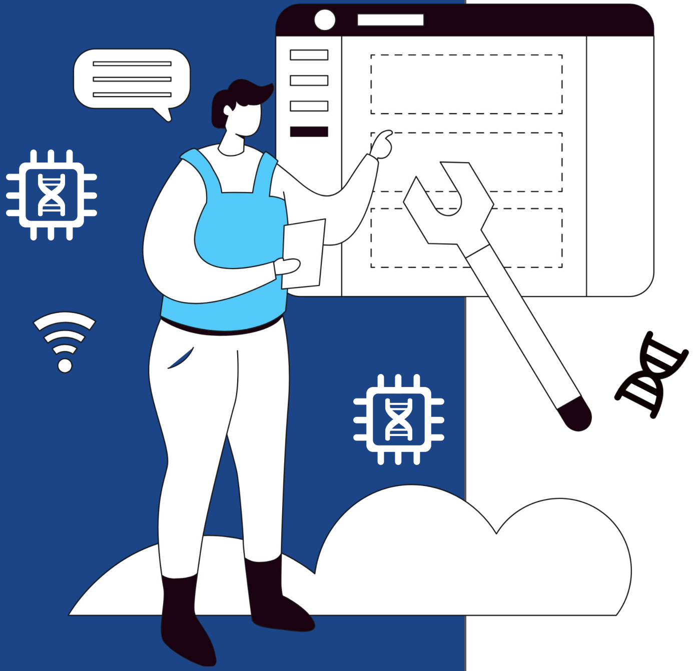


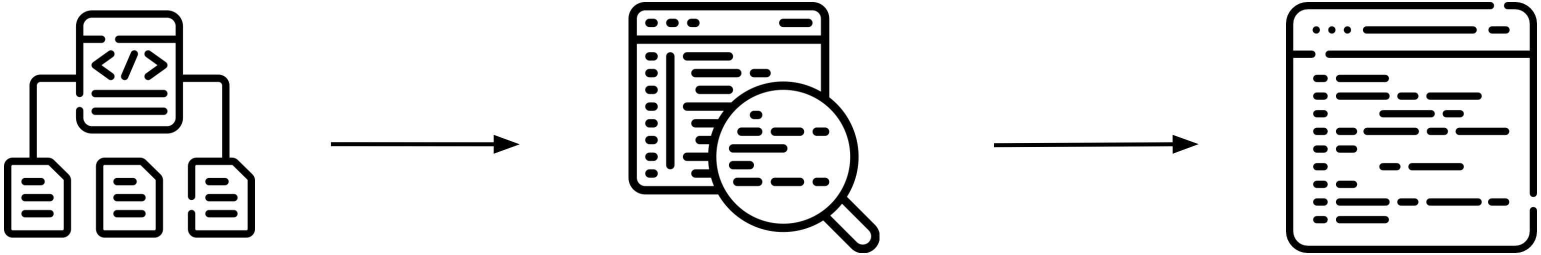
Developing a Proxy Application Based on a Parallel Pangenome Mapping Tool

Jessica Imlau Dagostini[1], Scott Beamer[1], Tyler Sorensen[1]
Joseph B. Manzano[2], Andres Marquez[2]
1. {jessica.dagostini, sbeamer, tyler.sorensen}@ucsc.edu
2. {joseph.manzano, andres.marquez}@pnnl.gov

Scientific Applications are hard to profile...

- **Complex inputs, many dependencies, complex code.**
  - Testing **software optimizations** is **hard**.
- **Hard to port** the code to different architectures.

