Understanding and Characterization of Pangenomics

Noah Kaplan, Yufeng Gu, Reetuparna Das

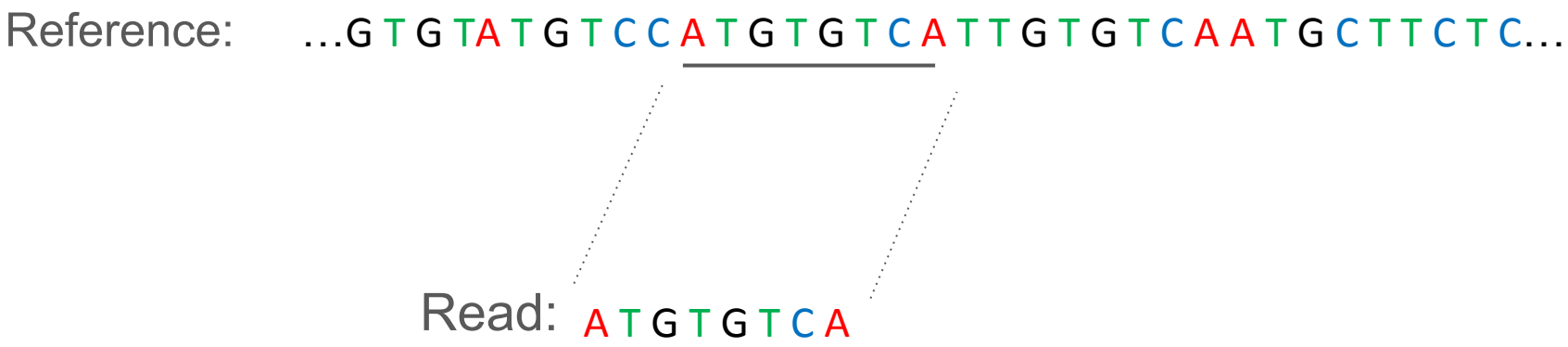
UNIVERSITY OF MICHIGAN

Linear Alignment

Reference: …G T G TA T G T C C A T G T G T C A T T G T G T C A A T G C T T C T C…

Read: A T G T G T C A

Linear Alignment

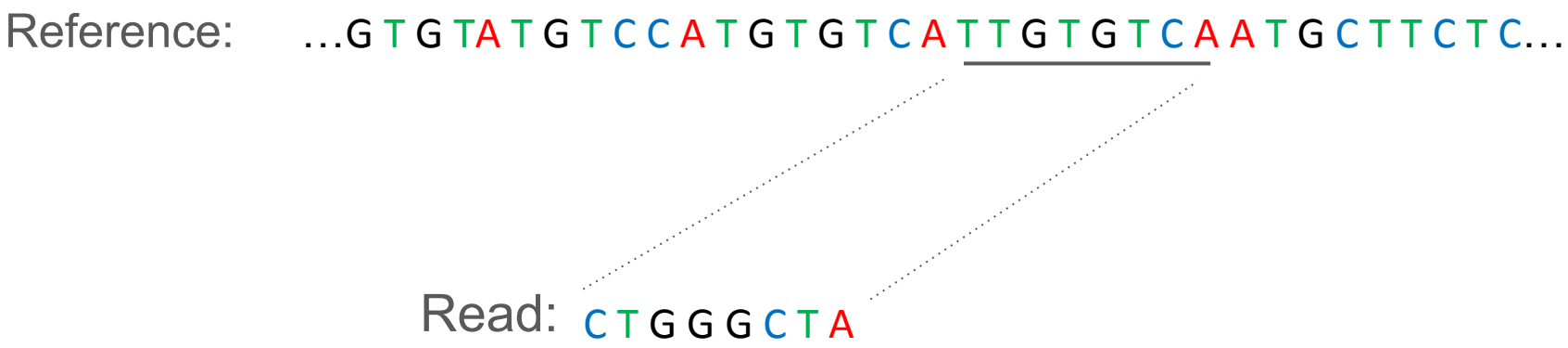


The Variant Problem

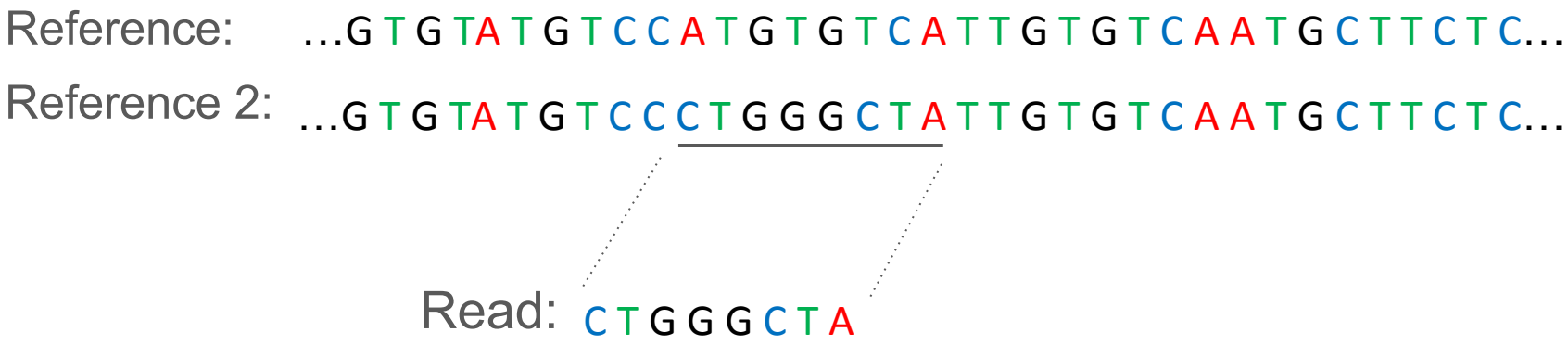
Reference: …G T G TA T G T C C A T G T G T C A T T G T G T C A A T G C T T C T C…

Read: C T G G G C T A

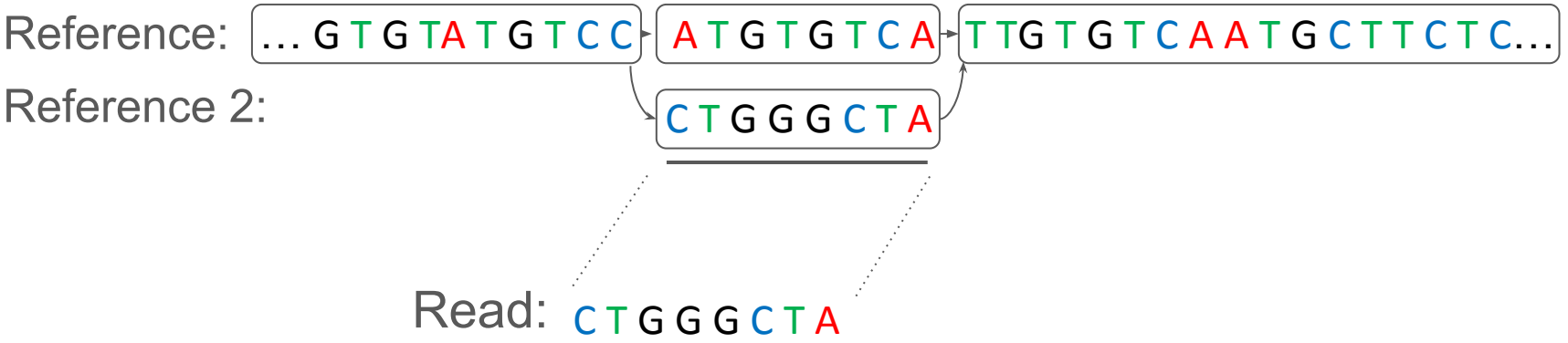
The Variant Problem



Pangenomes


Reference: …G T G TA T G T C C A T G T G T C A T T G T G T C A A T G C T T C T C…
Reference 2: …G T G TA T G T C C C T G G G C T A T T G T G T C A A T G C T T C T C…
Read: C T G G G C T A

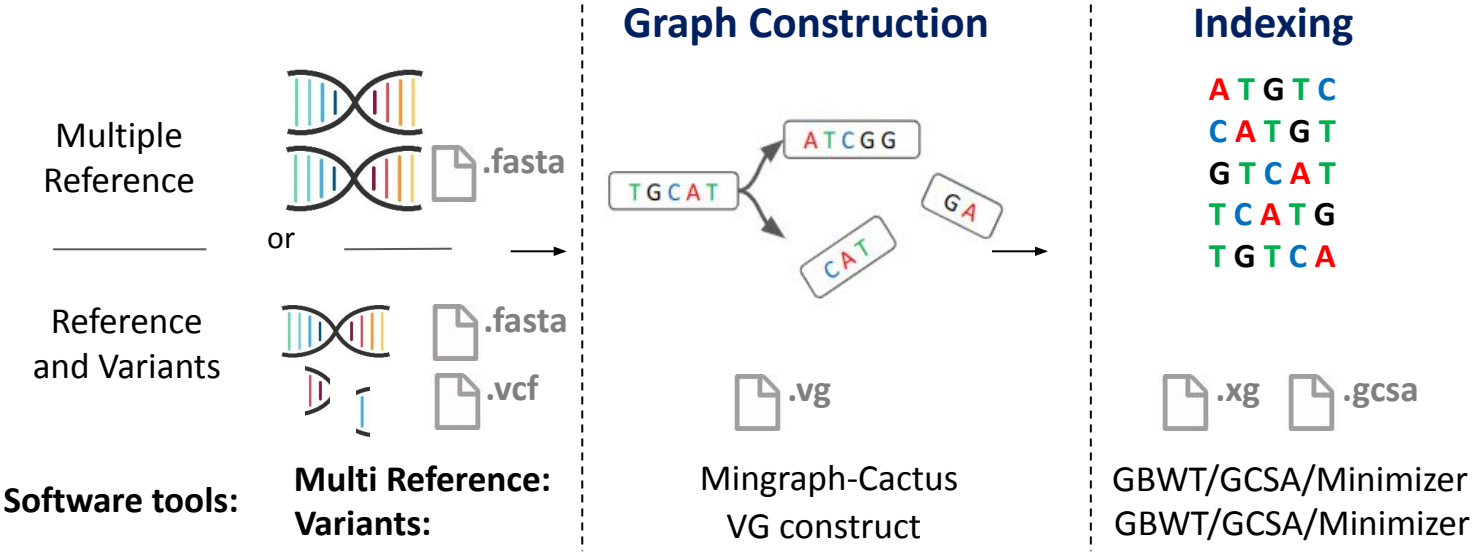
Pangenomes


Reference: … G T G TA T G T C C A T G T G T C A T TG T G T C A A T G C T T C T C…
Reference 2: C T G G G C T A
Read: C T G G G C T A

