Benchmarking read mapping on pangenomic variation graphs

Hajar Bouhamout, Benjamin Linard, Matthias Zytnicki

Benchmarking read mapping on pangenomic variation graphs

Hajar BOUAMOUT[1], Benjamin LINARD[1] and Matthias ZYTNICKI[1]

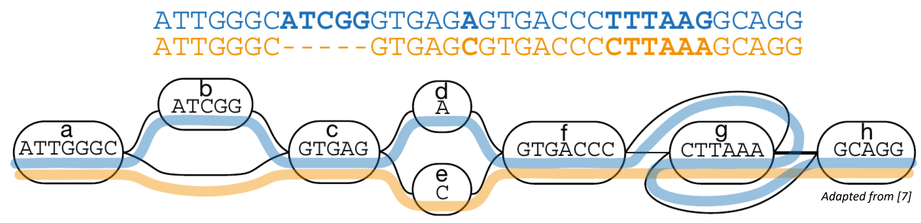
1 Unité de Mathématiques et Informatique Appliquées, INRAE,Castanet-Tolosan, France
Corresponding Author: benjamin.linard @inrae.fr

CONTEXT

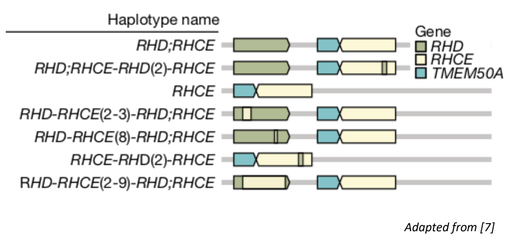
A pangenome represents the total genetic diversity of a species or a species complex. One can describe pangenomes in terms of gene presence / absence variations (PAVs) but a more recent alternative aims to integrate full length genomes in a sequence graph [1].In particular, pairwise alignments of a genome set can be used to build a "Variation Graph" (VG) in which nodes represent words of genome fragments and edges represent the contiguity of the words in one to several genomes (e.g., edges are associated to genome subsets). Each input genome consequently corresponds to a particular path in the graph. It has been showed that VGs can improve variant calling and genotyping processes [2]. Indeed, variant calling is often based on the mapping of linear query sequences to linear reference genomes, thus biasing the prediction to regions present in the reference and limiting the identification of structural variations that are specific to other genomes. In particular, biases are reduced when large structural variations (>50bp) are targeted.

Input : high-quality assemblies or full genomes, aligned.

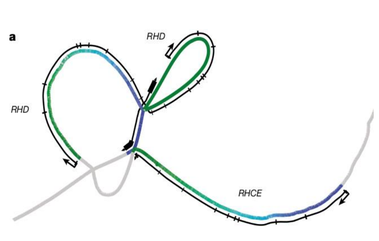
Output : a graph modelling the full diversity of the genomes, e.g. a non-biased reference



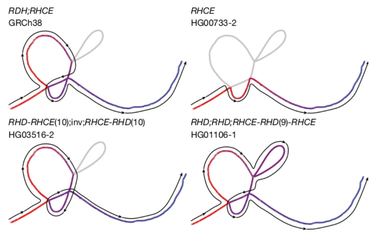
Classic linear representation.
Haplotypes of the RHD/RHCE genes.



Corresponding variation subgraph of the locus.
Annotated with RHD/RHCE genes.



Alternative paths correspond to different haplotypes.