We need some way to address such challenges



Full Application

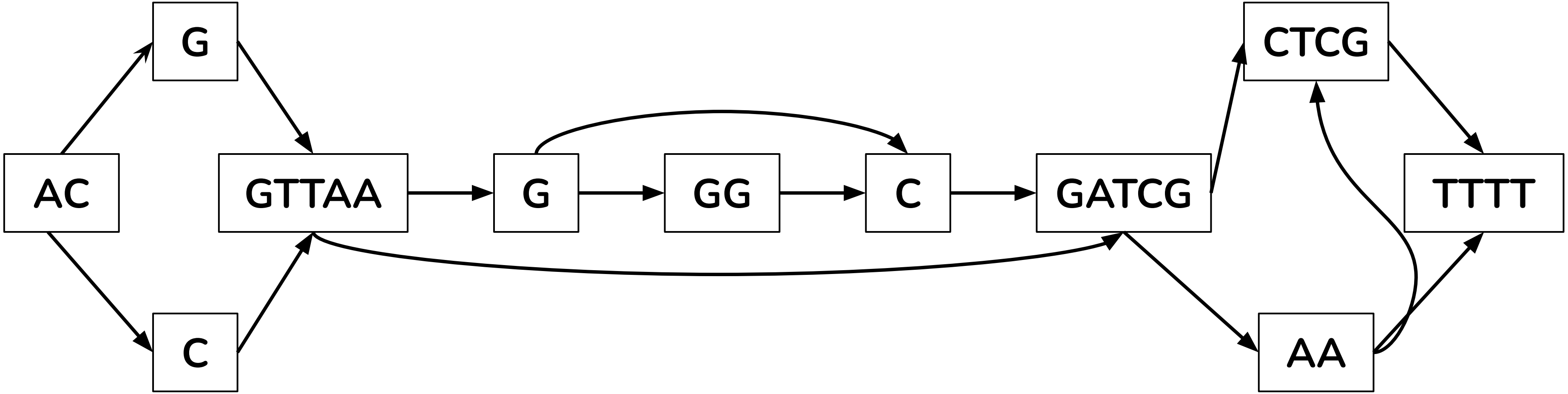
Identify Main Kernel

Proxy App

Proxy Applications

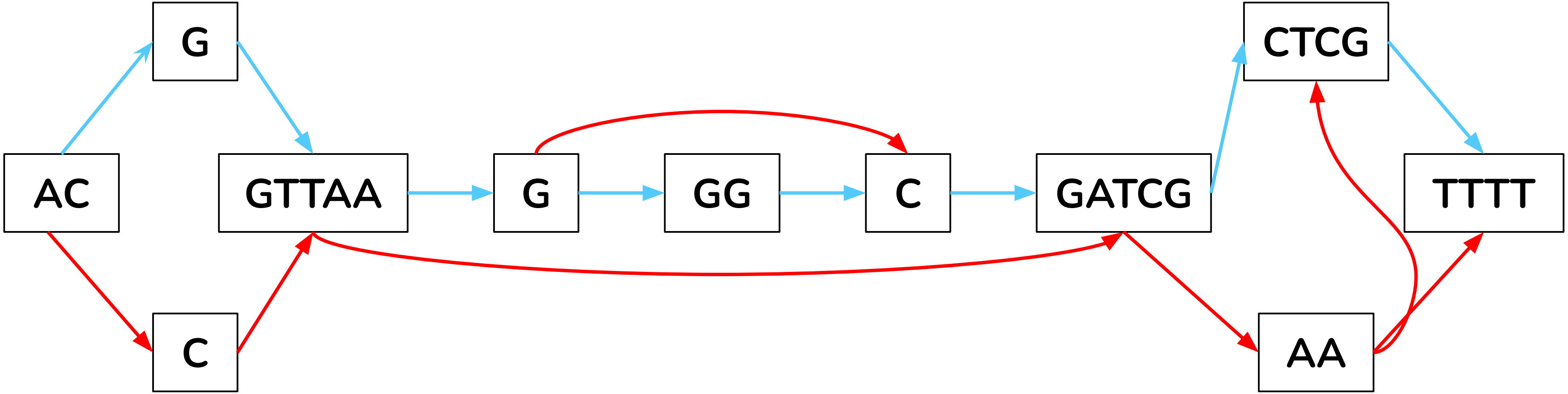
Lightweight representation of real workloads that ease algorithmic exploration, analysis, optimization, and even porting.

This work takes the first steps, i.e. performance characterization, to create a **proxy application** that can **facilitate the exploration** of the original, more complex code that **maps sequence reads to pangenomes**.



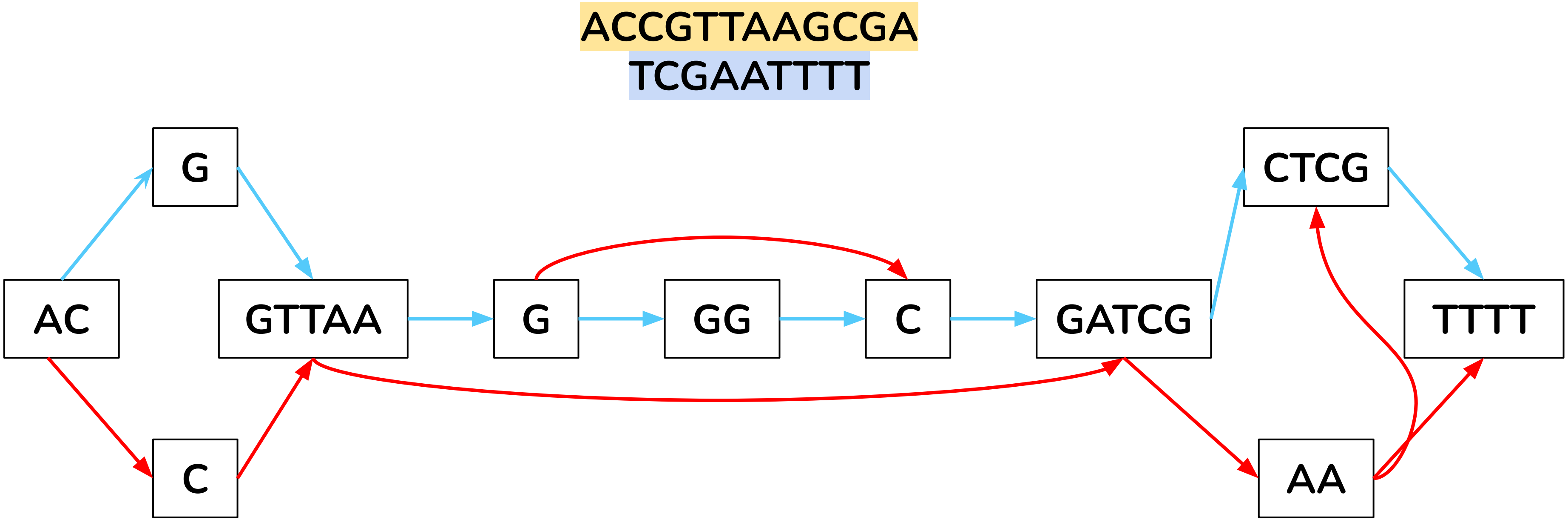
Source: Baaijens, J.A., Bonizzoni, P., Boucher, C. et al., 2022

This work takes the first steps, i.e. performance characterization, to create a **proxy application** that can **facilitate the exploration** of the original, more complex code that **maps sequence reads to pangenomes**.



