Separating strains in metagenomic long-read assemblies

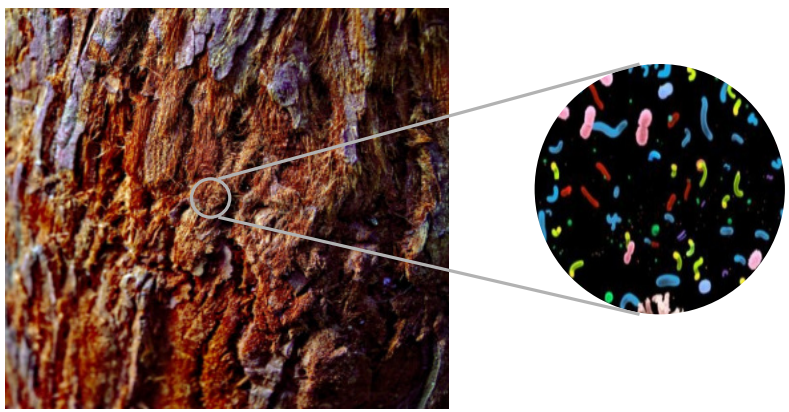
Roland Faure[1,2], Jean-François Flot[1], Dominique Lavenier[2]

[1]Université libre de Bruxelles (ULB)

[2]Université de Rennes, IRISA

June 2023

Microbiomes



- Microbiomes play crucial roles in organisms and ecosystems

State of the art: studies at species-level

Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression

Charles K. Fisher, Pankaj Mehta

dozens of microbial species could modulate or contribute to cancer

N. Cullin *et al.* , *Microbiome and cancer,* 2021

Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules

Roie Levy and Elhanan Borenstein Authors Info & Affiliations

bacteria are rare. Of the 639 species identified in a population study of 1135 Dutch individuals, 469 (73%) were present in

R.K. Weersma *et al.* , *Interaction between drugs and the gut microbiome,* 2020

difficulties to attribute the species that produce the identified metabolite.

M.S. Afridi *et al.* , *Plant Microbiome Engineering : Hopes or Hypes,* 2022

invasions into soil communities [50]. Similarly, low abundance bacterial species largely contributed to the production of antifungal volatile compounds that protect the plant against soil-borne

S. Compant *et al. A review on the plant micobiome : Ecology, functions, and emerging trends in microbial application,* 2019

State of the art: studies at species-level

Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries Using Sparse Linear Regression

Charles K. Fisher, Pankaj Mehta

Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules

Roie Levy and Elhanan Borenstein Authors Info & Affiliations

dozens of microbial species could modulate or contribute to cancer

N. Cullin *et al.* , *Microbiome and cancer,* 2021

bacteria are rare. Of the 639 species identified in a population study of 1135 Dutch individuals, 469 (73%) were present in

R.K. Weersma *et al.* , *Interaction between drugs and the gut microbiome,* 2020

difficulties to attribute the species that produce the identified metabolite.

M.S. Afridi *et al.* , *Plant Microbiome Engineering : Hopes or Hypes,* 2022

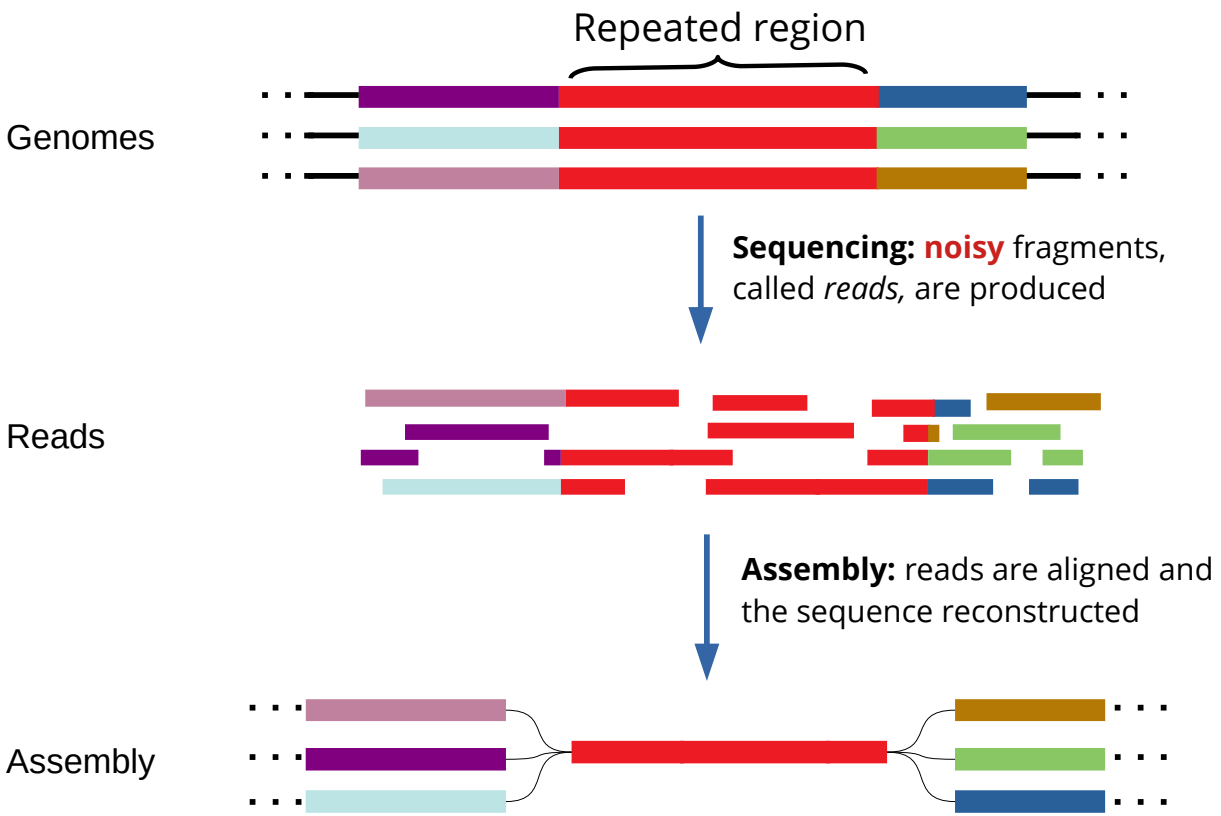
invasions into soil communities [50]. Similarly, low abundance bacterial species largely contributed to the production of antifungal volatile compounds that protect the plant against soil-borne

