Separating strains in metagenomic long-read assemblies

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

[1]Université libre de Bruxelles (ULB) - Belgium

[2]Université de Rennes, IRISA - France

ISMB/ECCB Lyon 2023

Microbiomes are studied 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

Microbiomes are studied 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

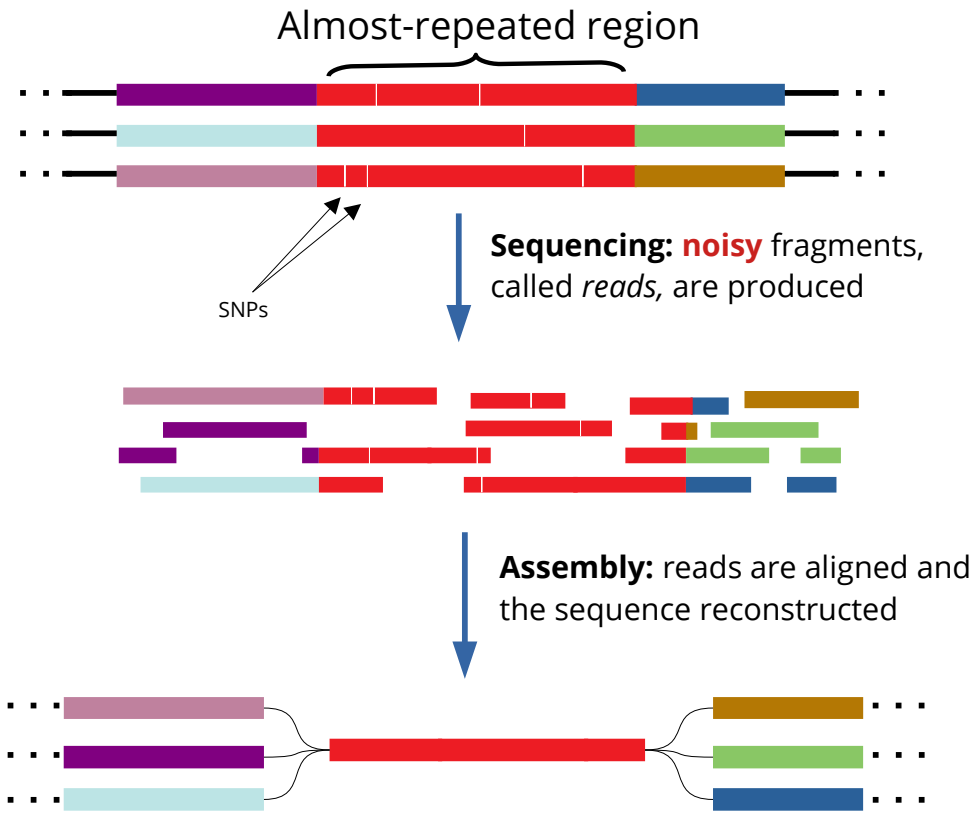
invasions into soil communities [50]. Similarly, low abundance bacterial species largely contributed to the production of antifungal volatile compounds that protect the plant against soil-borne

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

Knowledge gaps remain: strain diversity

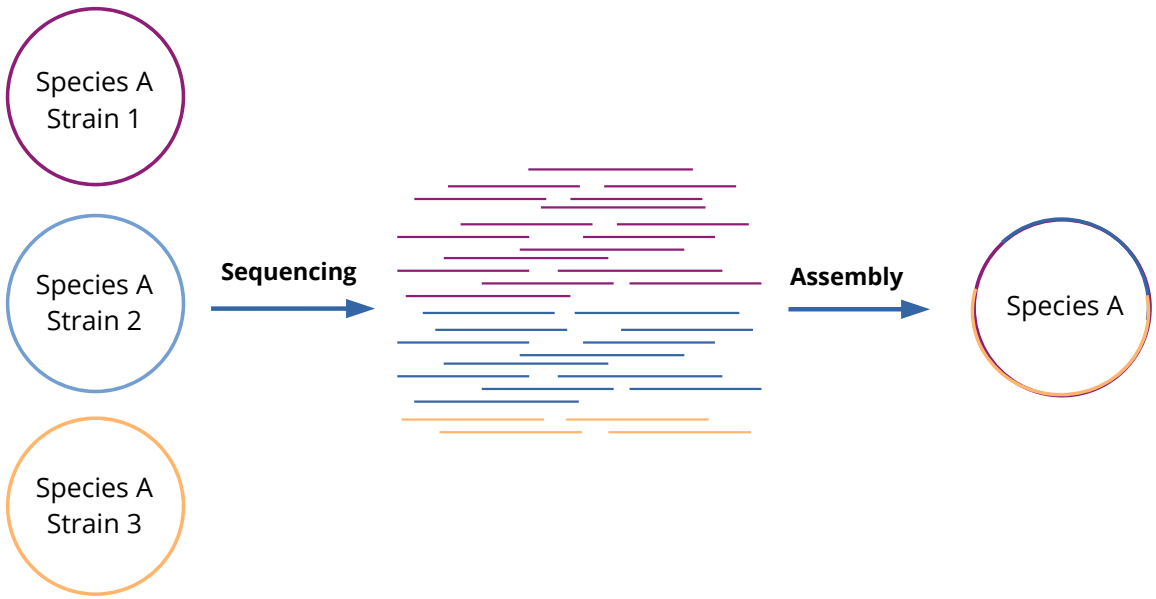
Ryan Caldwell, Wei Zhou, Julia Oh, *Strains to go: interactions of the skin microbiome beyond its species, 2022*

