Paper summary

Thursday, February 18, 2021 4:12 PM

Result(Part 1)： **Assess the performance of NanoSplicer with known ground truth using Sequins**

In the main text, mention what we did in this part and report the accuracy to show：

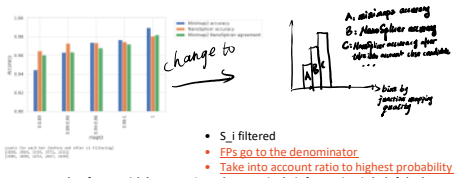
- **Squiggle has information to distinguish close splice site.**
- **If some of the candidate squiggles is similar to the true candidate, squiggle won't have power to distinguish them.**
  - Ratio to the highest probability (show most of the time if we are not correct, the probability of the true one is close to the chosen one.)

(the whole part may go to supplementary)

Result(Part 2): Sequins data analysis with real data pipeline

Local/**junction alignment Q**

- **Main argument:** NanoSplicer is better than minimap2 when the basecalling/ mapping quality near the splice junctions in low **(define the mapping quality, talk with Mike, there is perhaps a better name.)**
  - It makes sense that when the basecalling/mapping quality in the JWR is relatively low, the mapped splice junction from minimap2 is less reliable, so that NanoSplicer has higher potential to get correct splice site with the help of squiggle information



- S_i filtered
- FPs go to the denominator
- Take into account ratio to highest probability

- **Example of JWR with low mapQ to show squiggle information is helpful when mapQ is bad.**
- **Examples of why the mapQ is high (bin mapQ = 1) but NanoSplicer make it wrong**
  - When the true one identified as the second likely one, show how close the likelihood is.
  - When NanoSplicer completely missed the true one (with huge difference in likelihood), check the **distinguish point(sup).**

Supplementary:

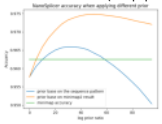
- Empirical distribution of S_i
- Show how is the result is sensitive to S_i threshold selection and choose appropriate S_i threshold
- How S_i can distinguish FP and tru

Result(Part 3): Real data analysis

- **Main argument: Same conclusion as Part 2 but with real data**
  - (How to frame this part depends on the results of junction mapQ analysis)

- S_i filtered
- FPs go to the denominator
- Take into account ratio to highest probabiltiy
- With seq prior

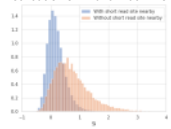
  - **Adding sequence prior in real data analysis is helpful （drop minimap2 prior）**
    - The sequence prior has been implemented in minimap2 already
    - The pattern is conservative (site paper)



  - **Quantification** (depends on whether or not get good result):
    - Categorized by mapQ.
    - and use seq prior
    - And examples

Supplementary:

- Empirical distribution of S_i-> appropriate threshold for S_i
  - Distribution of FP and Truth known JWR in real dataset



- **Example of JWR with low mapQ to show squiggle information is helpful when mapQ is bad.**
- **Examples of why the mapQ is high (bin mapQ = 1) but NanoSplicer make it wrong**
  - When the true one identified as the second likely one, show how close the likelihood is.
  - When NanoSplicer completely missed the true one (with huge difference in likelihood), check the distinguish point.
- Empirical distribution of S_i-> appropriate threshold for S_i
- How S_i can distinguish FP and truth known

Part 4

Software introduction:

Function/option blabla…

Real data analysis example of how people can use it on single gene (or specific region)…to show what people can do with NanoSplicer.