S. Compant *et al. A review on the plant micobiome : Ecology, functions, and emerging trends in microbial application,* 2019

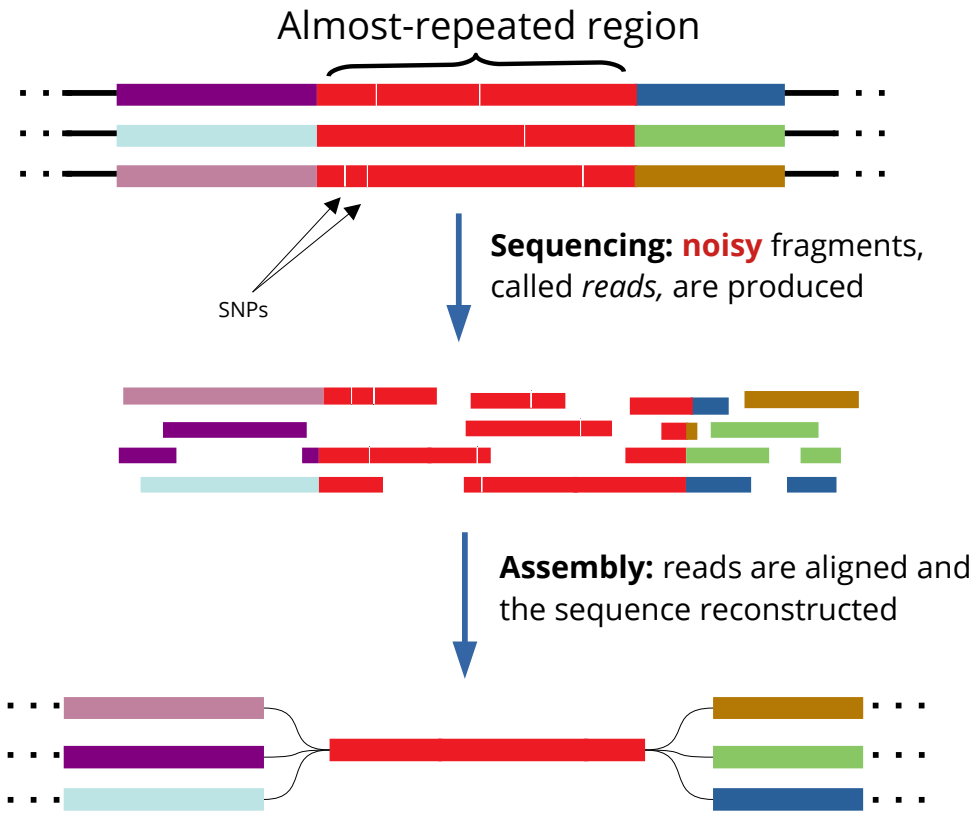
Knowledge gaps remain: strain diversity

Ryan Caldwell, Wei Zhou, Julia Oh, *Strains to go: interactions of the skin microbiome beyond its species, 2022*

Genome assembly

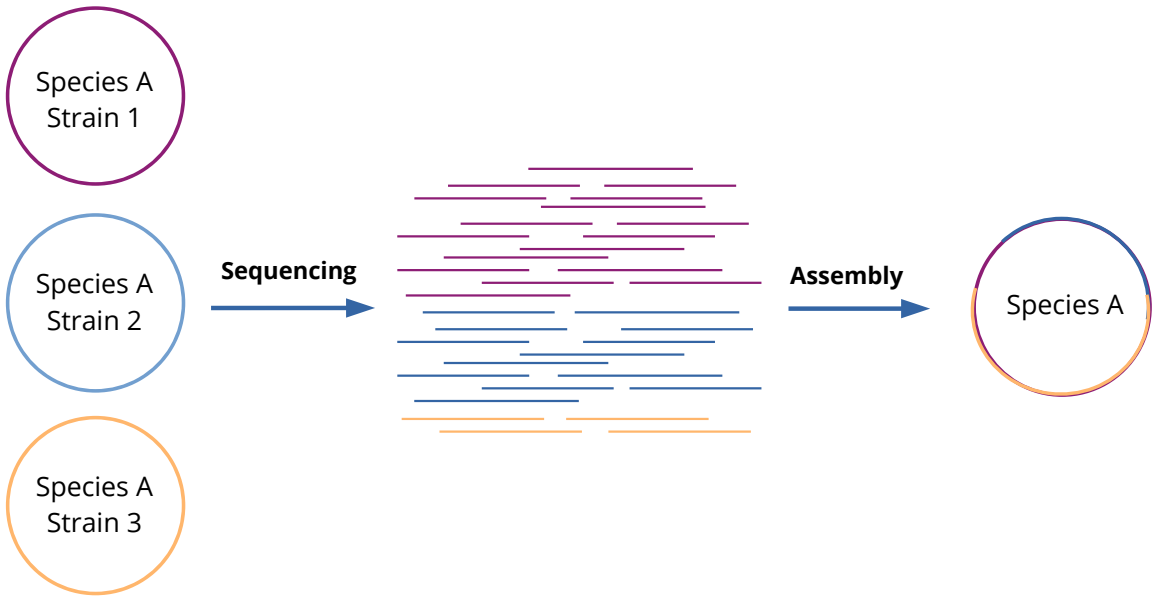


Genome assembly: similar regions get collapsed



- When divergence is small compared to the error rate of reads, it is discarded as sequencing errors

