



HAL
open science

Benchmarking read mapping on pangenomic variation graphs

Hajar Bouhamout, Benjamin Linard, Matthias Zytnicki

► **To cite this version:**

Hajar Bouhamout, Benjamin Linard, Matthias Zytnicki. Benchmarking read mapping on pangenomic variation graphs. jobim 2023, Jun 2023, Nice, France. hal-04150552

HAL Id: hal-04150552

<https://hal.science/hal-04150552>

Submitted on 4 Jul 2023

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 4.0 International License

Hajar BOUAMOUT¹, Benjamin LINARD¹ and Matthias ZYTNICKI¹

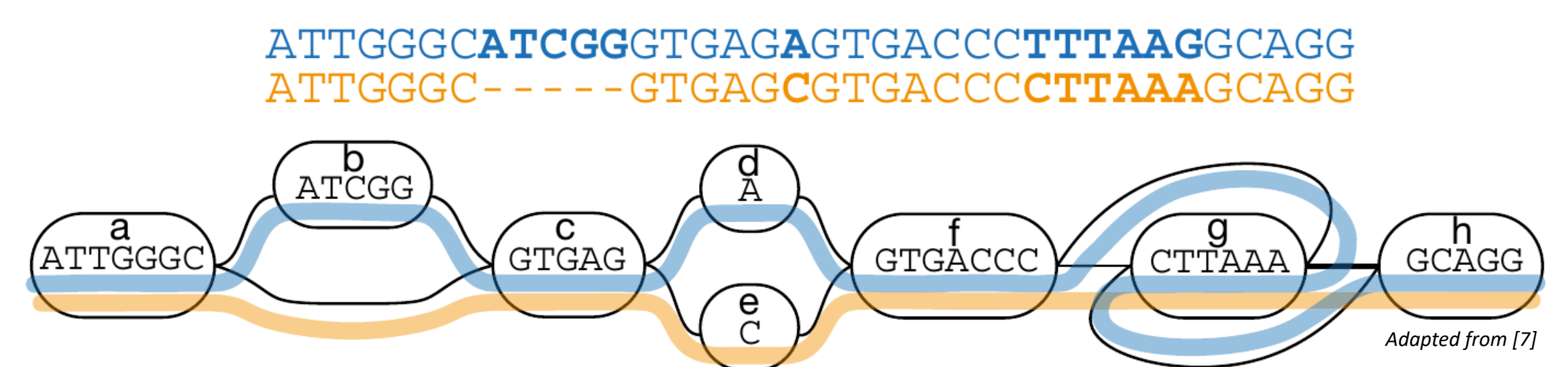
¹ Unité de Mathématiques et Informatique Appliquées, INRAE, Castanet-Tolosan, France
Corresponding Author: benjamin.linard@inrae.fr

CONTEXT

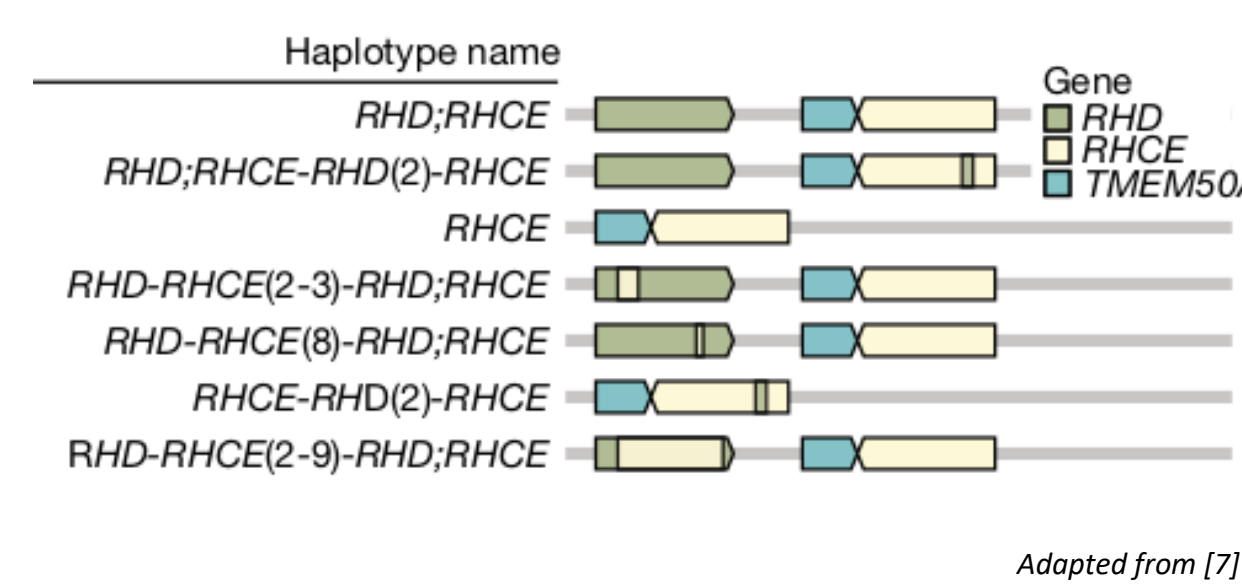
A pangenome represents the total genetic diversity of a species or a species complex. One can describe pangenomes in terms of gene presence / absence variations (PAVs) but a more recent alternative aims to integrate full length genomes in a sequence graph [1]. In particular, pairwise alignments of a 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.

