



HAL
open science

Benchmarking pangenome graph mapping with strobemer-based seeding

Marouane Boumlik, Benjamin Linard, Matthias Zytnicki

► **To cite this version:**

Marouane Boumlik, Benjamin Linard, Matthias Zytnicki. Benchmarking pangenome graph mapping with strobemer-based seeding. JOBIM 2024, Jun 2024, JOBIM 2024, France. hal-04656363

HAL Id: hal-04656363

<https://hal.science/hal-04656363>

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0
International License

Marouane BOUMLIK¹, Benjamin LINARD¹ and Matthias ZYTNIKI¹

¹ Unité de Mathématiques et Informatique Appliquées, INRAE, Castanet-Tolosan, France
Corresponding Author: benjamin.linard@inrae.fr

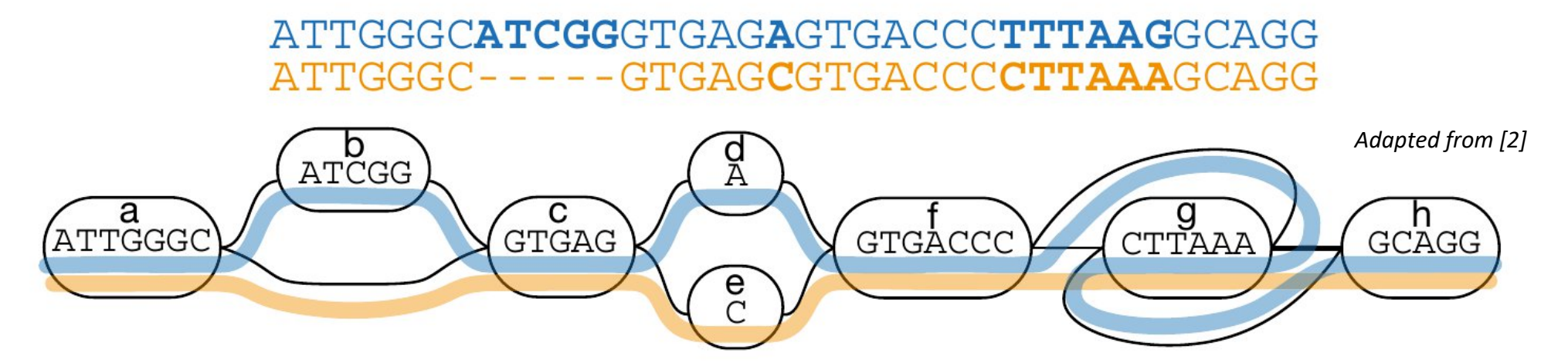
CONTEXT

A pangenome represents the total genetic diversity of a species or a species complex. A recent model, the variation graph, aims to integrate full length genomes [1]. Contrary to typical mapping approaches which rely on a single reference genome and are intrinsically biased, mapping query sequences to the full genetic variability of a species leads to more accurate mappings and improved genotype/phenotype analyses.

Mapping to a graph is similar to classic genome mapping and the first step involves a seed & extend approach, to target regions in the graph involving similar sequences before the steps of path selection and alignment. When divergent query sequences are analyzed, this step can be problematic: not enough k-mer seeds are detected, and the following steps do not happen.

