectively solve some of the insurmountable problems in short-read sequencing. With the emergence of PacBio HiFi sequencing technology and Nanopore ***Q20*** sequencing technology, long-read sequencing will become one of the main methods in the future.

2 Whole Genome Sequencing Technolog

PacBio human whole-genome resequencing products, using PacBio (Pacific Bioscience) for long-read sequencing, which can effectively detect large structural variations such as insertions, deletions, inversions, and duplications.

2.1 PacBio Sequencing

The PacBio library preparation process is shown as **Figure1**:

1. gDNA shearing and clean-up（HiFi）;

2．Remove Single-Strand Overhangs;

3．DNA Damage repair , End Repair, A-Tailing;

4．Adapter connection: the double-stranded positive and negative chains to obtain a dumbbell-like ("Tortle horse Ring") structure, called SMRT Bell;

5．Size-Selection;

6．Anneal and Bind SMRTbell Library;

7．Prepare for Sequencing.

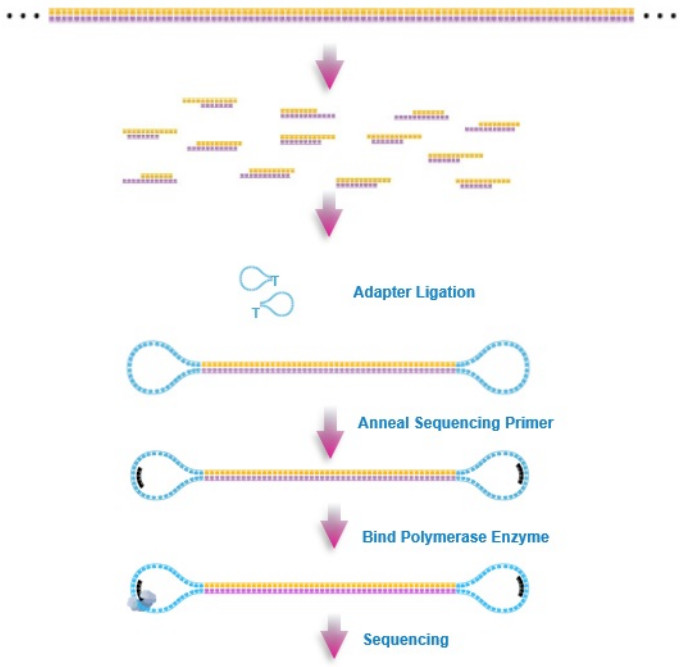


Figure 1 PacBio Library Construction.

The sequencing process on PacBio is shown as **Figure2**:

The PacBio library sequence is a circular structure, and the double-stranded structure is connected into a loop by the SMRT linker. In the sequencing process, the circular molecule will be tested many times, so the original data of the sequencer is to retain the sequence fragments of the sense strand and antisense strand, 3' and 5' linkers, and the Polymerase of the SMRT linker sequence reads. Subreads are to obtained after removing the linker, which is the insert sequence. If a circular molecule is detected more than once, then a circular molecule will produce multiple Subreads. Subreads belonging to the same ring molecule will be clustered and corrected, and then represented by a Read Of Insert (ROI).

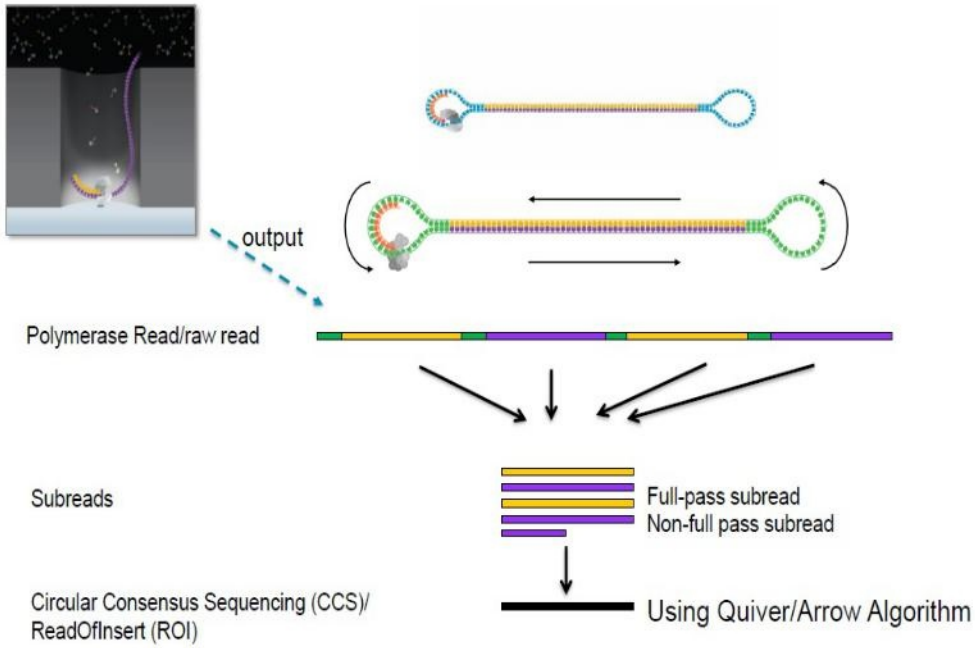


Figure 2 PacBio Sequencing Process.