(Meta)genome assembly: similar regions get collapsed



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

Assembling close strains is difficult



- Unknown (potentially high) number of strains
- Uneven coverage
- One of the reasons microbiomes are studied at species-level

State of the art (long reads)

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

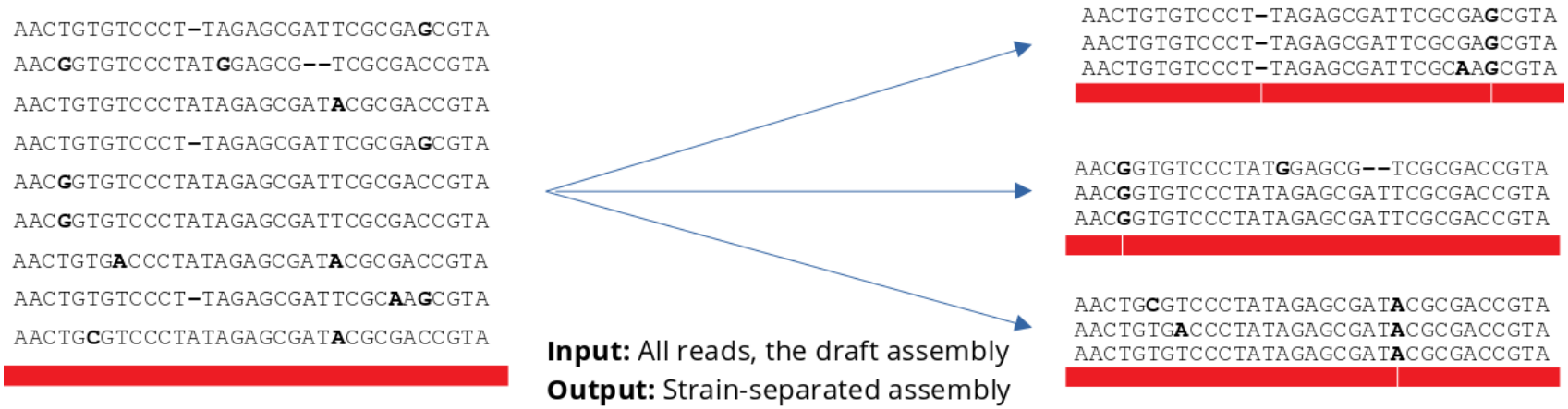
stRainy: assembly-based metagenomic strain phasing using long reads

Ekaterina Kazantseva, Ataberk Donmez, Mihai Pop, Mikhail Kolmogorov

doi: https://doi.org/10.1101/2023.01.31.526521

HairSplitter

- From an assembly, separate the contigs that are present in several versions



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

Let's try to split the reads

Draft assembly	AACTGTGTCCCTATAGAGCGATTCGCGACCGTACCTCGGAAGCTGAAGTGT
read_1	AACTGTGTCCCT**-**TAGAGCGATTCGCGA**G**CGTA**T**CTCGGAAGCTGAA**-**TGT
read_2	AAC**G**GTGTCC**A**TAT**G**GAGCG**--C**CGCGACCGTA**T**CTCG**A**AAGC**A**GAAGTGT
read_3	AACTGTGTCCCTATAGAGCGATTCGCGACCGTACCTCGGAAGCTGAAGTGT
read_4	AACTGTGTCC**A**T**-**TAGAGCGATTCGCGA**G**CGTA**T**CTCGGAAGCTGAA**-**TGT
read_5	AAC**G**GTGTCC**A**TATAGAGCGAT**C**CGCGACCGTACCTCG**A**AAGCTGAAGTGT
read_6	AAC**G**GTGTCCCTATAGAGCGAT**C**CGCGACCGTACCTCG**A**AAGC**A**GAA**A**TGT
read_7	AACTGTG**A**CCCTATAGAGCGATTCGCGACCGTACCTCGGAAGC**A**GAAGTGT
read_8	AACTGTGTCC**A**T**-**TAGAGCGATTCGC**A**A**G**CGTACCTCGGAAGCTGAA**-**TGT
read_9	AACTG**C**GTCCCTATAGAGCGATTCGCGACCGTACCTCGGAAGC**A**GAAGTGT

