Evaluating Short-Read RNA-seq Tools Using Long-Read Minimap2 Alignment

Alvin Zhang
8/29/24

Background: STAR, HISAT2, and Minimap2

STAR: published in 2013, originally meant for 50-100 base pair sequences

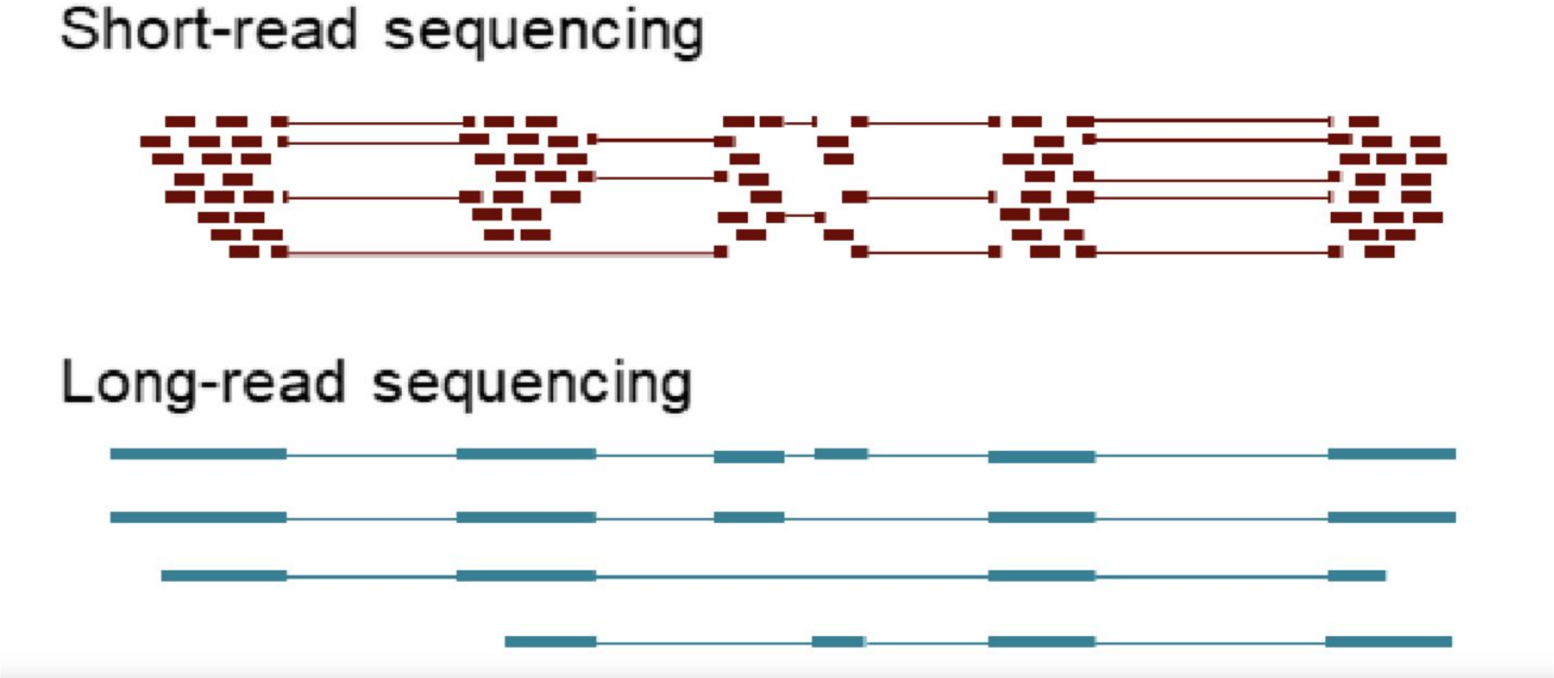
HISAT2: published in 2015, designed for 50-150 base pair sequences

Minimap2: published in 2018, viable for shorter and longer sequences (thousands to millions of base pairs)

Now: 100-150 bp is the norm, but there is a trend towards using longer reads up to 200-300 bp.