Assembling a metagenome is difficult



- Unknown (potentially high) number of strains
- Uneven coverage
- Highly similar strains
- One of the main reason microbiomes are studied at species-level

State of the art

Article | Open Access | Published: 23 July 2021

Strainberry: automated strain separation in low-complexity metagenomes using long reads

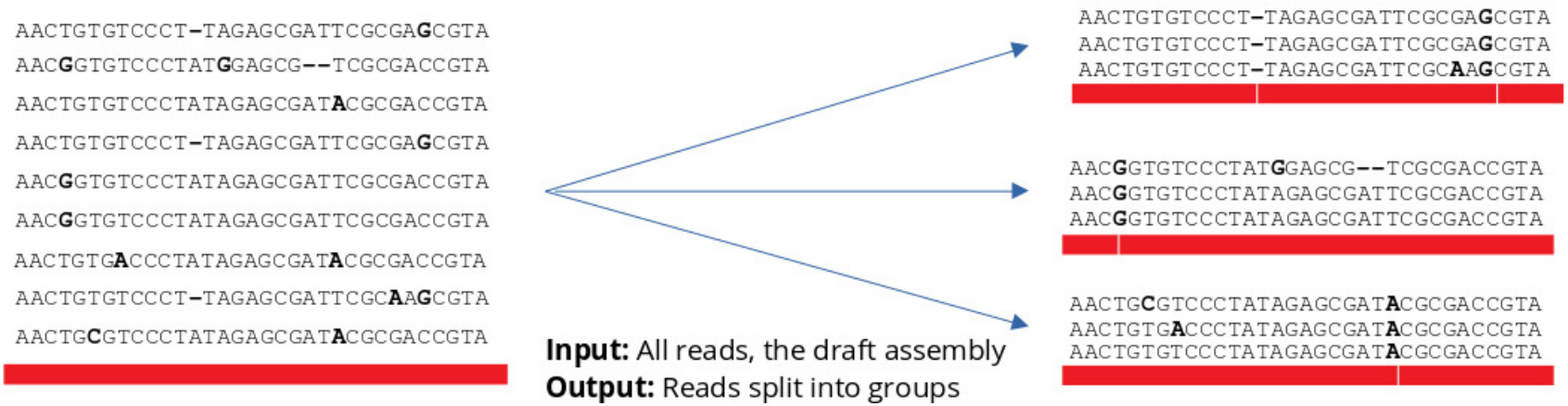
Riccardo Vicedomini, Christopher Quince, Aaron E. Darling & Rayan Chikhi

Nature Communications **12**, Article number: 4485 (2021) | Cite this article

- Unsatisfactory when many strains present

HairSplitter

- HairSplitter recovers the lost differences between strains



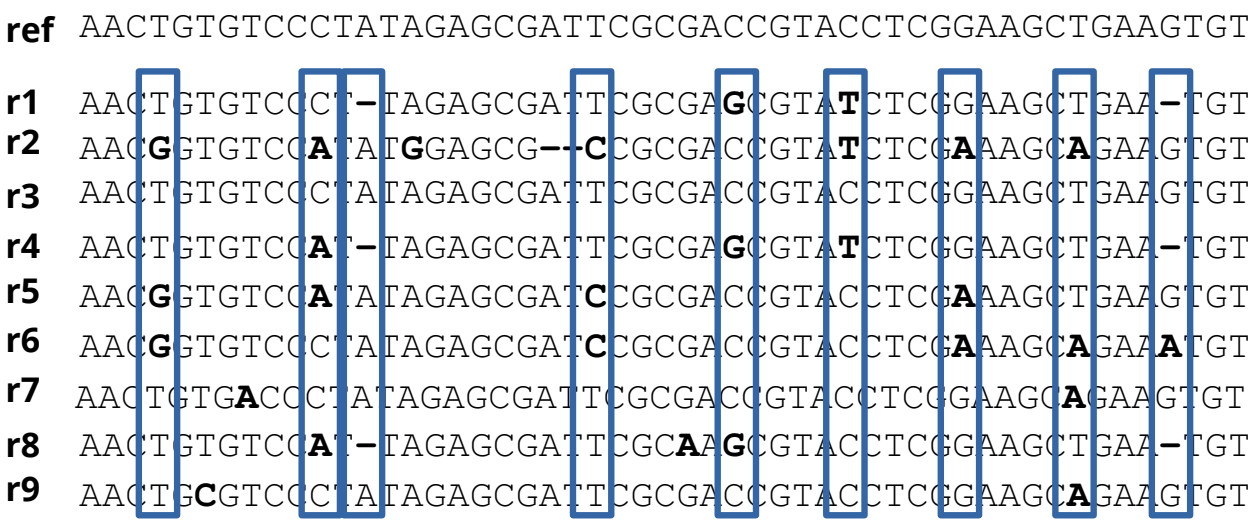
- *Hairsplitter:* One given to hair-splitting or making sophistical distinctions in reasoning. - *The Century Dictionary.*

