AMSI VACATION RESEARCH SCHOLARSHIPS 2022–23

SET YOUR SIGHTS ON RESEARCH THIS SUMMER

Accurate identification of splice junctions using nanopore direct RNA sequencing

Patrick Grave

Supervised by Heejung Shim

University of Melbourne

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Contents

Abstract

NanoSplicer (Yupei You et al, 2022) is a program that accurately identifies splice junctions using Oxford Nanopore Sequencing data. It performs well on cDNA data but suffers when direct RNA (dRNA) data is used. This project identifies failure modes and frequencies of NanoSplicer on dRNA by classifying NanoSplicer alignments. Future research will examine methods to tackle these errors.

1 Introduction

Splicing is the process of joining exons together after transcription to form mature messenger RNA. This allows genes to code for multiple different mRNA strings which is known as “alternative splicing”. Almost 95% of human genes undergo alternate splicing (Pan et al., 2008), allowing a diverse array of transcript isoforms. These diverse isoforms are translated into different proteins, controlling cell function.

Although short reads can be used to identify some forms of alternate splicing (LeGault and Dewey, 2013; Steijger et al., 2013), this approach is inherently challenging; a ‘bag’ of exon short reads with limited coverage of splice junctions makes it difficult to quantify the isoforms present or even identify them.

Alternatively, methods based on long-read sequencing data have found some success in identifying expressed isoforms. Oxford Nanopore Sequencing is a long-read sequencing technique that works by measuring changes in electric current density as a DNA or RNA molecule passes through a nanoscopic pore in a membrane. This can be completed without needing PCR amplification, producing a raw electrical signal (squiggle) that is then basecalled in software – the magnitude of electric current density corresponds to a specific polynucleotide sequence occupying the pore. However, nanopore sequencing has a higher basecalling error rate (~1-10%) than short-read sequencing, making it challenging to differentiate true splice junctions from mapping errors. (Yupei You et al, 2022)

Methods to identify isoforms from long-read sequencing data include FLAIR (Tang et al, 2020) and TranscriptClean (Wyman et al, 2018), which require a set of splice junction candidates from annotations or matched short reads. However, these may not be practically available. Other methods, including StringTie2 (Kovaka et al, 2019), TAMA (Kuo et al, 2020), and 2passtools (Parker et al, 2021), use information like nearby splice junctions supported by high read counts. However, this may lead to suppression of rare splice junctions. (Yupei You et al, 2022)

This project works with NanoSplicer, a method that generates candidate alignments for 'Junction within Reads' (JWRs) in a set of mapped long-reads. A Junction within Read is a subsequence in the mapped reads that splits and maps to different exons, NanoSplicer operates as follows:

1. A JWR is selected
2. A theoretical squiggle is generated for each candidate
3. Dynamic Time Warping is then applied to best align these to the raw JWR
4. A mixture model is used to calculate which candidate best fits the JWR