3 Overview of Bioinformatics Analysis

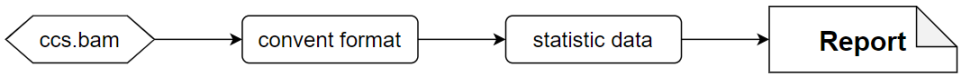


Figure 3 Bioinformatics Analysis Pipeline.

Figure3 shows the pipeline of whole genome sequencing bioinformatics analysis. For the download data of PacBio sequencing, first smrtlink[1] was used to change data to ***FASTQ*** format, then seqtk[2] software was used to filter sequences below 500bp and NanoPlot[3] was used to do statistics with the filtered data. In addition, in order to ensure high-quality sequencing data, a strict data quality control system (QC) is set up throughout the analysis process. The software and parameters involved in each step are described as follows:

3.1 Data Filtering

smrtlink software was used to change bam file to ***FASTQ*** format, and seqtk was used to filter the sequences less than 500 bp, and finally NanoPlot software was used to stat the result.The following are the command line parameters:

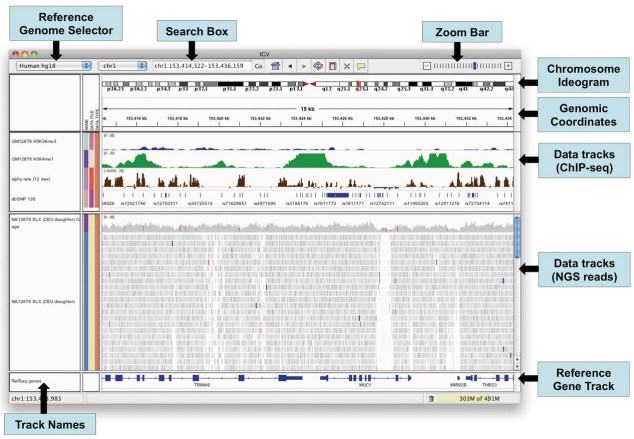
```
bam2fastq -o fastq_pass. FASTQ .gz subreads.bam
seqtk seq -L 500 fastq_pass. FASTQ .gz |gzip -c - > input. FASTQ .gz
NanoPlot -t 1 --color red --N50 -- FASTQ input. FASTQ .gz --outdir ./
```

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 usr: suggest winRAR software

Mac user: shell command: tar -zxvf *.tar.gz锛宻uggest: 'stuffit expander'

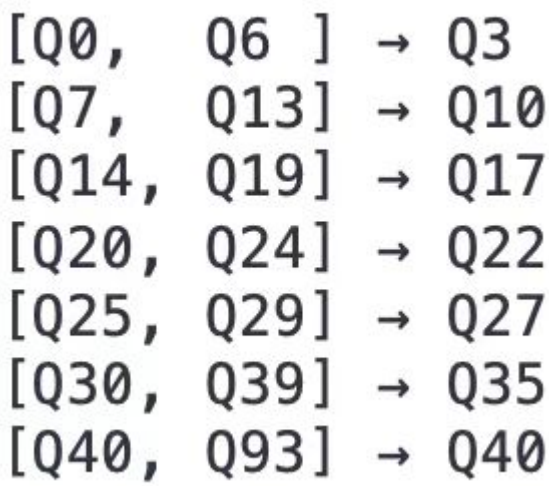
3 List Of Delivery Data File

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. Hifi_reads.bam files representing the ccs-reads will be produced natively by the PacBio Revio instrument.

```
cell1:
    |-m54160_170829_063305_s1.hifi_reads.bc1003.bam          ---ccs.bam file
    |-m54160_170829_063305_s1.hifi_reads.bc1003.bam.pbi      ---ccs.bam file index
    |-m54160_170829_063305_s1.hifi_reads.bc1003.fastq.gz     ---ccs.fastq file
    |-HistogramReadlength.png                                ---histogram of ccs read lengths
    |-LengthvsQualityScatterPlot_kde.png                     ---ccs read lengths vs average read quality plot
    |-LogTransformed_HistogramReadlength.png                 ---histogram of ccs read length after log transformation
    |-NanoStats.tsv                                          ---ccs data statistics
    \-md5check                                               ---md5sum check file
```

4 Revio platform data quality value

In order to save storage resources, PacBio-revio platform merges and simplifies the base quality values after the CCS consistency analysis. The following table is the simplified corresponding table. Therefore, the overall quality of the CCS data quality values is lower than the real data quality values.



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] smrtlink
[2] seqtk
[3] Wouter De Coster, Svenn 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.

Technical Support E-mail:info@bgitechsolutions.com
Website: www.bgitechsolutions.com