# Haplotype assembly from long reads

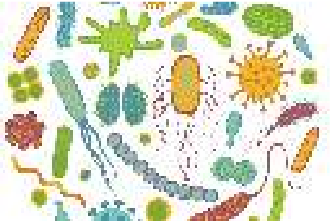
**Roland Faure**[1,2]

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

[2]Université de Rennes, IRISA - France

Public Ph.D. defence - November the 27th, 2024

# Microbiota



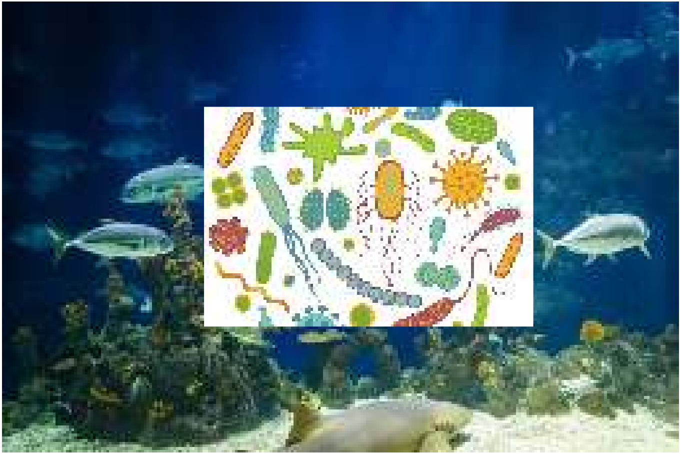
credits: microbiozindia.com

- Microbiota are mixes of bacteria, virus, archea and eukaryota

## Microbiota are everywhere

# Microbiota are everywhere

# Microbiota are everywhere



# Microbiota are everywhere

# Microbiota are important

The gut microbiome and mental health: advances in research and emerging priorities - Shoubridge et al.

# Microbiota are important

The gut microbiome and mental health: advances in research and emerging priorities - Shoubridge et al.

The Role of Soil Microorganisms in Plant Mineral Nutrition—Current Knowledge and Future Directions - Jacoby et al.

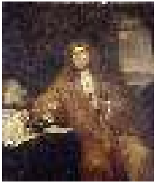
# Microbiota are important

The gut microbiome and mental health: advances in research and emerging priorities - Shoubridge et al.

The Role of Soil Microorganisms in Plant Mineral Nutrition—Current Knowledge and Future Directions - Jacoby et al.

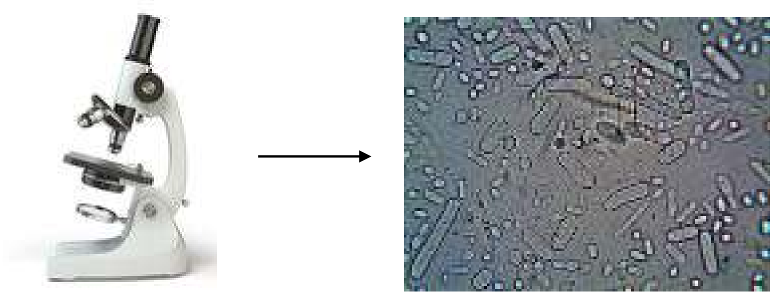
Core microbiota drive functional stability of soil microbiome in reforestation ecosystems - Jiao et al.

# How to study microbiota?

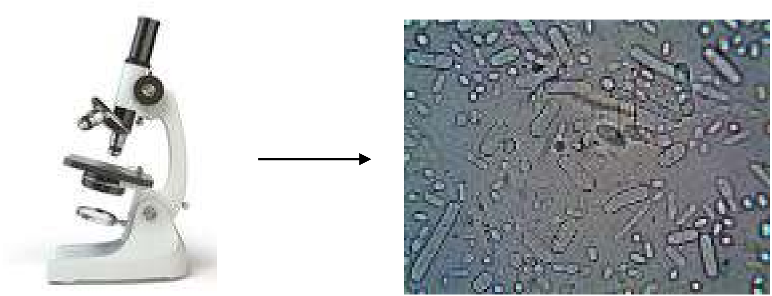


Anthonie van Leeuwenhoek
1673

# How to study microbiota?

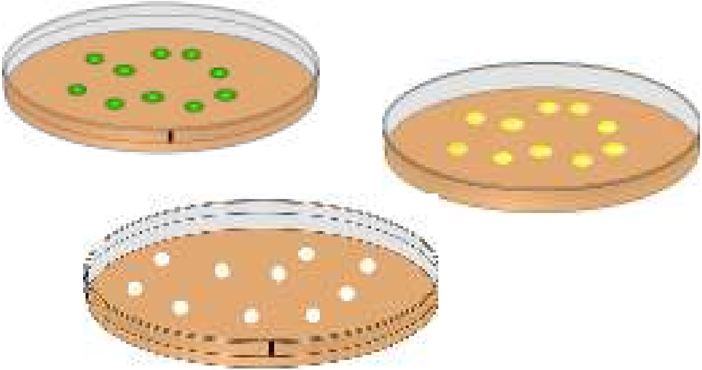


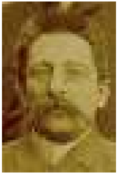
## How to study microbiota?



- But many microorganisms are undistiguishable under a microscope

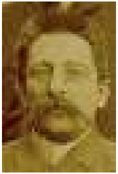
# How to study microbiota?



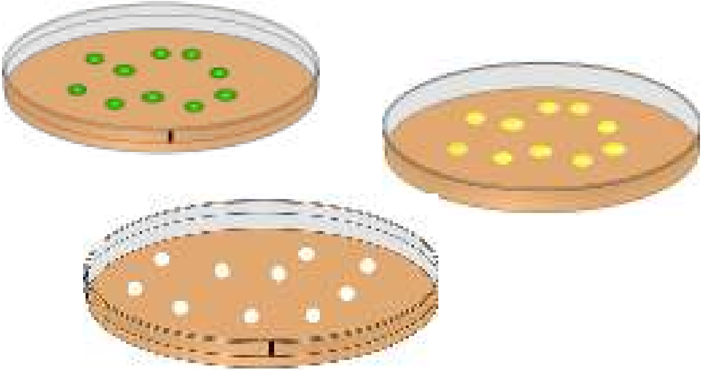


Julius Petri
1887

# How to study microbiota?



Julius Petri
1887



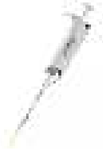
- But most microorganisms are not cultivable

# How to study microbiota?

Frederick Sanger
1975

Ecosystem

DNA extraction



Sequencing

CGTAGCTAGGAT
GTGCTAATCACGT
AAAGCTAATCACTT
TGCGAGCGATCAG
TGTCTGAAACCACA
CTCTGGGGTGACA

Long reads

Assembly

...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...