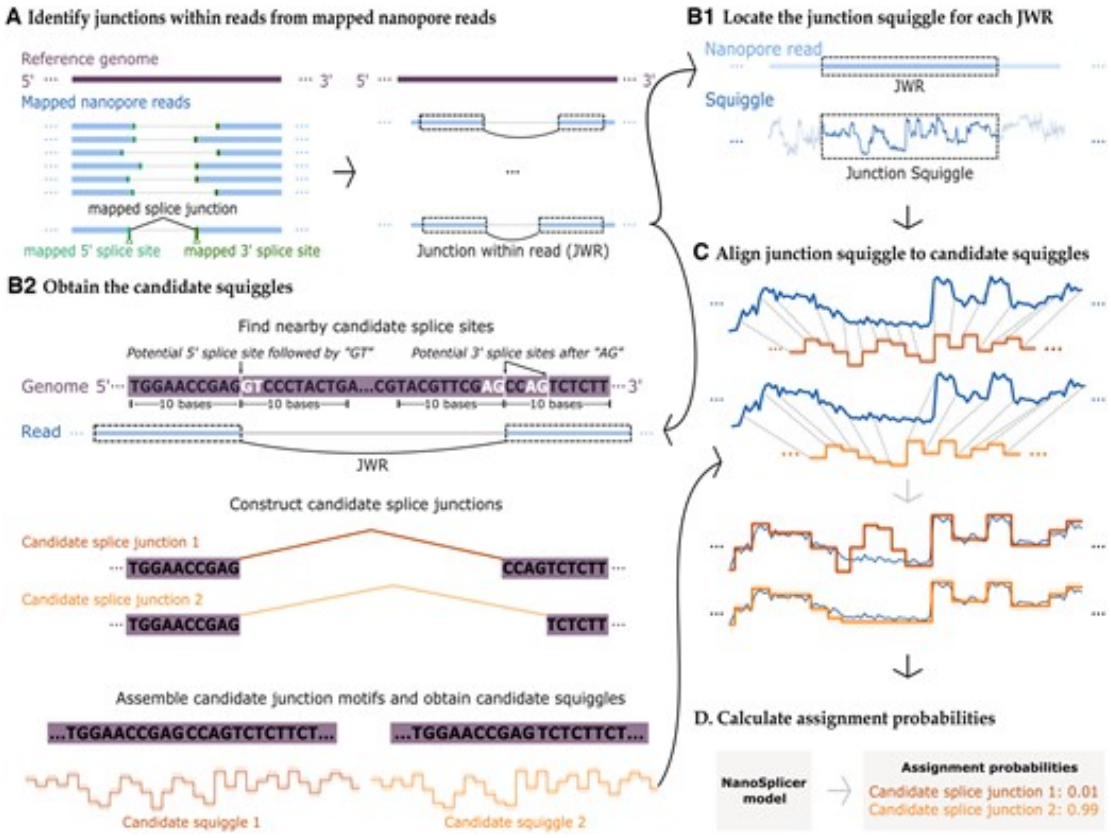

Figure 1: NanoSplicer workflow (Yupei You *et al*, 2022)

Although NanoSplicer performs well on complementary DNA (cDNA - DNA produced by reverse-transcribing RNA), the project examines its performance on direct RNA (dRNA). dRNA sequencing can involve less library preparation and chemistry (Wongsurawat et al., 2022), removing a source of experimental variation.

1.1 Statement of Authorship

The project idea and approach was formulated by Heejung Shim based on preliminary investigations done by Yupei You. The analysis was performed by Patrick Grave under supervision by Heejung Shim with additional guidance given by Yupei You. Some analysis code was contributed by Yupei You. Project funding was provided by AMSI. Computing resources owned by the University of Melbourne were used in conducting the analysis.

2 Method

2.1 Data

A dataset of direct RNA Sequins reads obtained from human SH-SY5Y cells by Gleeson et al, 2020.

2.2 Data Preparation

To prepare the data, the following steps were taken:

1. The raw Fast5s were aligned to a Sequins reference genome using minimap2
2. NanoSplicer was run on the mapped reads
    a. JAQ=0.8 setting for JWR selection
3. Analysis was run on the NanoSplicer output

2.3 Analysis

Three kinds of analysis were conducted:

1. Squiggle information quality (SIQ) calculations for NanoSplicer on dRNA data

- Squiggle information quality (SIQ) is calculated as in the original NanoSplicer paper. Alignment quality is calculated for each candidate squiggle by taking the average log-likelihood over the nucleotides. SIQ is the maximum average log-likelihood across all the candidates tested

2. Comparison with ground truth
    - As Sequins data is comprised of artificial, known nucleotide sequences, the 'true' mappings of Sequins reads to the Sequins reference genome are known
    - Performance is evaluated by counting the cases in which minimap2
3. Visual analysis of failure modes
    - Based on NanoSplicer's operation and the workflow used, 5 failure modes were hypothesized:
        - Poor base-calling for direct RNA nanopore data
        - Poor 'resquiggle' performance by tombo
        - NanoSplicer choosing poor candidates
        - NanoSplicer failing to align successfully
        - NanoSplicer selecting the wrong candidate

