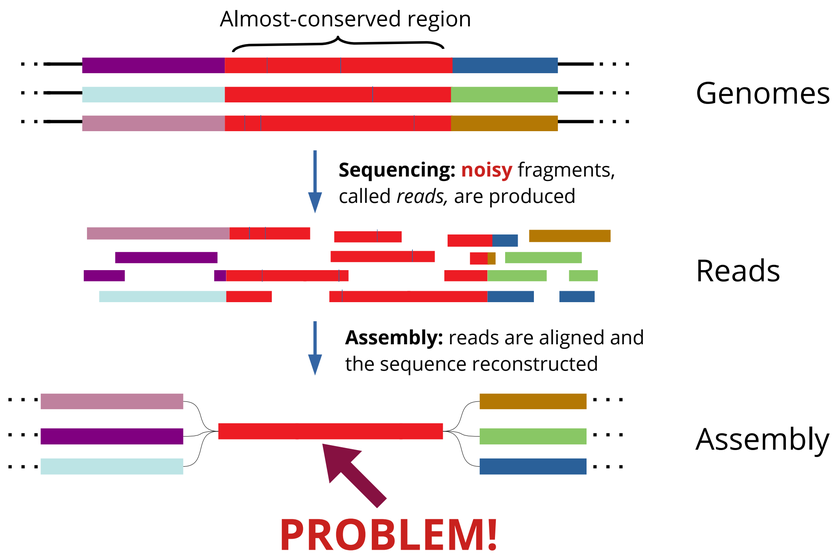
HairSplitter: separating similar strains in metagenome assembly

Roland Faure[1,2], Jean-François Flot[1], Dominique Lavenier[2]