Genomes

Understand the microbiome

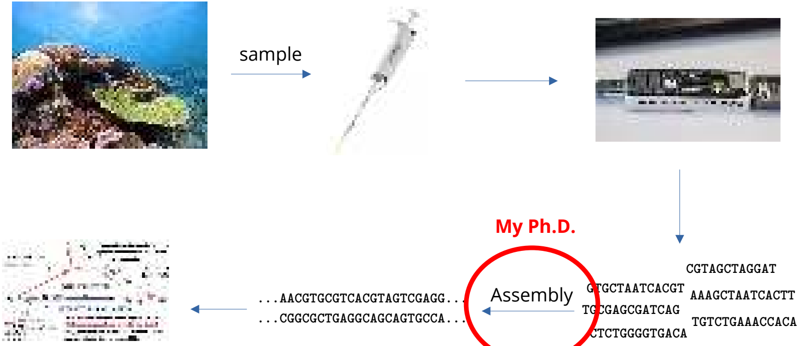
# How to study microbiota?

Frederick Sanger
1975

Microbiome

DNA extraction
& preparation

Sequencing



sample

My Ph.D.

Assembly

CGTAGCTAGGAT
GTGCTAATCACGT
AAAGCTAATCACTT
TGCGAGCGATCAG
TGTCTGAAACCACA
CTCTGGGGTGACA

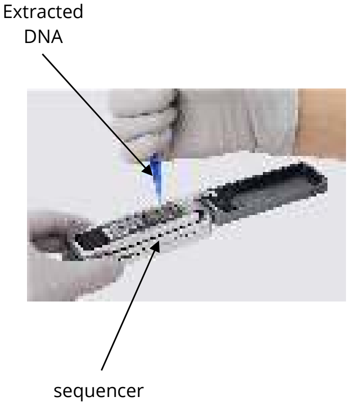
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...

Understand the microbiome

Genomes

Long reads

# DNA sequencing



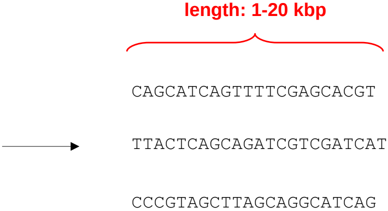
CAGCATCAGTTTTCGAGCACGT

TTACTCAGCAGATCGTCGATCAT

CCCGTAGCTTAGCAGGCATCAG

**Reads**

# DNA sequencing: difficulties



# DNA sequencing: difficulties

CAGC**A**TCAGTTTTCGAGCACGT

TTACTCAGCAGATCG**T**CGATCAT

CCCGTAGC**TT**AGCAGGCATCAG

**sequencing errors: 0.1 – 10 %**

# Genome sequencing

**Mysterious genome**

Sequencing →

```
                              TTCGGCGCTGAGGCAG
               CAGCGCTGAGGCAGCAGTGCCA   GGCAGCAGTGCCAG
CAGCATTGCCAGGC     TCGGC
                                 CGGCGCTGAGGCAGCATTGCC
```

**Mysterious picture**

Sequencing →

# Genome assembly

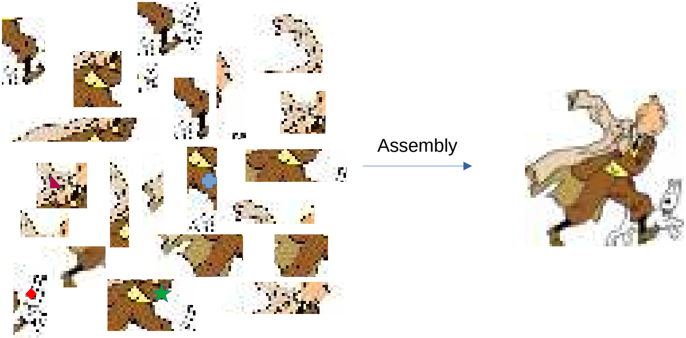
TTCGGCGCTGAGGCAG

CAGCGCTGAGGCAGCAGTGCCA GGCAGCAGTGCCAG

CAGCATTGCCAGGC TCGGC CGGCGCTGAGGCAGCATTGCC

Assembly →

...CGCTGAGGCAGCATGTGCCAGGCT...



# Genome assembly

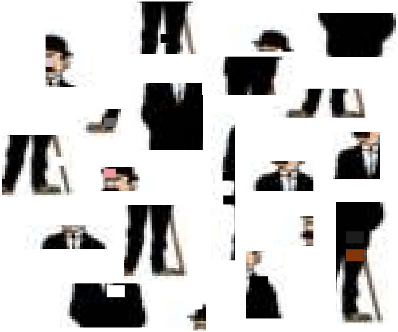
```
CGATGCTGGCTAGCATAGTCGATTTATCT
     CTGGCTAGCTTAGTCGATTTATCTGACAGT
           AGCATAGTCGATTTATCTGACAGTCATAT
                AGTCGATTTATATGACAGTCATATTGCT
                      TTTATCTGACAGTCAGATTGCTACACAC
```

genome assembly: stitching reads
correcting errors

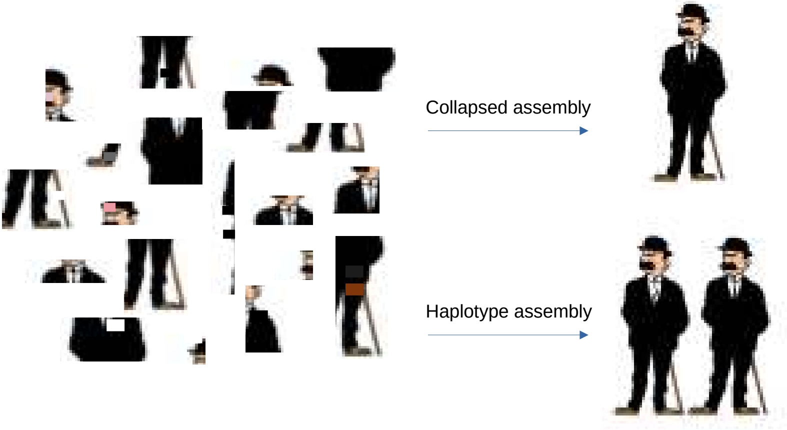
**CGATGCTGGCTAGCATAGTCGATTTATCTGACAGTCATATTGCTACACAC**

- Many software: Flye, wtdbg2, metaMDBG, hifiasm...

# Imagine you are an assembler



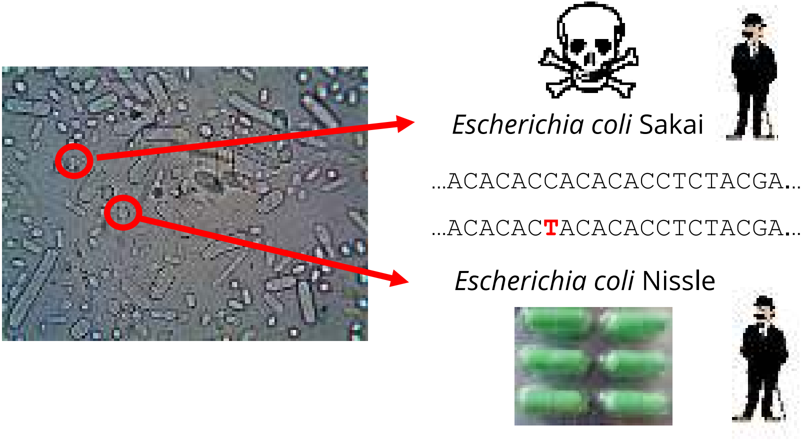
# *Haplotype* assembly



Collapsed assembly

Haplotype assembly

# Dupont & Dupond exist in microbiota!



*Escherichia coli* Sakai

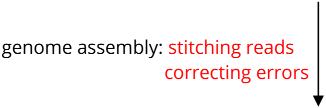
...ACACACCACACACCTCTACGA...