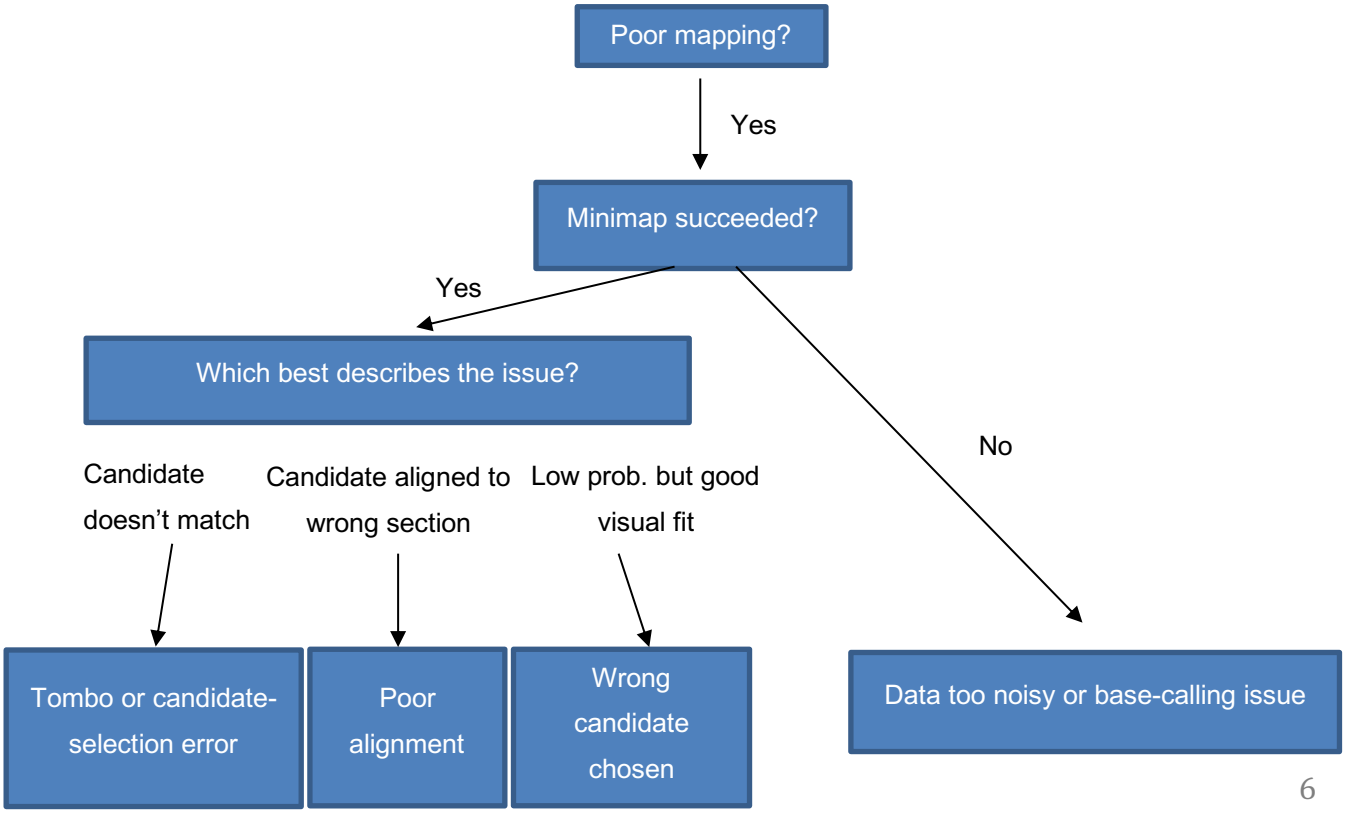

Figure 2: Flowchart used during visual analysis to label images. Five main categories were identified: Simply bad candidates, candidates with a good visual fit, very flat junction squiggles, fits with very dense regions, and JWRs with good-looking but low-scoring candidates

