Workflow Diagram (version for real data analysis)

Thursday, January 14, 2021 11:49 AM

Input:

Basecalled reads (.fastq)
Squiggle (.fast5)
Reference genome (.fa)

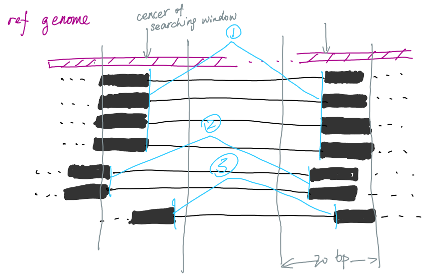
1. Minimap2 analysis (same as the previous pipeline)

To assess how accurate the splice sites can be identified by minimap2, I mapped the nanopore reads to the reference genome. Although minimap2 can take annotations as actual information to increase the mapping quality, we decide not to do so since there is no real world example of a perfect annotation as sequins, which may inflate the accuracy of splice site identification from minimap2. Mapping all 2191749 reads to the reference genome of sequins results in 1958069 alignment, in which 1957845 are primary alignment. Only primary alignments will be considered in this analysis, since we will need to take only one alignment for each read and the primary alignment will be the best one.

NanoSplicer Analysis

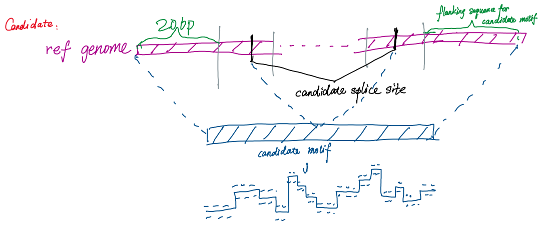
Step 1: Exon junctions identification.

To simulate the situation in real data analysis, in this version, the transcript annotation and reference transcript were not used until the accuracy assessment. In this case, the determination of junction within read is based on the minimap2 result(it was based on the true splice site in previous pipeline). After the genome mapping, a list of mapped splice sites can be extract from BAM file with their number of read supports. The following figure shows an example of an exon junction. In this exon junction, 3 splice sites are supported by 4, 2 and 1 junction within read respectively. I start from the most supported splice site in the whole BAM file, and group the junctions within reads supporting all the **mapped splice site**s in a window (size = 20) centered at the most supported splice site. In the example below, the candidate window will be centered at splice site 1. The same process will be repeat over rest junctions within reads until no more junction within reads left. Given that the dataset is cDNA sequencing data, the reads could come from the reverse strand of a transcript. I got the direction of the transcript by following the minimap2 output (Minimap2 will try to mapping the read in both direction and find which one is better). The candidate splice sites are then obtained by searching "GT" pattern in the donor site candidate searching window and searching "AG" in the acceptor site candidate searching window.



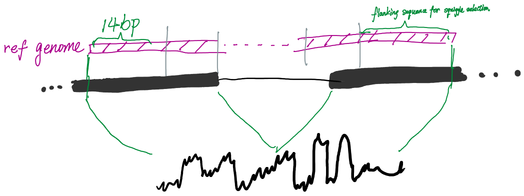
Step 2: obtaining candidate splice sites and candidate squiggles (same as the previous pipeline)

After obtaining the candidate splice site, candidate motifs are obtained from reference genome. To make sure we have enough number of bases included in the candidate motif, flanking sequences of size 20 are included in both sides of the candidate searching window. Candidate squiggles are then obtained from candidate motif using scrappie model (version1.4.0). The following figure shows an example of obtaining a candidate squiggle given one specific candidate splice site.



Step 3. Obtaining junction squiggle

In this version of NanoSplicer, junction squiggles are obtained from genome mapping, the start and end positon of a junction within read is first determine from reference genome. They are the boundary of candidate searching window (step 1) in both sides of splice sites plus flanking sequence of size 14. The flanking size is 20 in step 2, which is slightly larger than the one used here. The purpose of doing this is to ensure the shape of the junction squiggle is captured inside the candidate squiggle. After the determination of the coordinate of the junction within reads, I used tombo (v1.5) to map the squiggles to basecalls and subset the squiggles corresponding to the junction with read.



Step 4: Dynamic time warping