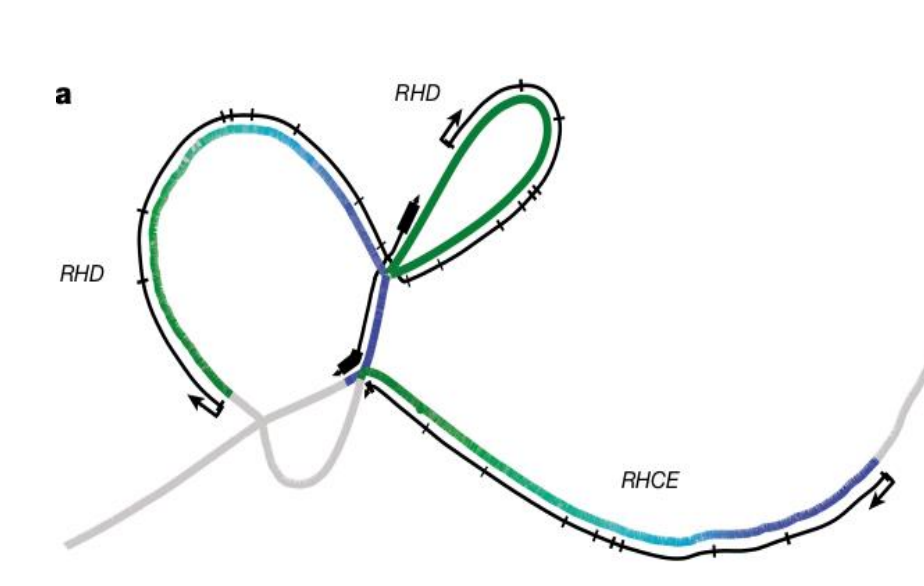
Output : a graph modelling the full diversity of the genomes, e.g. a non-biased reference



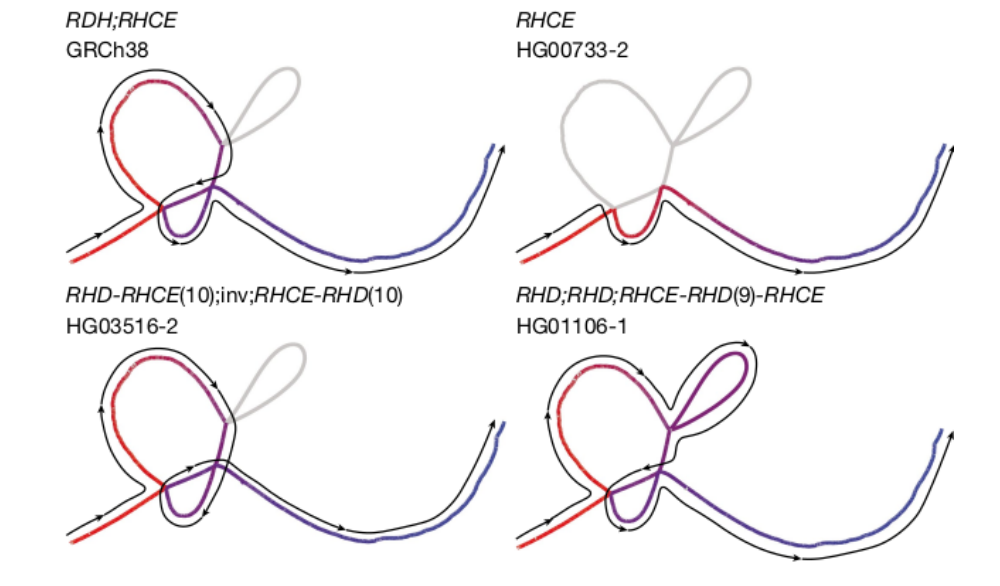
Classic linear representation.
Haplotypes of the RHD/RHCE genes.