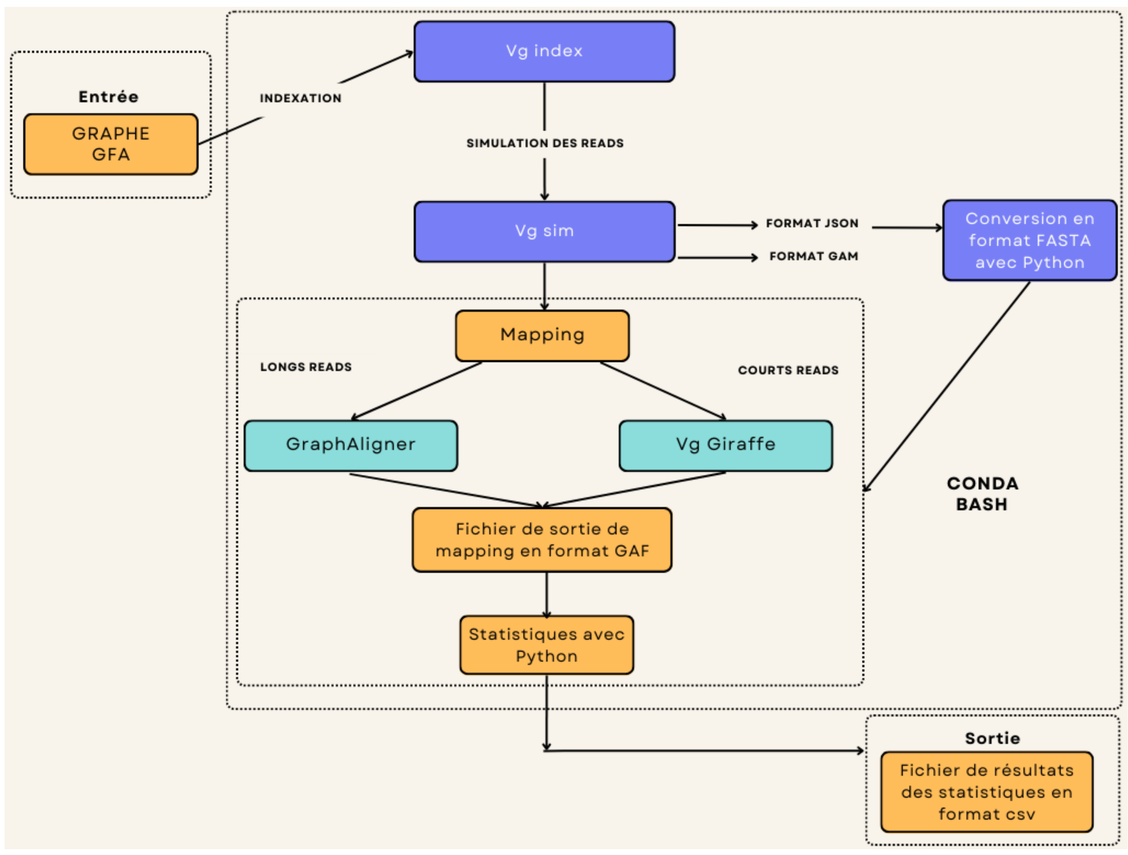
MAPPING TOOLS

	GrapAligner [4]	VG giraffe [5]	VG map [5]	Minichain [6]
Read length	Long	Short	Long+ Short	Long
Compatible graph formats	Vg, gfa	vg	vg	gfa
Output	Gam, json, gaf	Gam, json, gaf	Gam	Paf, gaf
Speed	Slow	Fast	Slow	Fast
OS	Bioconda	Unix, mac	Unix, mac	Unix, mac, windows

Other untested tools :
- V-map (private, not released publicly)
- HiSAT2 (not compatible with variation graph formats)

METHODS

- Variation graph of the human Y chromosome
- Reads simulated from the graph with VG sim
- Varying parameters
  1. Read length
  2. Substitution ratio
  3. Indels ratio
- 1000 simulated reads re-mapped to the graph
- Mapping output in GAF format
- Descriptive statistics (matched path, CIGAR-based …)