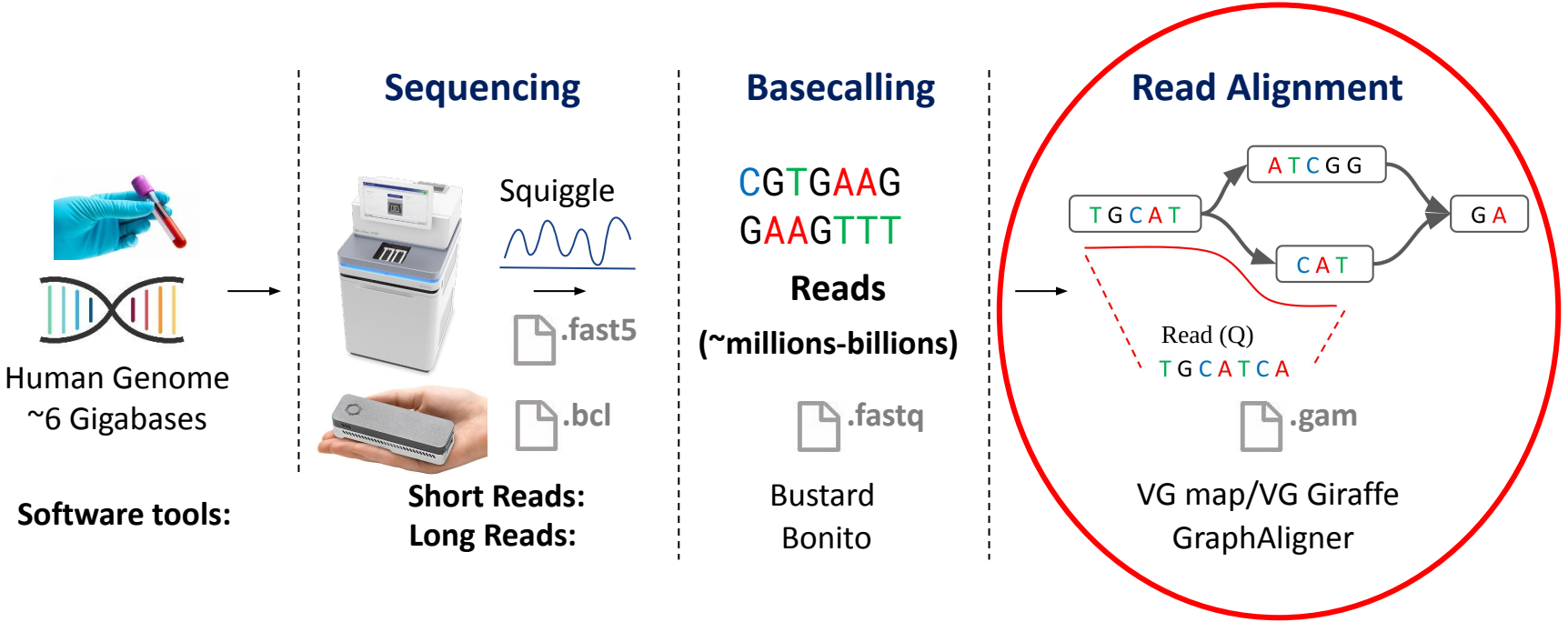
Outline

- **Workflows in Pangenomics**
- Common Tools
  - Vg Map (Short Reads)
  - GraphAligner (Long Reads)
- Profiling Results
- Conclusion

Pangenomics Reference Graph Building



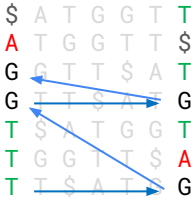
Pangenomics Alignment Workflow



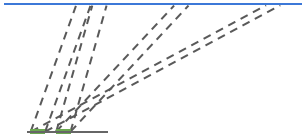
Outline

- Workflows in Pangenomics
- **Common Tools**
  - **Vg Map (Short Reads)**
  - GraphAligner (Long Reads)
- Profiling Results
- Conclusion

Algorithms in Vg Map

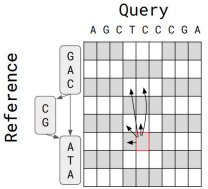


Seeding
(GCSA)



Clustering
(Distance Based)

Cluster Filtration
(Overlap Heuristic)



Extension
(GSSW)

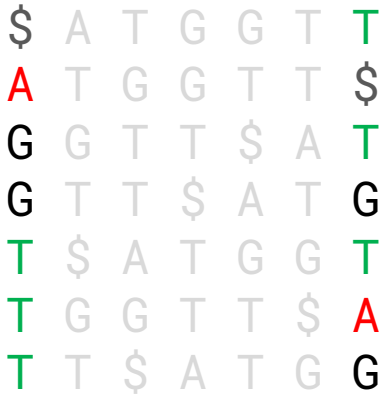
Seeding: GCSA

Reference: A T G G T T $

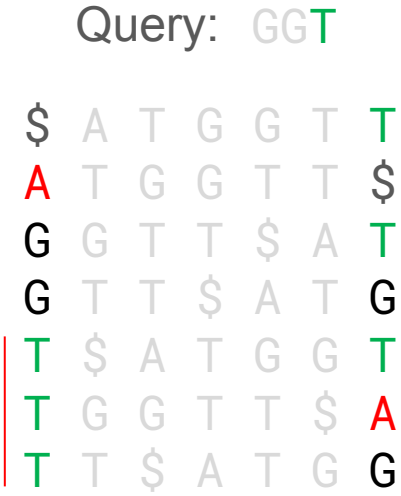
$ A T G G T T
A T G G T T $
G G T T $ A T
G T T $ A T G
T $ A T G G T
T G G T T $ A
T T $ A T G G

