



HAL
open science

HairSplitter: separating similar strains in metagenome assembly

Roland Faure, Jean-François Flot, Dominique Lavenier

► **To cite this version:**

Roland Faure, Jean-François Flot, Dominique Lavenier. HairSplitter: separating similar strains in metagenome assembly. ISMB/ECCB 2023 - 31st Annual Intelligent Systems For Molecular Biology and the 22nd Annual European Conference on Computational Biology, Jul 2023, Lyon, France. pp.1-1, 2023. hal-04272480

HAL Id: hal-04272480

<https://hal.science/hal-04272480>

Submitted on 6 Nov 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 4.0 International License

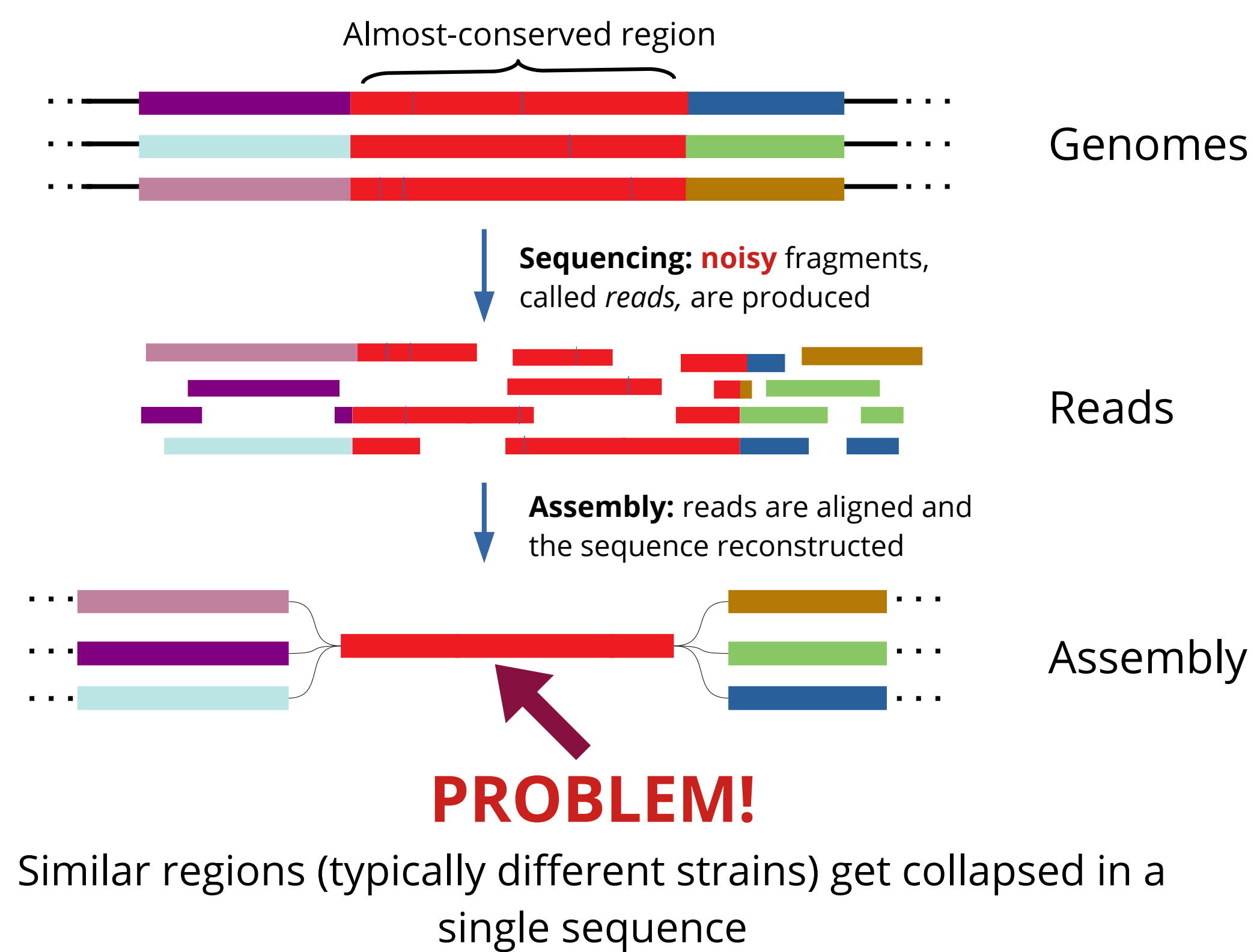
HairSplitter: separating similar strains in metagenome assembly