...ACACACTACACACCTCTACGA...

*Escherichia coli* Nissle

# Problem: assembling several haplotypes

```
CGATGCTGGCTAGCATAGTCGATTTATCT
     CTGGCTAGCTTAGTCGATTTATCTGACAGT
           AGCATAGTCGATTTATCTGACAGTCATAT
                AGTCGATTTATATGACAGTCATATTGCT
                      TTTATATGACAGTCAGATTGCTACACAC
```

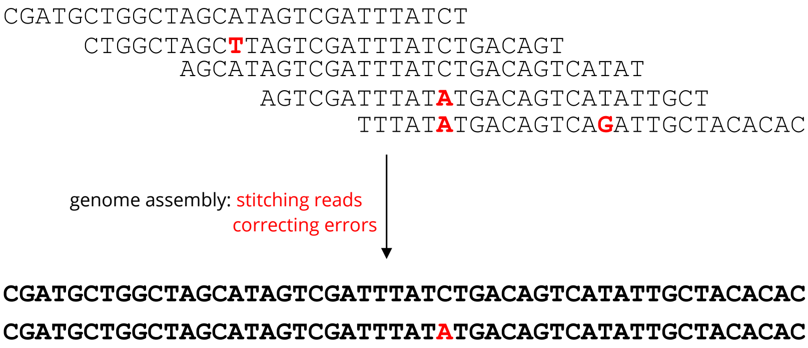


genome assembly: stitching reads
correcting errors

**CGATGCTGGCTAGCATAGTCGATTTATCTGACAGTCATATTGCTACACAC**

**CGATGCTGGCTAGCATAGTCGATTTATATGACAGTCATATTGCTACACAC**

# Problem: assembling several haplotypes



CGATGCTGGCTAGCATAGTCGATTTATCT

CTGGCTAGC**T**TAGTCGATTTATCTGACAGT

AGCATAGTCGATTTATCTGACAGTCATAT

AGTCGATTTAT**A**TGACAGTCATATTGCT

TTTAT**A**TGACAGTCA**G**ATTGCTACACAC

genome assembly: stitching reads
correcting errors

**CGATGCTGGCTAGCATAGTCGATTTATCTGACAGTCATATTGCTACACAC**

**CGATGCTGGCTAGCATAGTCGATTTATATGACAGTCATATTGCTACACAC**

- Not so many software!

# Haplotype assembly from long reads

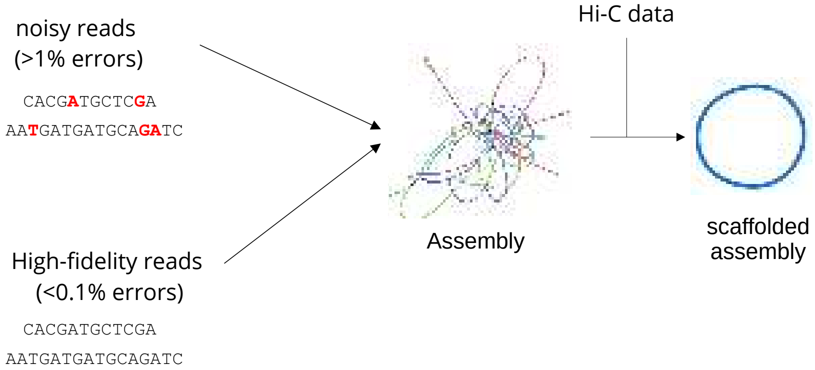
**Roland Faure**[1,2]

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

[2] Université de Rennes, IRISA - France

Public Ph.D. defence - November the 27th, 2024

# Overview of (meta)genome assembly



noisy reads (>1% errors)

CACG**A**TGCTC**G**A

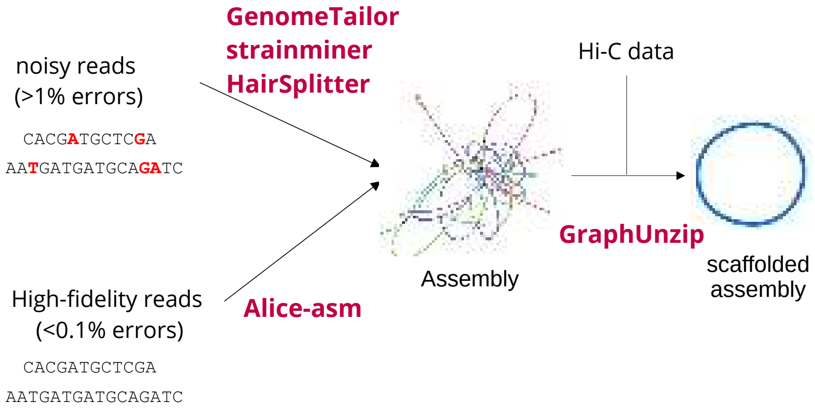
AA**T**GATGATGCA**GA**TC

High-fidelity reads (<0.1% errors)

CACGATGCTCGA

AATGATGATGCAGATC

# Overview of the Ph.D.



noisy reads
(>1% errors)

CACG**A**TGCTC**G**A

AA**T**GATGATGCA**GA**TC

**GenomeTailor**
**strainminer**
**HairSplitter**

High-fidelity reads
(<0.1% errors)

CACGATGCTCGA

AATGATGATGCAGATC

**Alice-asm**

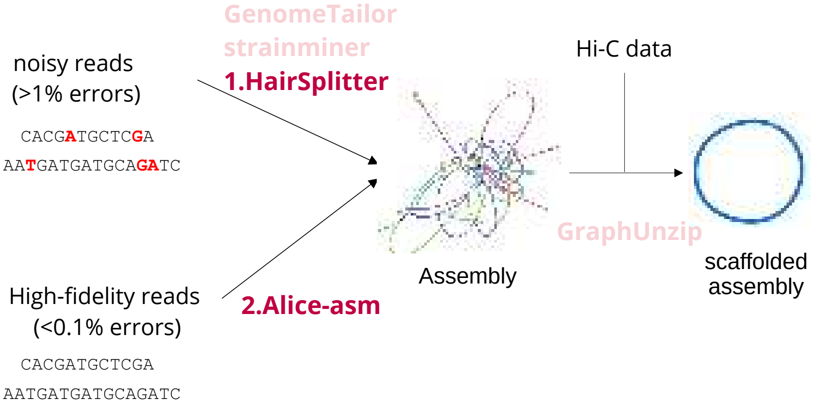
Assembly

Hi-C data

**GraphUnzip**

scaffolded
assembly

# Overview of the Ph.D.



GenomeTailor

strainminer

**1.HairSplitter**

**2.Alice-asm**

GraphUnzip

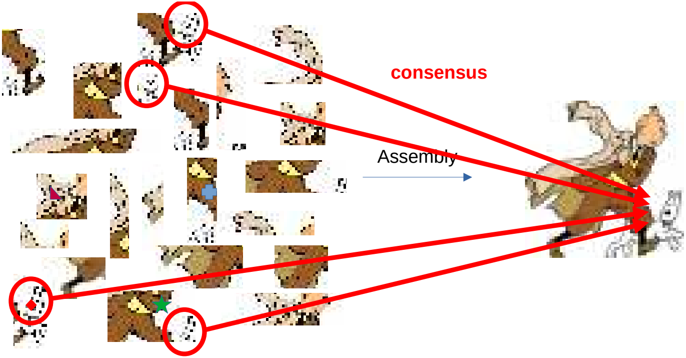
# Distiguishing haplotypes with noisy reads - HairSplitter