Let's try to split the reads

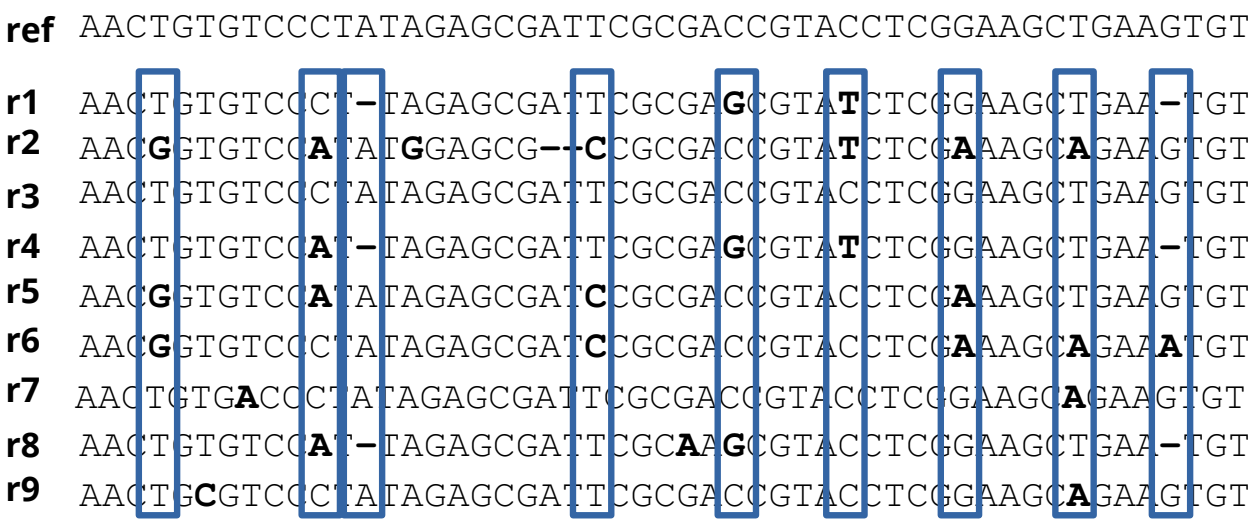
```
ref  AACTGTGTCCCTATAGAGCGATTCGCGACCGTACCTCGGAAGCTGAAGTGT

r1   AACTGTGTCCCT-TAGAGCGATTCGCGAGCGTATCTCGGAAGCTGAAGTGT
r2   AACGGTGTCCATATGGAGCG--TCGCGACCGTATCTCGAAAGCAGAAGTGT
r3   AACTGTGTCCCTATAGAGCGATACGCGACCGTACCTCGGAAGCTGAA-TGT
r4   AACTGTGTCCAT-TAGAGCGATTCGCGAGCGTATCTCGGAAGCTGAAGTGT
r5   AACGGTGTCCATATAGAGCGATTCGCGACCGTACCTCGAAAGCTGAAGTGT
r6   AACGGTGTCCCTATAGAGCGATTCGCGACCGTACCTCGAAAGCAGAAGTGT
r7   AACTGTGACCCTATAGAGCGATACGCGACCGTACCTCGGAAGCAGAA-TGT
r8   AACTGTGTCCAT-TAGAGCGATTCGCAAGCGTACCTCGGAAGCTGAAGTGT
r9   AACTGCGTCCCTATAGAGCGATACGCGACCGTACCTCGGAAGCAGAA-TGT
```