First intuition: calling variants

Draft assembly AACTGTGTCCCTATAGAGCGATTCGCGACCGTACCTCGGAAGCTGAAGTGT

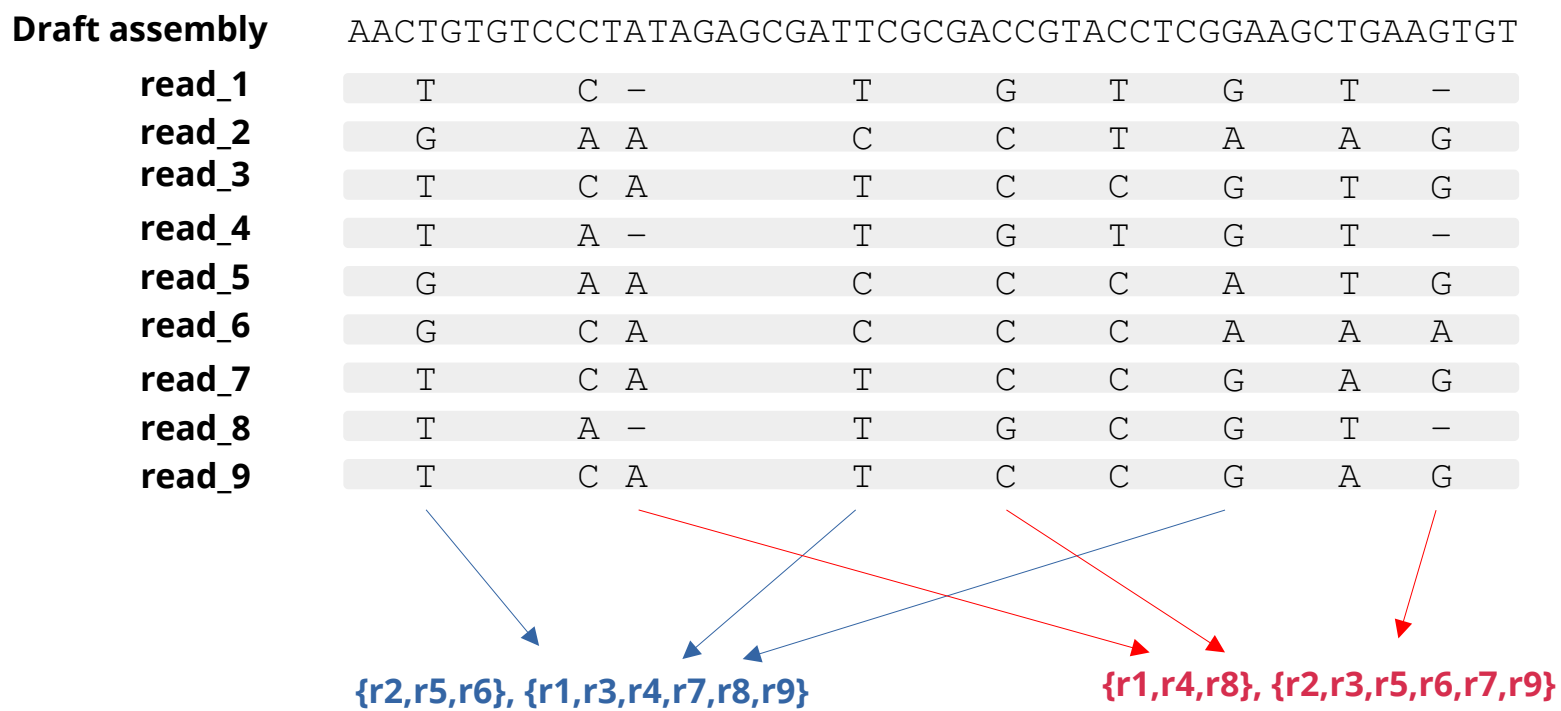
read_1	T	C	–	T	G	T	G	T	–
read_2	G	A	A	C	C	T	A	A	G
read_3	T	C	A	T	C	C	G	T	G
read_4	T	A	–	T	G	T	G	T	–
read_5	G	A	A	C	C	C	A	T	G
read_6	G	C	A	C	C	C	A	A	A
read_7	T	C	A	T	C	C	G	A	G
read_8	T	A	–	T	G	C	G	T	–
read_9	T	C	A	T	C	C	G	A	G

