Aligning pangenomes with hierarchical wavefront algorithm

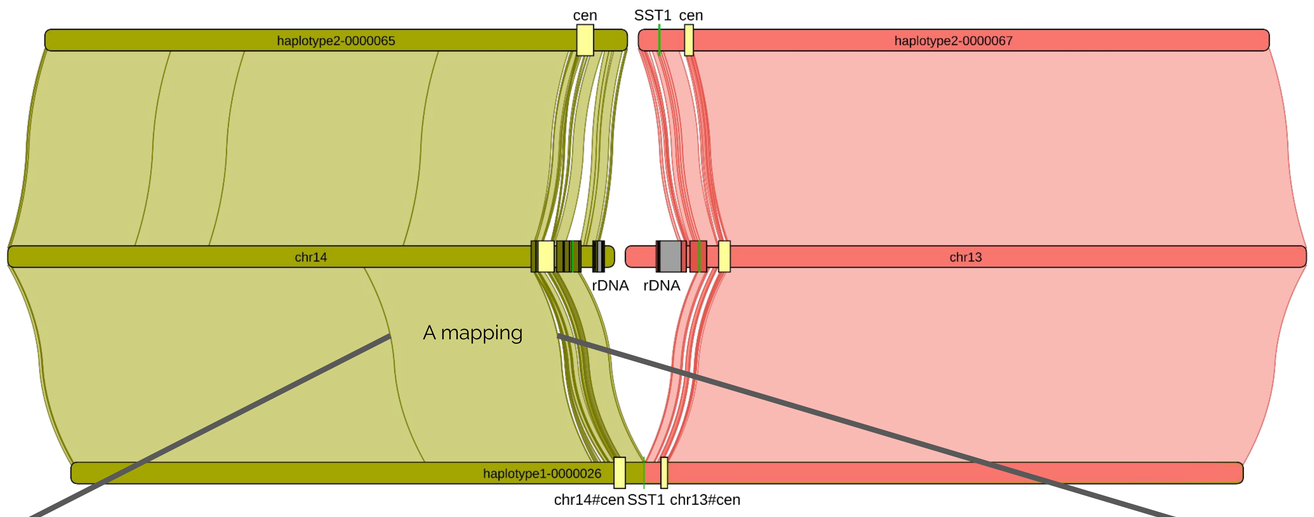
Andrea Guarracino[1], Zhigui Bao[2], Bryce Kille[3], Njagi Mwaniki[4], Santiago Marco-Sola[5], and Erik Garrison[1]

1 Department of Genetics, Genomics and Informatics, University of Tennessee Health Science Center, Memphis, TN 38163, USA; 2 Department of Molecular Biology, Max Planck Institute for Biology Tübingen, Max-Planck-Ring 9, 72076 Tübingen, Germany; 3 Department of Computer Science, Rice University, Houston, TX, United States; 4 Department of Computer Sciences, University of Pisa, Pisa, 56127, Italy; 5 Department of Computer Sciences, Barcelona Supercomputing Center, Barcelona, 08034, Spain.

1) Introduction

- **Sequence alignment** is a core task in bioinformatics.
- Due to the computational complexity of exact alignment, **heuristic methods** are closely tied to the sequence types being considered.
- **Seeding and extension** strategies are driven by seeds that are chained and filtered to find candidate regions for base-level alignment.
- Increasing read lengths and availability of big whole-genome assemblies rendered such approaches **overly sensitive**.
- The dramatic **increase in scale** poses significant challenges for current methods.