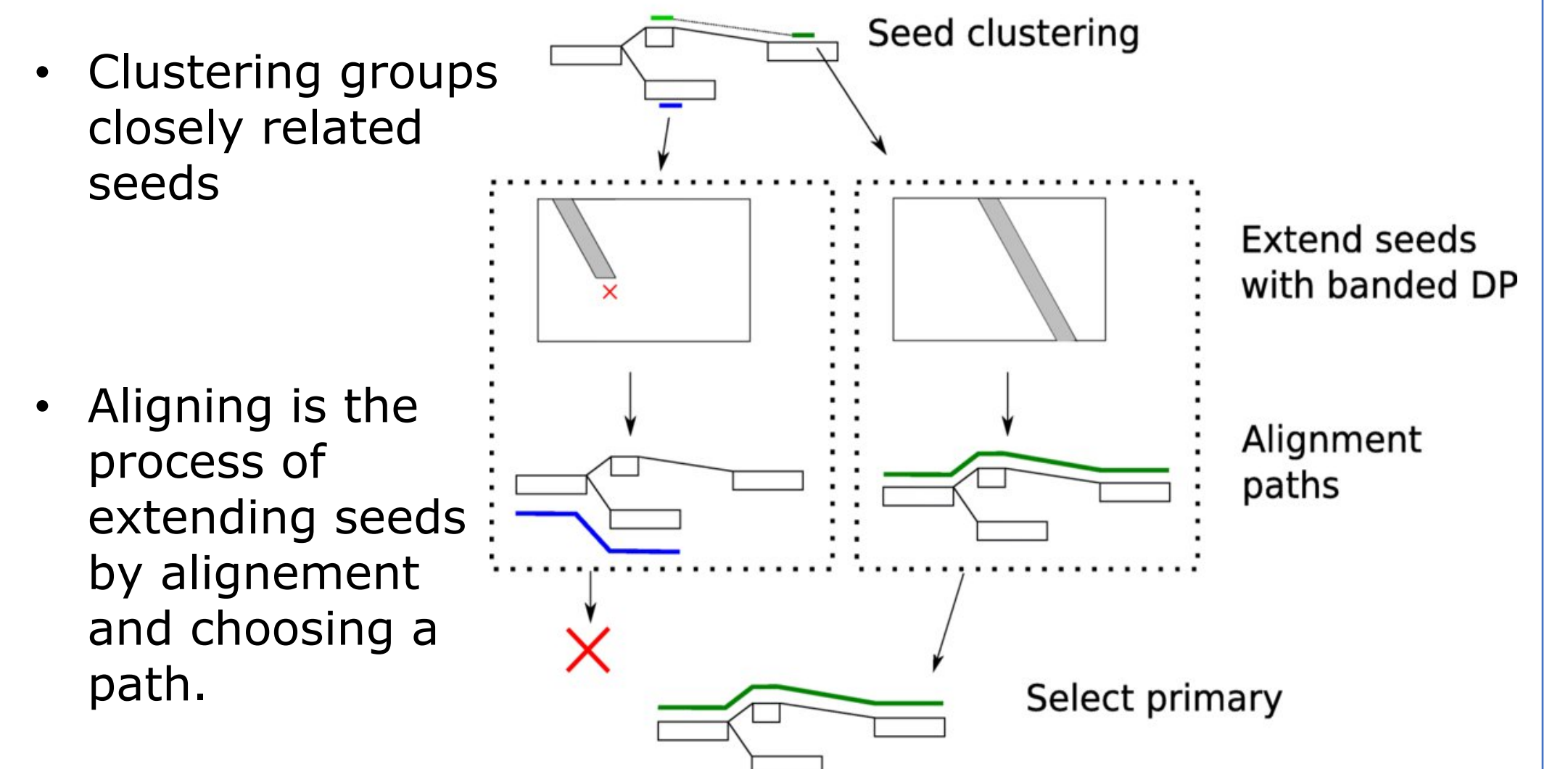
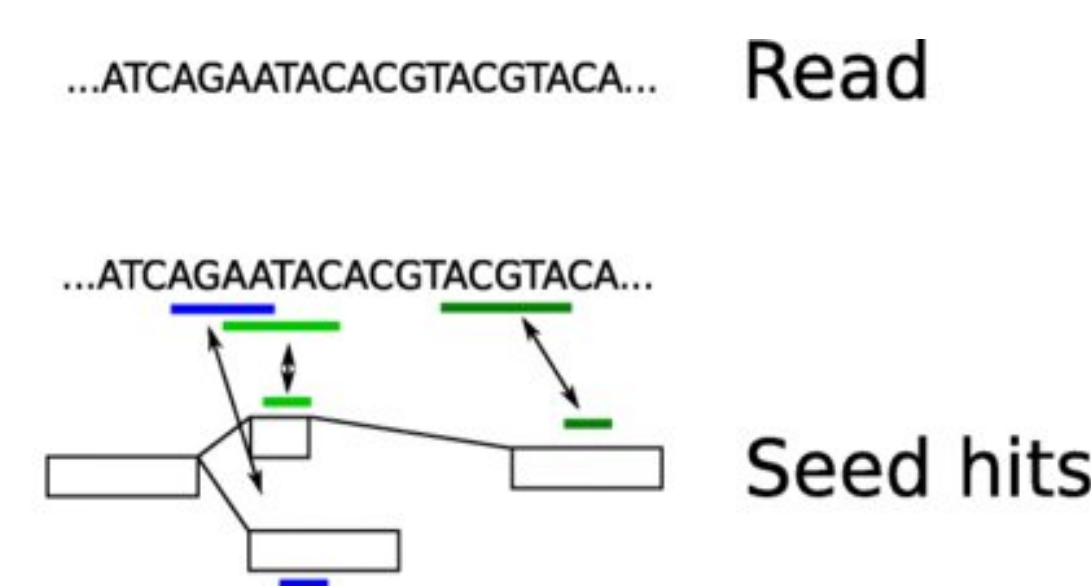
This work aims to test the potential of strobemers, an alternative to classic k-mer+minimizer seeding that allows more sequence divergence [4], in the context of mapping to pangenome graphs.

