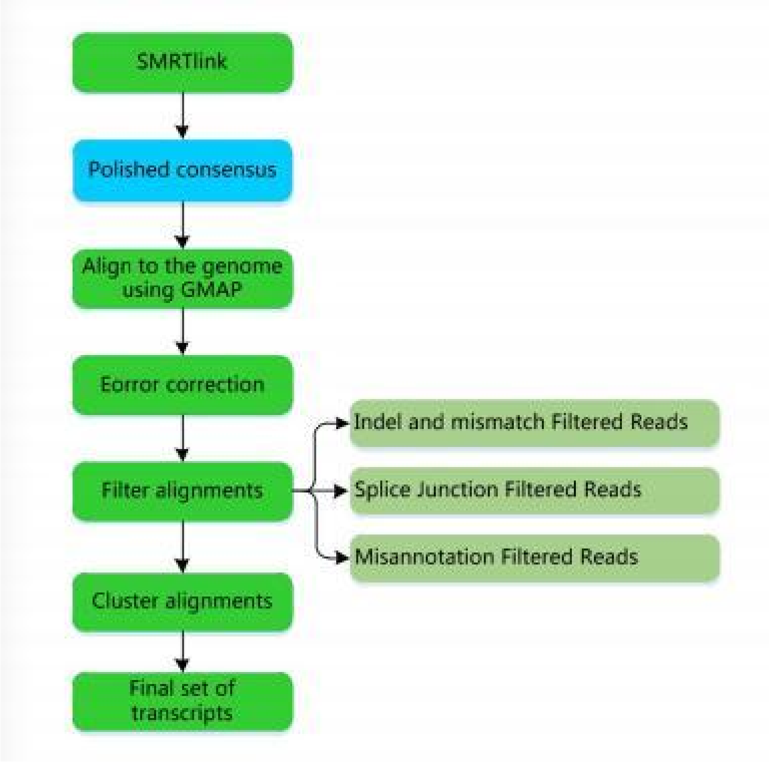
Report data flow combing with reference questions


SMRTlink
Polished consensus
Align to the genome using GMAP
Eorror correction
Filter alignments
Indel and mismatch Filtered Reads
Splice Junction Filtered Reads
Misannotation Filtered Reads
Cluster alignments
Final set of transcripts

Statistics in final report	sample data	Data Sources	Data description
Polymerase Reads	423875	Table 3.1.1 Statistical results of Polymerase read	1
Polymerase Read Bases(G)	23.99**G**	Table 3.1.1 Statistical results of Polymerase read	2
Subreads number	13784046	Table 3. 1.2.1 Subreads statistical results	3
Subreads base(G)	22.95**G**	Table 3. 1.2.1 Subreads statistical results	4
CCS_number	374105	Table 3.1.3.1 CCS statistical results	5
FLNC_number	315540	Table 3.1.4.1 Statistical results of FLNC	6
Consensus_number	32540	Table 3. 1.6.1 Polished consensus statistical results	7
Total_number	32540	Table 3.2.1 Statistical table of transcript length distribution before and after correction	8
Total mapped	31850	Table 3.3.1.1 Statistics of GMAP comparison results	9
Isoforms number	16027	Table 3.4.1.1 Statistics of classification results of full-length transcripts	10

1 The number of high-quality sequencing reads produced by a single molecule during the sequencing process, each read is a multi-copy sequence containing adapters obtained by cycle sequencing;

2 Polymerase Read data volume, obtained by multiplying Polymerase Reads by Polymerase Read Length (mean);

3 Remove the linker in Polymerase Read and the number of reads with a length less than 50bp to obtain the number of subreads. After removing the linker, the multi-copy sequence is interrupted, so the number of subreads is much greater than the number of Polymerase Reads;

4 Subreads data volume (same as Polymerase Read data volume);

5 The number of a consistent sequence obtained by self-calibration of multiple Subreads sequences in each ZMW (zero-mode waveguide hole) hole. In principle, each hole has and only one CCS;

6 Full-length non-chimeric sequences in CCS sequences (CCS also includes full-length chimeric sequences and non-full-length sequences);

7 FLNC sequence clustering to remove redundancy and correct the number of Consensus;

8 The number of reads before and after Consensus correction of the second-generation data, since the bases in the reads are corrected, the number of reads before and after correction remains unchanged;

9 Total mapped is the number of reads aligned to the reference genome. After the aligned reads are corrected and filtered, the resulting reads are shown in the result file 04.Structure/01.transcripts/sample/sample.id2id.xls (see the table below)

first column of;

Read_ id	Isoform_ id	Gene_ id
transcript_HQ_mix_pool_transcript 13753/f5p0/2040	TEA013673_novel01	TEA013673
transcript_HQ_mix_pool_transcript 15412/f2p0/ 1948	TEA013673_novel01	TEA013673
transcript_HQ_mix_pool_transcript 16337/f2p0/ 1889	TEA013673_novel01	TEA013673
transcript_HQ_mix_pool_transcript 12836/f6p0/2095	TEA013673_novel02	TEA013673

10 In the above table, one Gene_id corresponds to multiple Isoform_ids, and one Isoform_id corresponds to multiple Read_ids. The isoform is to cluster the filtered reads according to the reference genome to obtain the final Isoform, see the second column in the above table, go to The number of repetitions is the Isoforms number, and Isofrom is used for subsequent analysis of AS, TF, and LncRNA; since the first comparison only selects transcripts with a transcript source number of 1, the obtained isoform cannot be used for fusion analysis. The result of alignment of the Consensus sequence with the reference genome again is used for fusion analysis.