First intuition: calling variants

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

- Many of these positions are not actual variants

HairSplitter's key idea: keeping only robust variants



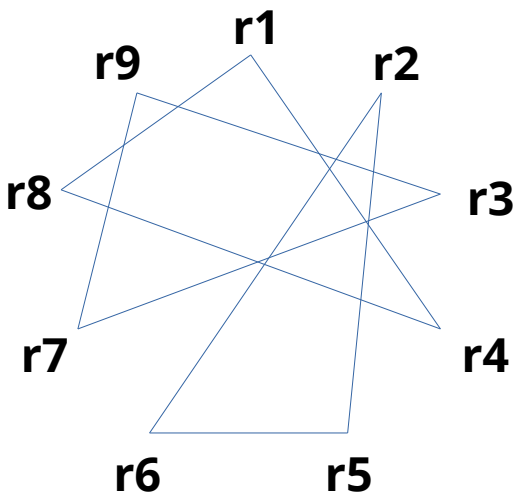
- A variant is robust if the partition induced is present ≥ 3 times

Separating the reads

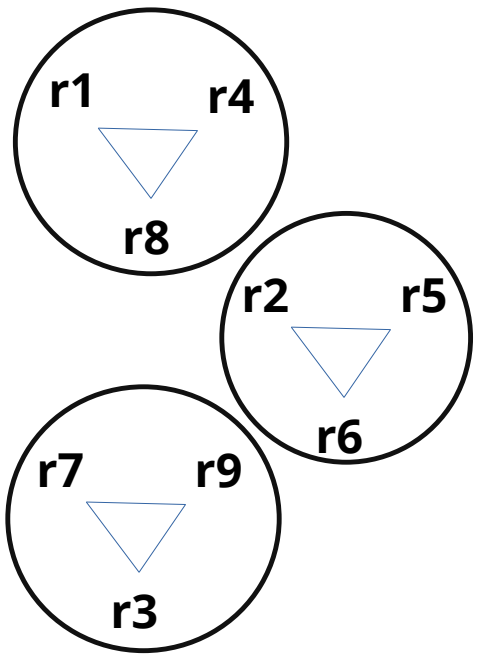
1. Compute pairwise distance between reads

	read_9	read_8	read_7	read_6	read_5	read_4	read_3	read_2	read_1
read_1	0.3	0.1	0.9	0.6	0.5	0.0	0.6	0.9	-
read_2	0.6	0.7	0.9	0.0	0.1	0.8	0.6	-	-
read_3	0.1	0.7	0.1	0.9	0.5	0.8	-	-	-
read_4	0.7	0.0	0.9	0.9	0.9	-	-	-	-
read_5	0.8	0.6	0.8	0.1	-	-	-	-	-
read_6	0.9	0.6	0.8	-	-	-	-	-	-
read_7	0.0	0.4	-	-	-	-	-	-	-
read_8	0.7	-	-	-	-	-	-	-	-
read_9	-	-	-	-	-	-	-	-	-