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



Alternative paths correspond to different haplotypes.



MAPPING TOOLS

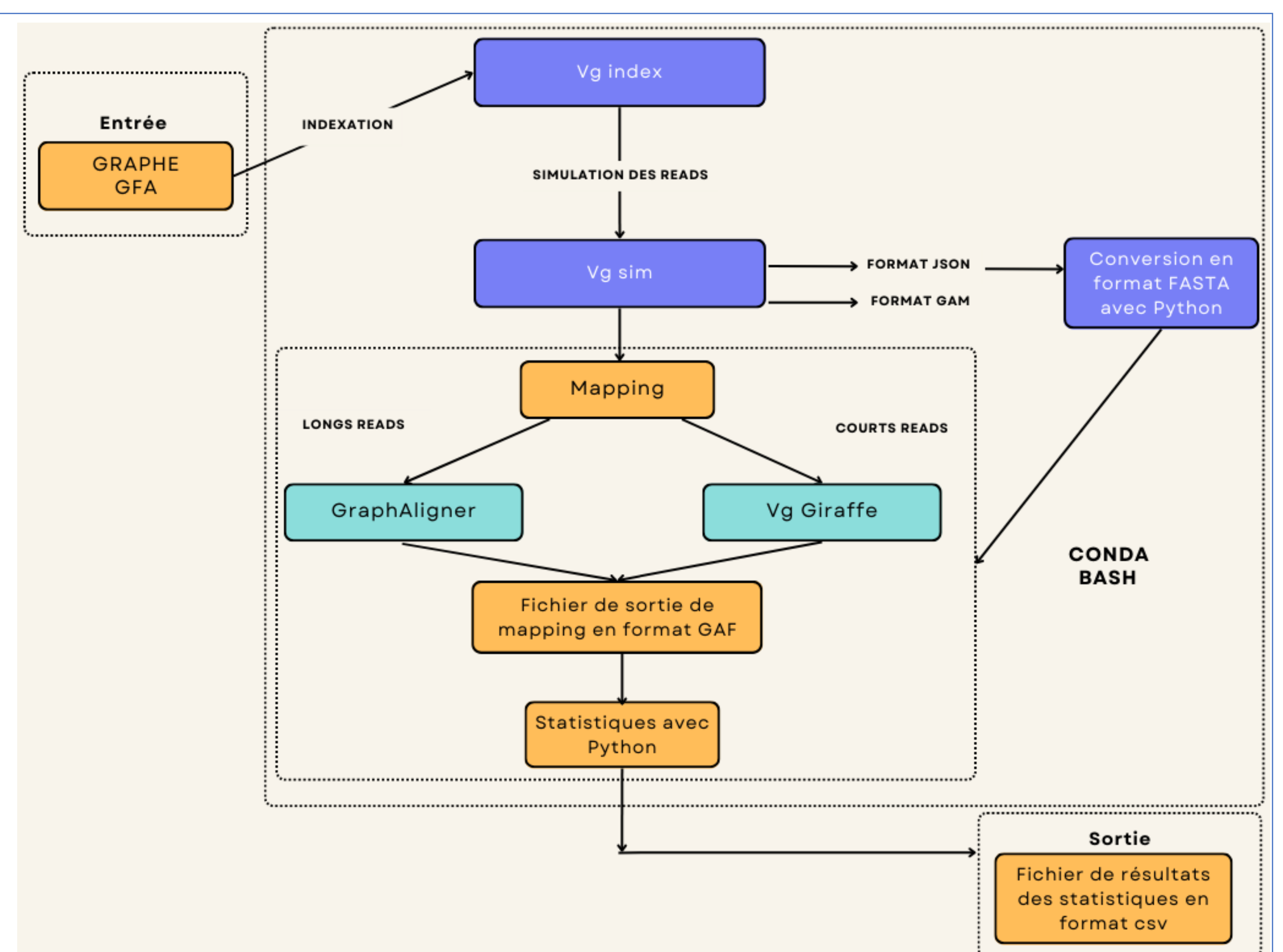
	GrpAligner [4]	VG giraffe [5]	VG map [5]	Minichain [6]
Read length	Long	Short	Long+ Short	Long
Compatible graph formats	Vg, gfa	vg	vg	gfa
Output	Gam, json, gaf	Gam, json, gaf	Gam	Paf, gaf
Speed	Slow	Fast	Slow	Fast
OS	Bioconda	Unix, mac	Unix, mac	Unix, mac, windows