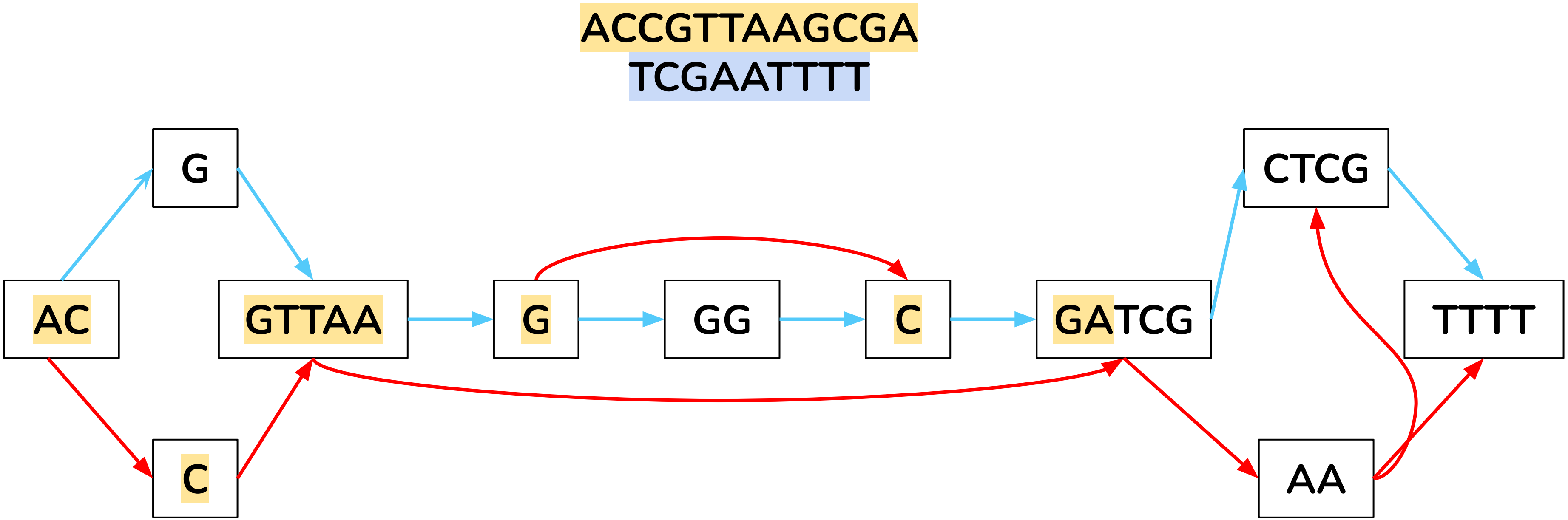
Source: Baaijens, J.A., Bonizzoni, P., Boucher, C. et al., 2022

This work takes the first steps, i.e. performance characterization, to create a **proxy application** that can **facilitate the exploration** of the original, more complex code that **maps sequence reads to pangenomes**.



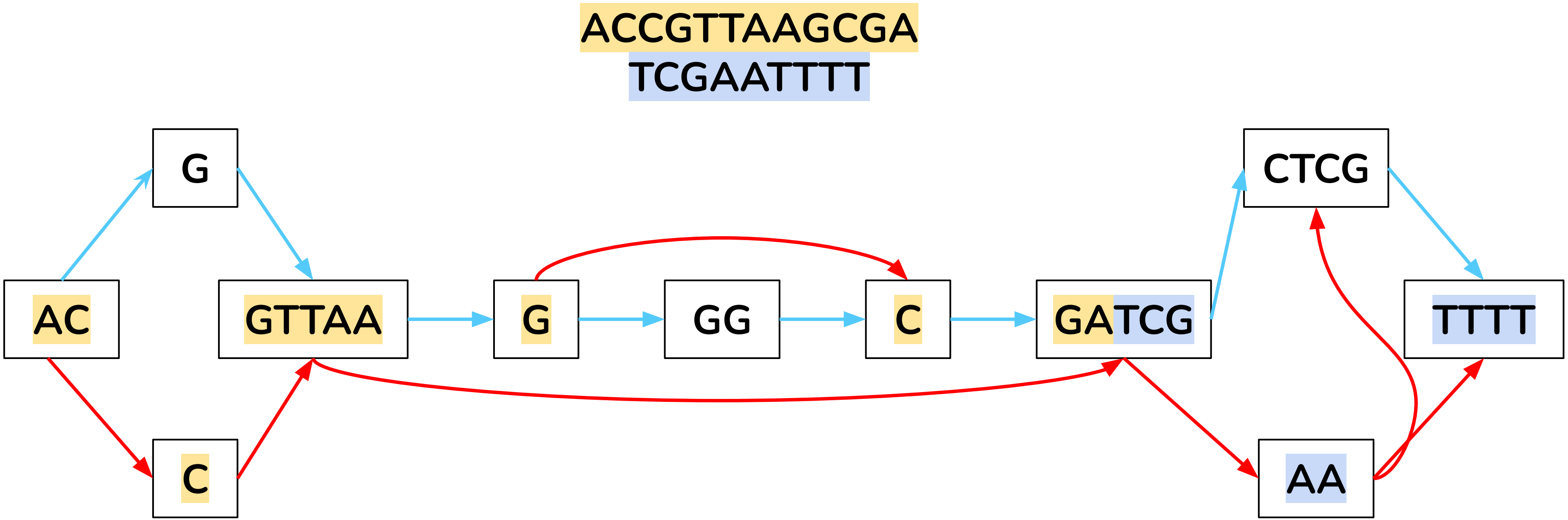
Source: Baaijens, J.A., Bonizzoni, P., Boucher, C. et al., 2022

This work takes the first steps, i.e. performance characterization, to create a **proxy application** that can **facilitate the exploration** of the original, more complex code that **maps sequence reads to pangenomes**.



Source: Baaijens, J.A., Bonizzoni, P., Boucher, C. et al., 2022

This work takes the first steps, i.e. performance characterization, to create a **proxy application** that can **facilitate the exploration** of the original, more complex code that **maps sequence reads to pangenomes**.