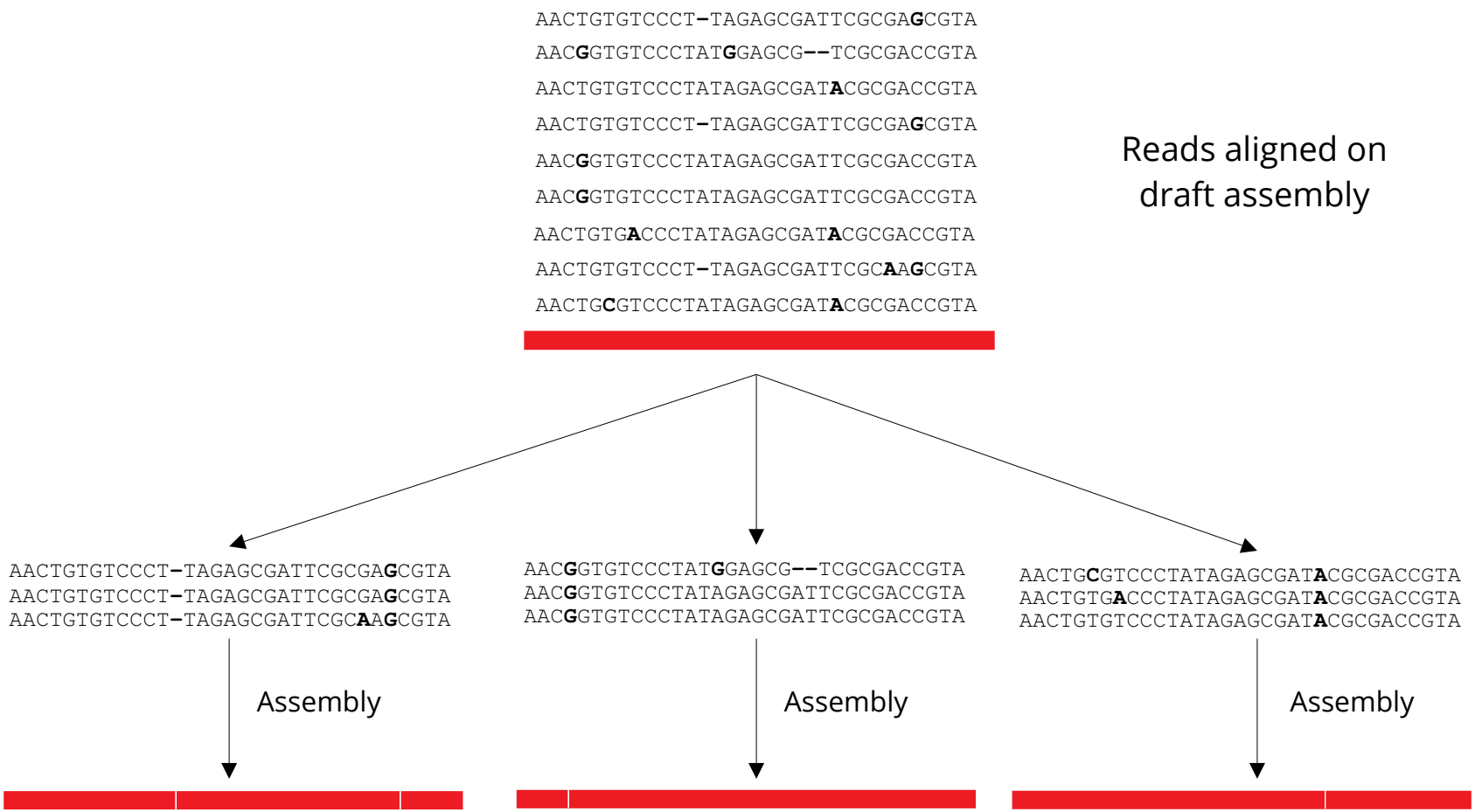
2.Build graph using KNN



3. Cluster graph



Re-assemble reads

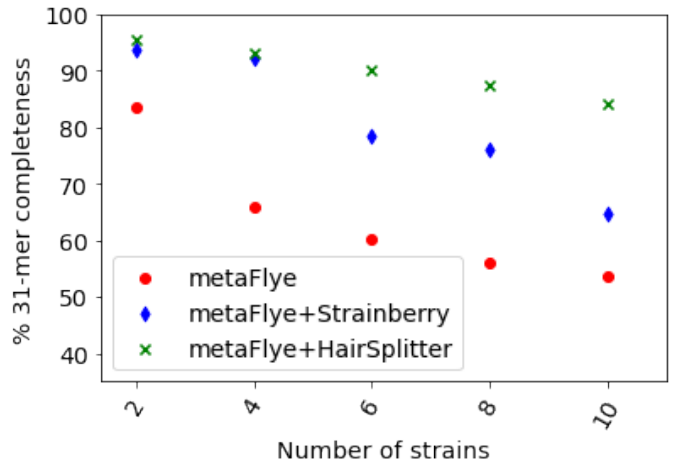


Simulated data

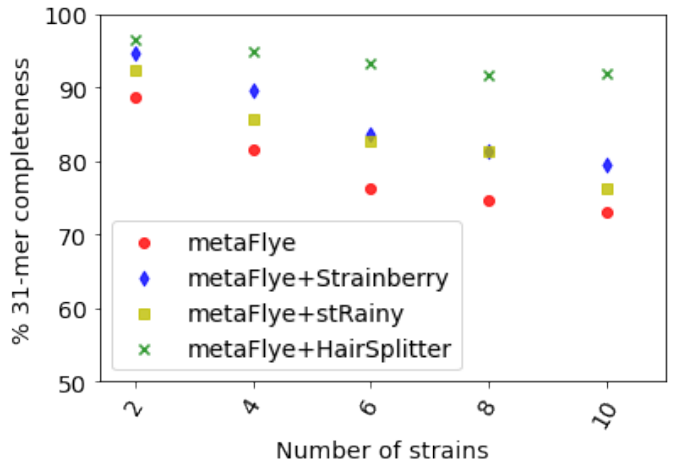
- Mix of 2 to 10 *E. coli* RefSeq genomes. Simulated reads
- Metrics: metaQuast completeness, error rate, **31-mer completeness**...

Simulated data

- Mix of 2 to 10 *E. coli* RefSeq genomes. Simulated reads
- Metrics: metaQuast completeness, error rate, **31-mer completeness**...



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



(b) HiFi sequencing, 10x per strain

- HairSplitter handles very well 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