4 Results

Total JWRs	1536
Mapped JWRs	1541

Table 1: Counts of JWRs analysed. (Produced by Yupei You's analysis code)

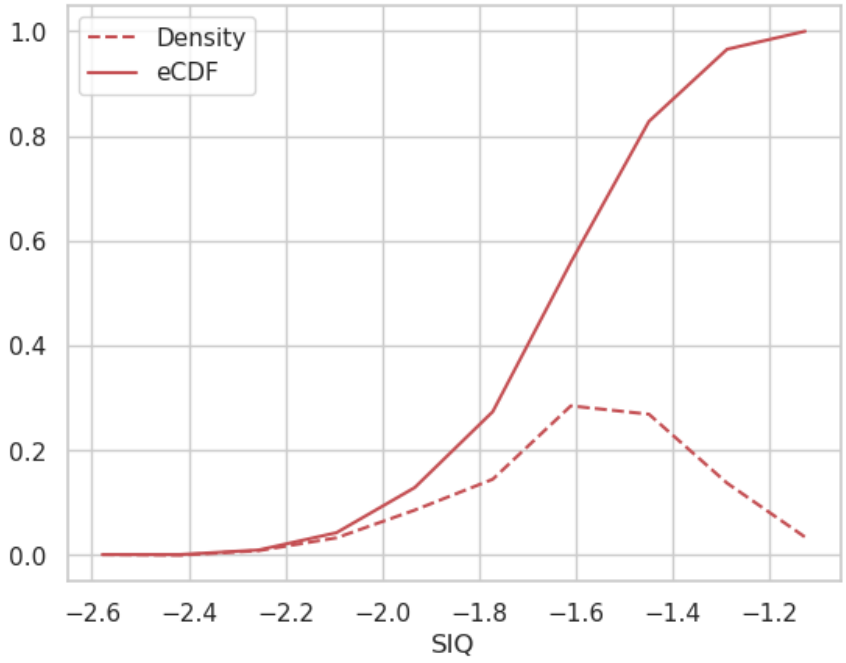

Figure 3: SIQ (Maximum average log-likelihood across candidates for a JWR) distribution for dRNA JWRs analysed with NanoSplicer. (Produced by Yupei You's analysis code)

	NanoSplicer correct	**NanoSplicer Incorrect**	**Total**
Minimap correct	191	189	380
Minimap Incorrect	5	113	118

Total	196	302	498

Unclassified	15

Table 1: JWR categorization of Sequins reads vs. Ground Truth

Bad Candidates	49
Good Visual Fit	114
'Semi-Flat' junction squiggle	5
Dense regions	18
Visually better candidate with low prob.	3
Total	189

Table 2: Counts of error type by category. Visual analysis was performed on candidates which minimap2 correctly labelled (vs. ground truth) but NanoSplicer failed on

5 Analysis

The results were notable in a few ways:

1. The very low fraction of JWRs for which minimap2 was incorrect but NanoSplicer was correct against the ground truth. This makes some sense as NanoSplicer includes the minimpa2 mapping as a candidate. However, the very large share (189 JWRs) for which NanoSplicer failed but minimap2 succeeded. In these instances, minimap2 selected the correct mapping (vs. ground truth) but NanoSplicer chose another candidate which didn't match the ground truth. These cases are the subject of further analysis
2. A large number of NanoSplicer candidates, which were incorrectly chosen over minimap2's mapping (according to the ground truth) had visually good fits (114). Naturally, many were simply bad candidates (49). However, many looked to match the junction squiggle. This could have a couple of explanations:
    a. My visual analysis could be biased to favour candidates that match long, flat, sections of the junction squiggle – rather than those that match short, varying sections. This may be exacerbated by dRNA's different dwell time properties

vs. cDNA and by Dynamic Time-Warping fitting long, flat, squiggle sections more aggressively.

b. Tombo's 're-squiggle' algorithm may be producing the wrong junction squiggle for many JWRs. This may occur more frequently for dRNA than for cDNA.