First intuition: some positions are suspicious

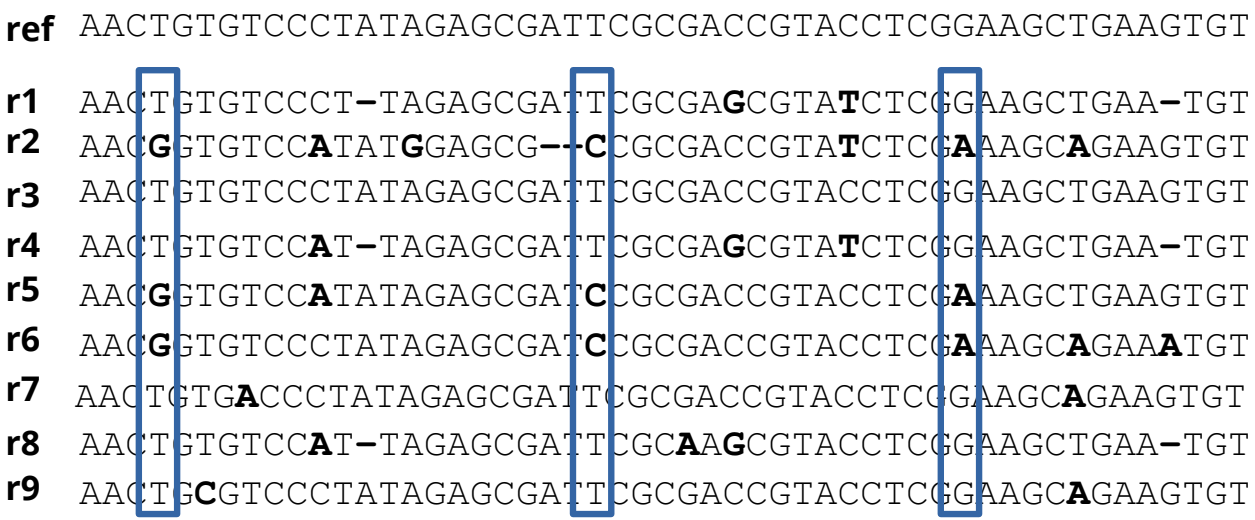


First intuition: some positions are suspicious

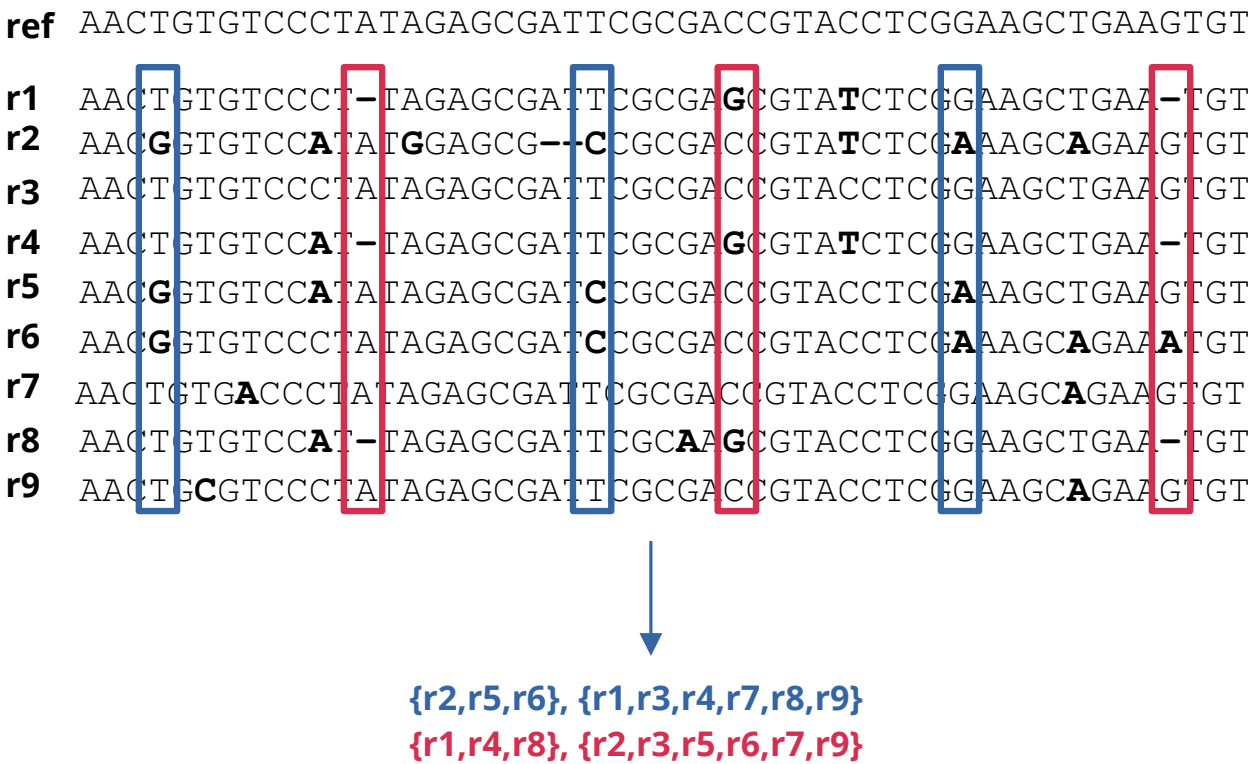


- Many of these positions are not actual variants