Source: Baaijens, J.A., Bonizzoni, P., Boucher, C. et al., 2022

Genome References

- Pangenome references bring more computational complexity
- Map reads can be harder to do than in single reference
  - Much more data to verify
- Giraffe is one of the first mapping tool for pangenome
  - From VG Toolkit
  - First to have similar performance to single reference mappings

Many dependencies

Complex Source-Code

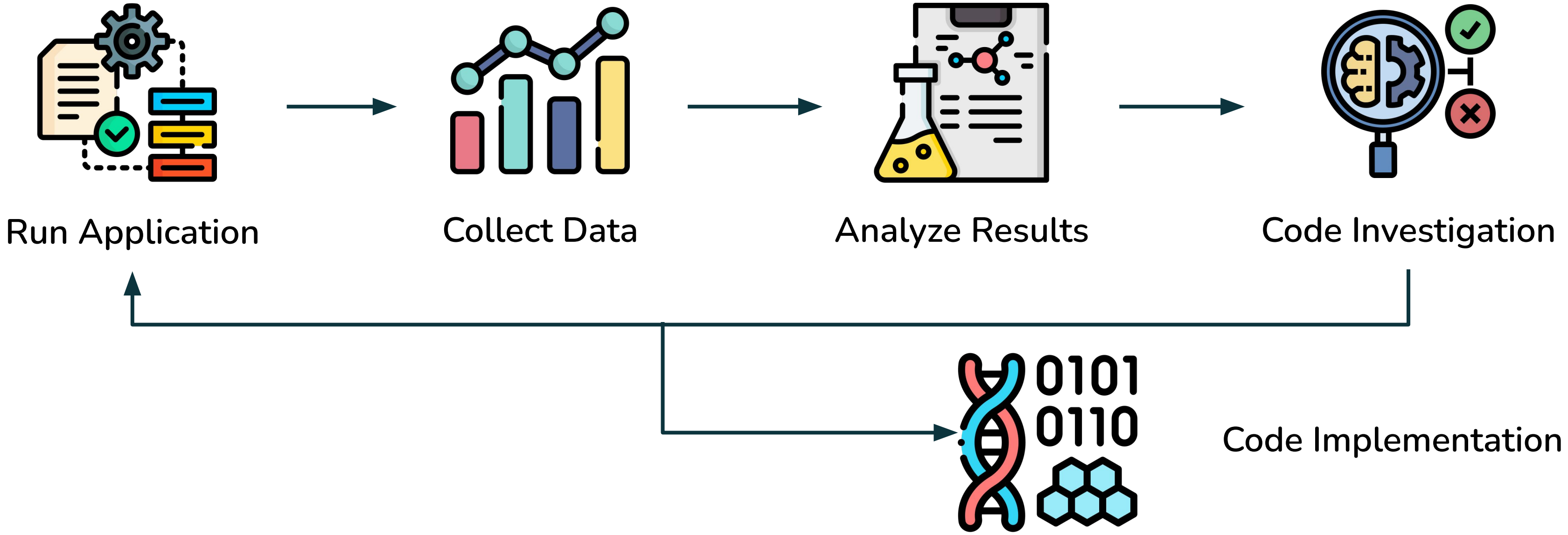
Complex Input

Giraffe Mapping and VG

- **VG**[1] is a toolkit to help in the **alignment** and **mapping** of **DNA sequences**
    - **Variation Graphs** (vg) is a **succinct encoding** of the **short-reading sequences** of **many genomes**
- **Giraffe** is the tool that makes fast, haplotype-based **mapping** to a **pangenome** graph
- They use Graph BWT (**GBWT**)
    - Multi-string FM-index for **indexing** large collections of similar paths
- VG has a **big source-code** and the tools inside it **share code and data structures**
    - More than 50k lines of code
    - Such dependencies make **hard to test** different computational strategies