# Assembling noisy reads: correcting errors by consensus

errors

```
CAGCGCTGAGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
```

# Assembling noisy reads: correcting errors by consensus

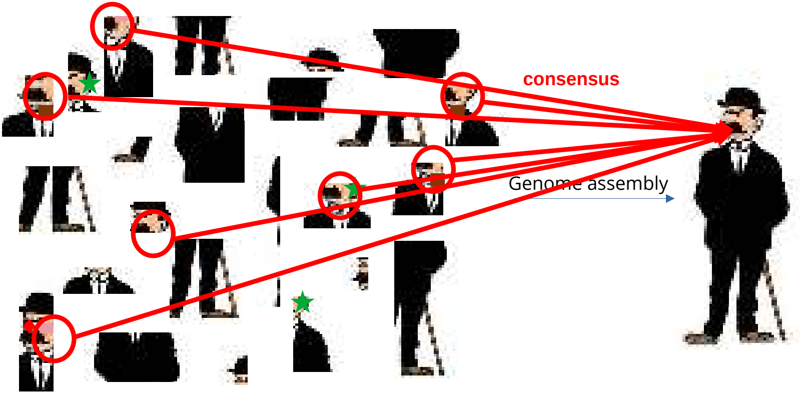


```
CAGCGCTGAGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
```

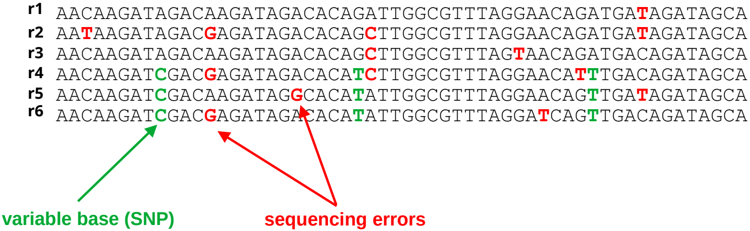
Assembly →

```
CAGCGCTGTGGCAGCAGTGCCA
```

# Consensus loses the variants

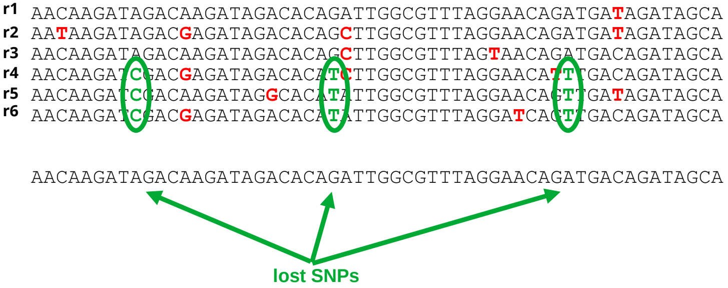


# Consensus loses the variants



```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

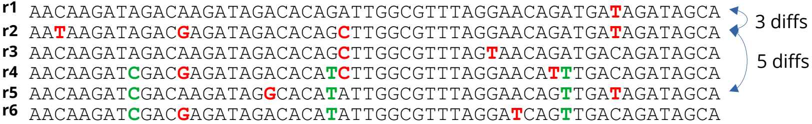
# Consensus loses the variants



```
r1  AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2  AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3  AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4  AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5  AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6  AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA

    AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGACAGATAGCA
```

lost SNPs

# Haplotype separation: state of the art [1]

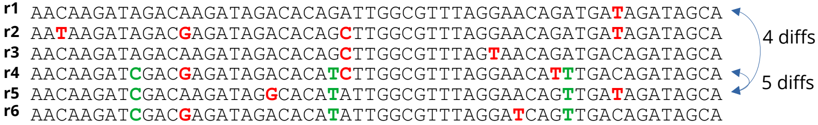


```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

**Reads from the same haplotype are more similar than reads from different haplotypes**

[1] WhatsHap, HapCut, Strainberry, stRainy...

# Haplotype separation: state of the art [1]



```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

4 diffs

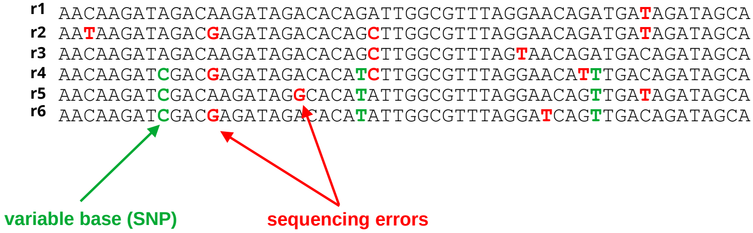
5 diffs

**Reads from the same haplotype are more similar than reads from different haplotypes**

**on average**

[1] WhatsHap, HapCut, Strainberry, stRainy...

# How to distinguish errors and SNPs?

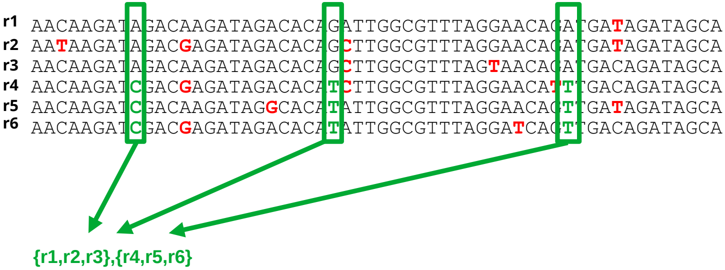