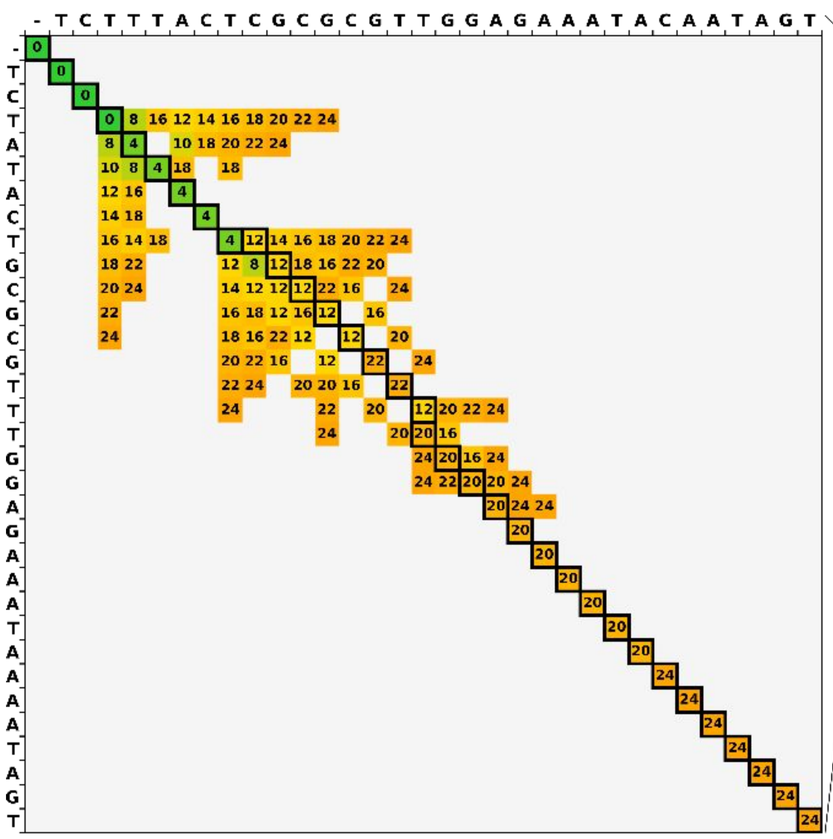
Sequence mappings between human acrocentric chromosomes