Long Read vs Short Read RNA-seq

- Long Reads: Higher accuracy and better effectiveness in complex regions
- Short Reads: Generally less expensive and produces large amounts of data

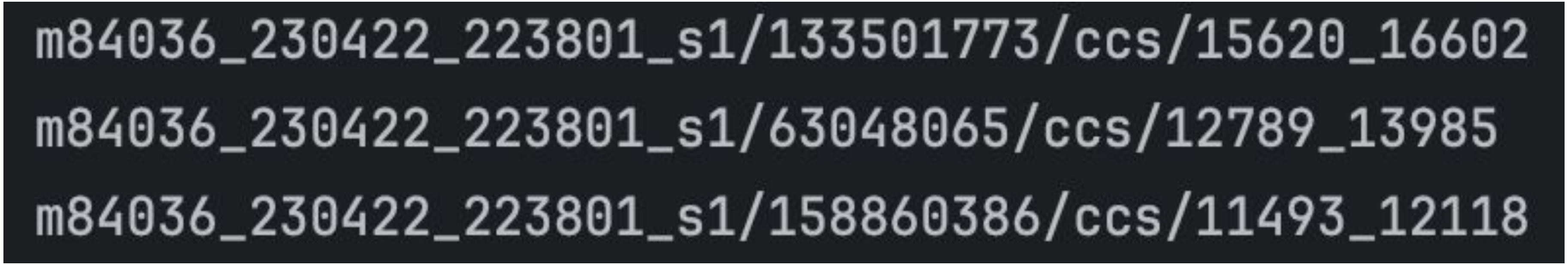


Project Overview

- Aim to compare short-read tools using Minimap2's long-read alignment as ground truth to assess accuracy
- Test Minimap2's capability in short-read RNA-seq alignment
- Extract subreads from the Minimap2 aligned long-reads and compare the original CIGAR string and CIGAR string from subread alignment to determine how well each short-read tool replicates Minimap2's long-read alignment
- Identify the most accurate short-read RNA-seq aligner

Minimap2 as Ground Truth

- Take long-read sequencing data from specific locations (chromosome 20)
- Align to reference genome (GRCh38) using Minimap2



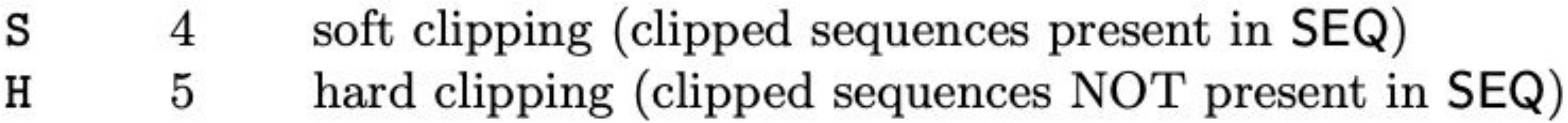
```
m84036_230422_223801_s1/133501773/ccs/15620_16602
m84036_230422_223801_s1/63048065/ccs/12789_13985
m84036_230422_223801_s1/158860386/ccs/11493_12118
```

Example long-read sequence IDs shown above

- Core ID is all the same between the different long-read sequences for this specific example

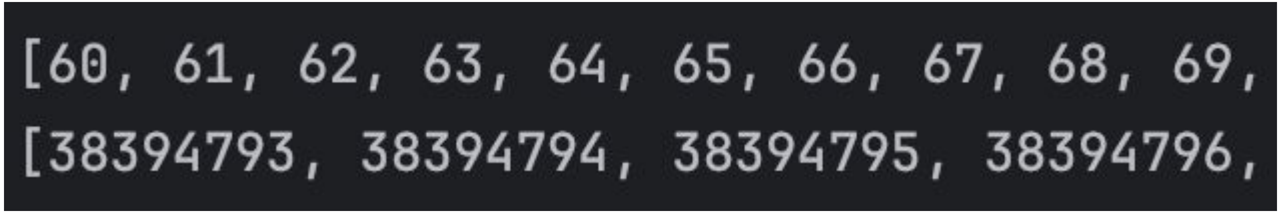
Query and Reference Positions Table

- Generate query and reference positions in Querytable script
- Account for clipping when generating query positions:



S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)

- Soft clip: query positions begin at clip length (consumes query)
- Hard clip: query positions begin at 0 (does not consume either query or reference)



```
[60, 61, 62, 63, 64, 65, 66, 67, 68, 69,
[38394793, 38394794, 38394795, 38394796,
```

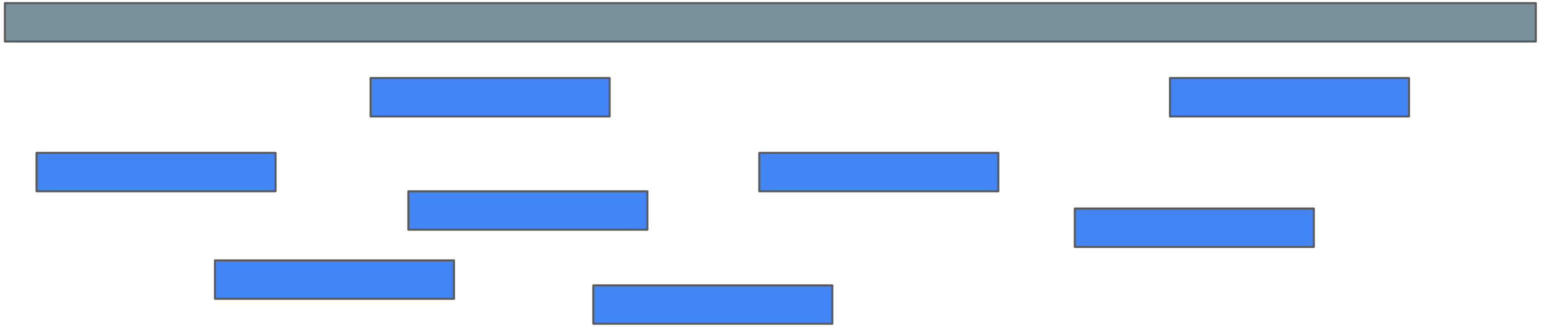
Query Positions

Ref Positions

Top image taken from: https://samtools.github.io/hts-specs/SAMv1.pdf

Random Sampling of Subreads From Long Reads

- Extract random subreads from the long-read alignment data using the table of positions generated
  - Choosing random start positions from the table

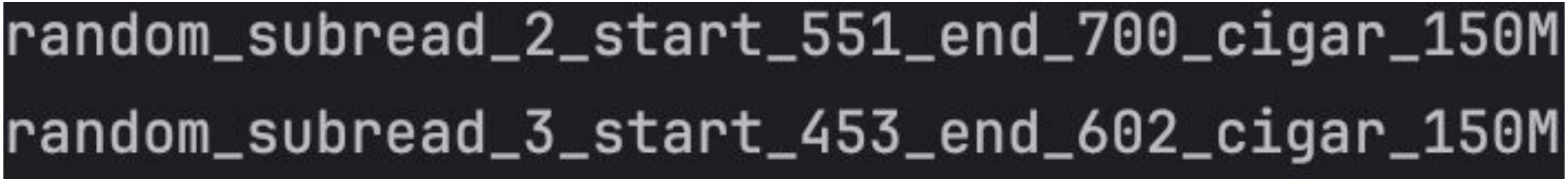


Random Sampling (Continued)

For better efficiency and memory optimization, the number of random subreads to sample:

- Long read length (~1500 bp) / subread length (150 bp)