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

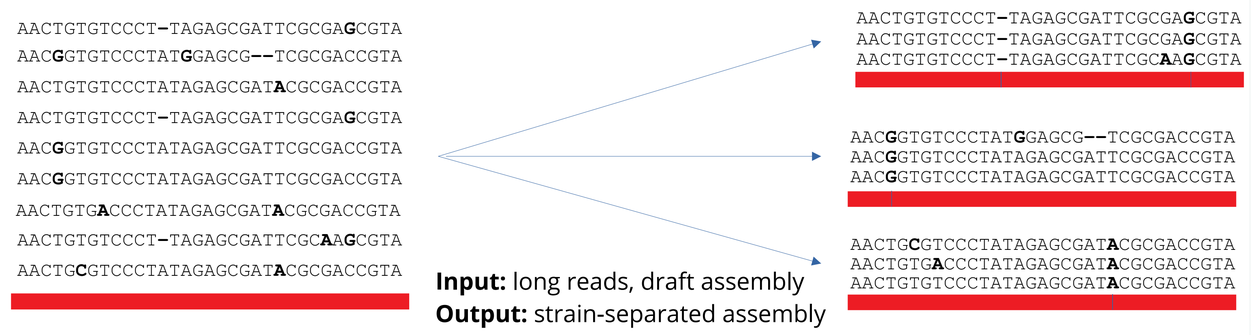
github.com/RolandFaure/Hairsplitter

Problem: assembling similar sequences



Similar regions (typically different strains) get collapsed in a single sequence

State of the art



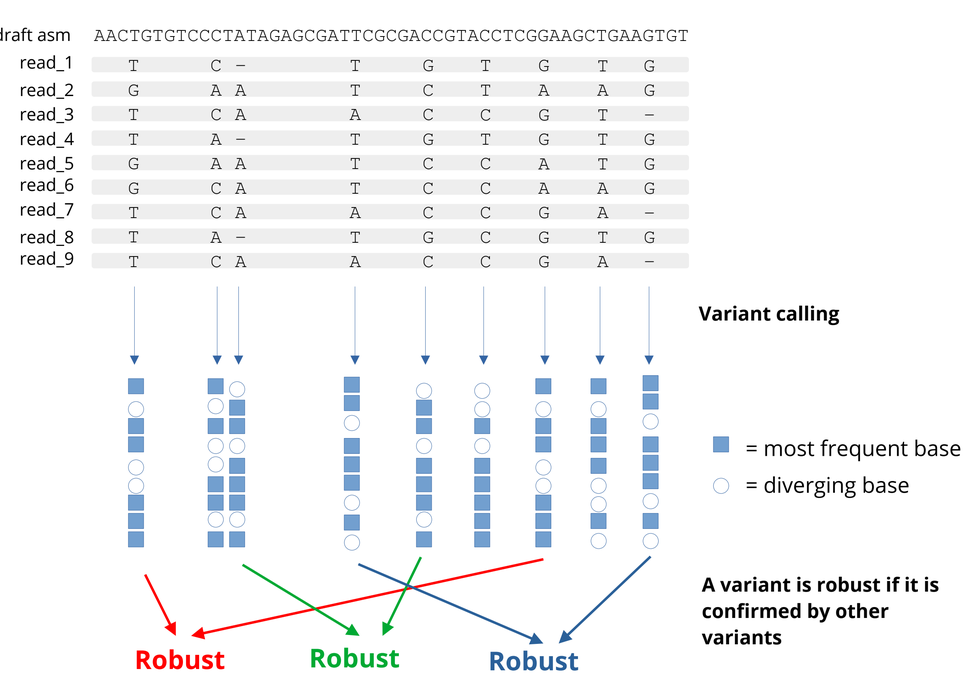
Difficulties: Unknown number of strains (potentially high), uneven coverage

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

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

Algorithm

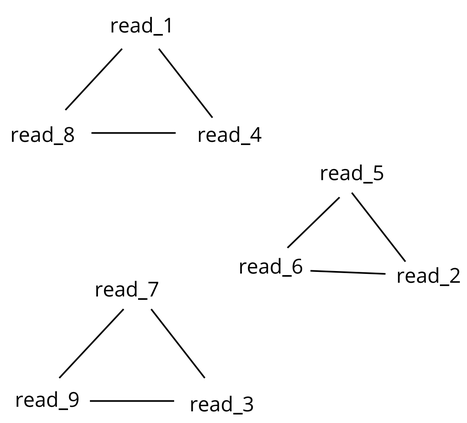
1. Robust variant calling



A variant is robust if it is confirmed by other variants

2. Read clustering

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

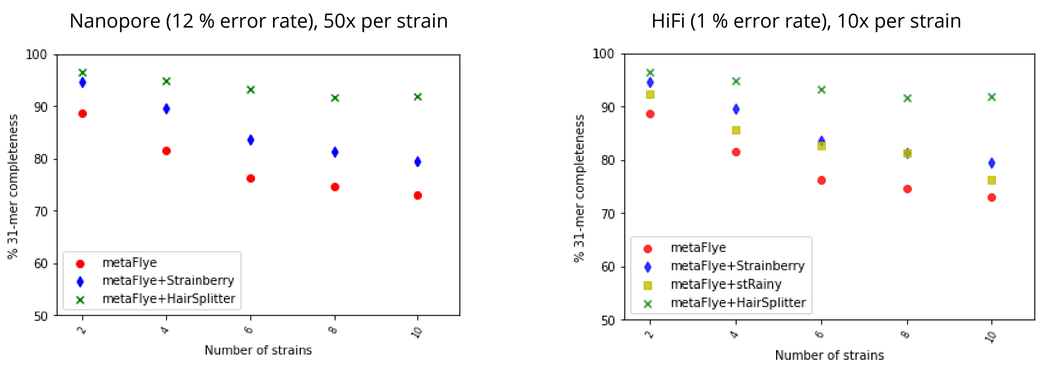


Graph is clustered by the Chinese Whispers algorithm

3. Reassembly

Results

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



- Mix of 5 *E. coli* strains, Zymobiomics 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 assemby, 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