Seeding: GCSA

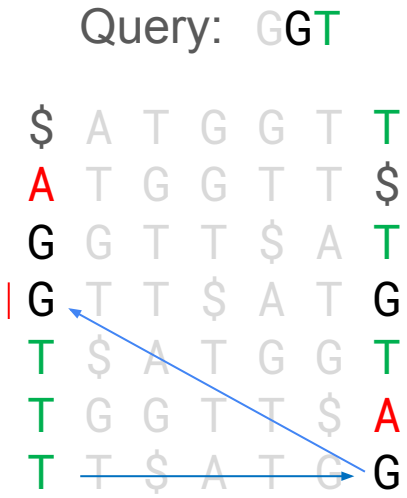
Query: GGT



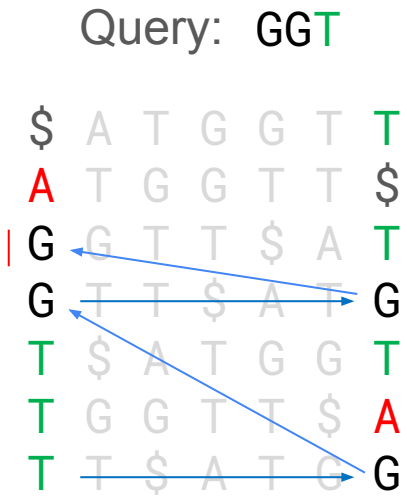
Seeding: GCSA



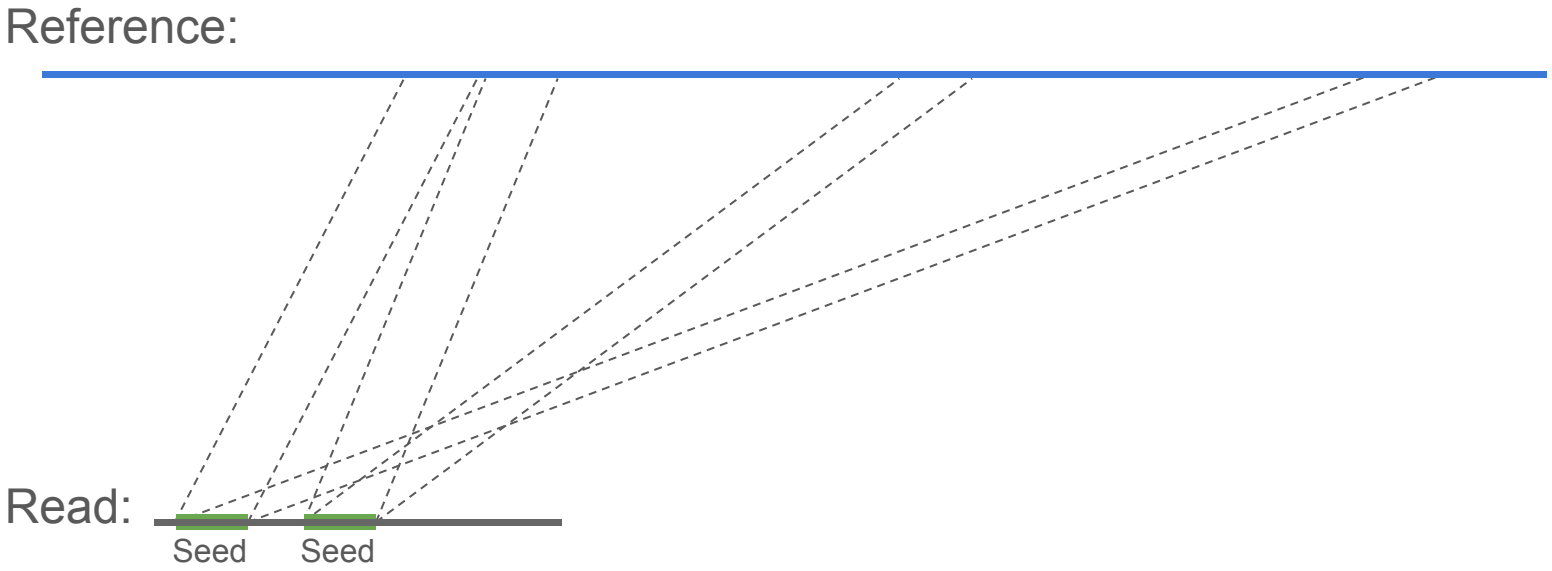
Seeding: GCSA



Seeding: GCSA



Clustering

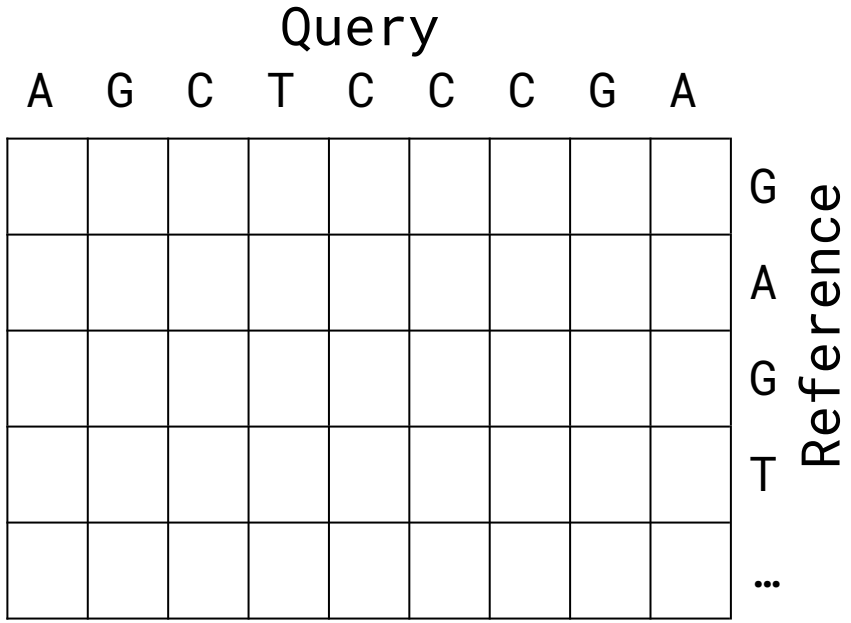


Cluster Filtration

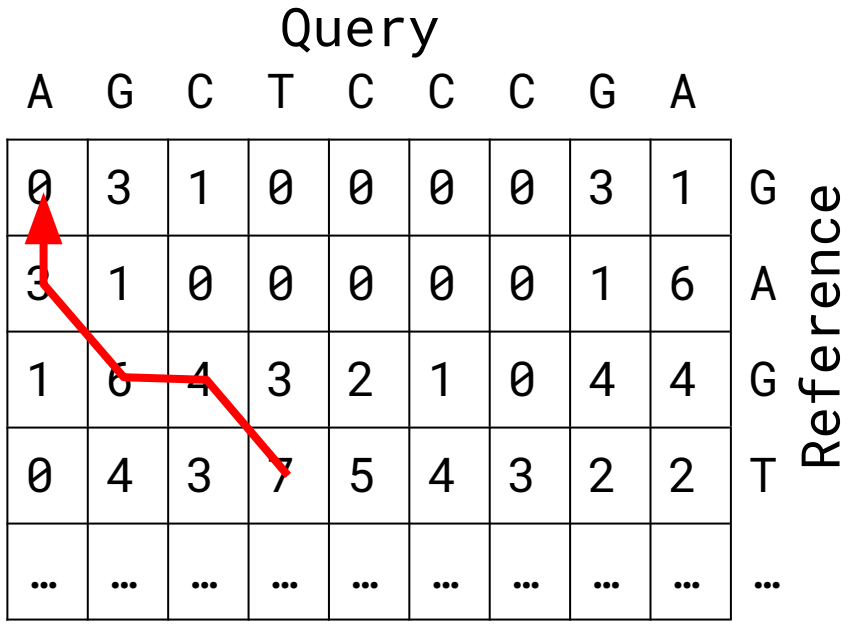
- If a cluster is too short, drop it
- If a cluster overlaps with another cluster, drop it

Clusters:

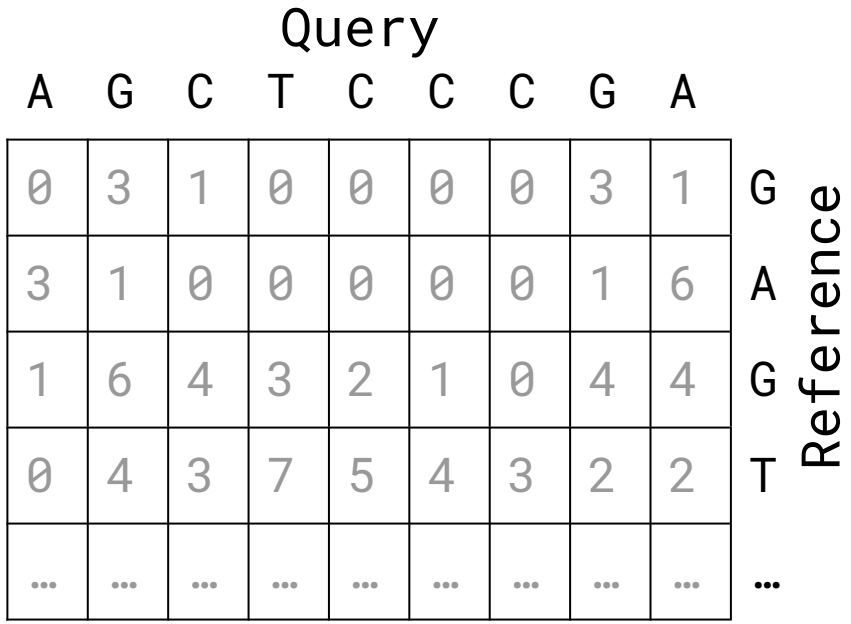
Extension: Graph Simd **Smith-Waterman** (GSSW)



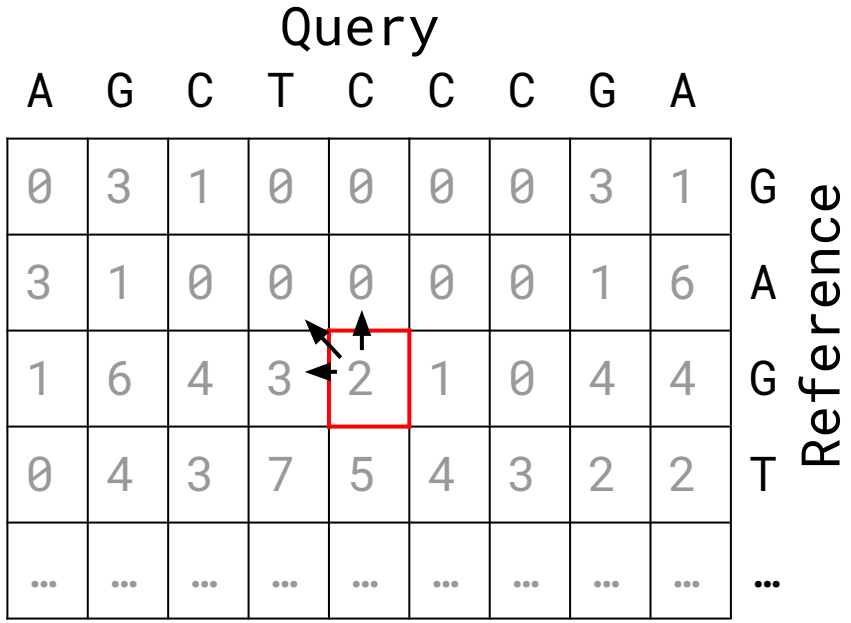
Extension: Graph Simd **Smith-Waterman** (GSSW)



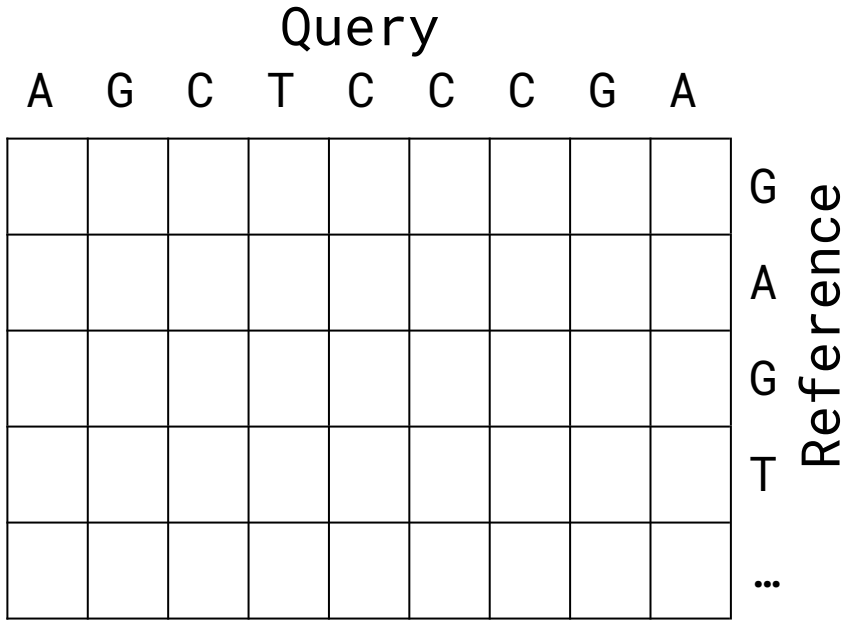
Extension: Graph Simd **Smith-Waterman** (GSSW)



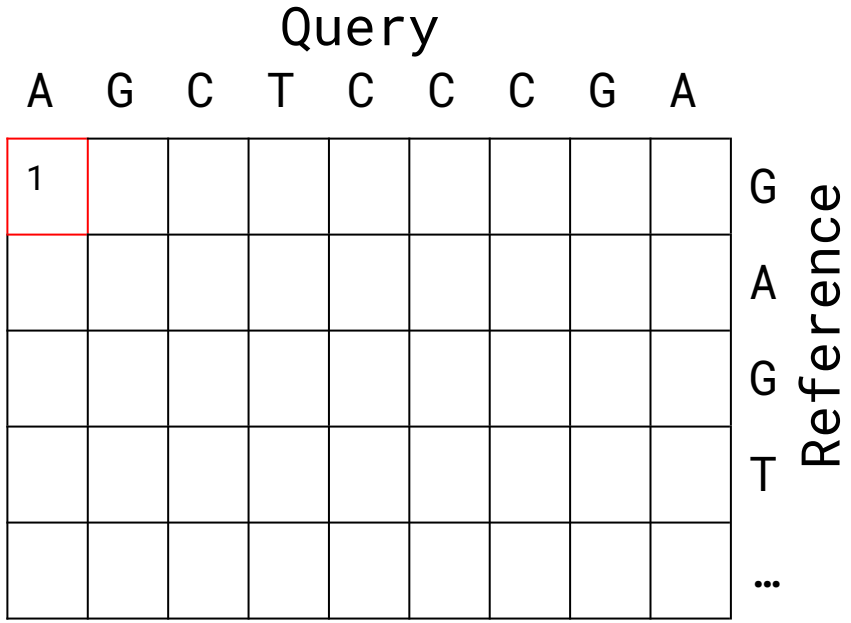
Extension: Graph Simd **Smith-Waterman** (GSSW)



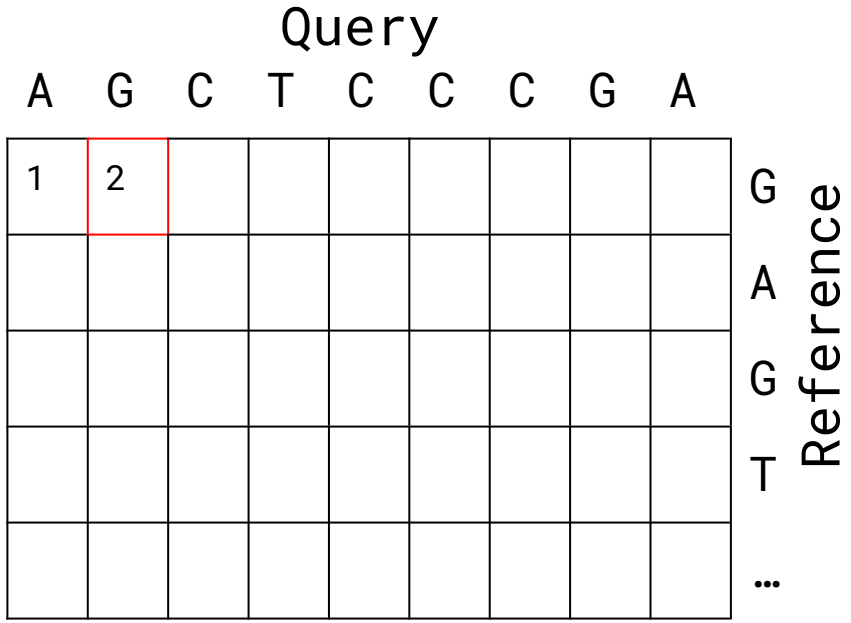
Extension: Graph Simd **Smith-Waterman** (GSSW)