Input: high-quality assemblies of full genomes, aligned.
Output: a graph modeling the full variation, e.g. a non-biased reference.



Seed & extend process in GraphAligner [3]

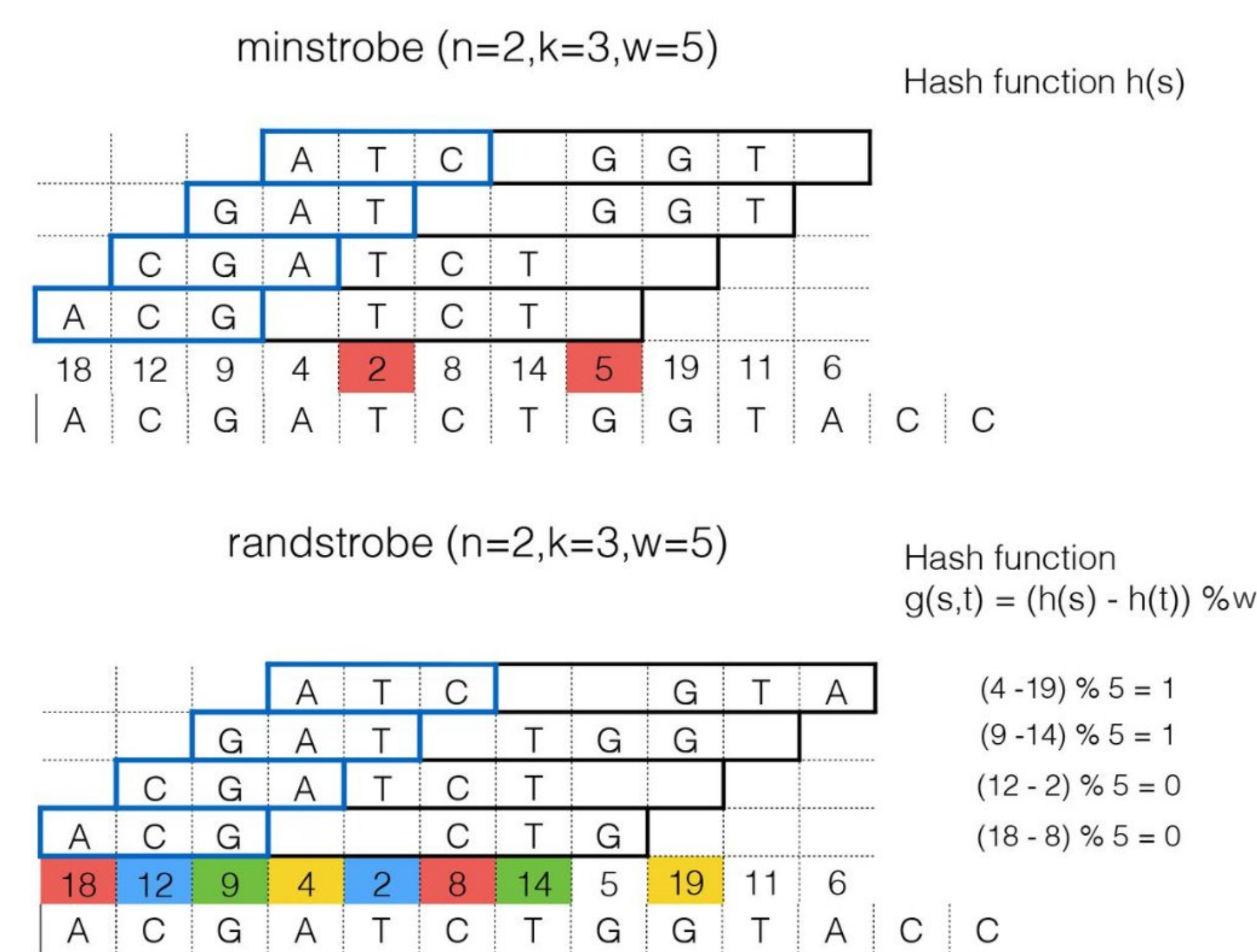
- Seeds are short sub-sequences that show a high degree of similarity and serve as anchors for following steps in the mapping process