```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

**variable base (SNP)**

**sequencing errors**

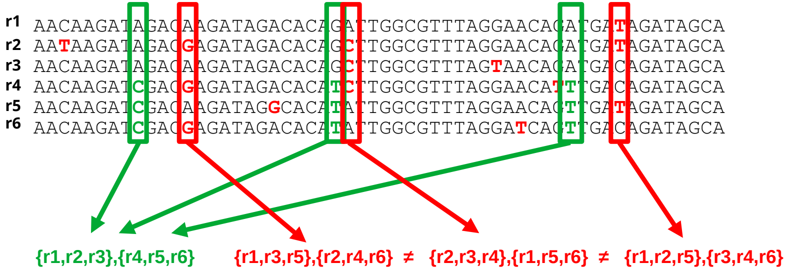
# My solution: looking at several positions simultaneously



r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA

{r1,r2,r3},{r4,r5,r6}

# My solution: looking at several positions simultaneously

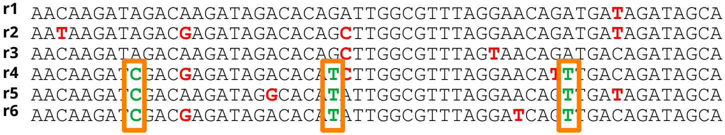


```
r1  AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2  AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3  AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4  AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5  AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6  AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

{r1,r2,r3},{r4,r5,r6}

{r1,r3,r5},{r2,r4,r6} ≠ {r2,r3,r4},{r1,r5,r6} ≠ {r1,r2,r5},{r3,r4,r6}

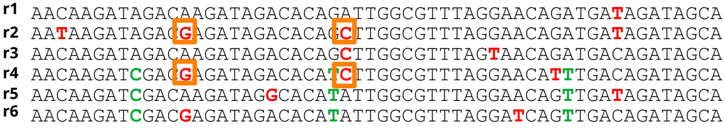
# Algorithm: 1) looking for variant patterns



```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

**variant pattern: subset of reads and positions containing minority bases**
**size: 3x3**

# Algorithm: 1) looking for variant patterns

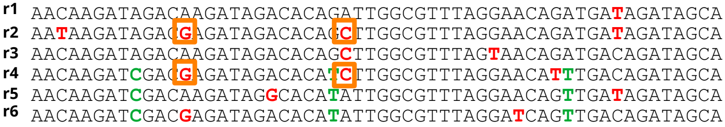


```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

**variant pattern: subset of reads and positions containing minority bases**

**size: 2x2**

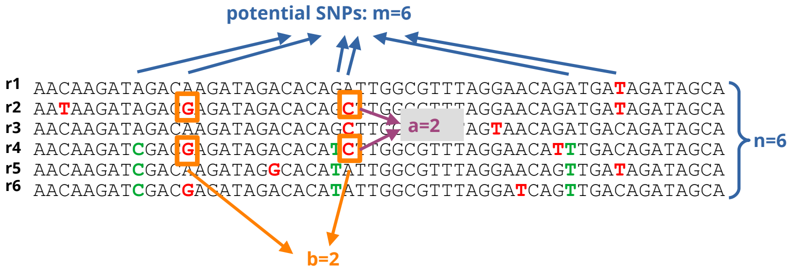
# Algorithm: 2) Statistical test



```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

Is this pattern too big to be due to errors ?

# Algorithm: 2) Statistical test



potential SNPs: m=6

```
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
```

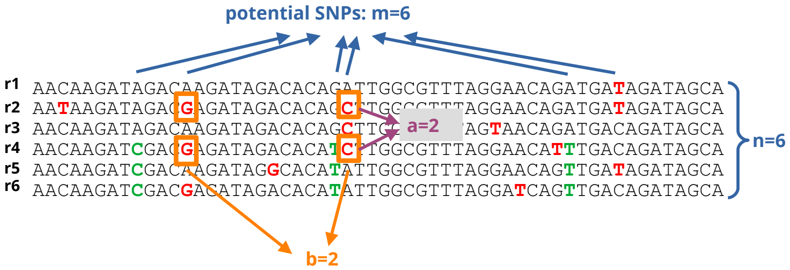
a=2

n=6

b=2

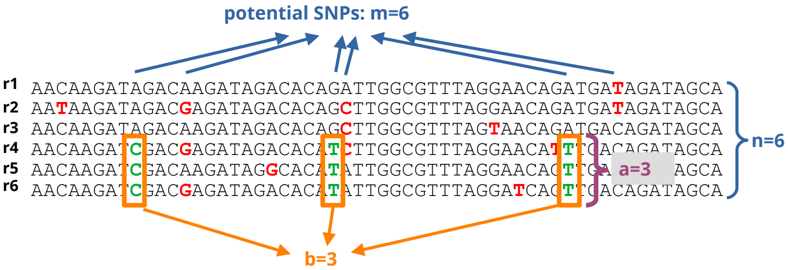
Is this pattern too big to be due to errors ?

# Algorithm: 2) Statistical test


potential SNPs: m=6
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGC
a=2
AGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGACATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
n=6
b=2

P(errors produce pattern of size **ab**) $\leq \binom{n}{a}\binom{m}{b} * \frac{a^{ab}}{n^{ab}} = 0.30$

# Algorithm: 2) Statistical test



potential SNPs: m=6

r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGA a=3 AGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA

n=6

b=3

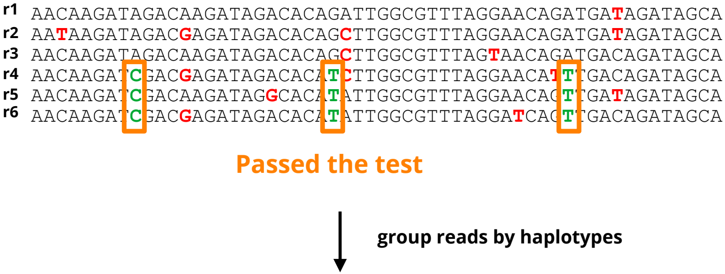
P(errors produce pattern of size **ab**) $\leq \binom{n}{a}\binom{m}{b} * \frac{a^{ab}}{n^{ab}} = 0.07$

## Statistical test: main result

$$\binom{n}{a}\binom{m}{b} * \frac{a^{ab}}{n^{ab}}$$

- No assumption on the number of haplotypes
- No assumption on balanced coverage
- No assumption on the error pattern of the reads
- Assumption: errors are independent

# Algorithm: 3) Group reads by haplotype



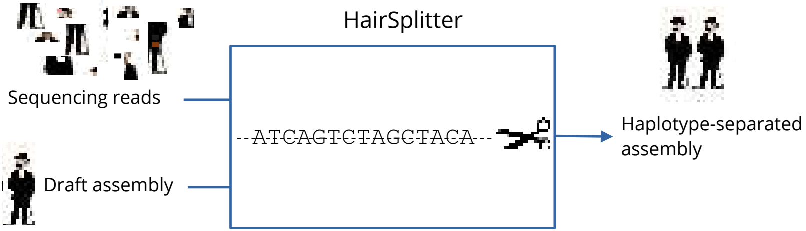
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AATAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA

**Passed the test**

**group reads by haplotypes**

**{r1,r2,r3}** **{r4,r5,r6}**

# The HairSplitter program



Sequencing reads

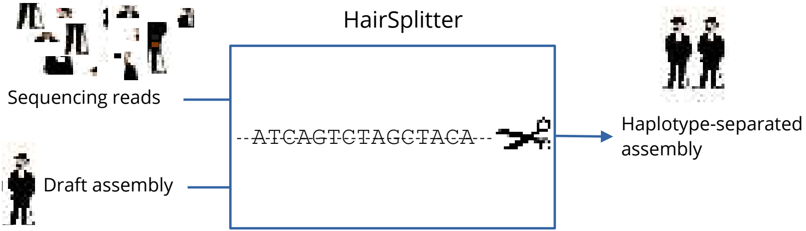
Draft assembly

HairSplitter

--~~ATCAGTCTAGCTACA~~--

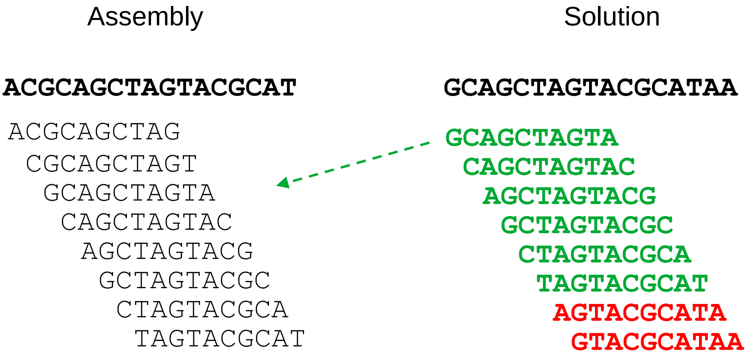
Haplotype-separated assembly

# The HairSplitter program



- *Hairsplitter:* A person who makes extremely, possibly excessively, fine distinctions (who would separate something as fine as a hair into two pieces and distinguish them) - *Wiktionary*

# Let's evaluate HairSplitter - k-mer completeness



Assembly

**ACGCAGCTAGTACGCAT**

ACGCAGCTAG
CGCAGCTAGT
GCAGCTAGTA
CAGCTAGTAC
AGCTAGTACG
GCTAGTACGC
CTAGTACGCA
TAGTACGCAT

Solution

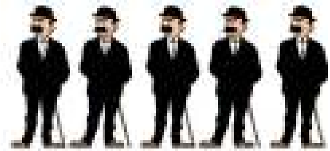
**GCAGCTAGTACGCATAA**

**GCAGCTAGTA**
**CAGCTAGTAC**
**AGCTAGTACG**
**GCTAGTACGC**
**CTAGTACGCA**
**TAGTACGCAT**
**AGTACGCATA**
**GTACGCATAA**

10-mer completeness:
6 out of 8 (75%)

# Evaluating HairSplitter - results

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



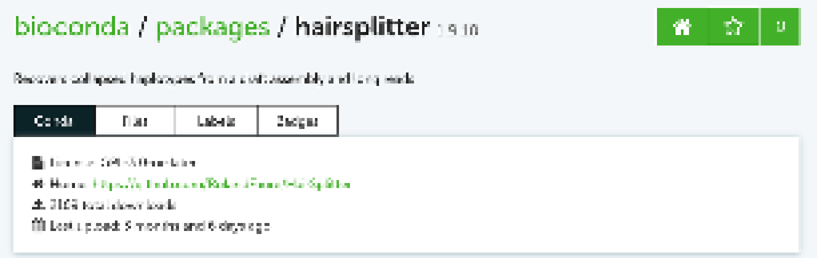
| | metaFlye | metaFlye+Strainberry | metaFlye+HairSplitter |
|---|---|---|---|
| Nanopore Q9 | 0.586 | 0.749 | **0.957** |
| Nanopore Q20 | 0.7524 | 0.9527 | **0.961** |
| PacBio HiFi | 0.9589 | 0.9793 | **0.9895** |

Table: 31-mer completeness of assemblies compared to the solution

- Improves over the state of the art on complex assemblies

# The HairSplitter project

- Presented in JOBIM, SeqBIM, ISMB/ECCB
- Published in *Peer Community Journal*



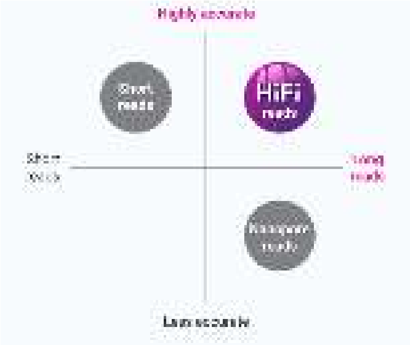
bioconda / packages / hairsplitter [illegible]

Recovers collapsed haplotypes from a contig assembly and long reads

License: GPL-3.0-or-later
Home: https://github.com/RolandFaure/HairSplitter
[illegible] total downloads
Last upload: 5 months and 6 days ago

# Distinguishing haplotypes with high-fidelity reads - Alice

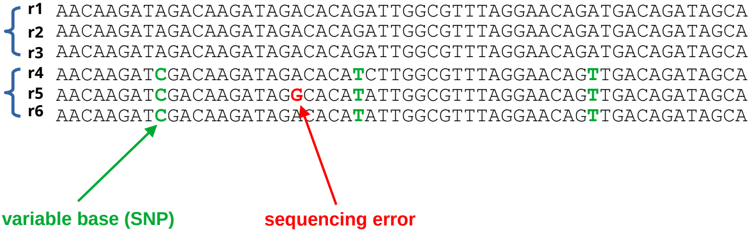
# New technology: high-fidelity long reads



pacb.com

- Emerged recently and are still emerging
- $<<$ 1% sequencing errors

# Assembly with high-fidelity long reads: easy!

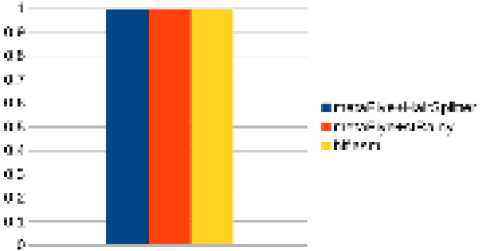


r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGACAGATAGCA
r2 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGACAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGACAGATAGCA
r4 AACAAGATCGACAAGATAGACACATCTTGGCGTTTAGGAACAGTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGACAGATAGCA
r6 AACAAGATCGACAAGATAGACACATATTGGCGTTTAGGAACAGTTGACAGATAGCA

variable base (SNP)

sequencing error

# Assembly with high-fidelity long reads: easy!



27-mer completeness of the assemblies of the Zymobiomics Gut Microbiome Standard

# Assembly with high-fidelity long reads: slow!

Table: CPU time

| | hifiasm | metaFlye+HairSplitter |
|---|---|---|
| Zymobiomics Gut Microbiome Standard | 20 days | 4 days |

# Assembly with high-fidelity long reads: slow!

Table: CPU time

| | hifiasm | metaFlye+HairSplitter |
|---|---|---|
| Zymobiomics Gut Microbiome Standard | 20 days | 4 days |
| human genome | 34 days | 25 days |

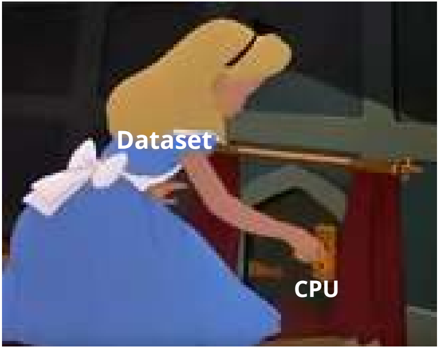
# Assembly with high-fidelity long reads: slow!