[1] https://doi.org/10.1038/nbt.4227

Methodology



Run Application

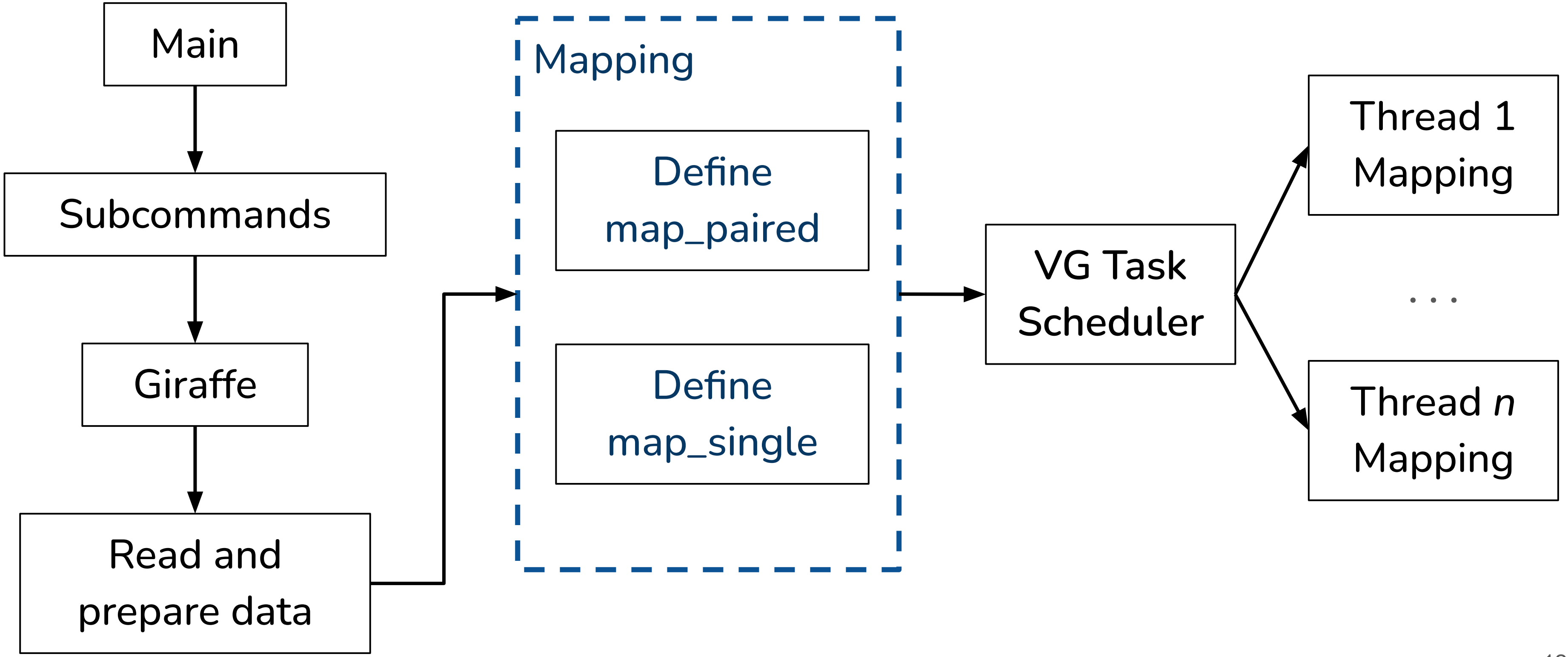
Collect Data

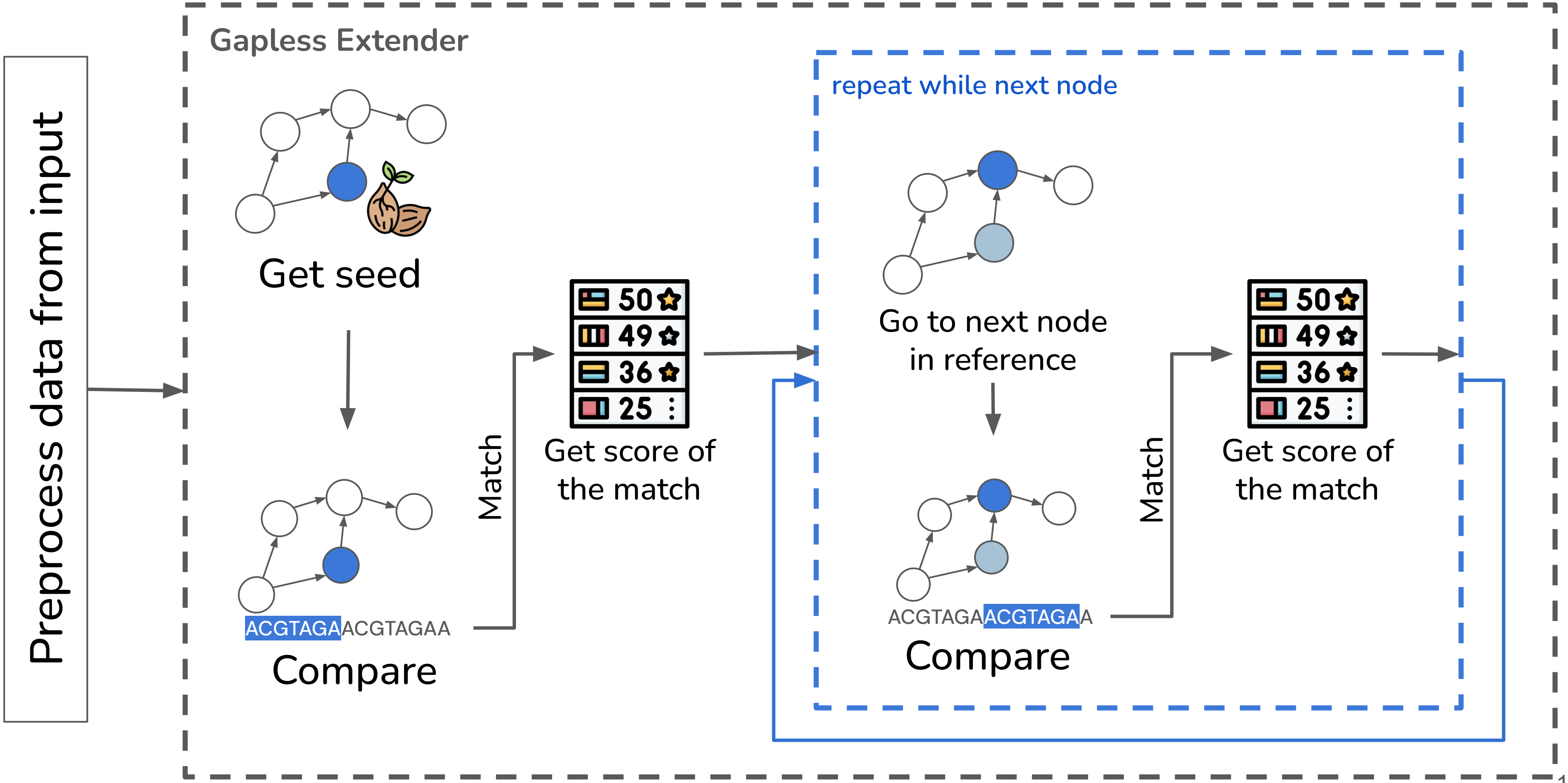
Analyze Results

Code Investigation

Code Implementation

Profiling and Kernel Identification





Preprocess data from input

Gapless Extender

Get seed

Compare

Match

Get score of the match

repeat while next node