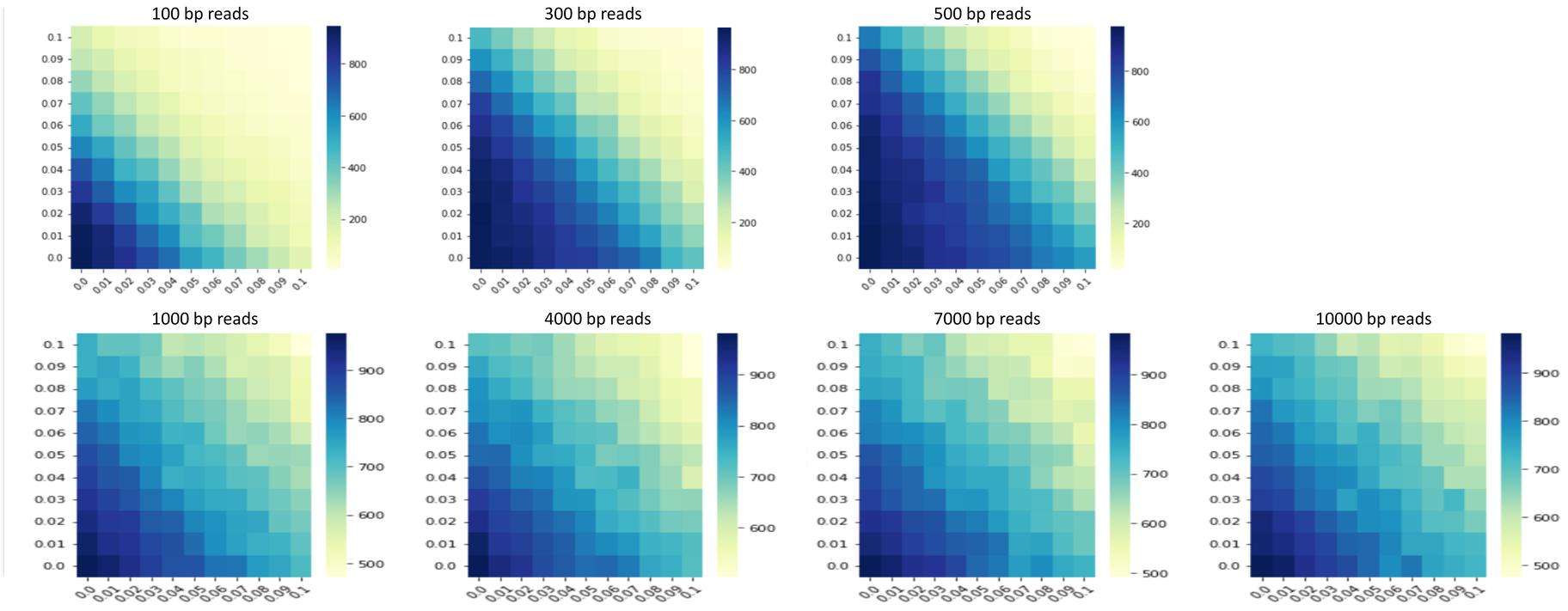


PRELIMINARY RESULTS

reads mapped to the correct graph path (e.g. nodes and offsets corresponding to the simulated read)

VG giraffe
(short reads)

GraphAligner
(long reads)



DISCUSSION

- Current mapping tools appear to perform poorly when more than 3-4% indels and/or substitution (>10% reads do not map to the correct graph path)
- For longer reads, indels have more impact than substitutions
- Mismatches remains to be characterized in more details (type of mismatches, local graph topology…)
- Results needs to be compared to performance of standard mapping on linear reference (ex: BWA)

[1] Paten B, Novak AM, Eizenga JM, Garrison E. Genome graphs and the evolution of genome inference. Genome Res. 2017 May;27(5):665-676. [2] Garrison E, Sirén J, Novak AM, Hickey G, Eizenga JM, Dawson ET, Jones W, Garg S, Markello C, Lin MF, Paten B, Durbin R. Variation graph toolkit improves read mapping by representing genetic variation in the reference. Nat Biotechnol. 2018 Oct;36(9):875-879. [3] Shuo Wang, Yong-Qing Qian, Ru-Peng Zhao, Ling-Ling Chen, Jia-Ming Song, Graph-based pan-genomes: increased opportunities in plant genomics, Journal of Experimental Botany, Volume 74, Issue 1, 1 January 2023, Pages 24–39 [4] Rautiainen M, Marschall T. GraphAligner: rapid and versatile sequence-to-graph alignment. Genome Biol. 2020 Sep 24;21(1):253. [5] Hickey G, Heller D, Monlong J, Sibbesen JA, Sirén J, Eizenga J, Dawson ET, Garrison E, Novak AM, Paten B. Genotyping structural variants in pangenome graphs using the vg toolkit. Genome Biol. 2020 Feb 12;21(1):35. [6] Ghanshyam Chandra, Chirag Jain. Sequence to graph alignment using gap-sensitive co-linear chaining. bioRxiv 2022.08.29.505691; [7] Liao, WW., Asri, M., Ebler, J. et al. A draft human pangenome reference. Nature 617, 312–324 (2023). https://doi.org/10.1038/s41586-023-05896-x