Roland Faure^{1,2}, Jean-François Flot¹, Dominique Lavenier²

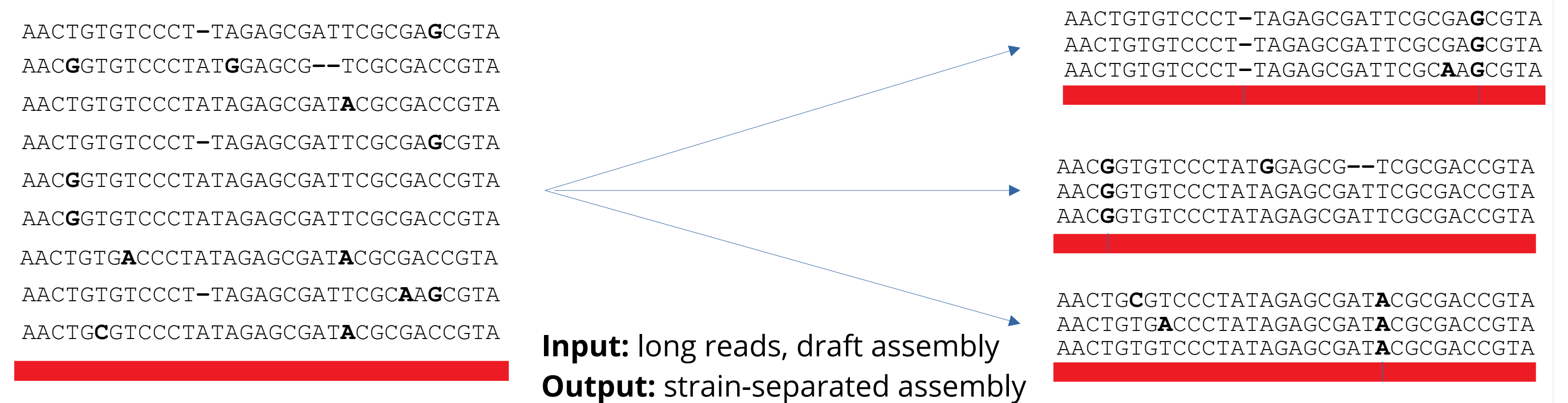
1. Service Evolution Biologique et Ecologie, ULB, Brussels, Belgium 2. Univ. Rennes, Inria RBA, CNRS UMR 6074, Rennes, France