Table: CPU time

| | hifiasm | metaFlye+HairSplitter |
|---|---|---|
| Zymobiomics Gut Microbiome Standard | 20 days | 4 days |
| human genome | 34 days | 25 days |
| human gut microbiome[1] | $\geq$ 60 days | $\geq$ 60 days |

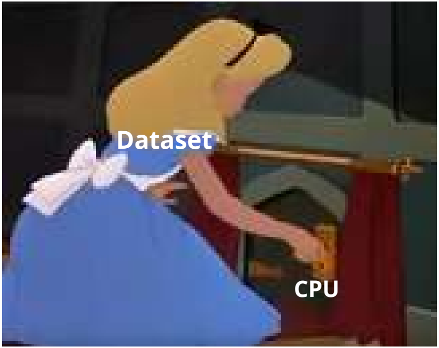
[1] Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods - BiorXiv

# How to perform fast assembly?

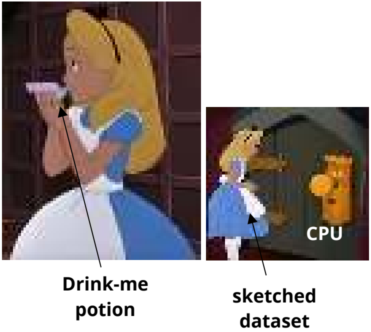


Credits: Alice in Wonderland, Lewis, Disney

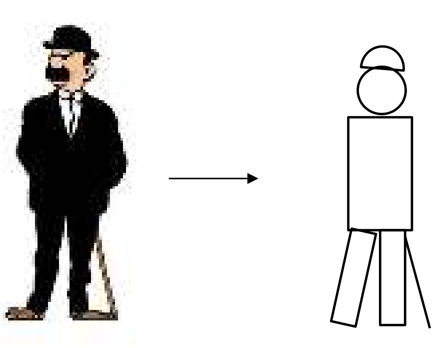
# Solution for fast assembly: sketching the reads



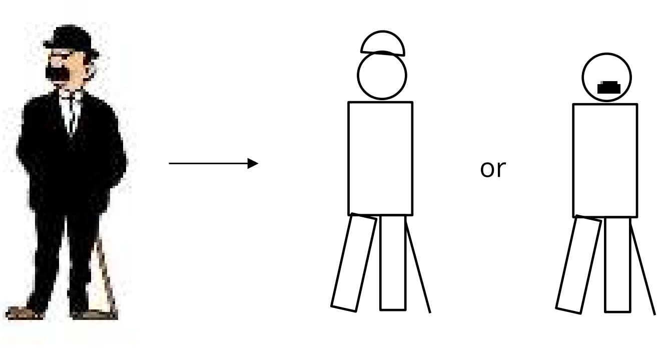
Credits: Alice in Wonderland, Lewis, Disney



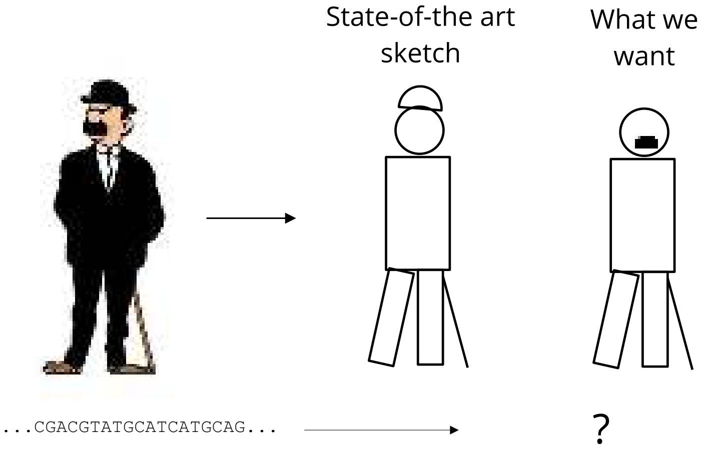
# Sketching: reducing the size of the data



# Sketching: reducing the size of the data



## Sketching: reducing the size of the data



State-of-the art sketch

What we want

...CGACGTATGCATCATGCAG...

?

# My contribution: MSR sketching

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

# MSR sketching

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

# MSR sketching

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\varnothing\}$$

$$f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]$$
$$f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]$$
$$f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]$$
$$f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]$$
$$f(10-mer)\rightarrow \varnothing \quad if \quad hash(10-mer)>0.2$$

sequence CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

# MSR sketching

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence **CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**CAGTATGGAT**)= 0.0023

f(**CAGTATGGAT**)= A

sketch A

# MSR sketching

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,Ø\}$$

$$\begin{aligned}
f(10-mer)&\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]\\
f(10-mer)&\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]\\
f(10-mer)&\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]\\
f(10-mer)&\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]\\
f(10-mer)&\rightarrow Ø \quad if \quad hash(10-mer)>0.2
\end{aligned}$$

sequence C**AGTATGGATA**CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**AGTATGGATA**)= 0.624

f(**AGTATGGATA**)= A

sketch A

# MSR sketching

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**GTATGGATAC**)= 0.124

f(**GTATGGATAC**)= G

sketch A G

# MSR sketching

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence CAG**TATGGATACA**GATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TATGGATACA**)= 0.88

f(**TATGGATACA**)= Ø

sketch A G

# MSR sketching

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,Ø\}$$

$$\begin{aligned}
&f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]\\
&f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]\\
&f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]\\
&f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]\\
&f(10-mer)\rightarrow Ø \quad if \quad hash(10-mer)>0.2
\end{aligned}$$

sequence `CAGT`**ATGGATACAG**`ATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

hash(**ATGGATACAG**)= 0.32

f(**ATGGATACAG**)= Ø

sketch `A G`

# MSR sketching

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \text{Ø}\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \text{Ø} \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence CAGTA**TGGATACAGA**TGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**TGGATACAGA**)= 0.19

f(**TGGATACAGA**)= T

sketch A G T

# MSR sketching

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, Ø\}$$

$$\begin{aligned}
f(10-mer) &\rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
f(10-mer) &\rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
f(10-mer) &\rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
f(10-mer) &\rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
f(10-mer) &\rightarrow Ø \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence `CAGTAT`**`GGATACAGAT`**`GGAGATATCATCGAGTAGGGGCACTGTACCAGAG`

hash(**GGATACAGAT**)= 0.214

f(**GGATACAGAT**)= Ø

sketch `A G  T`

# MSR sketching

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\varnothing\}$$

$$\begin{aligned}
&f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]\\
&f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]\\
&f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]\\
&f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]\\
&f(10-mer)\rightarrow \varnothing \quad if \quad hash(10-mer)>0.2
\end{aligned}$$

sequence CAGTATG**GATACAGATG**GAGATATCATCGAGTAGGGCACTGTACCAGAG

hash(**GATACAGATG**)= 0.678

f(**GATACAGATG**)= Ø

sketch A G T

# MSR sketching

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence CAGTATGG**ATACAGATGG**AGATATCATCGAGTAGGGGCACTGTACCAGAG

hash(**ATACAGATGG**)= 0.669

f(**ATACAGATGG**)= Ø

sketch A G T

# MSR sketching

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence

```
CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
```

hash(**TGTACCAGAG**)= 0.06

f(**TGTACCAGAG**)= C

sketch

```
    A G   T         T C       C    G    T    C
```

# MSR sketching

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\varnothing\}$$

order (l)