Go to next node in reference

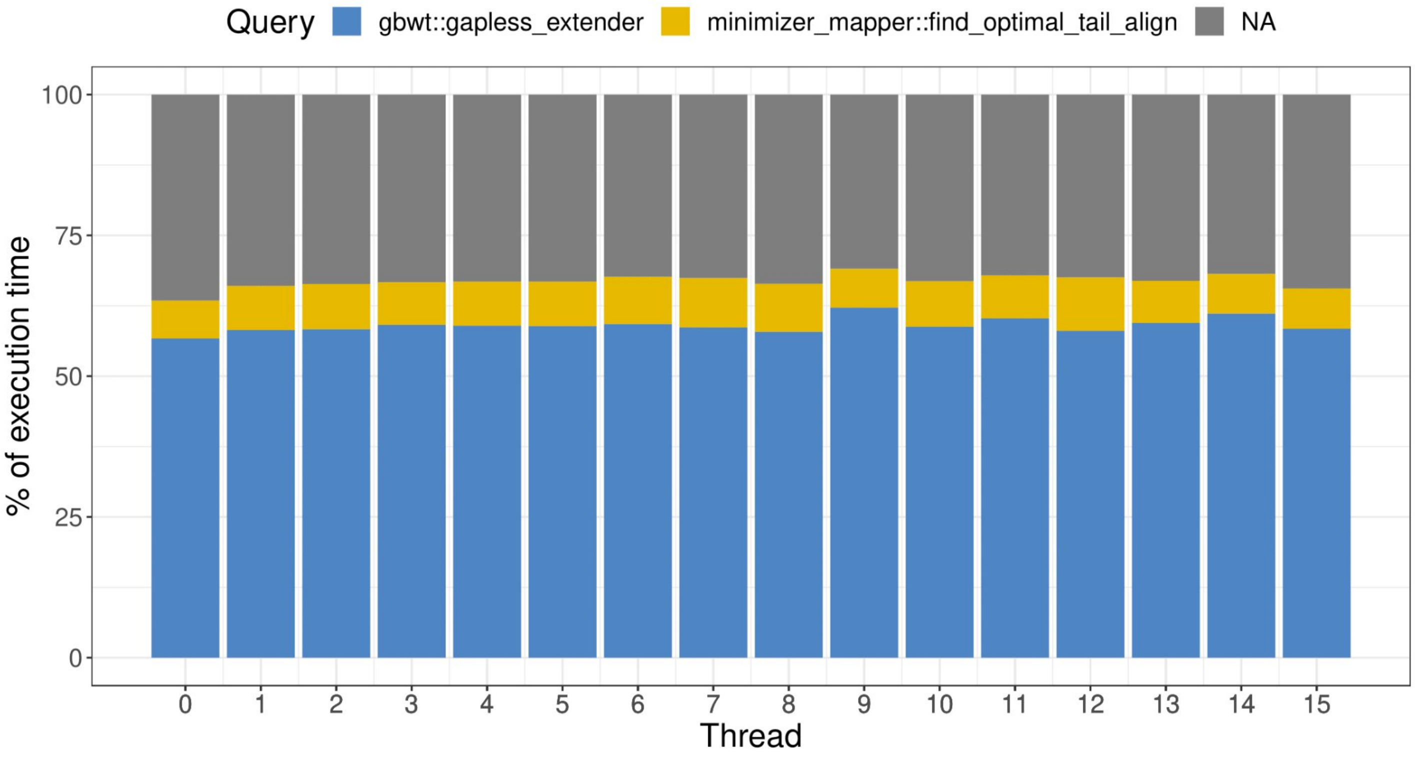
Compare

Match

Get score of the match

Profiling

gapless_ extender represents more than **60%** of the application's runtime in all threads.



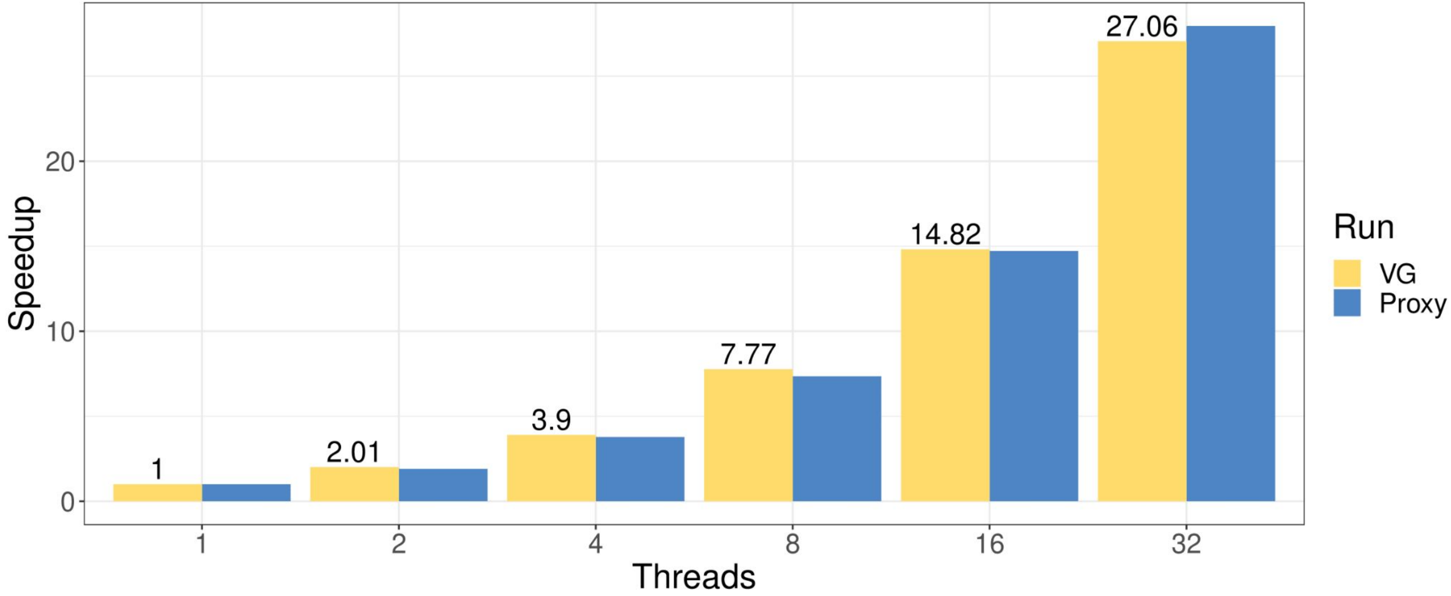
Giraffe *vs* Proxy

- ~50k lines of code
- ~350 source-files
- ~50 library dependencies
- Difficult to compile
- More than 60s to compile
- Hard to modify

