Multi-Seed Consensus Reference Mapper

The Undergrads

Pre-processing Reference Genome

>NC_045512.2

ATTAAAGGTTTATACCTTCCCAGGTAACAAACCAACCAACTTTCGATCTCTTGT
AGATCTGTTCTCTAAACGAACTTTAAAATCTGTGTGGCTGTCACTCGGCTGCA
TGCTTAGTGCACTCACGCAGTATAATTAATAACTAATTACTGTCGTTGACAGG
ACACGAGTAACTCGTCTATCTTCTGCAGGCTGCTTACGGTTTCGTCCGTGTT
GCAGCCGATCATCAGCACATCTAGGTTTCGTCCGGGTGTGACCGAAAGGTAA
GATGGAGAGCCTTGTC...

Pre-processing Reference Genome

>NC_045512.2

K-mers of length 13

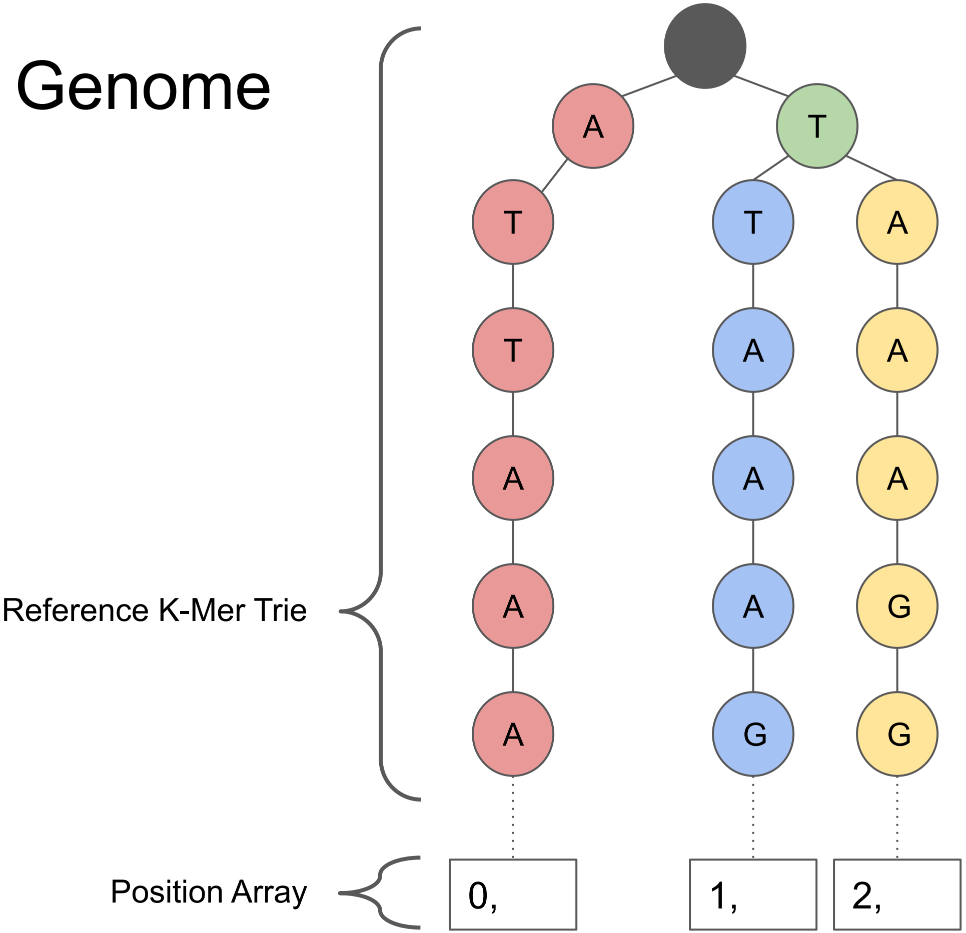
ATTAAAGGTTTATACCTTC...

