



HAL
open science

Hairsplitter: Separating noisy long reads into an unknown number of haplotypes

Roland Faure, Jean-François Flot, Dominique Lavenier

► **To cite this version:**

Roland Faure, Jean-François Flot, Dominique Lavenier. Hairsplitter: Separating noisy long reads into an unknown number of haplotypes. *Genome Informatics 2022*, Sep 2022, London / Virtual, United Kingdom. pp.1-1. hal-03817928

HAL Id: hal-03817928

<https://hal.science/hal-03817928v1>

Submitted on 17 Oct 2022

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.

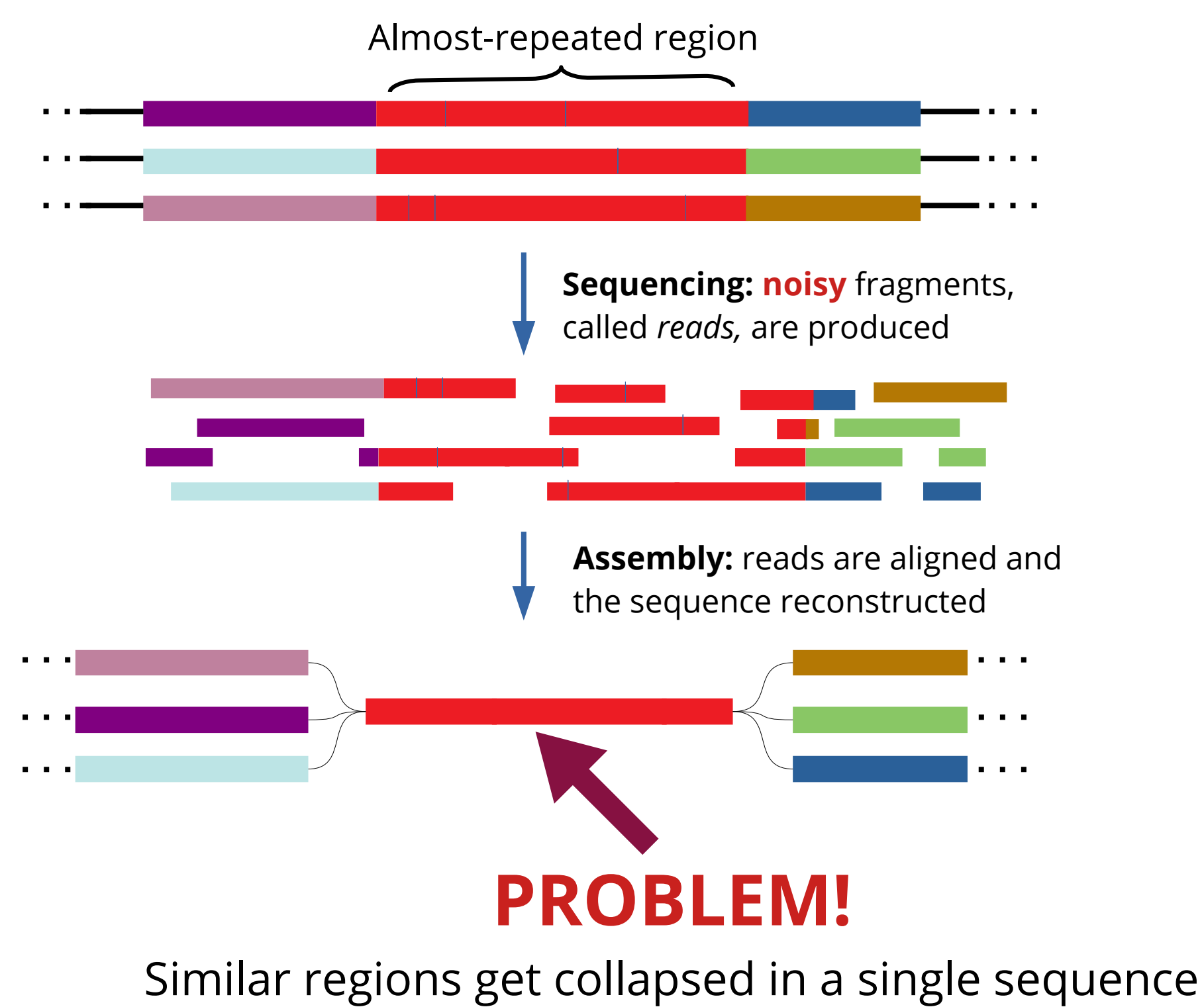
L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hairsplitter: Separating noisy long reads into an unknown number of haplotypes