HairSplitter's key idea: this can't be chance



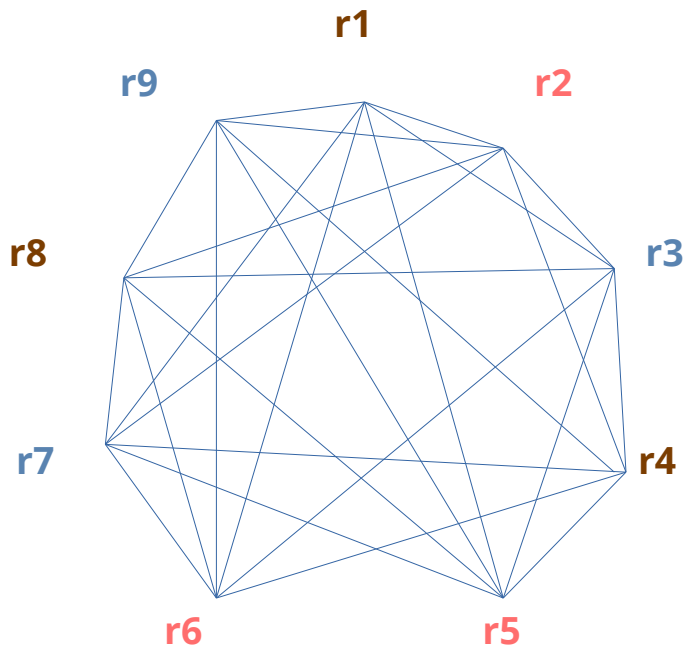
Data mining to highlight variants



Separating the reads

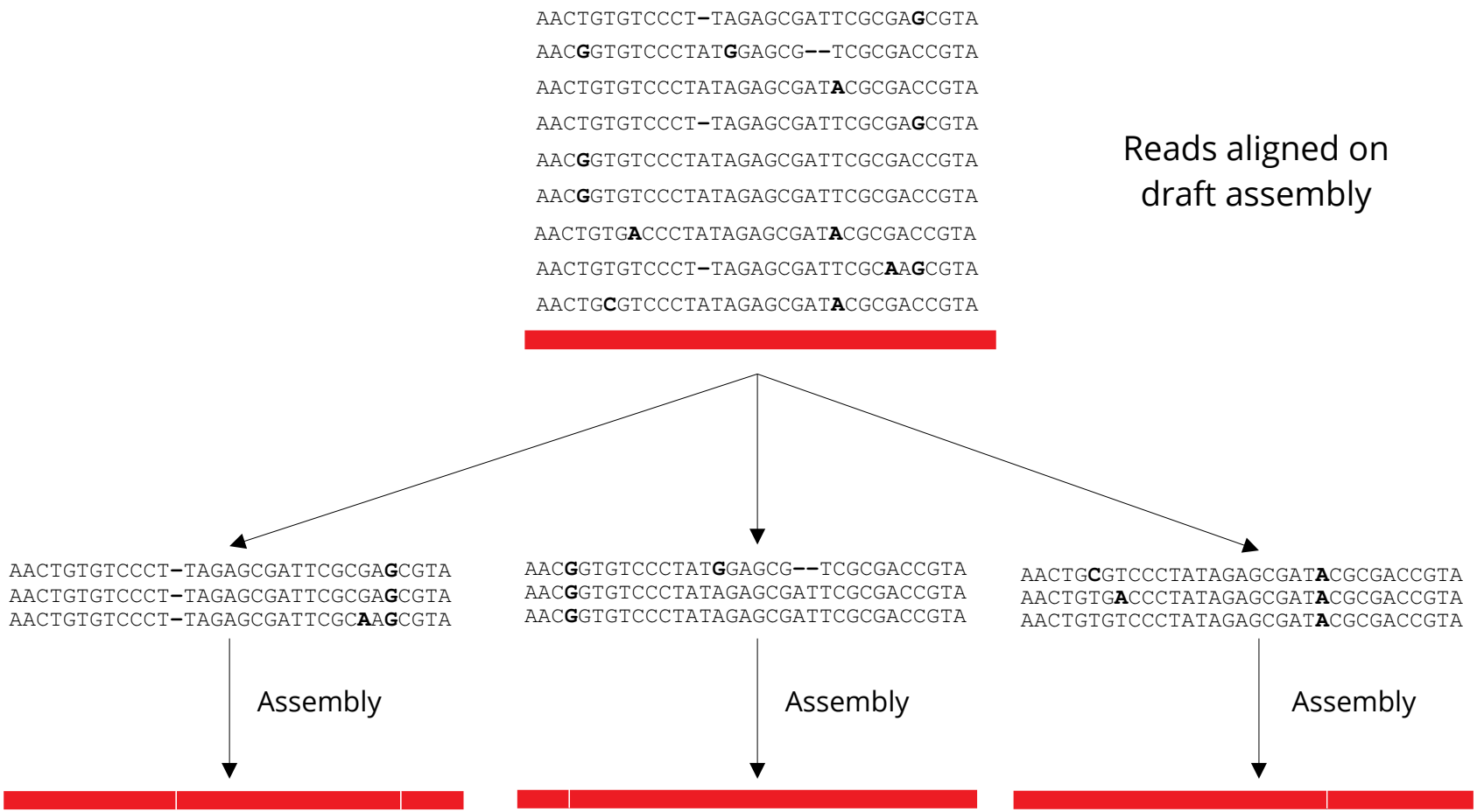
- Reads that should not be in the same strain are linked