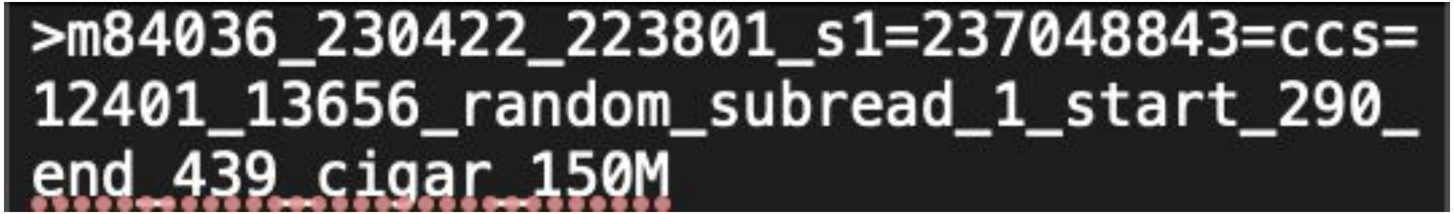
Add the start and stop positions and cigar string of the subread in the sequence ID:



```
random_subread_2_start_551_end_700_cigar_150M
random_subread_3_start_453_end_602_cigar_150M
```

Subread Sequence ID Modification

- Write the list of subreads into a FASTA file
- Replace all slashes (/) with equal signs (=) to ensure that the full sequence ID is retained



```
>m84036_230422_223801_s1=237048843=ccs=
12401_13656_random_subread_1_start_290_
end_439_cigar_150M
```

(Example modified subread sequence ID shown above)

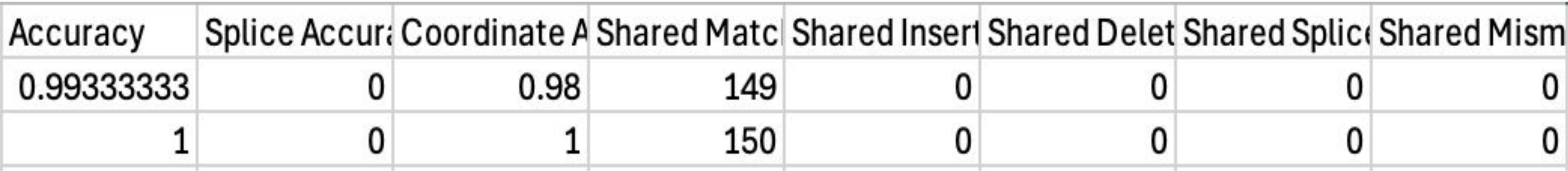
- Run HISAT2, STAR, and Minimap2 to align these sampled subreads

Output Modification and CIGAR String Processing

- Sort and index the BAM file from aligned subreads to visualize through IGV and to assess accuracy
- Extract the CIGAR string of each subread while keeping track of the original CIGAR string kept in the sequence ID
- Compare CIGAR string of subreads after having run short-read RNA-seq tools with “original” CIGAR string included in the sequence ID of the subread

CIGAR comparison and Accuracy/Precision counter

- Write information into TSV file
- Summarize overall data:
  - Accuracy, Coordinate Accuracy, and Splice Accuracy



Accuracy	Splice Accur	Coordinate A	Shared Matc	Shared Insert	Shared Delet	Shared Splic	Shared Mism
0.99333333	0	0.98	149	0	0	0	0
1	0	1	150	0	0	0	0

Example of two subreads from the TSV file (opened with Excel)