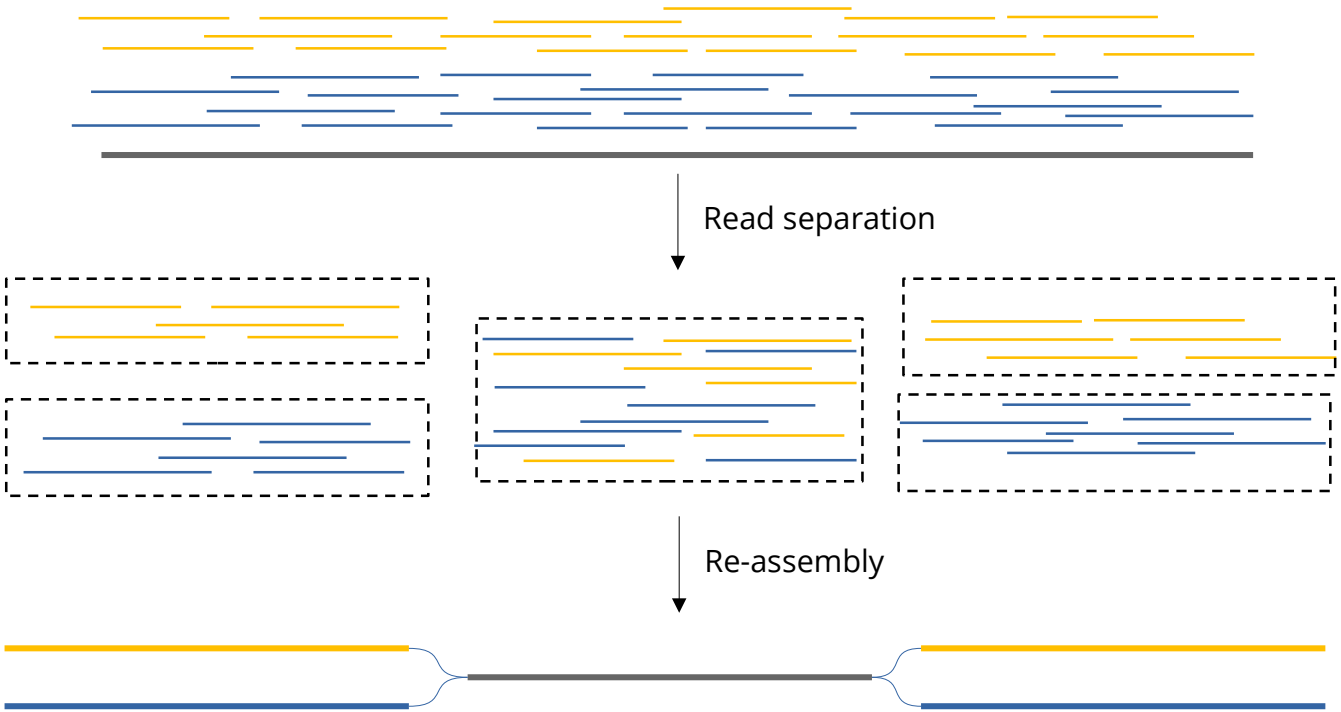
Real data

- 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
- N50 can decrease significantly

Take-home message

- HairSplitter reconstruct collapsed sequences from "draft" assemblies obtained by any means
- HairSplitter also works well on complex metagenomes and with high-error rate reads

Take-home message

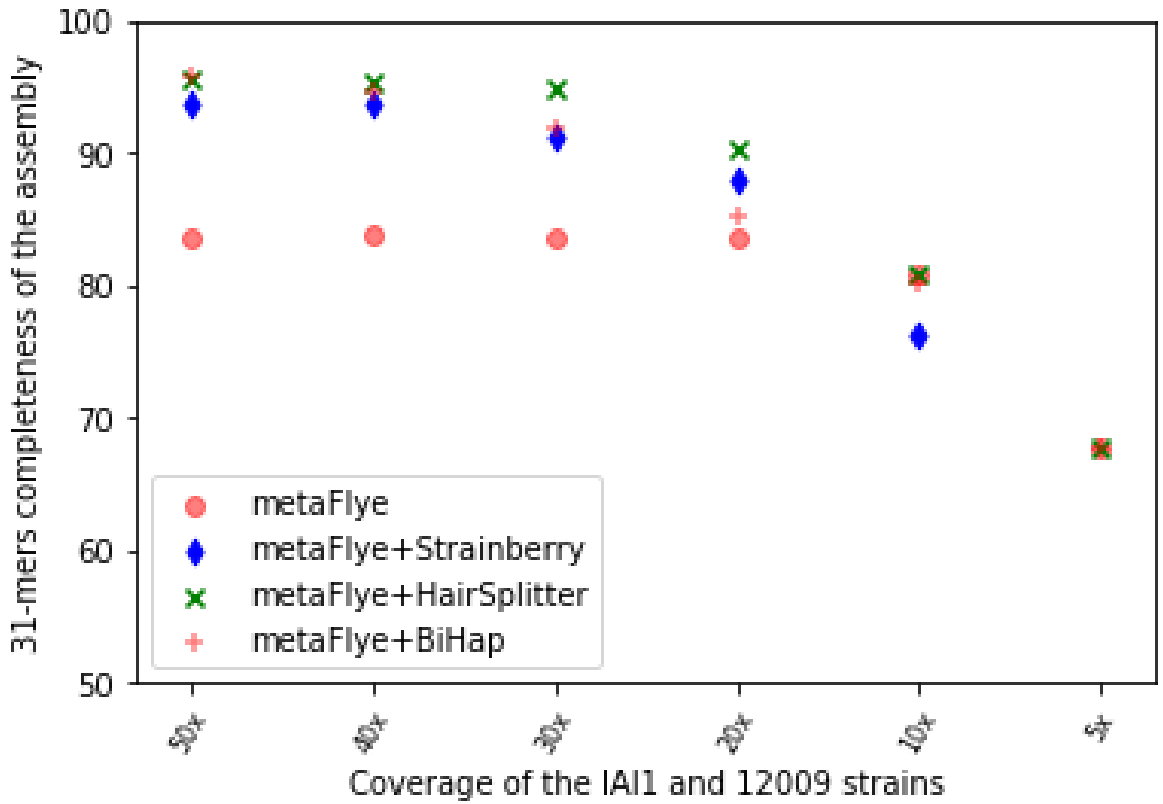
- HairSplitter reconstruct collapsed sequences from "draft" assemblies obtained by any means
- HairSplitter also works well on complex metagenomes and with high-error rate reads
- Available prototype ! github.com/RolandFaure/HairSplitter