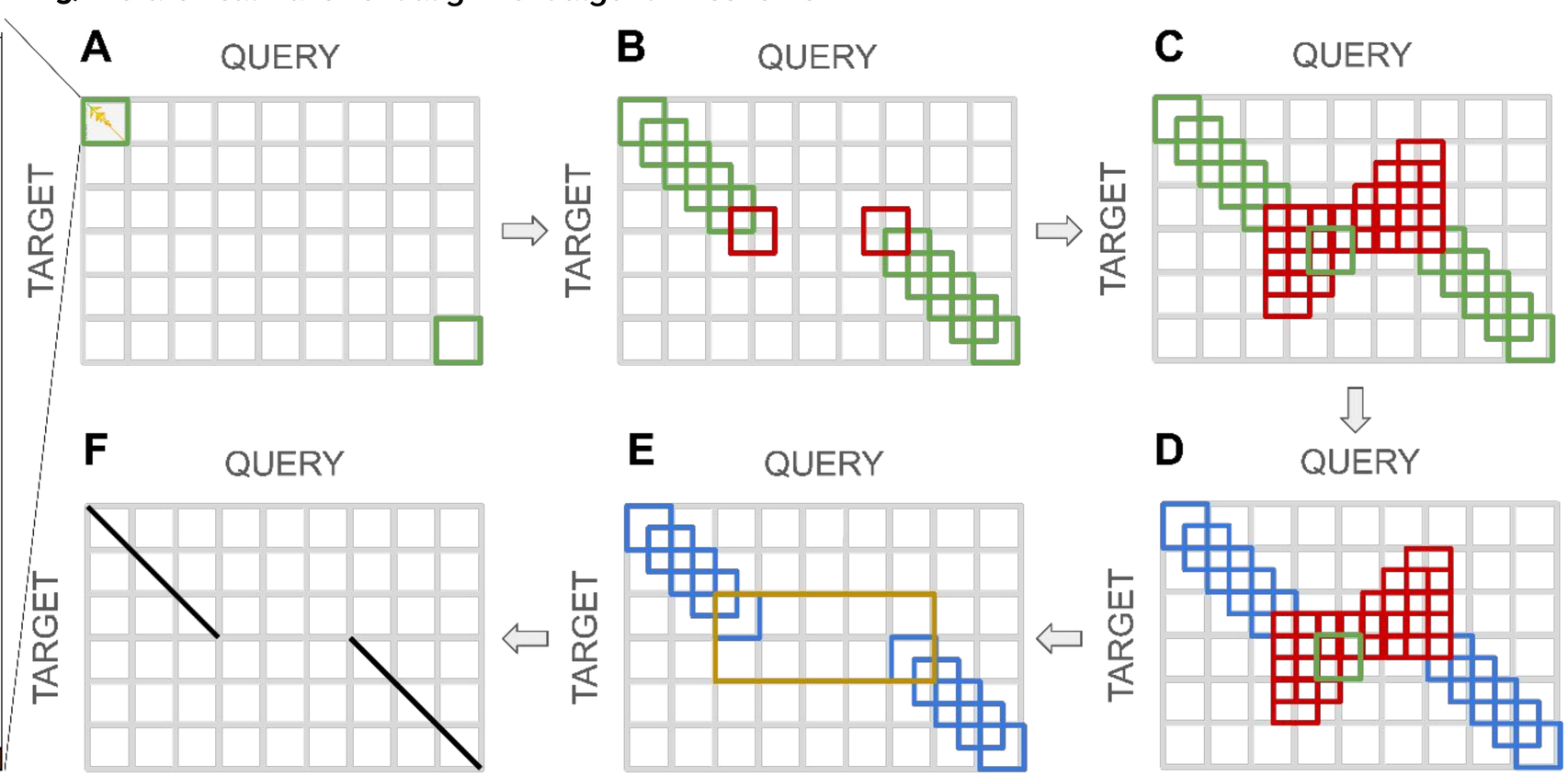


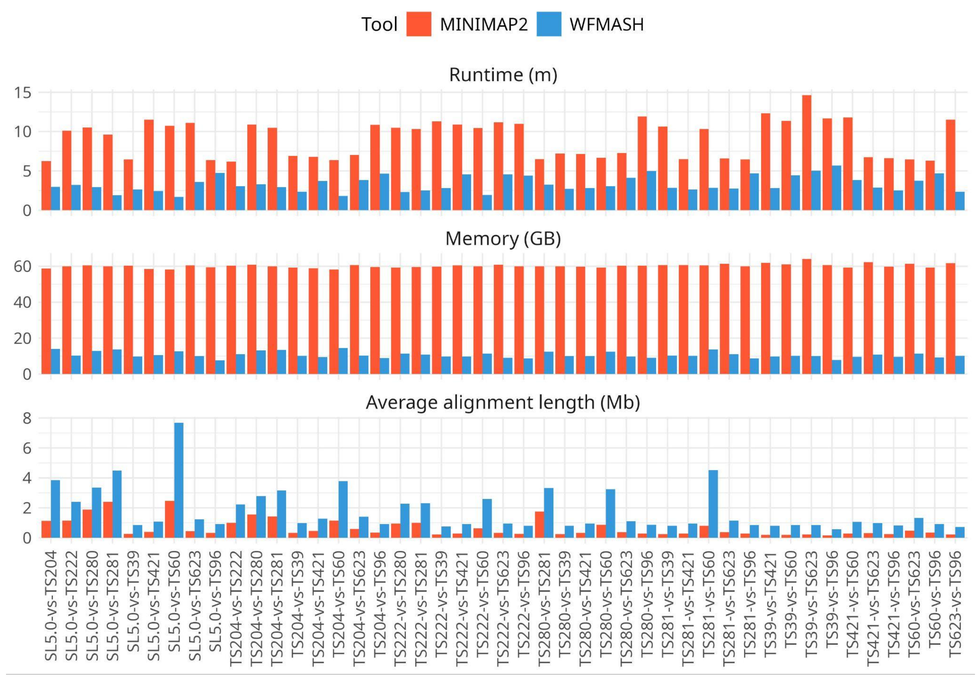
2) Method

- We improve alignment performance for this new scale by **sparsifying** key information that drives mapping and alignment.
- We present **WFMASH**[1], a **sequence aligner** for large pangenomes.
- WFMASH leverages a 2-step strategy:
  a) It first applies a locality-sensitive hashing algorithm, adapted from MashMap[2,3], to determine **syntenic region boundaries** (aka mappings) between long DNA sequences;
  b) Then, a **hierarchical implementation of the wavefront alignment (WFA) algorithm**[4,5] allows computing the base-level alignment of the identified regions.