Accuracy, Coordinate Accuracy, and Splice Junction Accuracy

Accuracy: Accurate Bases / Total Bases

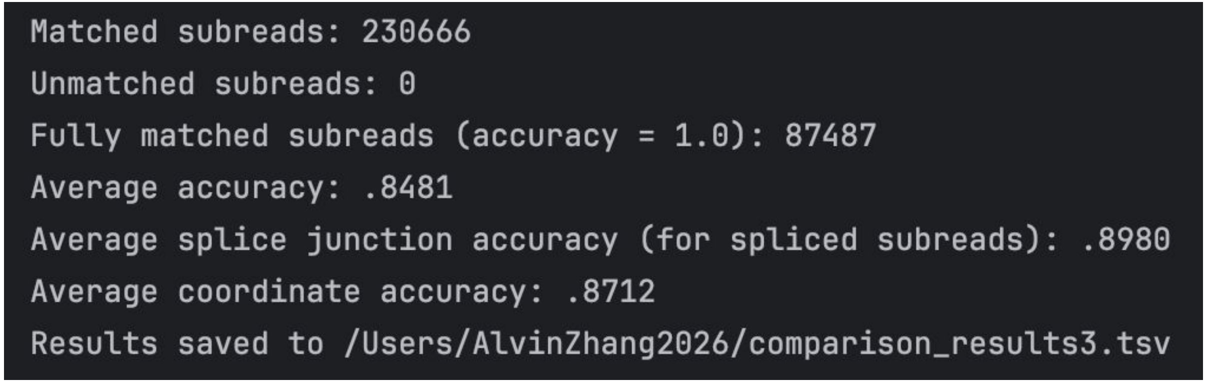
- Accurate Bases: Correctly identified bases (matches, insertions, deletions, splices, and mismatches)
- Total Bases: All bases present

Coordinate Accuracy: Intersection / Union

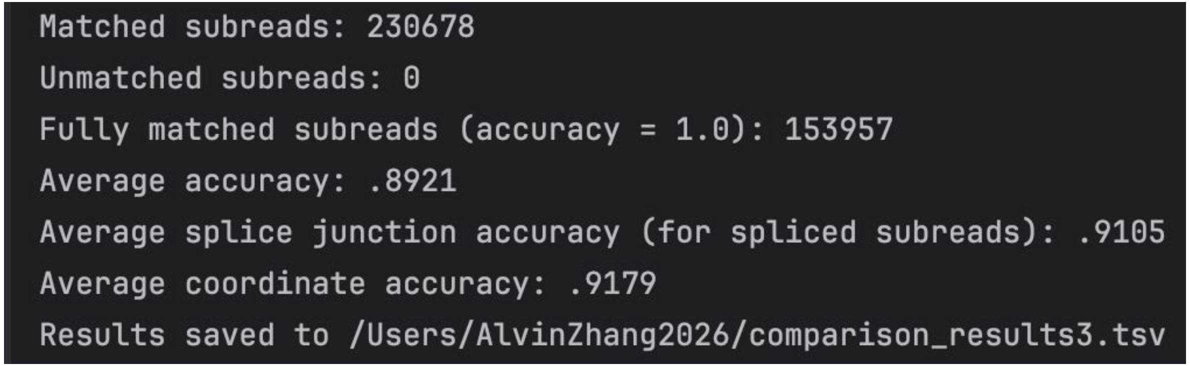
Splice Junction Accuracy: Shared Splice / (Shared Splice + Missing Splice + Additional False Positive Splice)

- False Positive Splice: Not found in Long-read but detected in short-read subreads