Acknowledgements

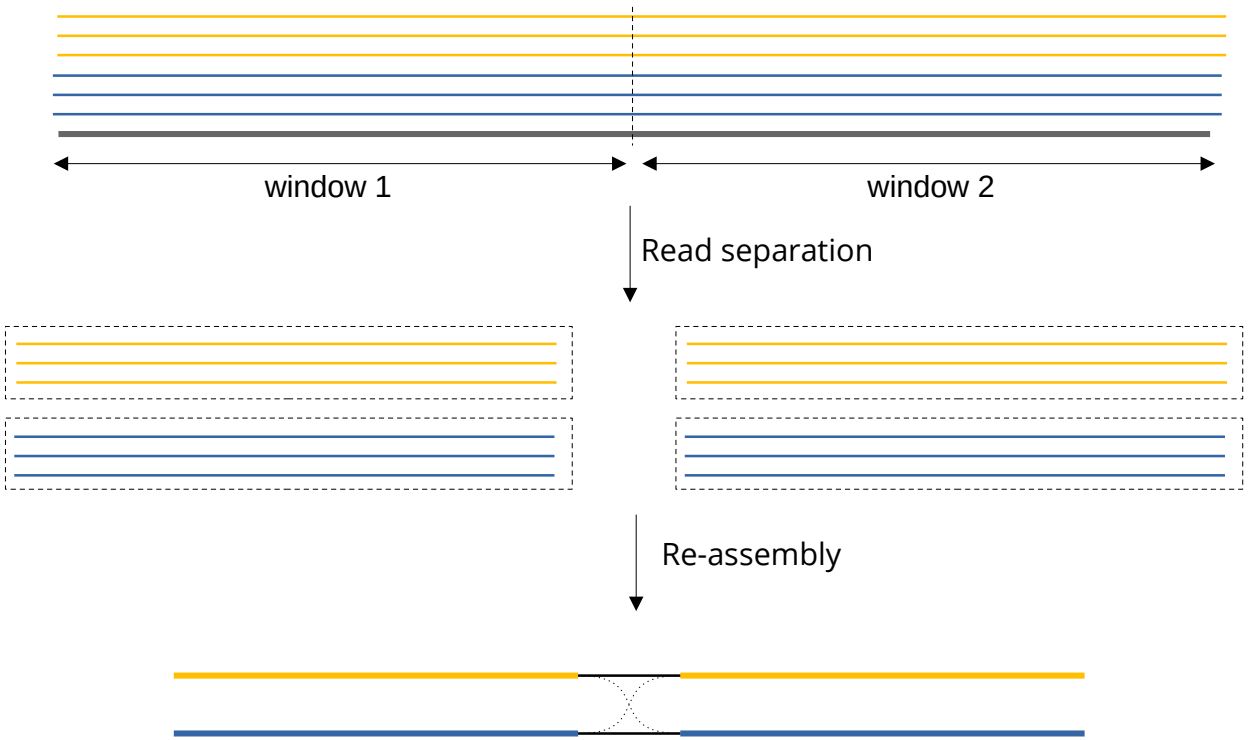
- Dominique Lavenier and Jean-François Flot for their supervision
- Rumen Andonov and Tam Truong for their help in formalizing the problem
- The EEB-EBE and GenScale teams

