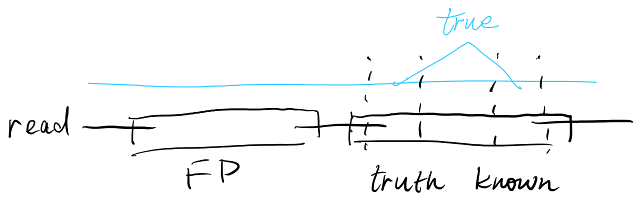
Primary site proportion and Junction Mapping quality analysis using Sequins data (Feb 17 meeting notes)

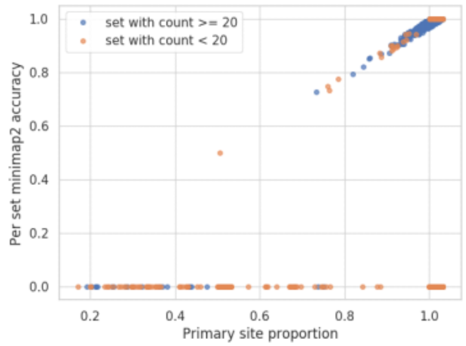
Tuesday, February 16, 2021 7:18 AM



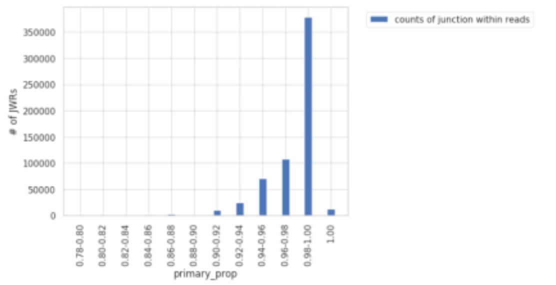
All the figure shown in this document was based on the truth known junction within read only

Primary site proportion analysis

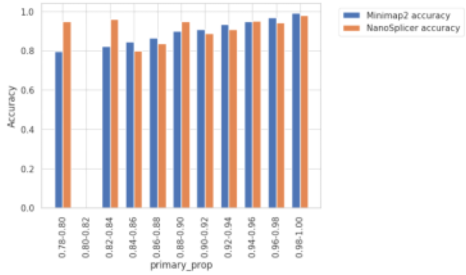
- Primary site proportion **definition:**
  - The proportion of junction within reads(JWR) that supporting the best supported splice junctions within each JWR set.
  - The reason of bringing the primary site proportion is that, when most (e.g. 100%) of the JWRs in each set supporting a same splice junction, we assume that there is the candidate supported is true and there is no other true candidate within the set. So there is no need to apply NanoSplicer.
  - To validate the setup of primary site proportion, I checked the following points using sequins data:
    - □ Whether or not PSP=100% means all of them are correct
    - □ How much it is affected by low count JWR set



Results



□ Accuracy of NanoSplicer and Minimap2

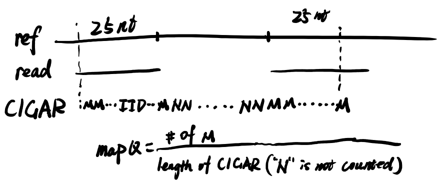


Limitation: when the number of JWRs is low, the estimation of primary site porportion is not reliable, for example 1 out of 1 means primary site proportion = 100%

Junction mapping quality

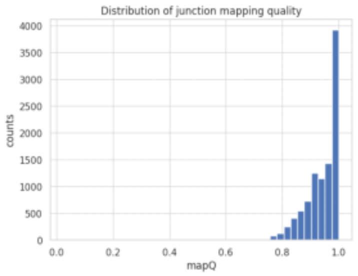
Get junction mapping quality

After discussion with Heejung, we have a intuition that NanoSplicer results will agree with minimap2 when the junction mapping quality (defined as the proportion of 'M's within the cigar string near the junctions, currently looking at **50** bases) is high. If so, we could filter out those JWRs before running NanoSplicer. Because querying a single read in a bam file using read id is pretty slow (bam is not indexed by read id), I subsampled **10000 JWR**.

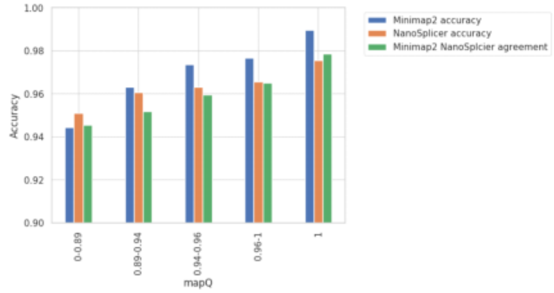

ref
read
CIGAR
25 nt
25 nt
MM···IID··MNN·····NNMM······M
mapQ =
# of M
length of CIGAR ("N" is not counted)

Distribution of junction mapping quality(mapQ)

To assess how the junction mapping quality(mapQ) associated with the minimap2/NanoSplicer accuracy, I am looking that the bin ed accuracy of minimap2/NanoSplicer

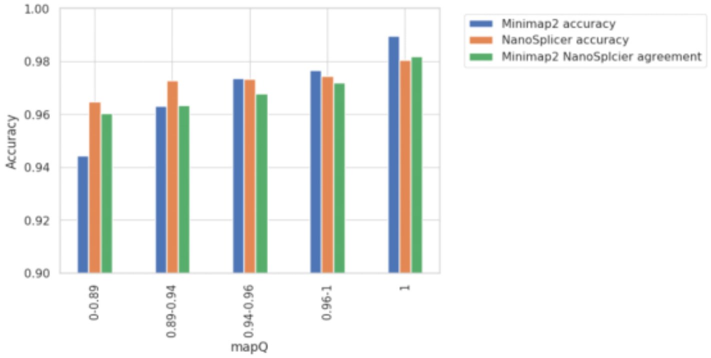


```
number of fails in getting junction mapping quality: 5/4000
```



```
counts for each bar
[1838, 2014, 1334, 2571, 2211]
```

For the bins with low mapQ, I observed that for a lot of cases that minimap2 are correct but NanoSplicer is wrong, the main reason is Si. It makes sense that when the mapQ is low, it is harder for tombo to get correct junction squiggle



```
counts for each bar (before and after si filtering)
[1838, 2014, 1334, 2571, 2211]
[1586, 1838, 1272, 2467, 2160]
```

Heejung's comments (to-do)
- Explain 1 bin
- FP analysis
- Run entire thing (also for real data analysis)
- Check mapQ definition
- Why is not 100% when perfect thing is given
- Minimap2 prior -》 mapQ prior
- Second likely candidate (showing how it is close to the best one)