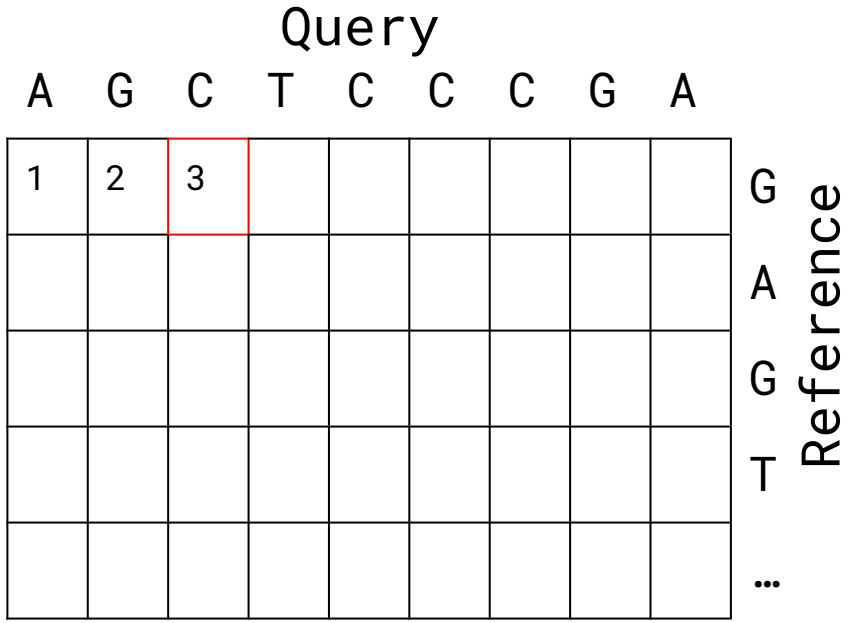
Extension: Graph Simd **Smith-Waterman** (GSSW)



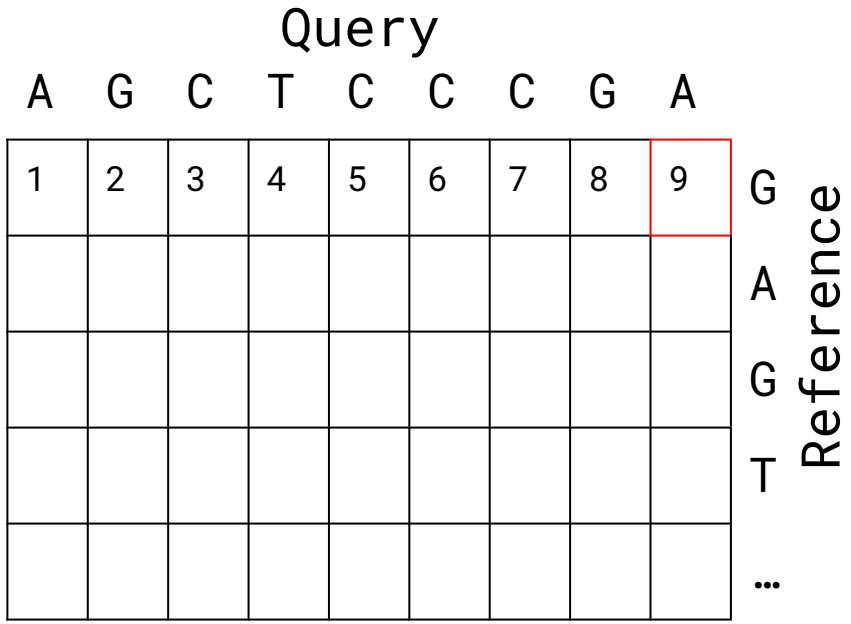
Extension: Graph Simd **Smith-Waterman** (GSSW)



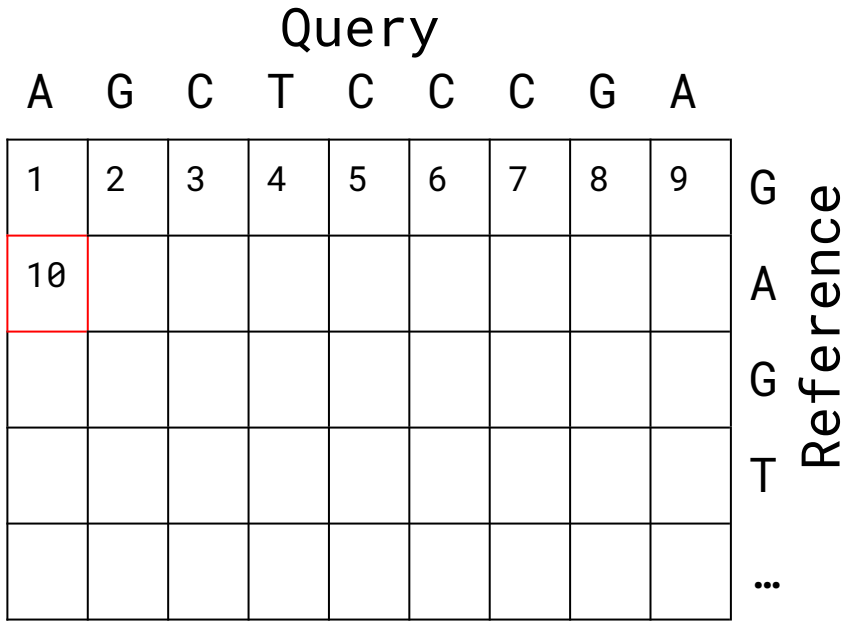
Extension: Graph Simd **Smith-Waterman** (GSSW)



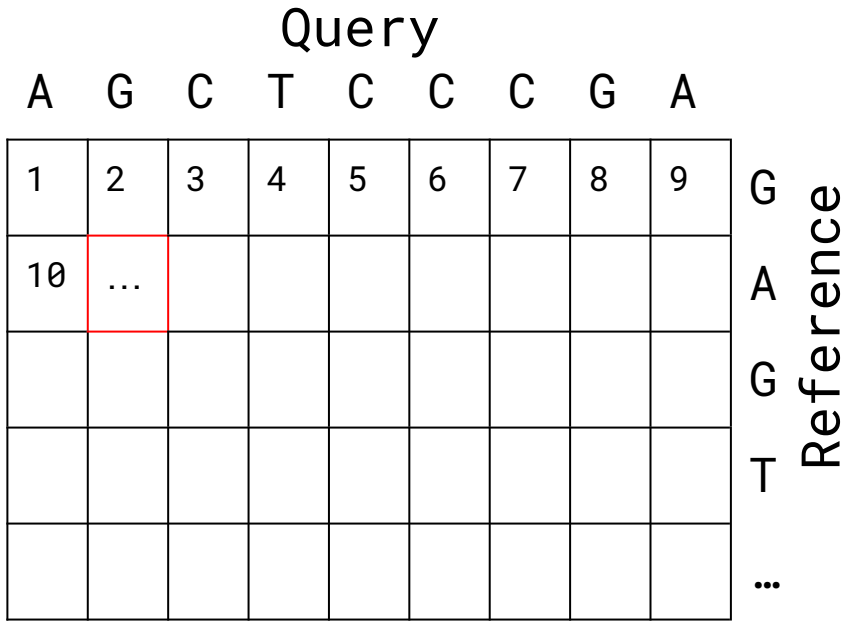
Extension: Graph Simd **Smith-Waterman** (GSSW)



Extension: Graph Simd **Smith-Waterman** (GSSW)

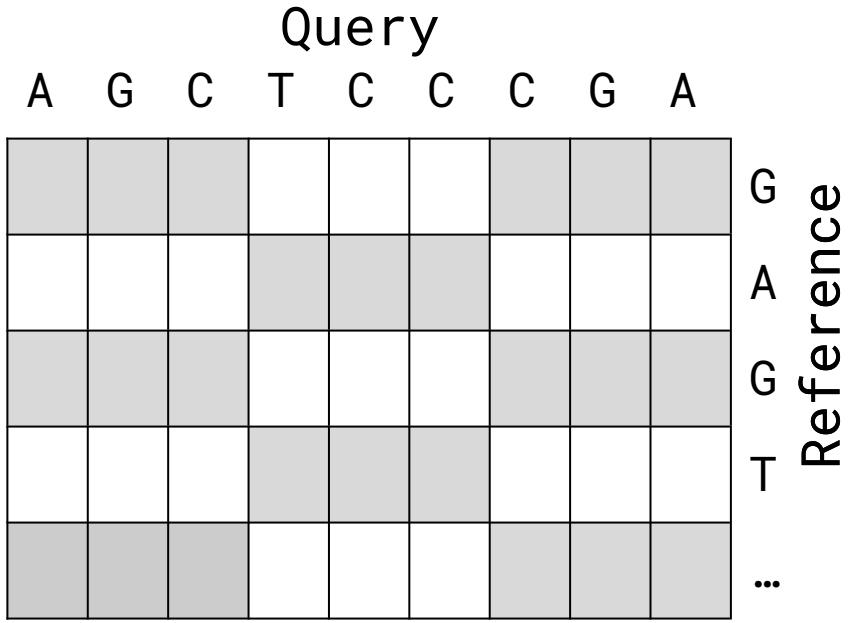


Extension: Graph Simd **Smith-Waterman** (GSSW)



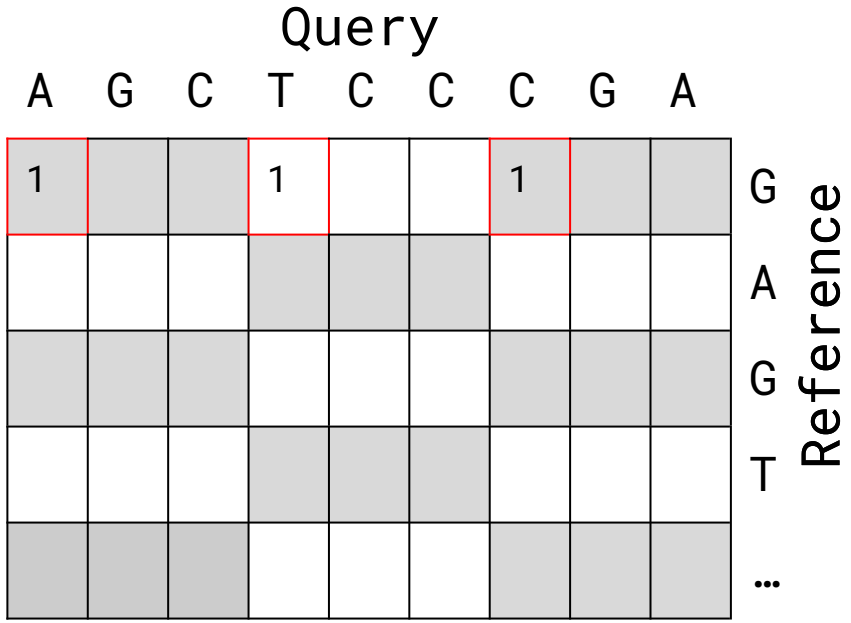
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3