{r2,r5,r6}, {r1,r3,r4,r7,r8,r9}
{r1,r4,r8}, {r2,r3,r5,r6,r7,r9}



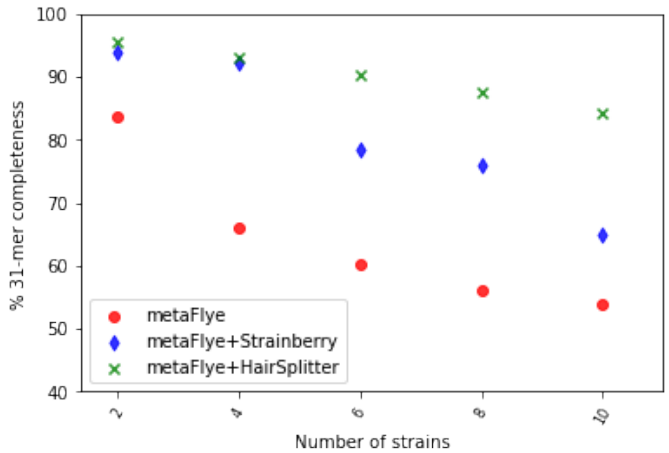
- Vertex coloring problem

Reads are re-assembled

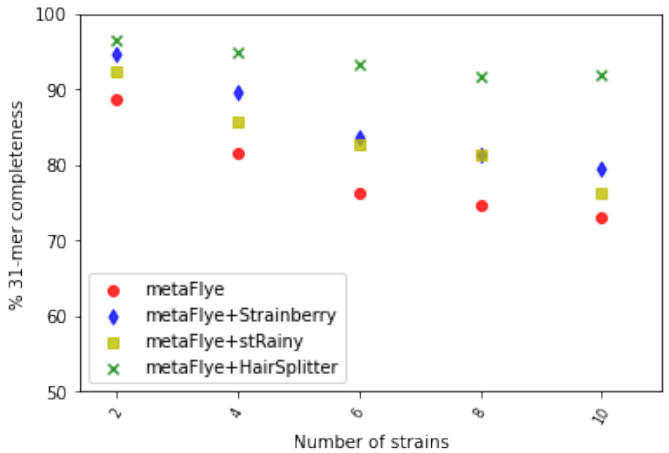


Simulated data

- Mix of 2 to 10 *E. coli* genomes. Reads simulated with Badreads



(a) Nanopore sequencing (12% error), 50x per strain



(b) HiFi sequencing, 10x per strain

- HairSplitter performs very well with high number of strains

Mock data

- Zymobiomics gut microbiome standard: contains a mix of 5 *E. coli strains*

	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

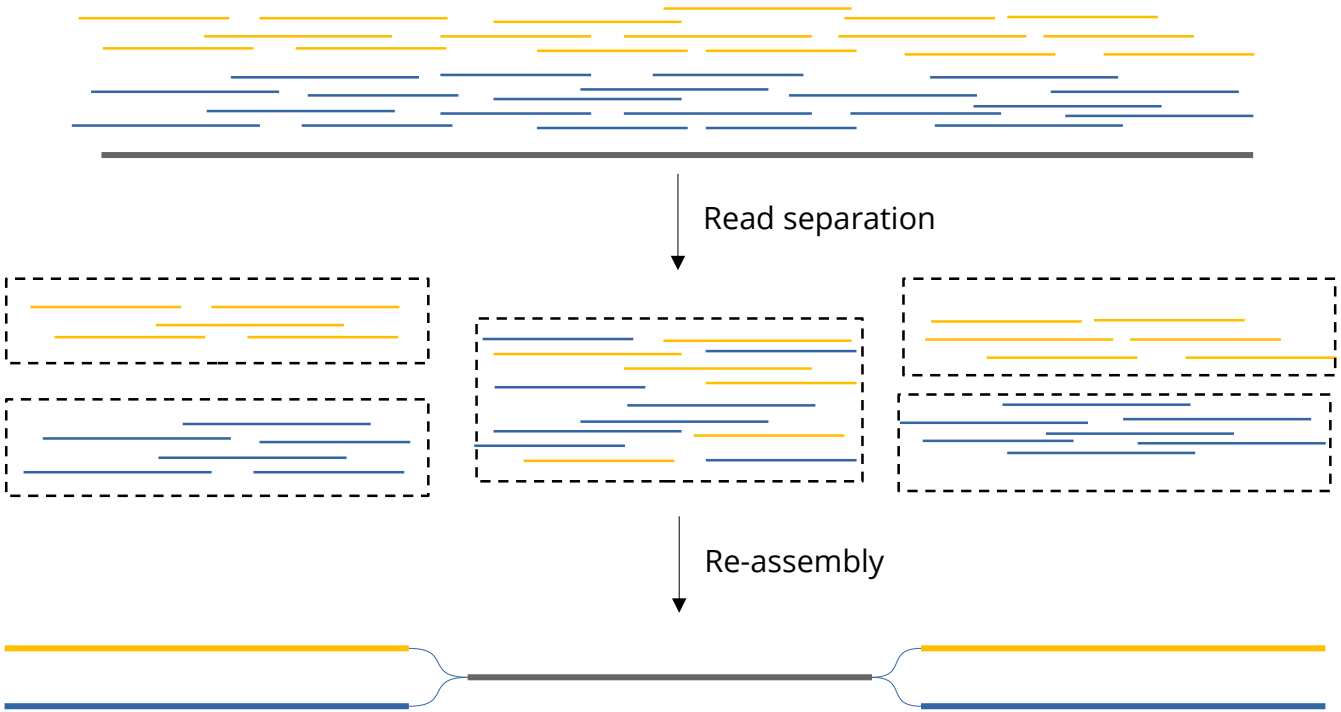
Real data

- 5 *Vagococcus fluvialis* strains sequenced with Nanopore **barcoded** reads (doi.org/10.1186/s12864-022-08842-9).

	metaFlye	metaFlye+Strainberry	metaFlye+HairSplitter
Nanopore	0.718	0.7398	**0.9042**

Table: 31-mer completeness of assemblies w.r.t. a Flye assembly where reads from different strains were separated.

Limitation: Contiguity



- Strains are separated only locally
- Contiguity can decrease significantly

Take-home message

- HairSplitter reconstruct collapsed sequences from "draft" assemblies obtained by any means
- HairSplitter can be used on metagenomic (and multiploid ?) assemblies
- Only needs sequencing reads, potentially error-prone