github.com/RolandFaure/Hairsplitter

Problem: assembling similar sequences



State of the art



Difficulties: Unknown number of strains (potentially high), uneven coverage

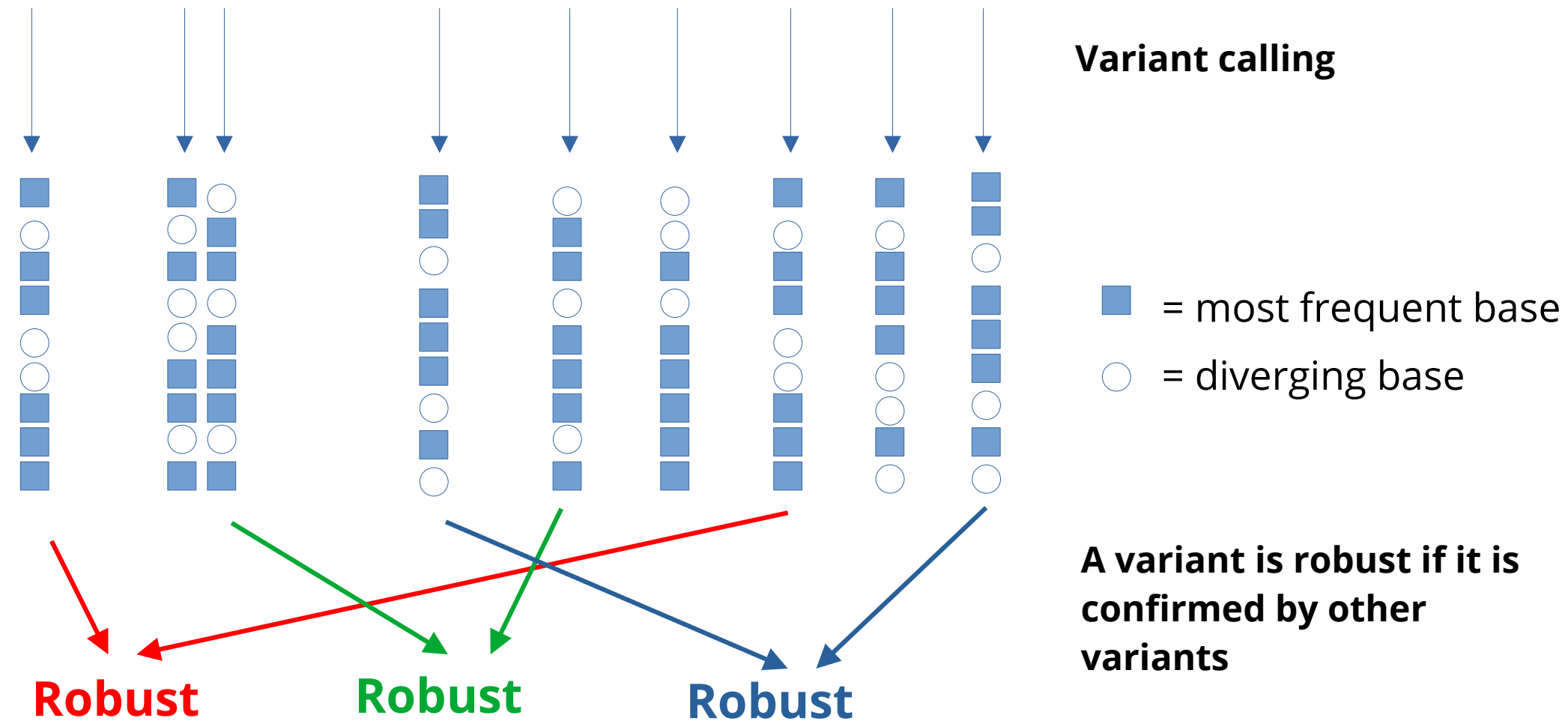
Existing software: Strainberry [1], stRainy (under development) [2], hifiasm [3]