cnrs GDR Groupement de recherche
BIM Bioinformatique moléculaire

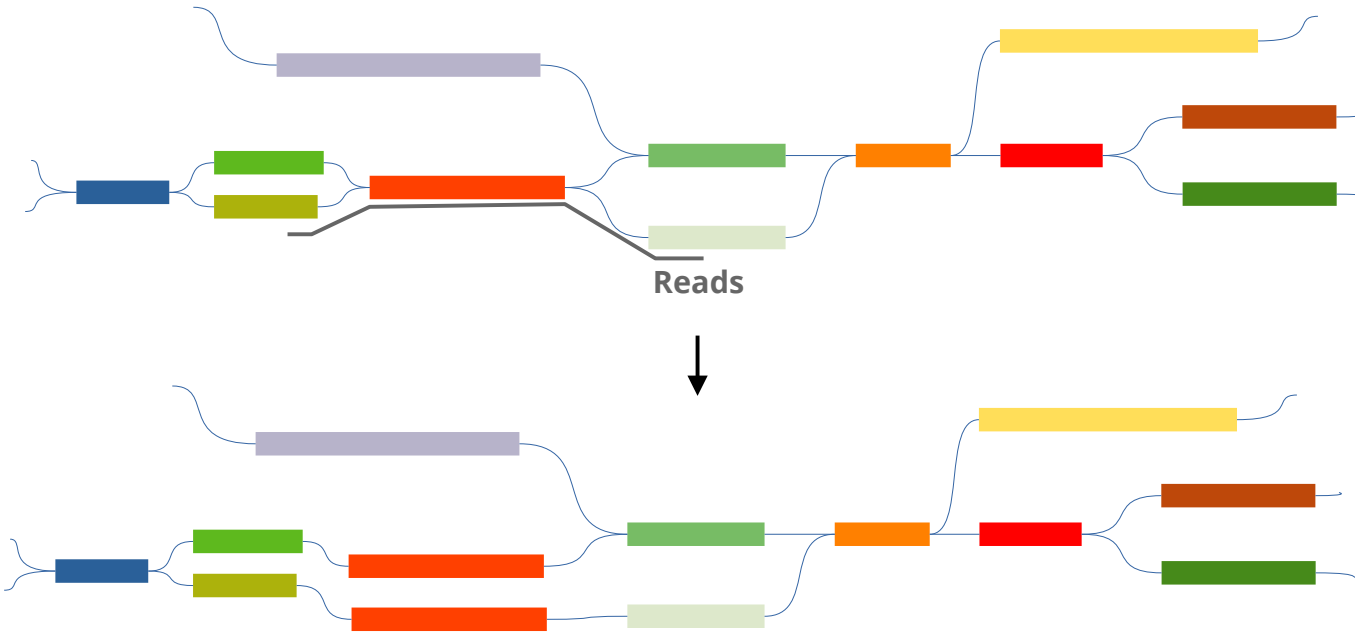
Behaviour of HairSplitter/BiHap: coverage



Local strain separation



Local strain separation



metaQuast evaluation

Genome statistics	flye	strainberry	hairsplitter
Genome fraction (%)	89.353	58.457	97.891
Duplication ratio	1.228	1.429	2.656
Largest alignment	330 446	312 915	272 305
Total aligned length	4 053 565	5 347 608	10 043 014
NGA50	...	...	...
LGA50	...	...	...
Misassemblies			
# misassemblies	50	51	141
Misassembled contigs length	2 412 308	3 237 621	5 265 749
Mismatches			
# mismatches per 100 kbp	1914.71	766.25	1887.03
# indels per 100 kbp	1488.89	1170.19	1274.4
# N's per 100 kbp	0	135.18	0
Statistics without reference			
# contigs	240	295	399
Largest contig	845 747	551 381	611 375
Total length	4 573 883	5 693 921	10 384 735
Total length (>= 1000 bp)	4 552 674	5 669 228	10 369 591
Total length (>= 10000 bp)	3 977 976	4 941 261	9 534 633
Total length (>= 50000 bp)	3 016 695	3 786 882	7 396 612

Extended report

metaQuast evaluation of *Vagococcus fluvialis* assemblies

Genome statistics	flye	strainberry	hairsplitter
Genome fraction (%)	80.324	95.595	95.321
Duplication ratio	1.133	1.976	2.002
Largest alignment	233 623	236 712	254 460
Total aligned length	7 186 249	16 406 302	17 127 330
NGA50	...	...	...
E_coli_ref	-	26 028	12 000
LGA50	...	...	...
Misassemblies			
# misassemblies	66	74	99
Misassembled contigs length	4 083 597	4 961 648	4 365 176
Mismatches			
# mismatches per 100 kbp	1771.49	1684.86	1474.75
# indels per 100 kbp	163.72	114.67	88.75
# N's per 100 kbp	0	40.1	0
Statistics without reference			
# contigs	1347	1814	7265
Largest contig	4 733 384	4 799 087	2 026 315
Total length	60 410 911	69 375 520	71 918 465
Total length (>= 1000 bp)	60 402 812	69 334 461	71 856 894
Total length (>= 10000 bp)	58 032 898	66 014 854	52 120 366
Total length (>= 50000 bp)	39 735 723	44 207 867	31 223 699

Extended report

metaQuast evaluation of Zymobiomics gut microbiome standard Nanopore Q20 assemblies