STROBEMERS

- Strobemers [4] is a combination of multiple discontinuous subsequences, known as "strokes".

Minstroke vs Randstroke with parameters n, k and w (number of strokes, stroke length and window length)

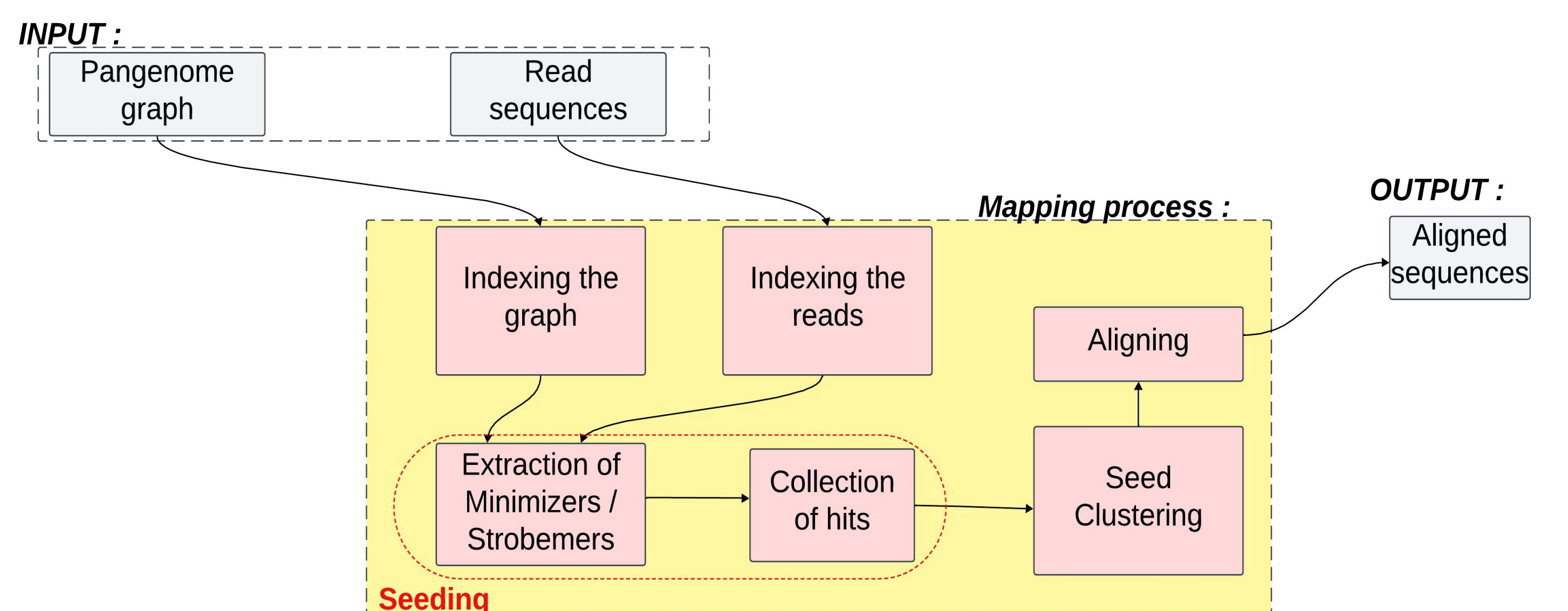


- Strobemers have two sub-schemes: the minstrokes and the randstrokes [4].