Pre-processing Reference Genome

>NC_045512.2

K-mers of length 13

ATTAAAGGTTTATACCTTC...



Mapping a Read

```
@S0R0/1

TTTACTTACAAAGTCCTCAGAAGACAAAGGTCCTATTACGGATGTTTTCTACA
AAGAAAACAGTTACACAACAACCATAAAACCAGTTACTTATAAATTGGATCGT
GTTGTTTGTACAGTAATTGACCCTAAGTTGGACAATTATTATAA

+

=C1GGGGGGGCGJGGGJJJJ$JJGJ1GCGGGGGGCCGJJJGJJGJGGJGJG
CCJ=CJCGCGGGGGGCCCJGCGGG=8GGCGCGG8G1$GGCC=GGGG1=G$G
GGGGGGGC$CGG=G$CGGG$GGGGGCGGGGGGCCGGG=C$CGGGC
```

Mapping a Read

@S0R0/1

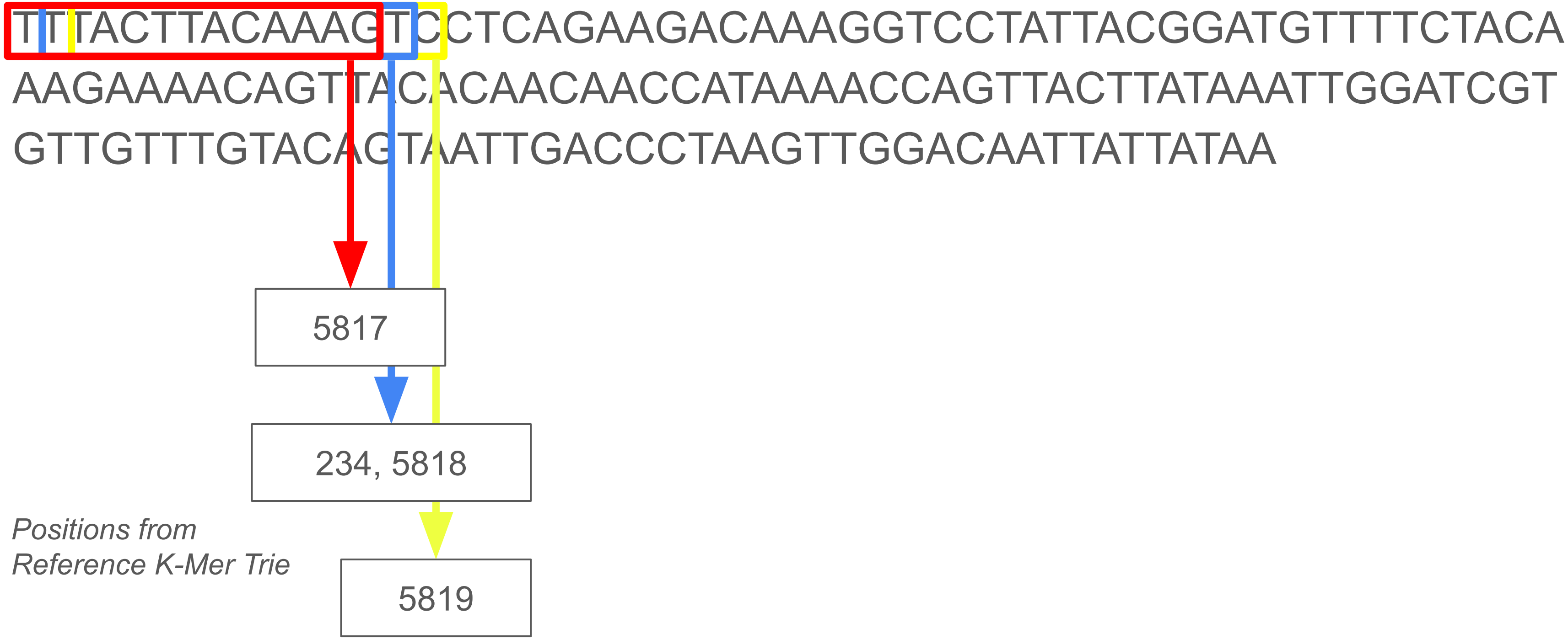
K-mers of length 13

TTTACTTACAAAGTCCTCAGAAGACAAAGGTCCTATTACGGATGTTTTCTACA
AAGAAAACAGTTACACAACAACCATAAAACCAGTTACTTATAAATTGGATCGT
GTTGTTTGTACAGTAATTGACCCTAAGTTGGACAATTATTATAA