Results - STAR

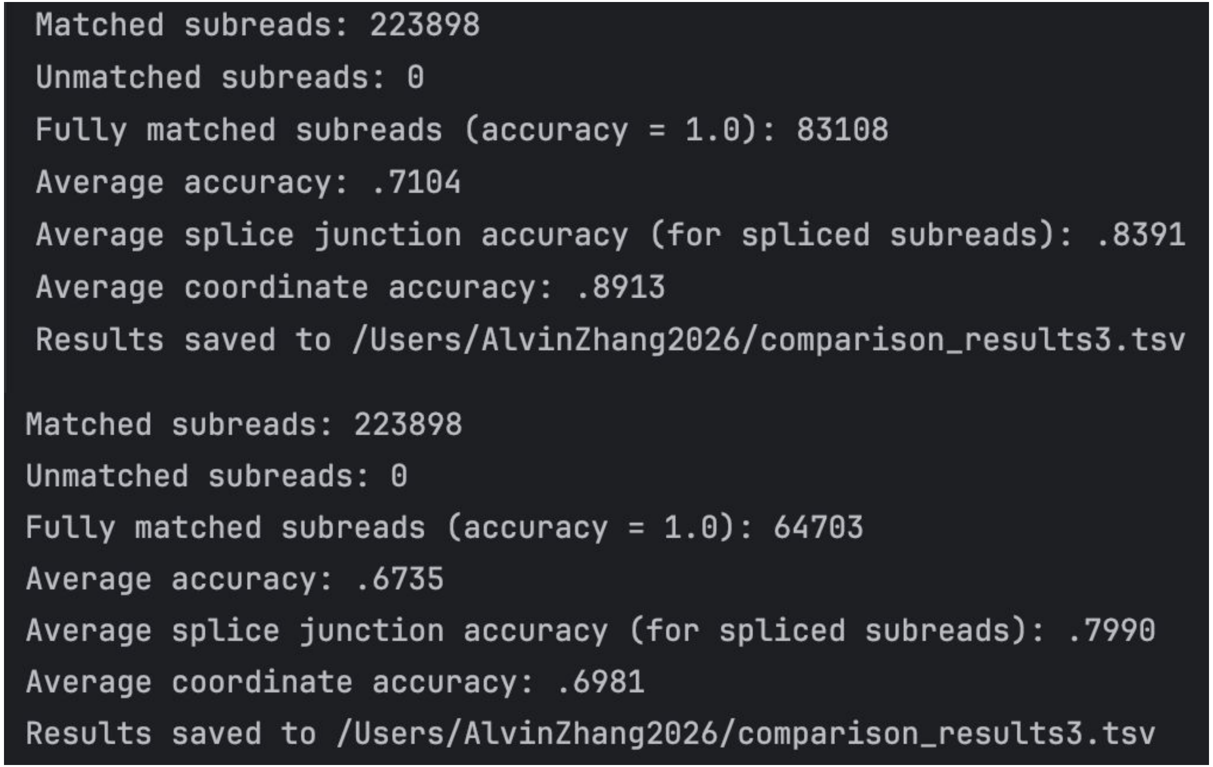


```
Matched subreads: 230666
Unmatched subreads: 0
Fully matched subreads (accuracy = 1.0): 87487
Average accuracy: .8481
Average splice junction accuracy (for spliced subreads): .8980
Average coordinate accuracy: .8712
Results saved to /Users/AlvinZhang2026/comparison_results3.tsv
```



```
Matched subreads: 230678
Unmatched subreads: 0
Fully matched subreads (accuracy = 1.0): 153957
Average accuracy: .8921
Average splice junction accuracy (for spliced subreads): .9105
Average coordinate accuracy: .9179
Results saved to /Users/AlvinZhang2026/comparison_results3.tsv
```