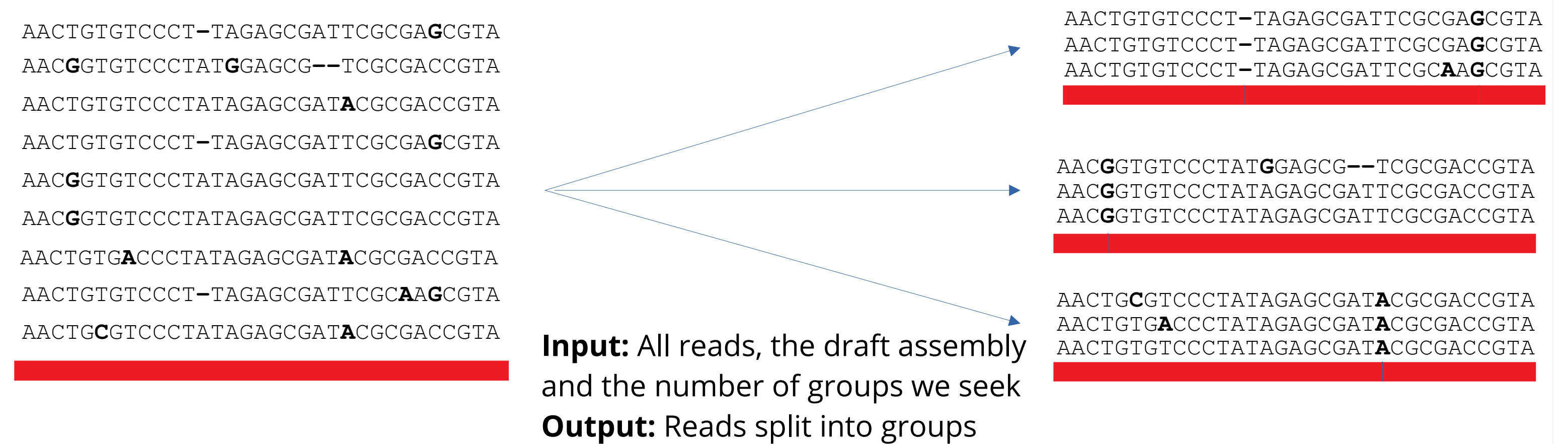
Roland Faure^{1,2}, Jean-François Flot¹, Dominique Lavenier²