Some examples of cases where NanoSplicer failed to choose the ground-truth candidate but minimap2 succeeded:

- A candidate with *good visual fit*. The candidate's predicted squiggle (dark blue) appears to fit the junction squiggle (light glue) much better than the minimap2 predicted squiggle (orange)

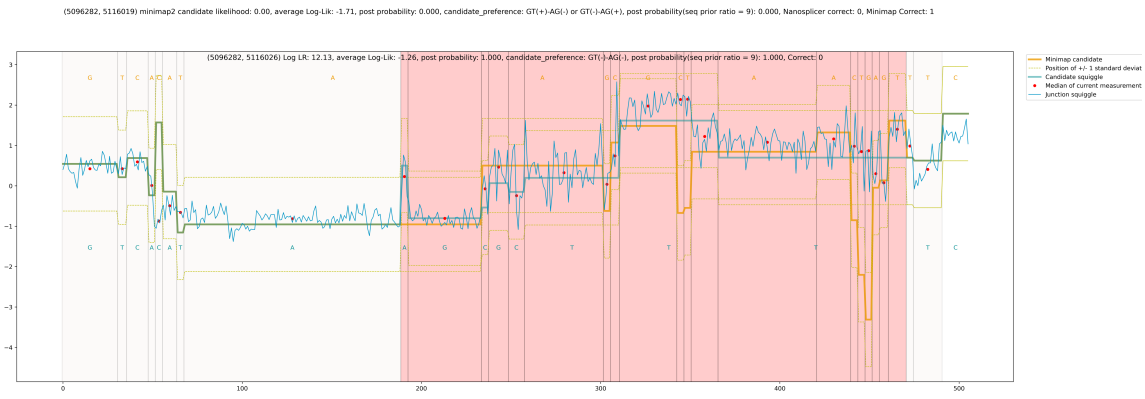

Figure 4: A NanoSplicer candidate with 'good' visual fit that doesn't match the ground truth

- Two candidates for the same JWR. One with a higher average log-likelihood and post probability (seq prior ratio) but a seemingly worse visual fit than the candidate with lower scores

Figure 5: NanoSplicer candidates with scores that don't relatively reflect their visual accuracy vs. junction squiggle

- A candidate with many nucleotides mapped to a relatively small region. The underlying junction squiggle does not reflect this high variation over such a small period.

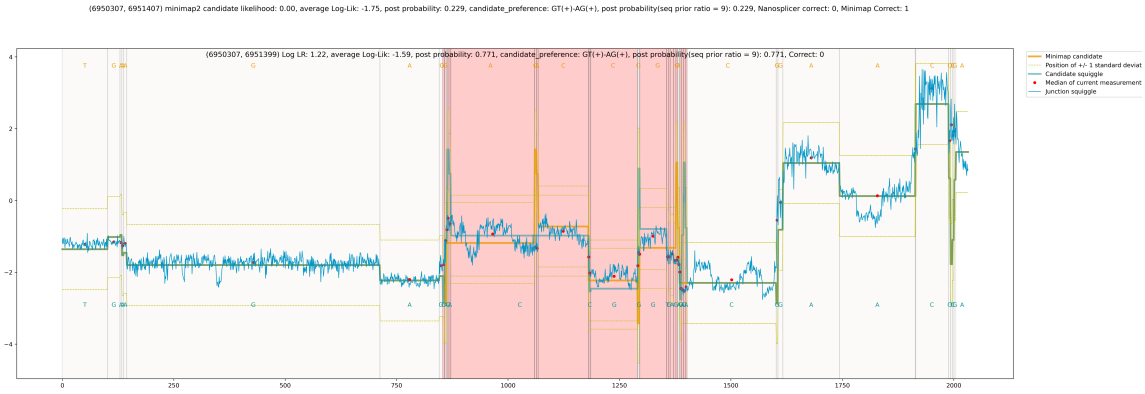

Figure 6: NanoSplicer candidate with lots of nucleotides mapped to a very small period