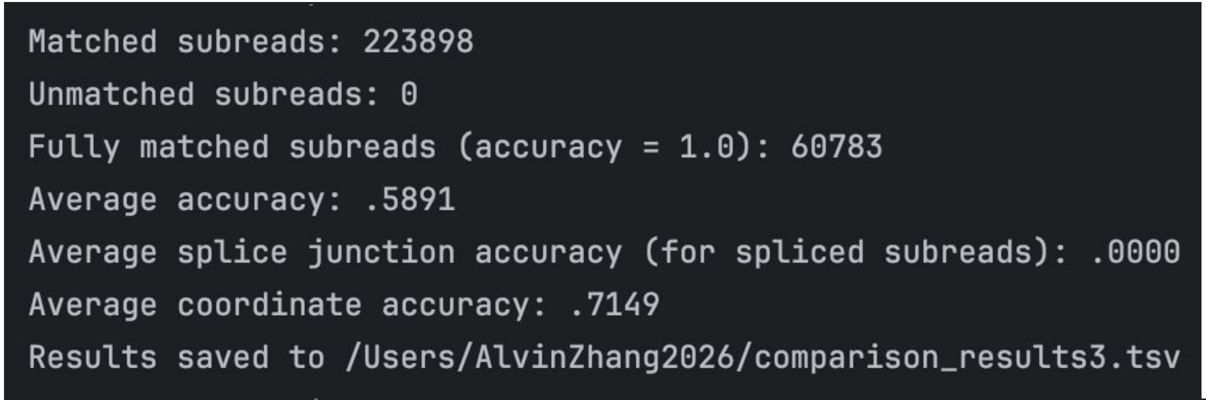
Results - HISAT2



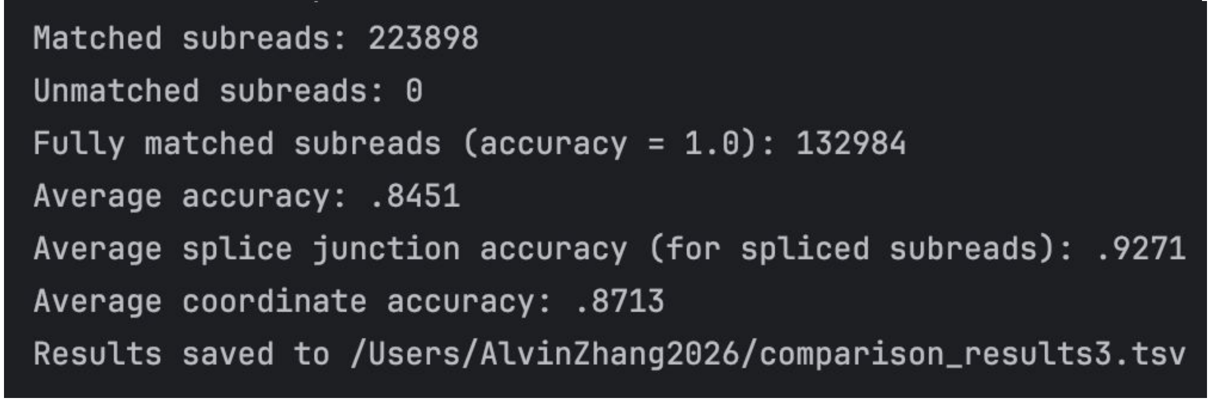
```
Matched subreads: 223898
Unmatched subreads: 0
Fully matched subreads (accuracy = 1.0): 83108
Average accuracy: .7104
Average splice junction accuracy (for spliced subreads): .8391
Average coordinate accuracy: .8913
Results saved to /Users/AlvinZhang2026/comparison_results3.tsv
```

```
Matched subreads: 223898
Unmatched subreads: 0
Fully matched subreads (accuracy = 1.0): 64703
Average accuracy: .6735
Average splice junction accuracy (for spliced subreads): .7990
Average coordinate accuracy: .6981
Results saved to /Users/AlvinZhang2026/comparison_results3.tsv
```

Results - Minimap2



```
Matched subreads: 223898
Unmatched subreads: 0
Fully matched subreads (accuracy = 1.0): 60783
Average accuracy: .5891
Average splice junction accuracy (for spliced subreads): .0000
Average coordinate accuracy: .7149
Results saved to /Users/AlvinZhang2026/comparison_results3.tsv
```



```
Matched subreads: 223898
Unmatched subreads: 0
Fully matched subreads (accuracy = 1.0): 132984
Average accuracy: .8451
Average splice junction accuracy (for spliced subreads): .9271
Average coordinate accuracy: .8713
Results saved to /Users/AlvinZhang2026/comparison_results3.tsv
```

Overall Summary

- STAR has much higher average accuracy, coordinate accuracy, and splice junction accuracy with more consistency
- HISAT2 is typically less accurate for the three measurements
- Minimap2 is much less consistent (perhaps a bug in my code), and is the extreme of both:
  - One result shows very low accuracy and coordinate accuracy, with a 0 for average splice accuracy
  - One result has high accuracy, coordinate accuracy, and average splice accuracy

In general: STAR is still the most viable short-read RNA-seq aligner (for 150 bp), but this may change as the default length for short reads gets longer and longer.

Acknowledgements:

- Huang Neng
- Professor Li Heng
- Ying Zhou

And:

Li Lab