- ~1k lines of code
- 2 source-files
- 3 library dependencies
- Easy to compile
- Around 10s to compile
- Easy to modify and play with different strategies

Initial Results

- First working version produce results with high accuracy
- Reproduce speedup values parallelizing with *OpenMP Dynamic Scheduler*



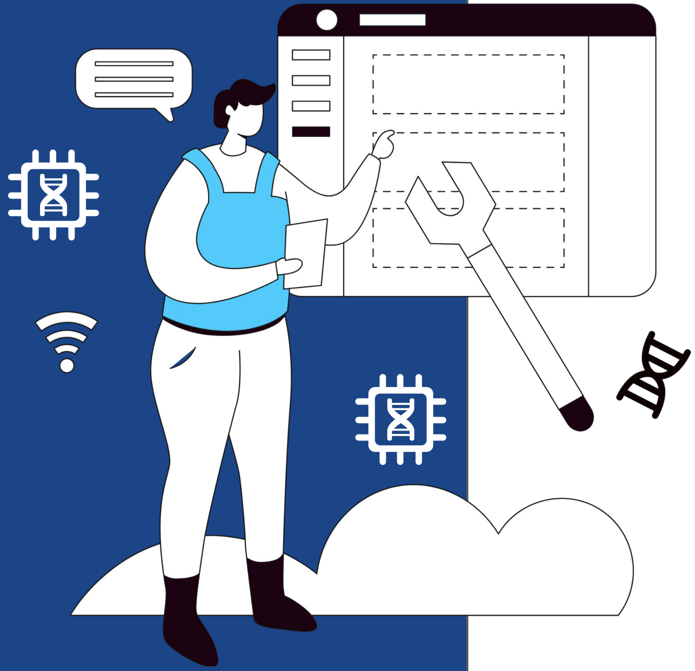
Pangenome Proxy Application

- Pangenome Proxy App can be a **valuable tool** for **testing new strategies** of performance improvement focused on **short-read mappings to pangenomes**
  - Parallelization strategies, workload characterization, hardware acceleration, etc
- In the near future:
  - Validate proxy's memory access pattern is representative of original application
  - Perform a more **complete workload characterization** of the mapping process
  - Propose **different parallelization strategies**
  - Run on **different architectures** (GPUs, FPGAs, etc)

Jessica Imlau Dagostini[1],
Scott Beamer[1],
Tyler Sorensen[1]
Joseph B. Manzano[2],
Andres Marquez[2]

1. {jessica.dagostini, sbeamer, tyler.sorensen}@ucsc.edu
2. {joseph.manzano, andres.marquez}@pnnl.gov

Scan for contact information



Developing a Proxy Application Based on a Parallel Pangenome Mapping Tool

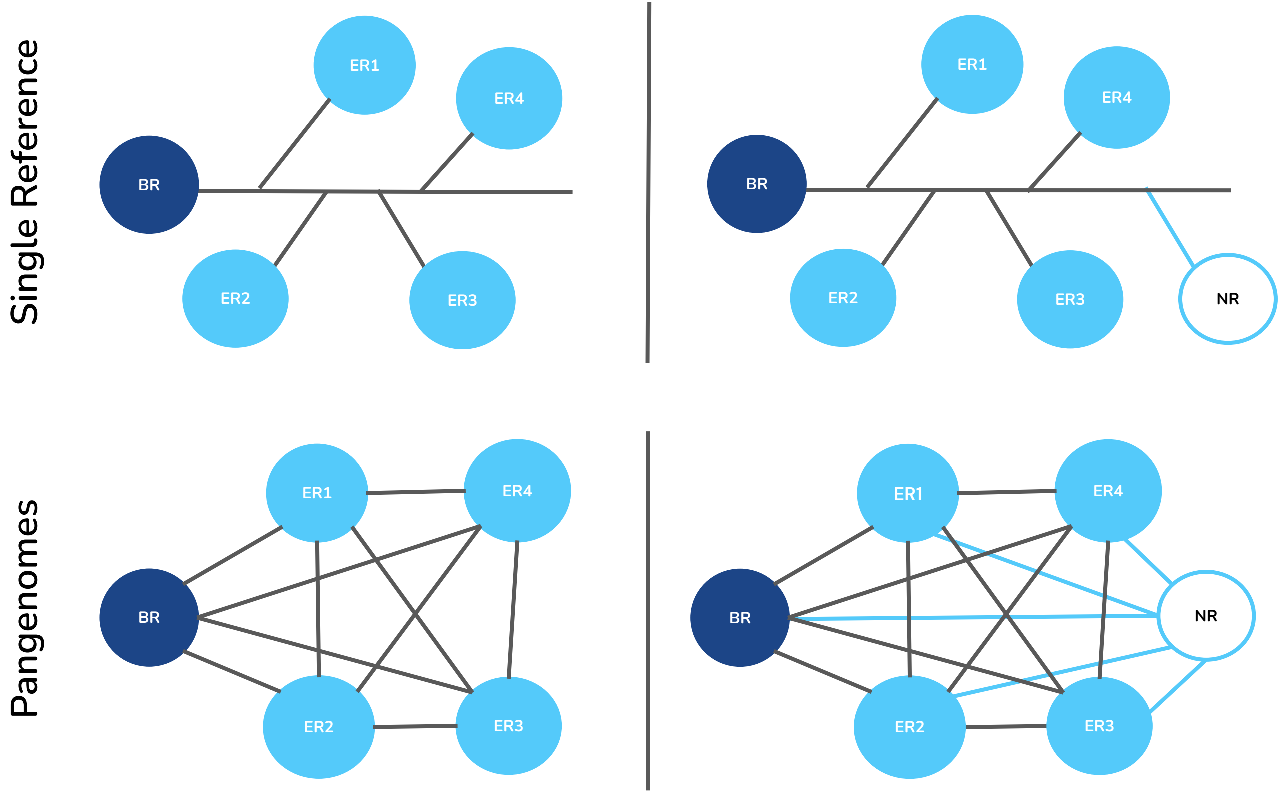
Jessica Imlau Dagostini[1], Scott Beamer[1], Tyler Sorensen[1]
Joseph B. Manzano[2], Andres Marquez[2]