Mapping a Read

@S0R0/1

K-mers of length 13



TTTACTTACAAAGTCCTCAGAAGACAAAGGTCCTATTACGGATGTTTTCTACA
AAGAAAACAGTTACACAACAACCATAAAACCAGTTACTTATAAATTGGATCGT
GTTGTTTGTACAGTAATTGACCCTAAGTTGGACAATTATTATAA

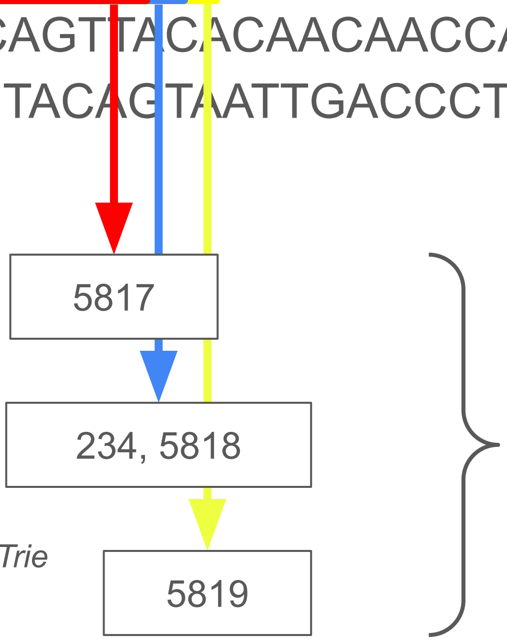
Positions from Reference K-Mer Trie

Mapping a Read

@S0R0/1

K-mers of length 13

TTTACTTACAAAGTCCTCAGAAGACAAAGGTCCTATTACGGATGTTTTCTACA
AAGAAAACAGTTACACAACAACCATAAAACCAGTTACTTATAAATTGGATCGT
GTTGTTTGTACAGTAATTGACCCTAAGTTGGACAATTATTATAA



Positions from Reference K-Mer Trie

Consensus Start Position: 5817

Optimizations for Speedup (1)

- For each paired read, only take reverse complement of one pair.
  - If read0 is not on the reversed strand, then read1 is guaranteed to be on the reverse strand, and vice versa.
- For each paired read, if first pair is not mapped, then second is not mapped.

These optimizations add 20-30% speedup.

Optimizations for Speedup (2)

- Introduce consensus threshold.
  - Minimum percentage of seeds that agree on a consensus start position.
  - Our experiments show value of 0.2 to be fairly accurate & efficient.
- Infer k-mers on reads on need basis.
  - Use Java Iterator<> instead of pre-computing all k-mers as List<>.
  - Avoids inferring k-mers that are not used.

These optimizations add a further 50-70% speedup.

Optimizations for Speedup (3)

- Map each read pair in parallel.
    - Each read pair can be mapped independently of other pairs.
    - Biggest bottleneck is IO (serialization of writing to .bam file).

This optimization adds further 800-1000% speedup depending on the processor.

Results from SARS-COV-2

Total Number of Reads: 6,657,204

Total Time Taken: 20.52 seconds
(in 12-core/24-thread Ryzen 9 5900X CPU)

Reads per Min: 19,465,509

Recall: 0.999

Precision: 1.000

Limitations

- MSCRM doesn’t account for indels, performs ungapped alignment.
- Recall & precision may go down with high read error rate.
  - Can be mitigated to some degree with increasing the consensus threshold.

Questions?