- A C++ library prototype was available to exploit them in another software as an alternative seeding technique.

METHOD

- Strobemers are injected as an alternative seeding scheme into GraphAligner.
- Other steps remain untouched.
- A basic strobemer filter "1 out of 5" is applied (for scalability).
- Long reads are simulated from the graph and mapped back to the graph.
- We measure mapping quality differences between classic k-mer + minimizers and strobemers approaches.

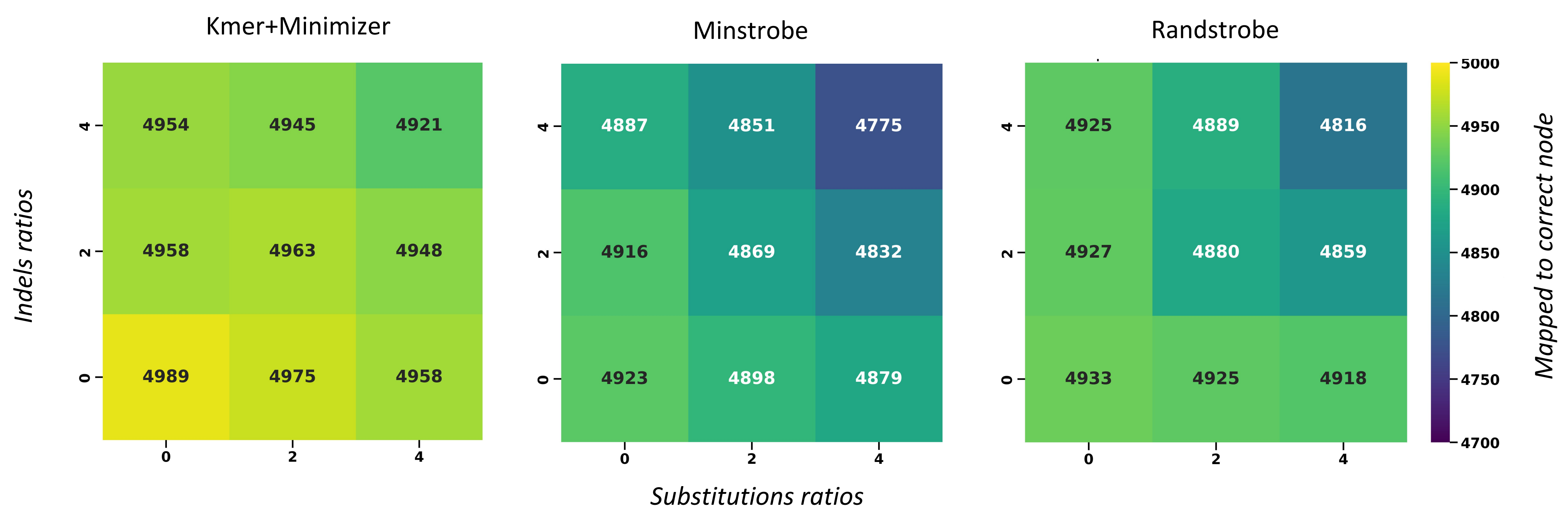


PRELIMINARY RESULTS

- Software engineering**:
- Improvement the C++ strobemer library: bug-fixes, cmake compilation, functionality ... (accepted in author's GIT repository).
 - Development of benchmarking prototype "GraphAligner-strobe": A hybrid software allowing to choose between k-mer + minimizers and strobemer seeding.
 - Allows to set different parameters: rand|minstroke approaches, and strobemer parameters k, n, w_{min}, w_{max} .

Benchmark:

Simulated reads (vg sim)
Variable SNP/indel ratio
5000 reads of 5kbp



DISCUSSION

- Randstroke performs better than Minstroke, but in our preliminary tests, both perform worse than kmer+minimizer.
- We used a basic strobemer filter (filter 1 out of 5), what about a strobemer+minimizers?
- Critical choice: we index k-mers or strobemers from node's sequences, which excludes nodes with length < k (GraphAligner's approach). What about path indexing?
- Strobemer parameters chosen for this early benchmark may be unadapted, and strobemer seeding may be improved with different parameters.
- The benchmark should be extended to more parameter testing before further conclusions.