1. {jessica.dagostini, sbeamer, tyler.sorensen}@ucsc.edu
2. {joseph.manzano, andres.marquez}@pnnl.gov

UC SANTA CRUZ

Pacific Northwest NATIONAL LABORATORY

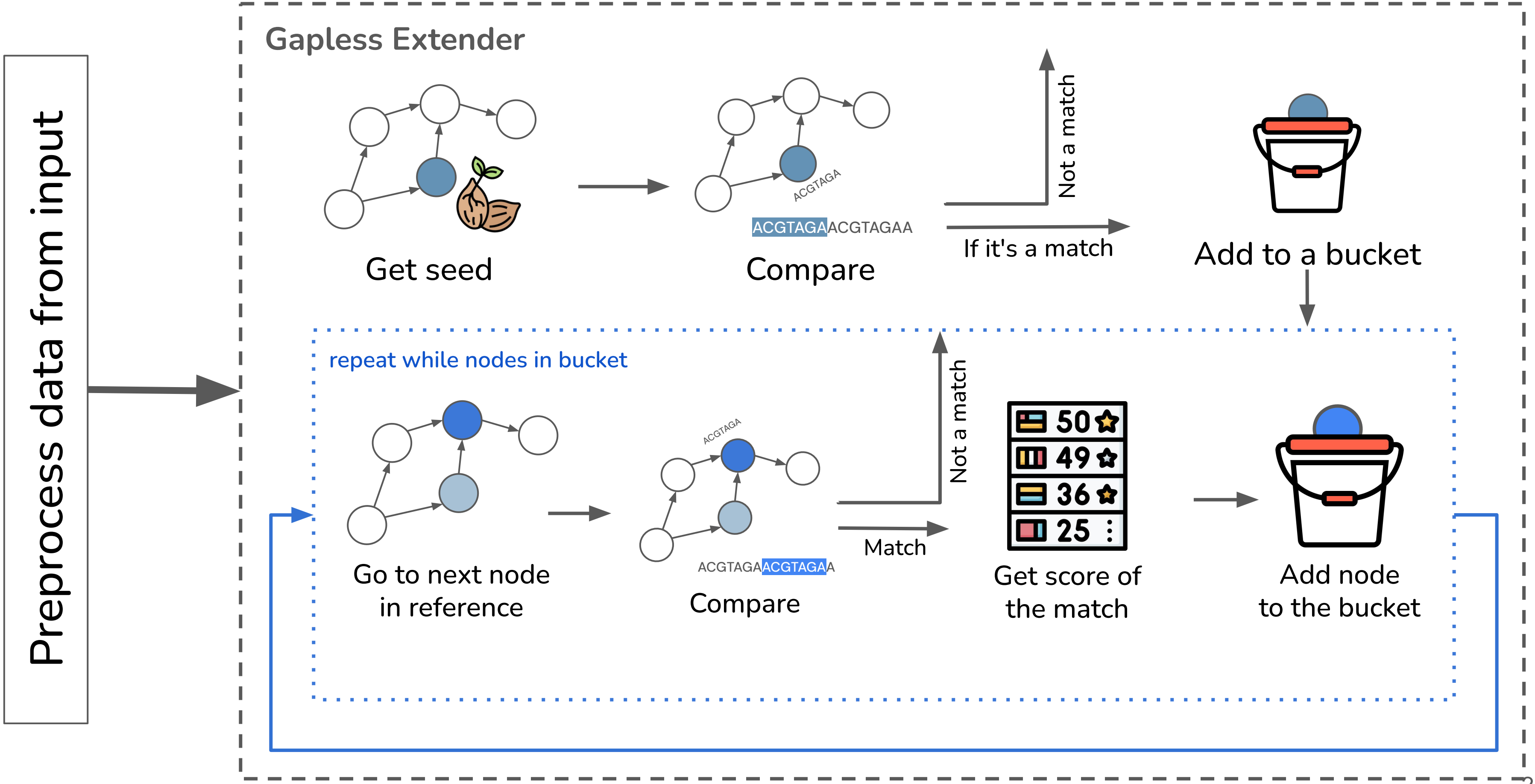
Genome References



Single Genomes can be augmented with variations but do not contain all possible cross-references

Pangenomes allow to relate all genomes or sequences directly to each other

Source: Practical Graphical Pangenomics [https://pangenome.github.io/]



Preprocess data from input

Gapless Extender

Get seed

Compare

If it's a match

Add to a bucket

repeat while nodes in bucket

Go to next node in reference

Compare

Match

Get score of the match

Add node to the bucket