But: To be improved for noisy reads and high number of strains

Algorithm

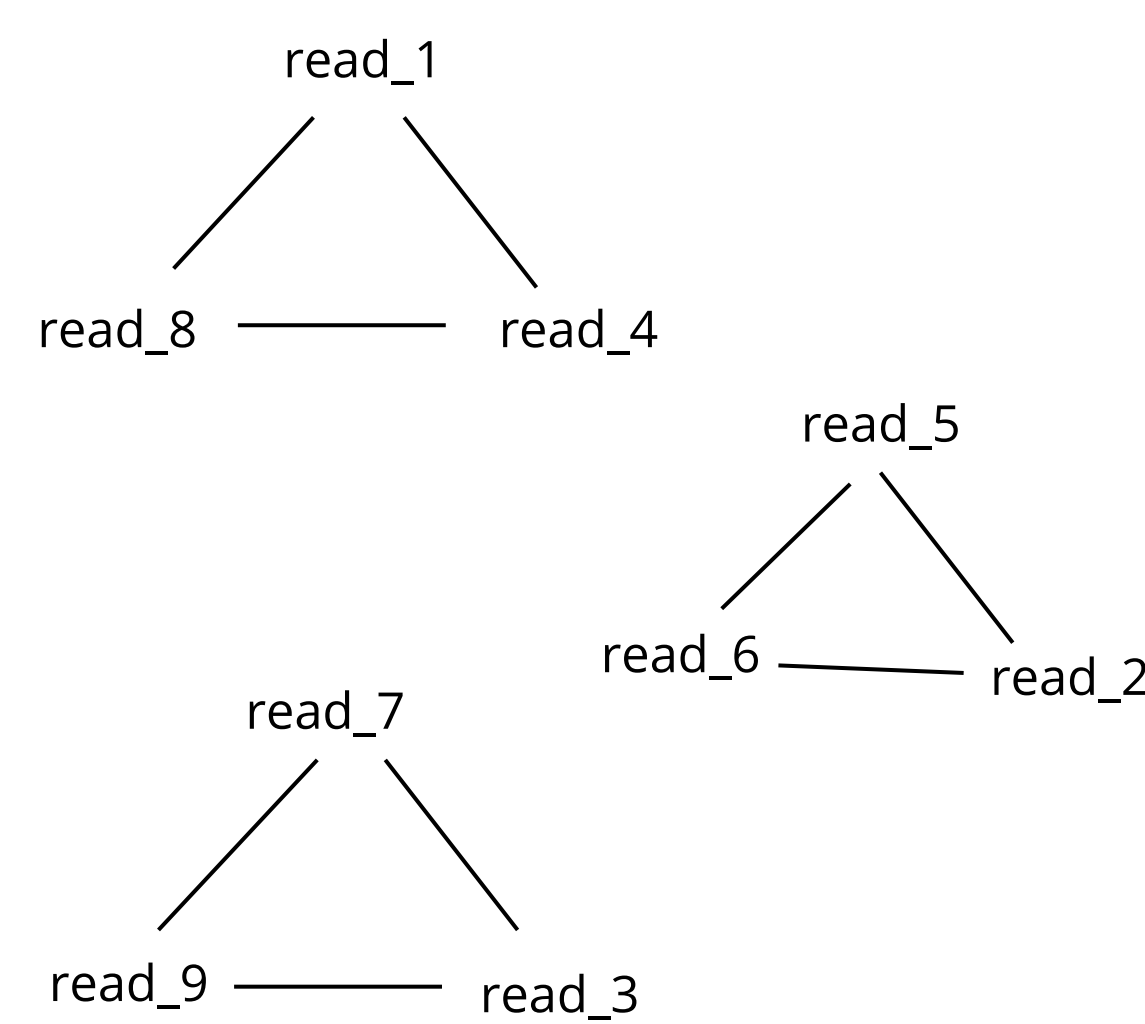
1. Robust variant calling

draft asm	A	A	C	T	G	T	G	T	G
read_1	T	C	-	T	G	T	G	T	G
read_2	G	A	A	T	C	T	A	A	G
read_3	T	C	A	A	C	C	G	T	-
read_4	T	A	-	T	G	T	G	T	G
read_5	G	A	A	T	C	C	A	T	G
read_6	G	C	A	T	C	C	A	A	G
read_7	T	C	A	A	C	C	G	A	-
read_8	T	A	-	T	G	C	G	T	G
read_9	T	C	A	A	C	C	G	A	-



2. Read clustering

Each read is linked to the k nearest reads ($k=2$ here), computed using robust variants