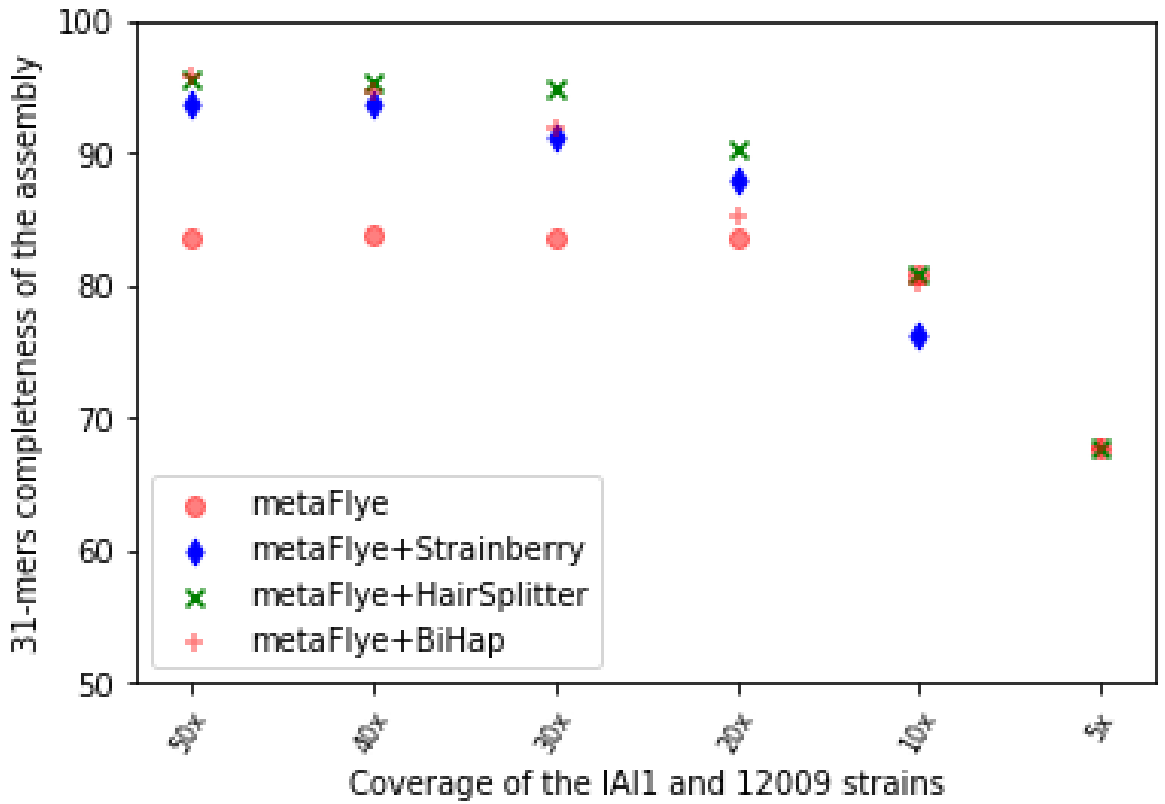
Take-home message

- HairSplitter reconstruct collapsed sequences from "draft" assemblies obtained by any means
- HairSplitter can be used on metagenomic (and multiploid ?) assemblies
- Only needs sequencing reads, potentially error-prone
- Available prototype ! github.com/RolandFaure/HairSplitter

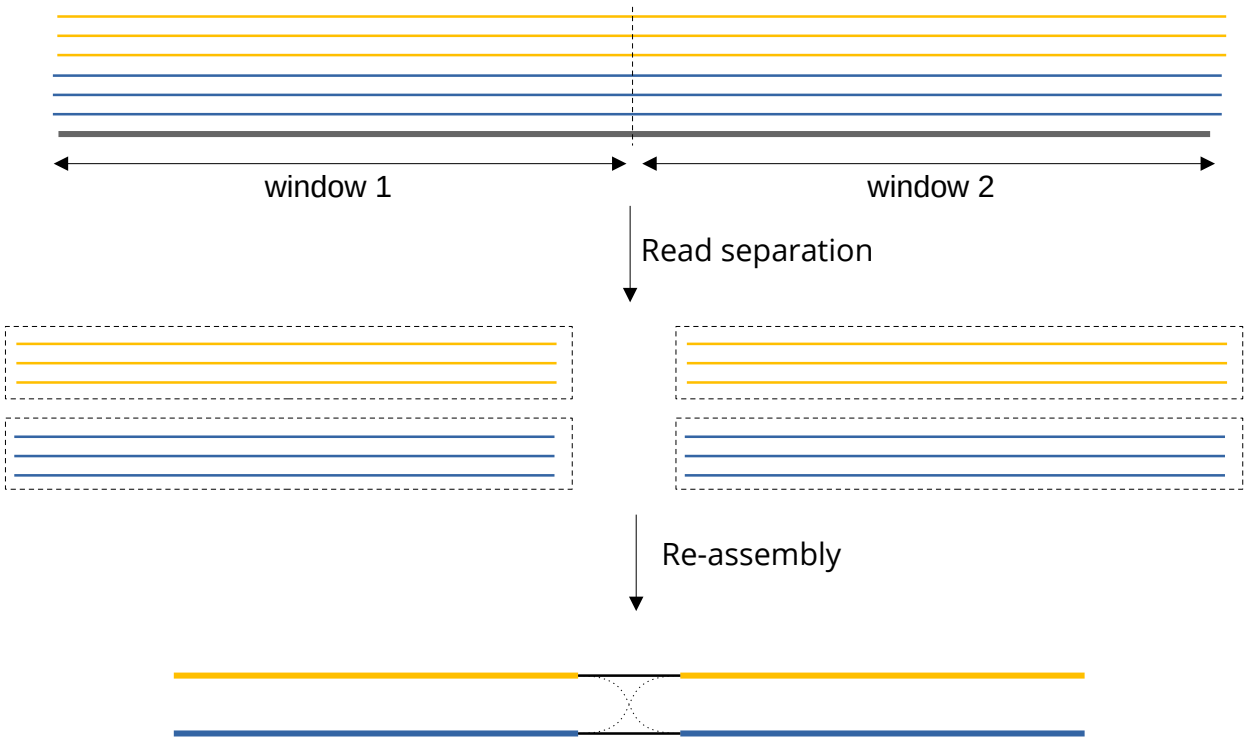
Acknowledgements

- Dominique Lavenier and Jean-François Flot for their supervision
- Rumen Andonov and Tam Truong for their help in formalizing the problem
- The EEB-EBE and GenScale teams

Behaviour of HairSplitter/BiHap: coverage



Local strain separation



Local strain separation