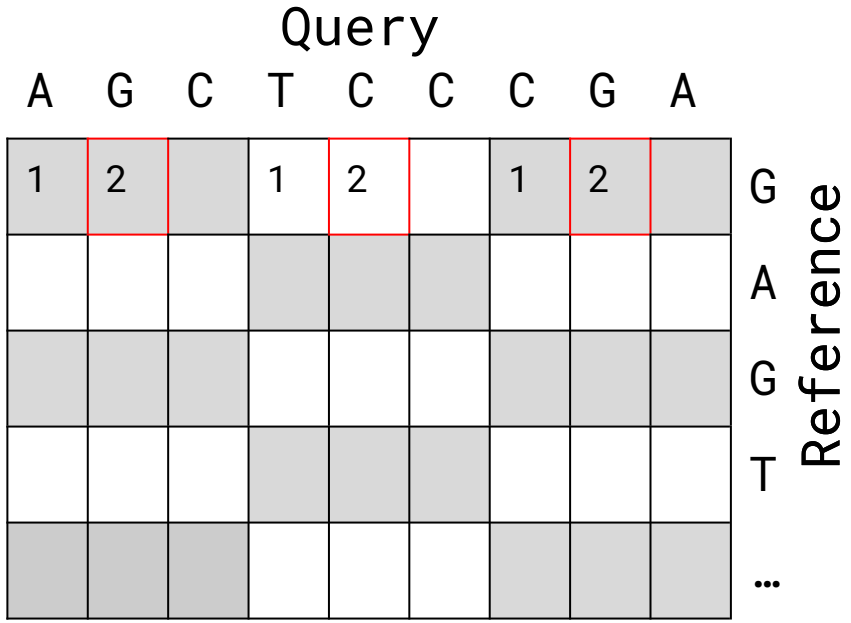
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3



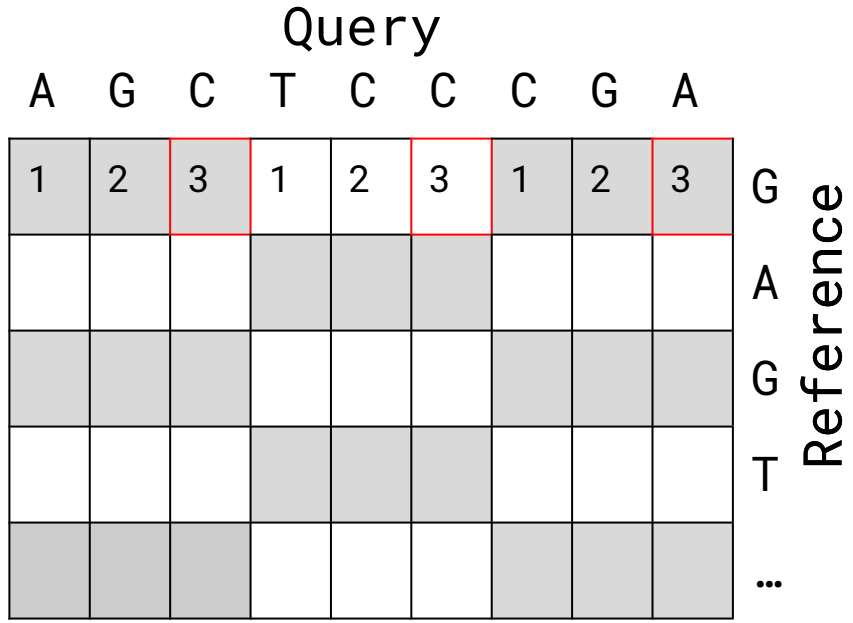
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3



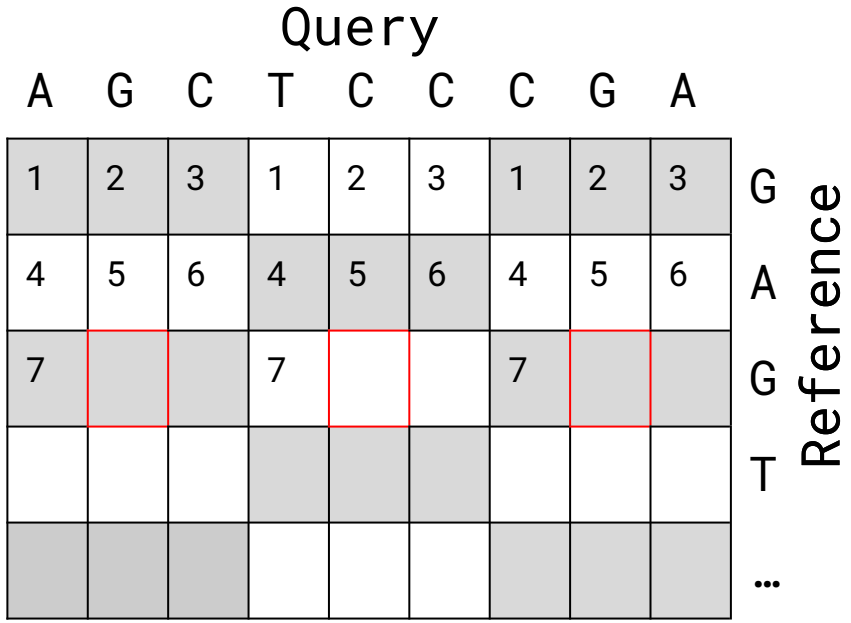
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3



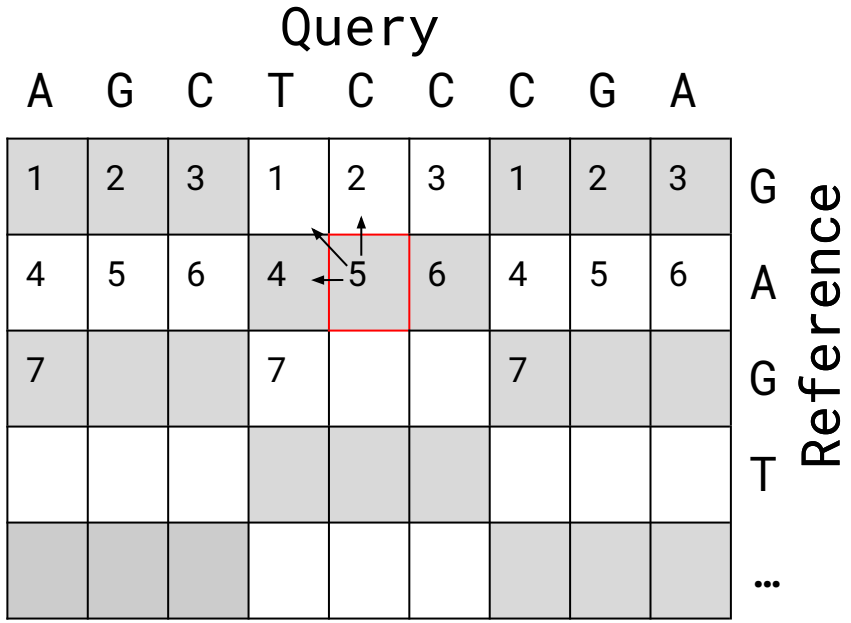
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3



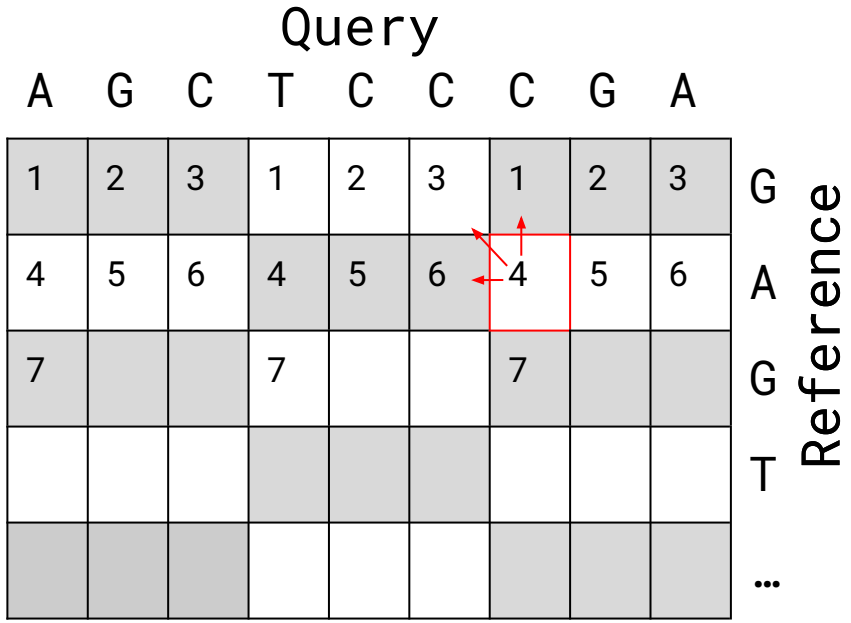
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3