Other untested tools :

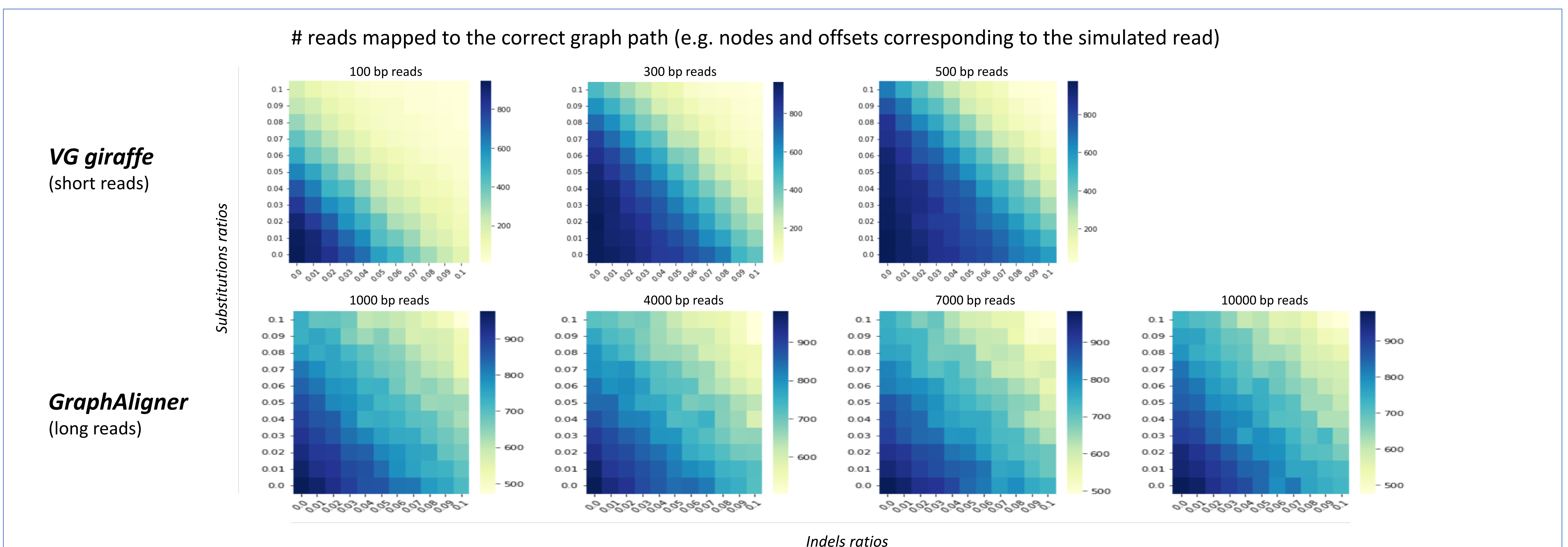
- V-map (private, not released publicly)
- HiSAT2 (not compatible with variation graph formats)

METHODS

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



PRELIMINARY RESULTS



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)