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

Problem: assembling similar sequences



State of the art

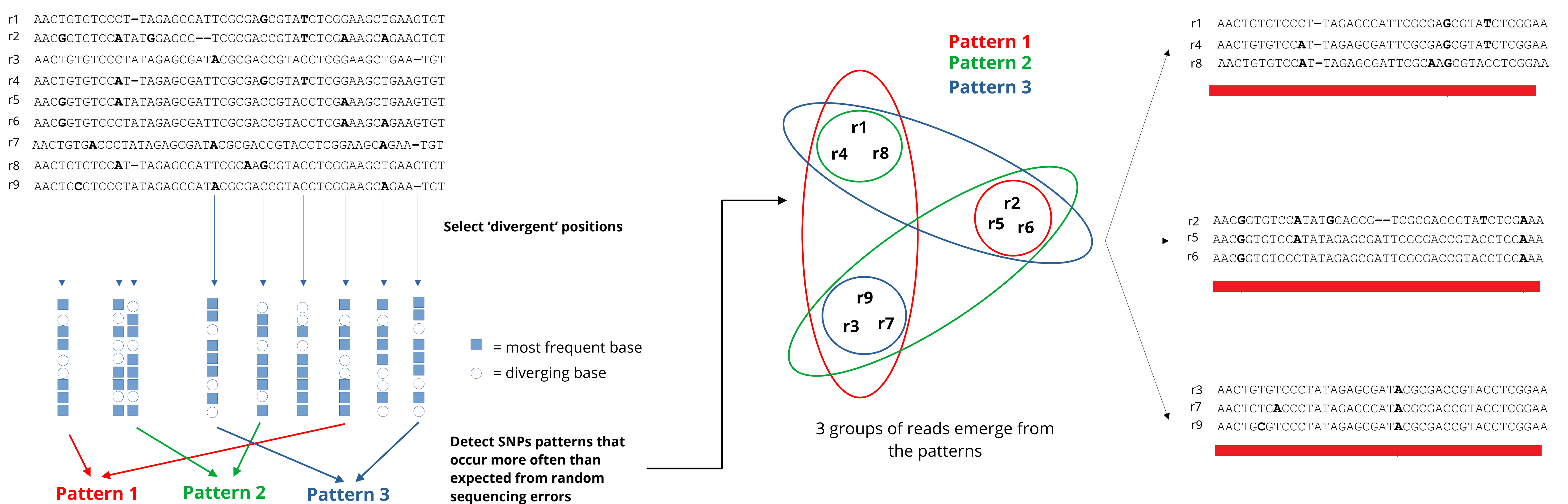


Existing software: WhatsHap (polyphase), Phasebook, Falcon-Unzip, HapCut, HapDup...

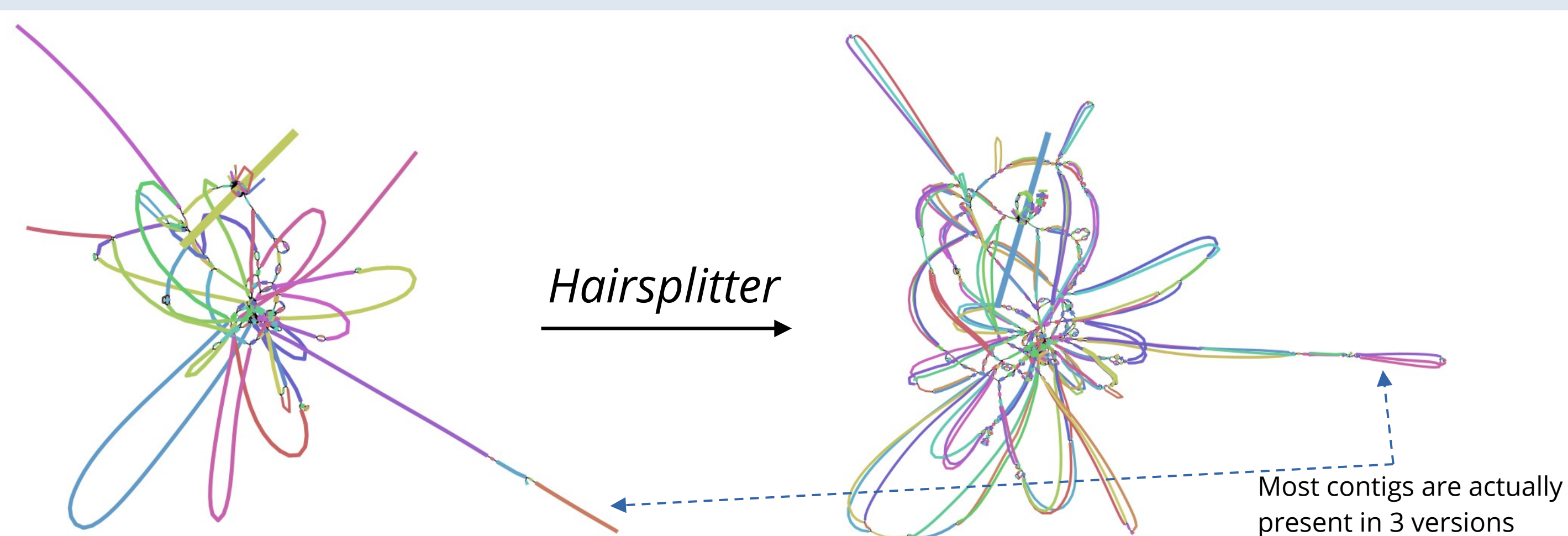
But: All software need to know the number of groups beforehand! We don't always know that → metagenomes, genomic repeats, insertions...