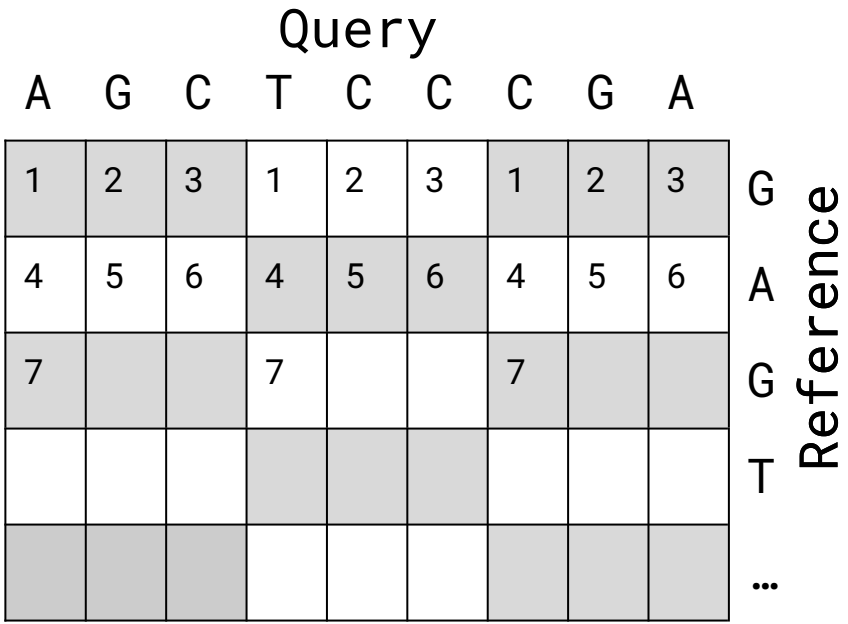


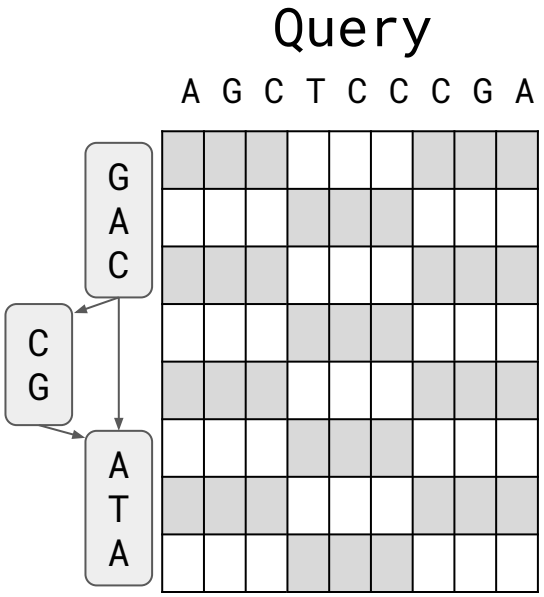
Extension: Graph **Simd Smith-Waterman** (GSSW)

wordSize=3

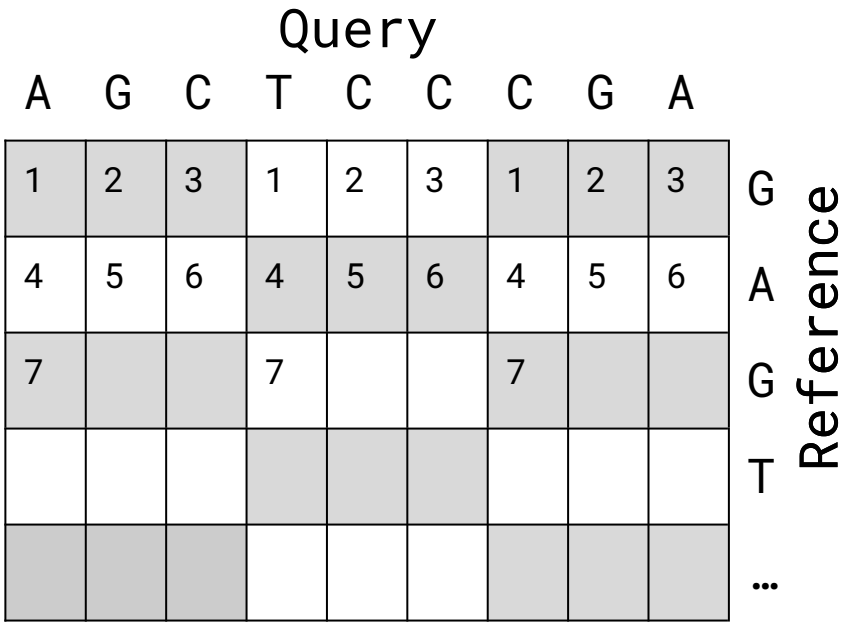


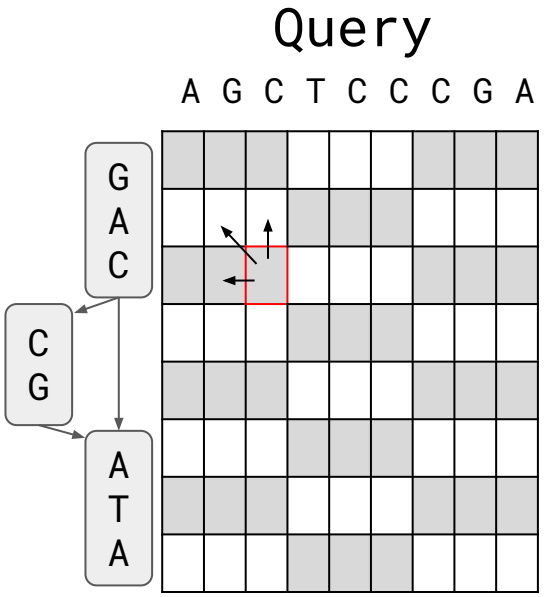
Extension: **Graph Simd Smith-Waterman** (GSSW)



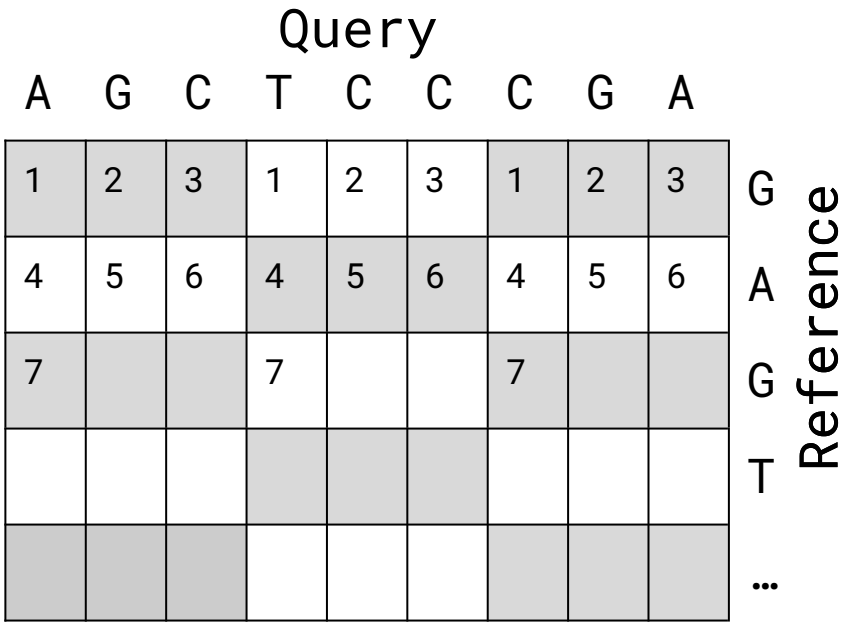


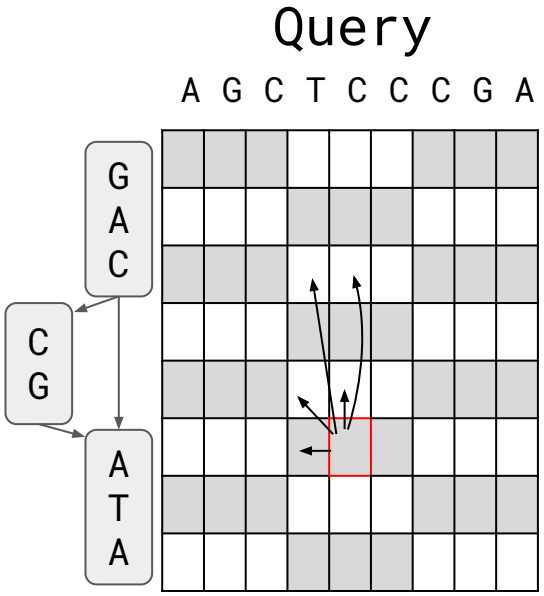
Extension: **Graph Simd Smith-Waterman** (GSSW)





Extension: **Graph Simd Smith-Waterman** (GSSW)

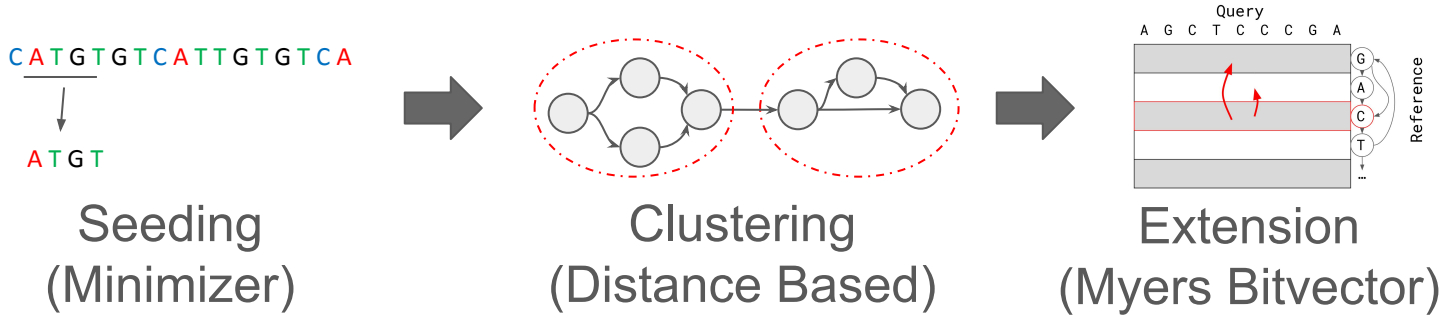




Outline

- Workflows in Pangenomics
- **Common Tools**
  - Vg Map (Short Reads)
  - **GraphAligner (Long Reads)**
- Profiling Results
- Conclusion

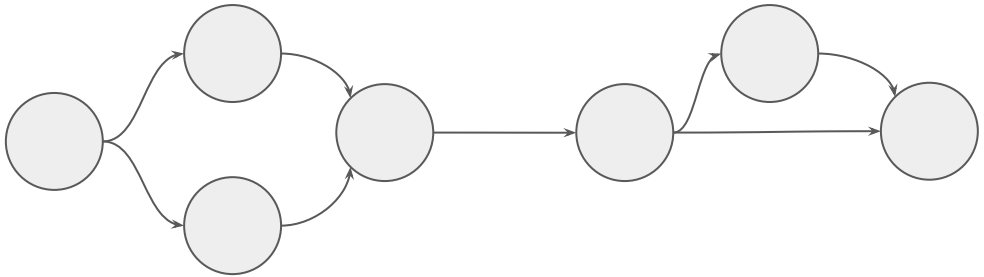
Algorithms in GraphAligner



Seeding: Minimizer

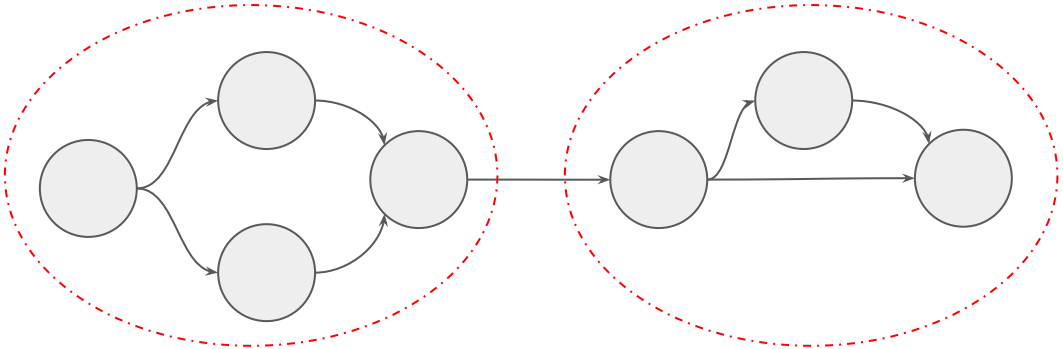
Clustering

Reference Graph:



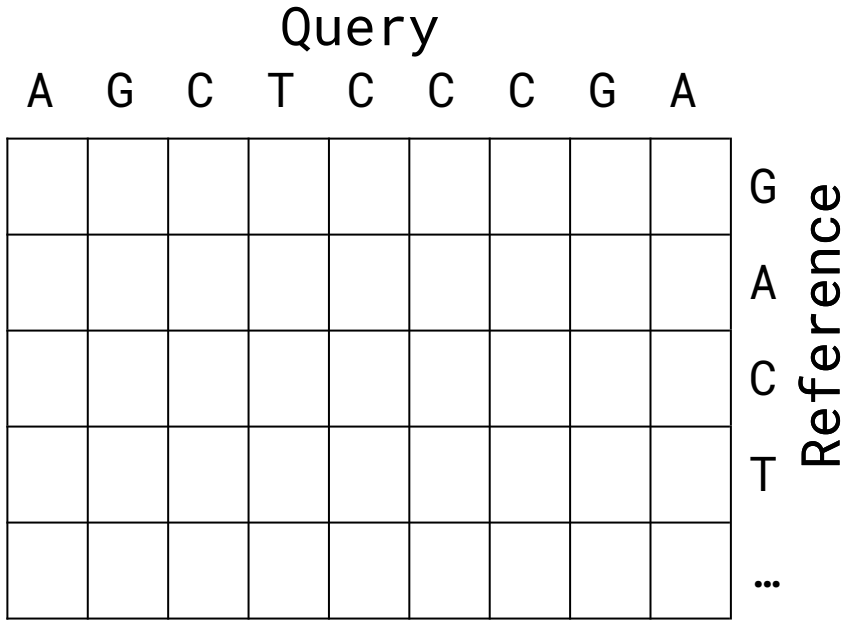
Clustering

Reference Graph:

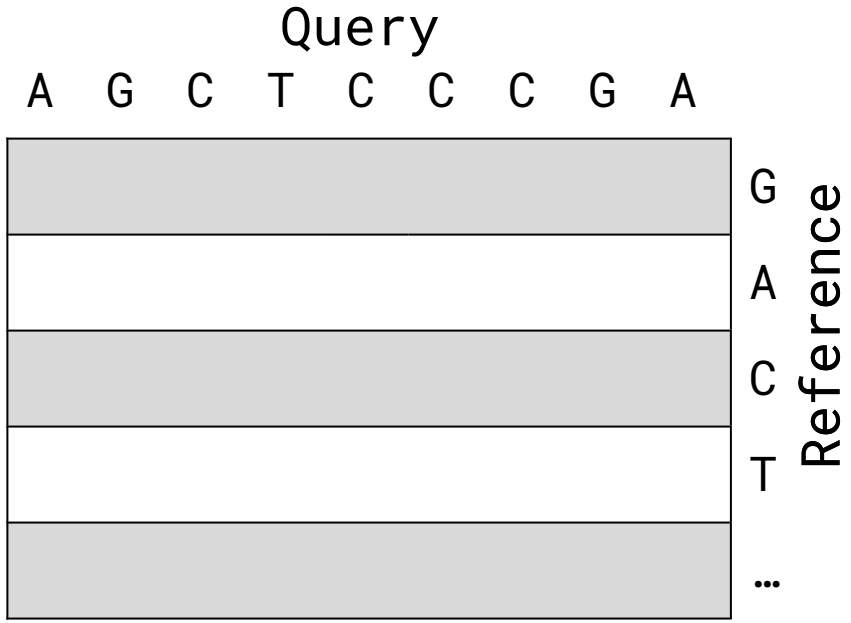


Create linear chains of
“superbubbles”

