SubOrder_Pacbio Sequencing Result Report

Experimental Methods

After preliminary preparation of the DNA, a SMRTbell library was constructed using the SMRTbell Express Template Prep Kit 2.0 kit. The library quality was then measured by Qubit3.1 and Qseq100 bioanalyzers. Subsequently, the constructed SMRTbell library was combined with the primers and polymerase of the SMRTbell Express Template Prep Kit 2.0 kit and was loaded to the SMRT cell of the PacBio platform in a free diffusion way for sequencing.

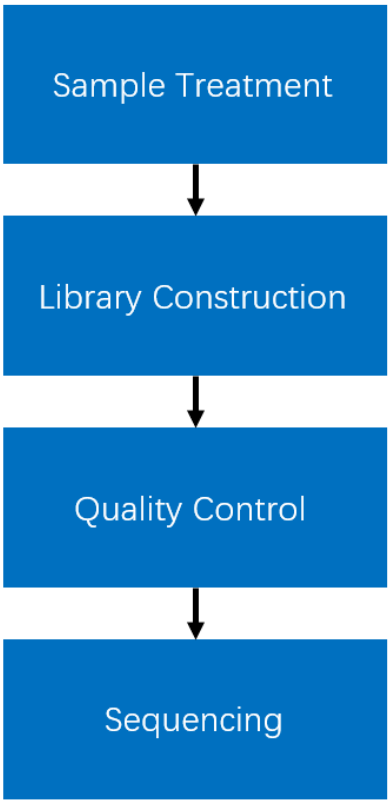

Figure 1. Experimental Procedure

Bioinformatics Analysis

（1） Circular Consensus Sequencing: CCS(4.2.0) [1] software was used to generate a consensus (CCS) for each strand within the same zero mode waveguide;

（2） Read Mapping: Mapping Circular Consensus Sequencing to the reference sequences was done with minimap2 (version 2.15-r905) [2] software;

（3） Validation analysis: The mapping rate, Coverage Analysis, and Variant Calling information were calculated based on the mapping results after quality control by Samtools (Version: 1.9) [3] and visualized for display thereafter;

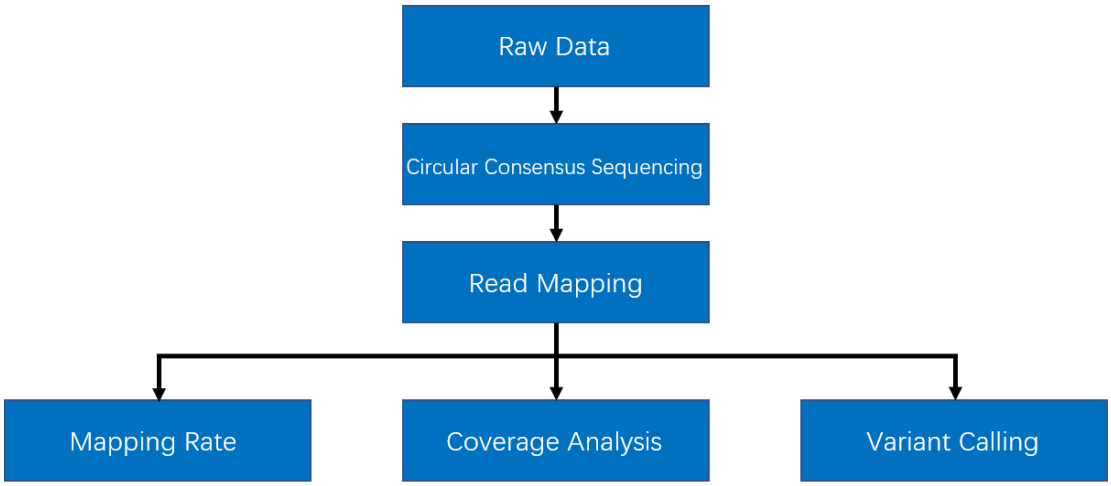

Figure 2. Bioinformatics Analysis Process

Results

According to the comparison results, the coverage and variation information by sequencing data was determined for the reference sequence, these statistical results are shown in the file " SubOrder_snp_indel.xlsx".

Abbreviations used in the file:

（1） Id: Reference sequence name;
（2） Pos: The reference position, with the 1st base having position 1. Positions are sorted numerically, in increasing order, within reference;
（3） Ref: Reference base;
（4） Depth: Read depth at this position for this sample;
（5） A: Number of reads that support adenine at this position;
（6） C: Number of reads that support cytosine at this position;
（7） G: Number of reads that support guanine at this position;
（8） T: Number of reads that support thymine at this position;
（9） MutRate(%): Percentage of reads that support non-reference base at this position;
（10） Mut: Mutation genotype and percentage;
（11） InsCov: Counts of reads that support insertion at this position;
（12） InsRate(%): Percentage of reads that support insertion at this position;
（13） Ins: Insertion genotype and percentage;
（14） DelCov: Counts of reads that support deletion at this position;
（15） DelRate(%): Percentage of reads that support deletion at this position;
（16） Del: Deletion genotype and percentage;

The coverage and variation information of the reference sequence by sequencing data were visualized according to the statistical results. The results are shown in Figure 3, where the abscissa is the position information of the reference sequence (unit bp), and the ordinate is the corresponding statistical value. Please note, the error rate within homopolymeric regions might be relatively high in PacBio sequencing. This type of region uses the genotype corresponding to the maximum allele frequency as the candidate result. When the candidate result is inconsistent with the reference sequence, it is recommended to use Sanger sequencing for validation and integrate the results. The positive mutation information detected will be marked on the upper left of the picture, includes the location, type and proportion of the variation. For example, 10362 G->A 99.77%, indicates that at base pair position 10,362 in the reference sequence there is a guanine nucleobase, while 99.77% of the sequencing data support an adenine nucleobase at this position in the sample. The insertion-deletion type variation only indicates the base type(s) supported by the sequencing data at the corresponding position in the sample. The result (Figure 3) below demonstrates that the synthesized sequence is accurate.

Figure 3. Overlay of Reference Sequence by Sequencing Results

Regarding the sample sequence, the consensus sequence of the genotype with maximal allelic frequency at each locus can be found in SEQ format in the file " SubOrder_cons_seq.seq ".

References

[1] Gorrieri R , Versari C . CCS: A Calculus of Communicating Systems[J]. Springer International Publishing, 2015:81-161.

[2] Heng. Minimap2: pairwise alignment for nucleotide sequences[J]. Bioinformatics, 2018, 34(18):3094-3100.

[3] Danecek P , Bonfield J K , Liddle J , et al. Twelve years of SAMtools and BCFtools[J]. GigaScience, 2021, 10(2):1-4.