- A high-scoring candidate which only seems to fit the long and flat sections of the junction squiggle

Figure 7: NanoSplicer candidate fitting only long and flat sections of the junction squiggle

- Finally, a JWR whose junction squiggle appears to be flat. This reflects a long sequence of one nucleotide, a very short sequence, or a re-squiggle error

Figure 8: JWR with a very short junction squiggle

6 Conclusion

In Conclusion, NanoSplicer produces very poor results when used on dRNA data. Notably, minimap2 was able to match the ground truth far more frequently than NanoSplicer. This may be due to the much larger variations in dwell-time observed during direct RNA sequencing (potentially causing issues with tombo's re-squiggle algorithm).

Further research should examine the performance of tombo against other 're-squiggle' tools on direct RNA data, in-case this is causing failed 're-squiggles which would lead to NanoSplicer generating poor candidates. Alternatively, a base-caller that also provides an alignment to the raw squiggle would eliminate the need for a 're-squiggle' tool like tombo so a search should be conducted for such a tool.

7 References

Yupei You, Michael B Clark, Heejung Shim, NanoSplicer: accurate identification of splice junctions using Oxford Nanopore sequencing, *Bioinformatics*, Volume 38, Issue 15, 1 August 2022, Pages 3741–3748, https://doi.org/10.1093/bioinformatics/btac359

Pan, Q., Shai, O., Lee, L. *et al.* Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–1415 (2008). https://doi.org/10.1038/ng.259

LeGault LH, Dewey CN. Inference of alternative splicing from RNA-Seq data with probabilistic splice graphs. Bioinformatics. 2013 Sep 15;29(18):2300-10. doi: 10.1093/bioinformatics/btt396. Epub 2013 Jul 11. PMID: 23846746; PMCID: PMC3753571.

Steijger, T., Abril, J., Engström, P. *et al.* Assessment of transcript reconstruction methods for RNA-seq. *Nat Methods* **10**, 1177–1184 (2013). https://doi.org/10.1038/nmeth.2714

Josie Gleeson, Tracy A. Lane, Paul J Harrison, Wilfried Haerty, Michael B Clark. Nanopore direct RNA sequencing detects differential expression between human cell populations, bioRxiv 2020.08.02.232785; doi: https://doi.org/10.1101/2020.08.02.232785

Tang, A.D., Soulette, C.M., van Baren, M.J. *et al.* Full-length transcript characterization of *SF3B1* mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nat Commun* **11**, 1438 (2020). https://doi.org/10.1038/s41467-020-15171-6

Dana Wyman, Ali Mortazavi, TranscriptClean: variant-aware correction of indels, mismatches and splice junctions in long-read transcripts, *Bioinformatics*, Volume 35, Issue 2, January 2019, Pages 340–342, https://doi.org/10.1093/bioinformatics/bty483

Kovaka, S., Zimin, A.V., Pertea, G.M. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 278 (2019). https://doi.org/10.1186/s13059-019-1910-1

Kuo, R.I., Cheng, Y., Zhang, R. *et al.* Illuminating the dark side of the human transcriptome with long read transcript sequencing. *BMC Genomics* **21**, 751 (2020). https://doi.org/10.1186/s12864-020-07123-7

Parker, M.T., Knop, K., Barton, G.J. *et al.* 2passtools: two-pass alignment using machine-learning-filtered splice junctions increases the accuracy of intron detection in long-read RNA sequencing. *Genome Biol* **22**, 72 (2021). https://doi.org/10.1186/s13059-021-02296-0

Wongsurawat, T., Jenjaroenpun, P., Wanchai, V. and Nookaew, I. (2022). Native RNA or cDNA Sequencing for Transcriptomic Analysis: A Case Study on Saccharomyces cerevisiae. *Frontiers in Bioengineering and Biotechnology*, 10. https://doi.org/10.3389/fbioe.2022.842299