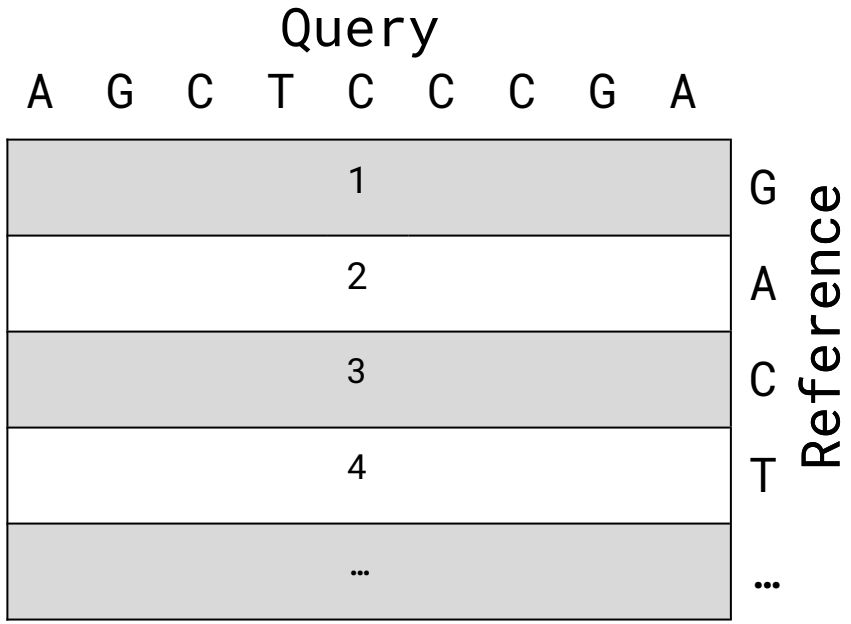
Extension: Myers Bitvector



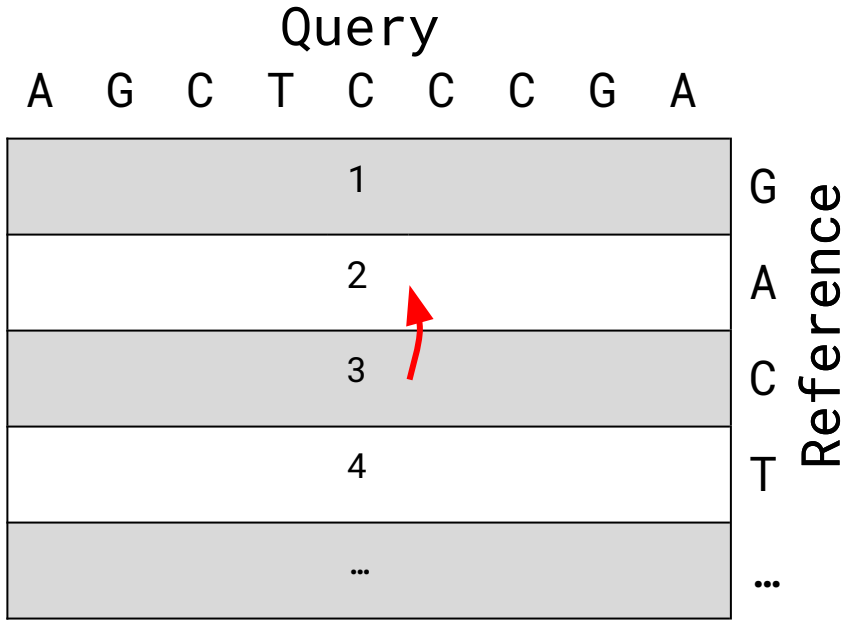
Extension: Myers Bitvector



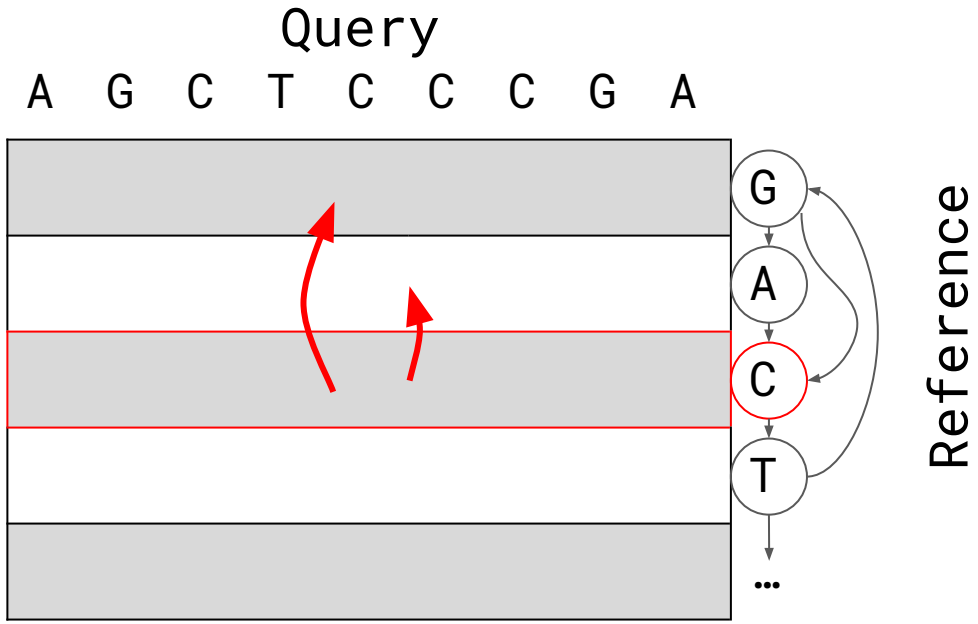
Extension: Myers Bitvector



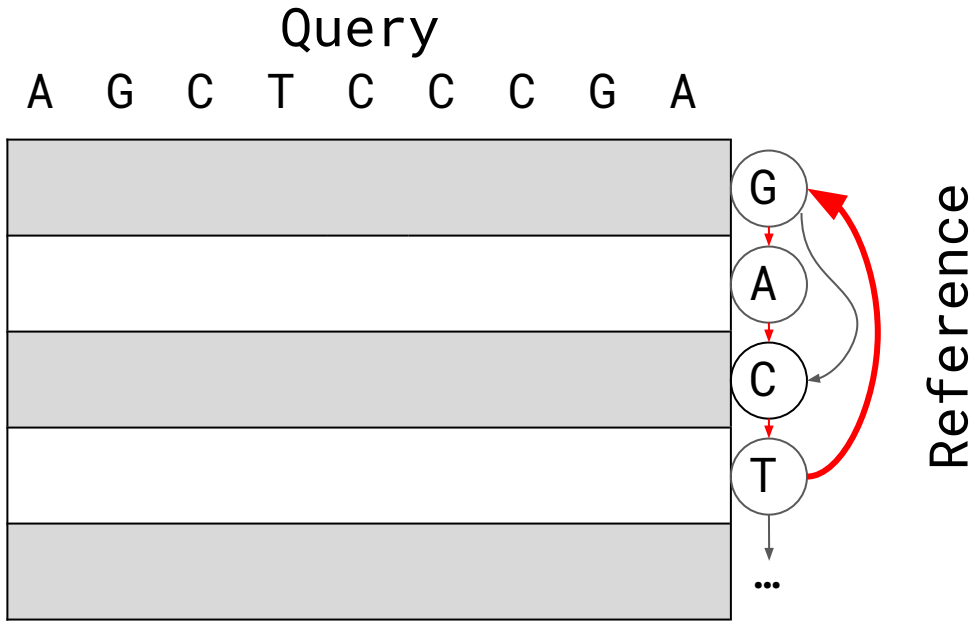
Extension: Myers Bitvector



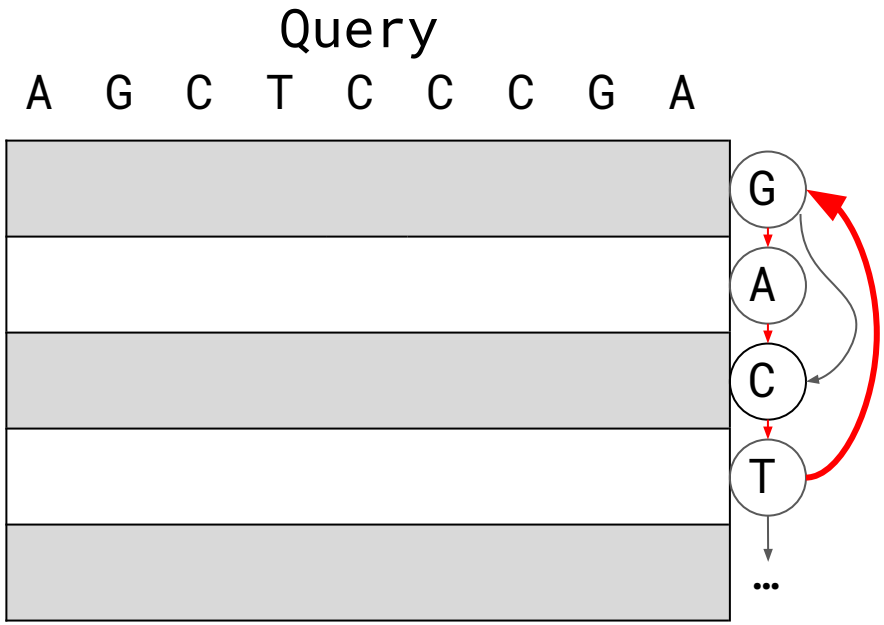
Extension: Myers Bitvector



Extension: Myers Bitvector



Extension: Myers Bitvector



Reference

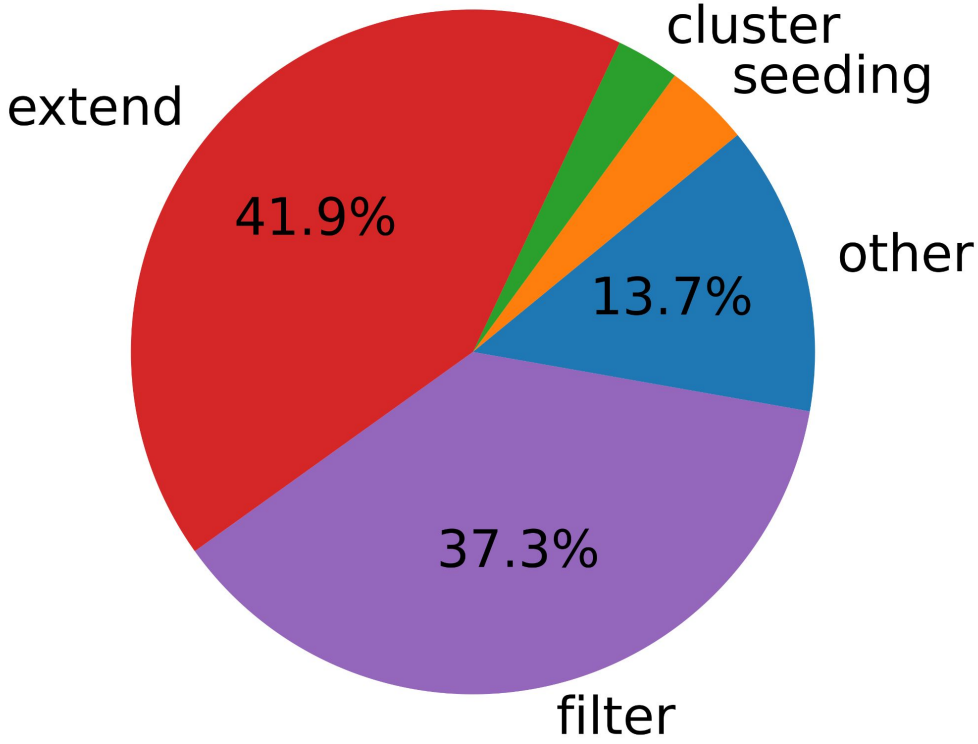
Priority Queue

- Pop when you process a node
- Push when a node changes
- Terminate when empty

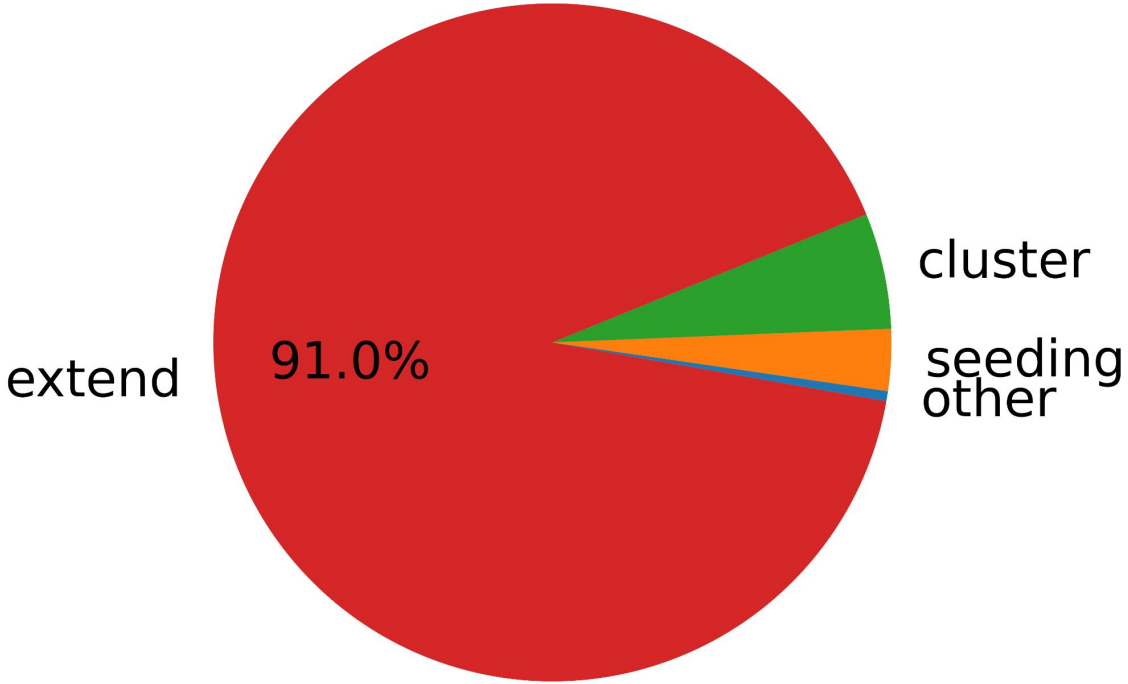
Outline

- Workflows in Pangenomics
- Common Tools
  - Vg Map (Short Reads)
  - GraphAligner (Long Reads)
- **Profiling Results**
- Conclusion

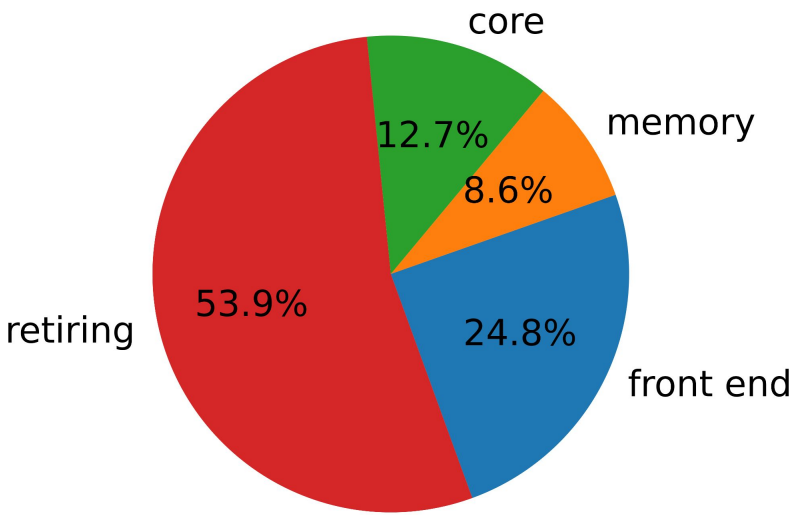
Vg Map Timing Breakdown



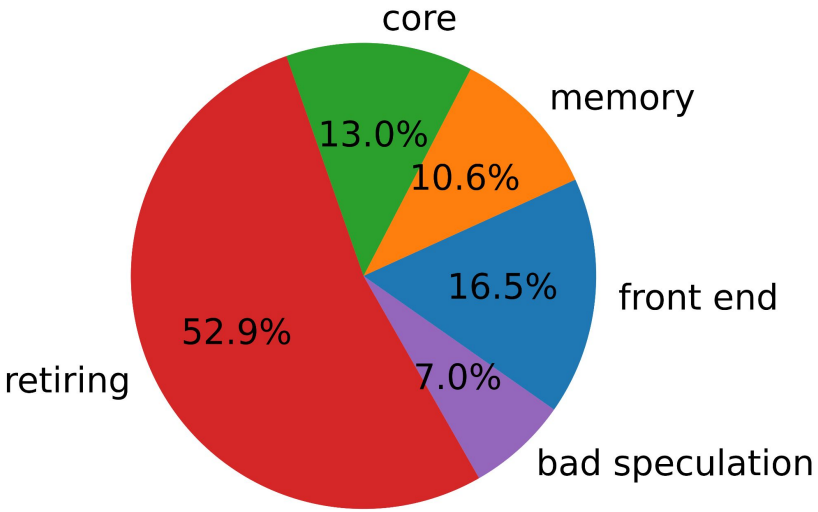
GraphAligner Timing Breakdown



Microarchitecture Utilization



Vg Map



GraphAligner

Outline

- Workflows in Pangenomics
- Common Tools
  - Vg Map (Short Reads)
  - GraphAligner (Long Reads)
- Profiling Results
- **Conclusion**

Conclusion

- **Seed extension** is the bottleneck for short and long read alignment
- **Cluster filtration** is also a bottleneck for Vg
- Multiple resource constraints restrict application performance without a single bottleneck

Understanding and Characterization of Pangenomics

Noah Kaplan, Yufeng Gu, Reetuparna Das

UNIVERSITY OF MICHIGAN