→ *HairSplitter* splits reads in an agnostic number of groups

Algorithm

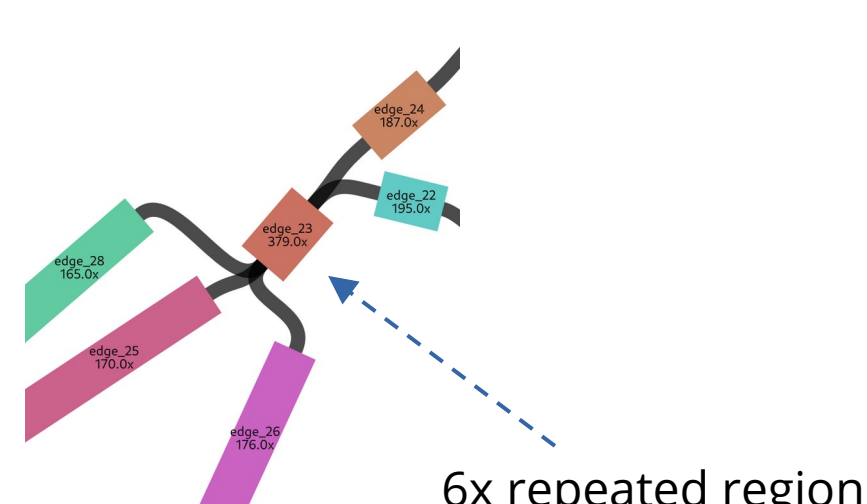


Results

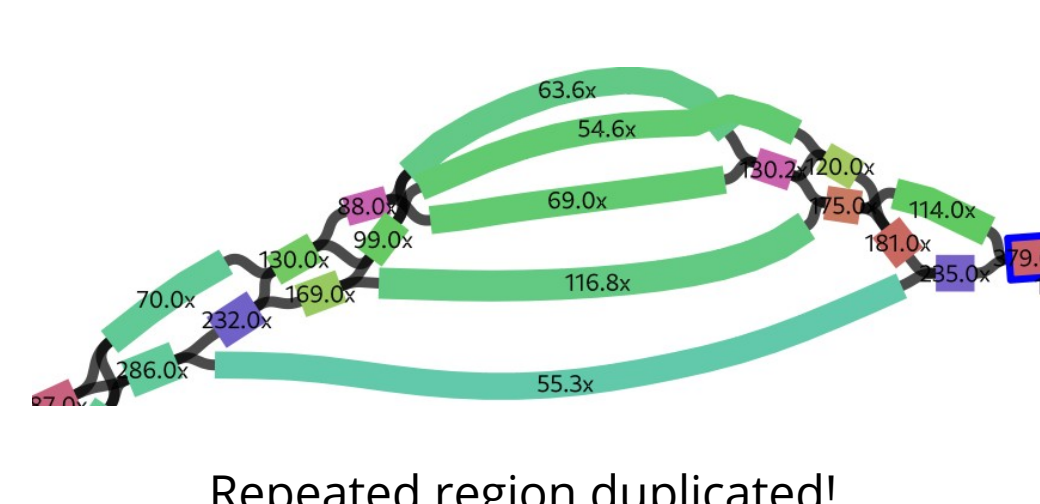


Mix of 3 haploid *Saccharomyces cerevisiae* strains
Nanopore ~ 6 % error rate, assembled with Flye
21% of 21-mers missing compared to separate strain assemblies

5 % of 21-mers missing compared to separate strain assemblies



Hairsplitter



Conclusion & Perspectives

- *Hairsplitter* splits a contig into an **agnostic number of groups**
- *Hairsplitter* can safely be **applied to all contigs** of the graph to split the contigs that need to be split
- *Hairsplitter* could be used to **improve the contiguity** of assemblies by overcoming almost-repeated regions

References

- [1] S.D. Schrunner, R.S. Mari, J. Ebler, M. Rautiainen, L. Seillier, J.J. Reimer, B. Usadel, T. Marschall, G. W. Klau. 2020. Haplotype threading: accurate polyploid phasing from long reads. *Genome Biology*. doi: 10.1186/s13059-020-02158-1
- [2] C. Chin, P. Peluso, F. Sedlazeck, M. Nattestad, G. Concepcion, A. Clum, C. Dunn, R. O'Malley, R. Figueroa-Balderas, A. Morales-Cruz, G. Cramer, M. Delledonne, C. Luo, J.Ecker, D. Cantu, D. Rank, M. Schatzl. 2016. Phased diploid genome assembly with single molecule real-time sequencing. *Nature Methods*. 10.1101/056887.
- [3] Luo, Xiao & Kang, Xiongbin & Schönhuth, Alexander. 2021. phasebook: haplotype-aware de novo assembly of diploid genomes from long reads. *Genome Biology*. 22. 10.1186/s13059-021-02512-x.
- [4] R. Wick, M. Schultz, J. Zobel, K. Holt. 2015. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics*. 31. 10.1093/bioinformatics/btv383.