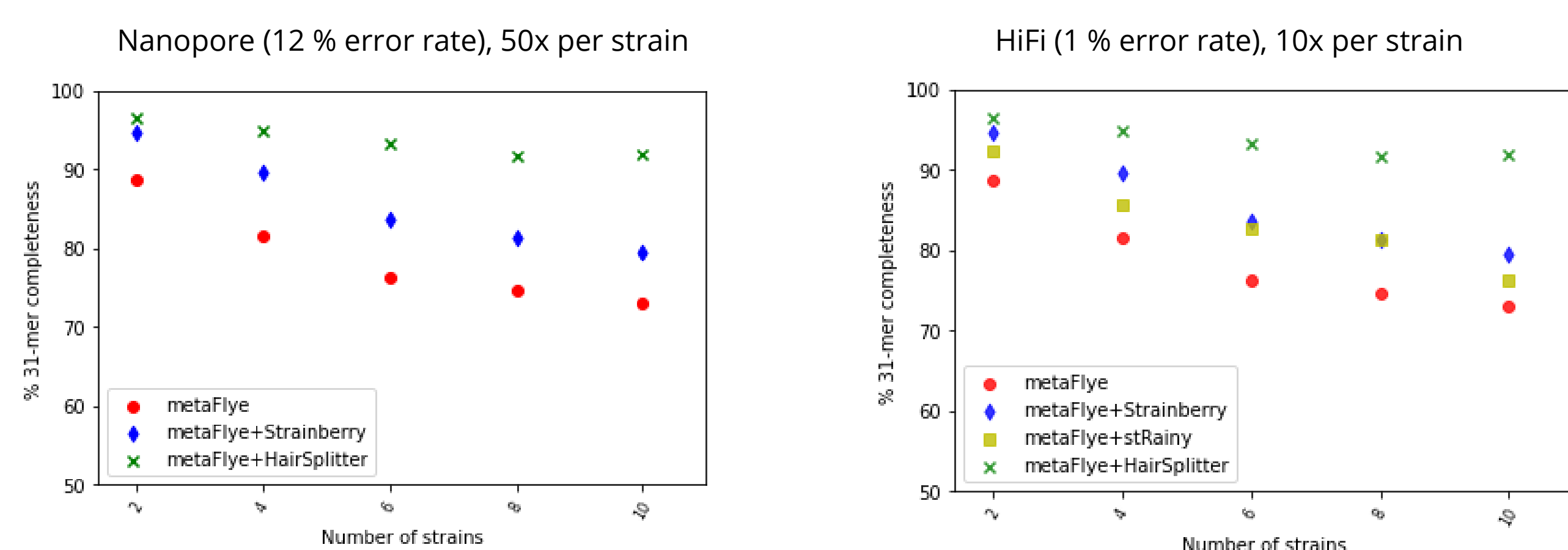


3. Reassembly

read_1	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_4	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_8	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_2	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_5	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_6	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_3	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_7	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A
read_9	A	A	C	T	G	T	C	C	C	T	A	T	A	G	A	G	C	G	A	T	C	G	G	A	A

Results

> Mix of 2 to 10 *E. coli* strains, simulated sequencing from RefSeq genomes



> Mix of 5 *E. coli* strains, Zymbiomics gut microbiome standard

	metaFlye	metaFlye+Strainberry	metaFlye+HairSplitter
Nanopore Q9	0.586	0.749	0.957
Nanopore Q20	0.7524	0.9527	0.961
PacBio HiFi	0.9589	0.9793	0.9895

Table: 31-mer completeness of assemblies w.r.t. the reference

Conclusion & Perspectives

- > *Hairsplitter* reconstructs a **strain-separated assembly** from a draft assembly, improving on the state-of-the-art
- > *Hairsplitter* uses **any type of long reads** (incl. high-error reads) on any type of assembly (incl. polyploid genome assemblies)
- > Perspective: improve the understanding of true microbiomes

References

- [1] Vicedomini, R., Quince, C., Darling, A.E. et al. Strainberry: automated strain separation in low-complexity metagenomes using long reads. *Nat Commun* 12, 4485 (2021). <https://doi.org/10.1038/s41467-021-24515-9>
- [2] Ekaterina Kazantseva, Ataberk Donmez, Mihai Pop, Mikhail Kolmogorov. stRainy: assembly-based metagenomic strain phasing using long reads. *BioRxiv* 2023.01.31.526521
- [3] Cheng, H., Concepcion, G.T., Feng, X. et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* 18, 170–175 (2021). doi.org/10.1038/s41592-020-01056-5
- [4] Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* 2017 May;27(5):737-746. doi: 10.1101/gr.214270.116.
- [5] Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics.* 2018 Sep 15;34(18):3094-3100. doi: 10.1093/bioinformatics/bty191