3) Hierarchical wavefront alignment algorithm scheme

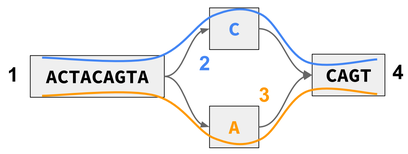




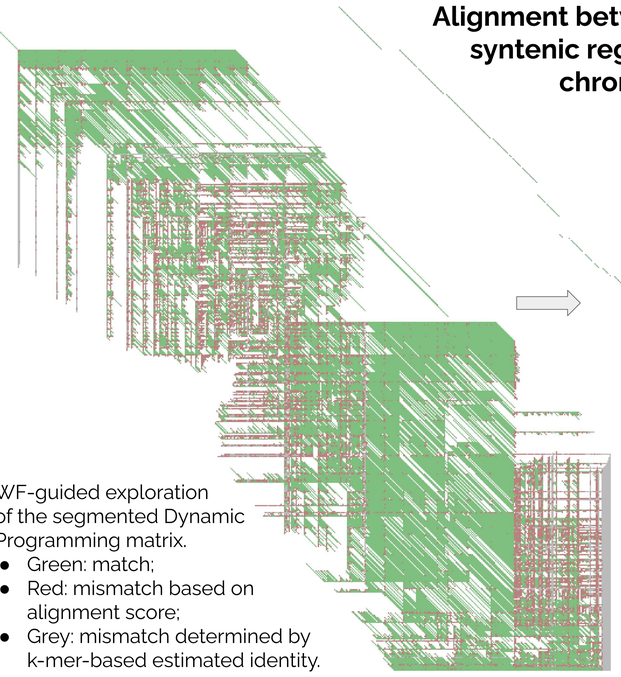


4) Results

- With a **plant pangenome** made with 10 **tomato** haploid assemblies, WFMASH computes the all-vs-all alignment ~2.7X faster than MINIMAP2[6], on average, while requiring ~5.6X less memory and aligning longer sequences.
- With an **inter-species pangenome** made with 16 **primates** assemblies (from *Homo sapiens*, *Gorilla gorilla*, *Pan paniscus*, *Pan troglodytes*, *Pongo abelii*, *Pongo pygmaeus*, *Symphalangus syndactylus*), WFMASH computes the all-vs-all alignment in 70 hours using 128 threads on an AMD EPYC 7742 64-Core processor, with a memory peak of 202 GB.
- We use WFMASH in the Pangenome Graph Building (PGGB) pipeline[7,8] to build unbiased **pangenome variation graphs**.



Alignment between two 70 kbp long syntenic regions of two human chromosomes 13



Segments belonging to the final optimal alignment.

WF-guided exploration of the segmented Dynamic Programming matrix.

- Green: match;
- Red: mismatch based on alignment score;
- Grey: mismatch determined by k-mer-based estimated identity.

1. WFMASH GitHub repository: https://github.com/waveygang/wfmash
2. Chirag Jain et al., 2018. A fast adaptive algorithm for computing whole-genome homology maps. Bioinformatics
3. Bryce Kille et al., 2023. Minmers are a generalization of minimizers that enable unbiased local Jaccard estimation. Bioinformatics
4. Santiago Marco-Sola et al., 2020. Fast gap-affine pairwise alignment using the wavefront algorithm
5. Santiago Marco-Sola, et al., 2023. Optimal gap-affine alignment in O(s) space. Bioinformatics
6. Heng Li, 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics
7. Erik Garrison and Andrea Guarracino, 2023. Unbiased pangenome graphs. Bioinformatics
8. Erik Garrison, Andrea Guarracino et al., 2024. Building pangenome graphs, bioRxiv

Cold Spring Harbor Laboratory

aguarra1@uthsc.edu
egarris5@uthsc.edu

@AndresGuarahino