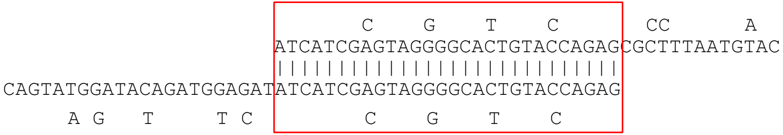
$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05] \\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1] \\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15] \\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2] \\
&f(10-mer) \rightarrow \varnothing \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

compression ratio (c)

```
sequence   CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch         A G  T         T C       C     G   T     C
```

# MSR=**Mapping-friendly** Sequence Reductions

- If two reads align, their sketchs align too

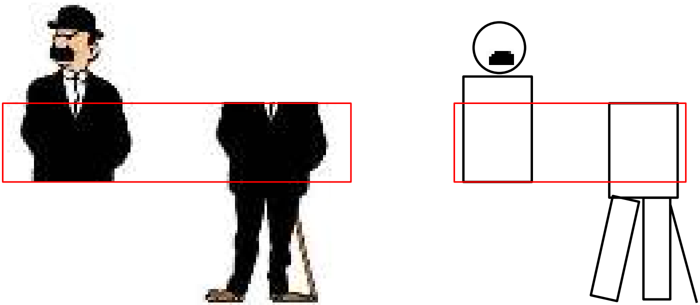


```
                             C    G    T    C       CC      A
                      ATCATCGAGTAGGGGCACTGTACCAGAGCGCTTTAATGTAC
                      ||||||||||||||||||||||||||||
CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
     A G   T     T C         C    G    T    C
```

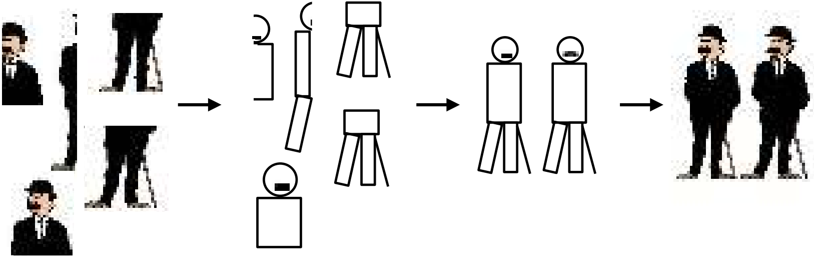
# MSR=**Mapping-friendly** Sequence Reductions

- If two reads align, their sketchs align too

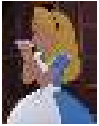
```
                             C    G    T    C       CC      A
                      ATCATCGAGTAGGGGCACTGTACCAGAGCGCTTTAATGTAC
                      ||||||||||||||||||||||||||||
CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
     A G   T     T C         C    G    T    C
```

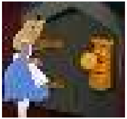


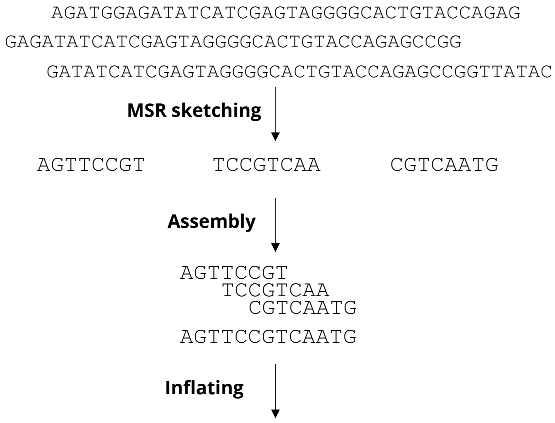
## Assembling using MSR sketches



# Very fast assembly: the Alice assembler







```
    AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG
GAGATATCATCGAGTAGGGGCACTGTACCAGAGCCGG
   GATATCATCGAGTAGGGGCACTGTACCAGAGCCGGTTATAC
```

**MSR sketching**

```
AGTTCCGT     TCCGTCAA     CGTCAATG
```

**Assembly**

```
AGTTCCGT
   TCCGTCAA
     CGTCAATG

AGTTCCGTCAATG
```

**Inflating**

```
AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAGCCGGTTATAC
```

# MSR sketching keeps SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

sequence1 CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sequence2 CAGTATGGATACAGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG

# MSR sketching keeps SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

| | |
|---|---|
| sequence1 | **CAGTATGGAT**ACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG |
| sketch1 | A |
| sequence2 | **CAGTATGGAT**ACAGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG |
| sketch2 | A |

# MSR sketching keeps SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

| | |
|---|---|
| sequence1 | C**AGTATGGATA**CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG |
| sketch1 | A |
| sequence2 | C**AGTATGGATA**CAGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG |
| sketch2 | A |

# MSR sketching keeps SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\varnothing\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]\\
&f(10-mer) \rightarrow \varnothing \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

| | |
|---|---|
| sequence1 | CA**GTATGGATAC**AGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG |
| sketch1 | A G |
| sequence2 | CA**GTATGGATAC**AGATGGAGATAT**G**ATCGAGTAGGGGCACTGTACCAGAG |
| sketch2 | A G |

# MSR sketching keeps SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$
\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}
$$

| | |
|---|---|
| sequence1 | CAGTATGGATACAG**ATGGAGATAT**CATCGAGTAGGGGCACTGTACCAGAG |
| sketch1 | A G T |
| sequence2 | CAGTATGGATACAG**ATGGAGATATG**ATCGAGTAGGGGCACTGTACCAGAG |
| sketch2 | A G T |

# MSR sketching keeps SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\varnothing\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \varnothing \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

```
sequence1    CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1          A G   T         T

sequence2    CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2          A G   T
```

# MSR sketching keeps SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\text{Ø}\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]\\
&f(10-mer) \rightarrow \text{Ø} \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

```
sequence1  CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1        A G   T         T

sequence2  CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2        A G   T          G
```

# MSR sketching keeps SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \varnothing\}$$

$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \varnothing \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

```
sequence1    CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1          A G   T        T C

sequence2    CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2          A G   T         G
```

# MSR sketching keeps SNPs

$$f : \{A, C, G, T\}^{10} \rightarrow \{A, C, G, T, \emptyset\}$$

$$\begin{aligned}
f(10-mer) &\rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
f(10-mer) &\rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
f(10-mer) &\rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
f(10-mer) &\rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
f(10-mer) &\rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$

```
sequence1   CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1         A G   T         T C

sequence2   CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2         A G   T          G
```

# MSR sketching keeps SNPs

$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

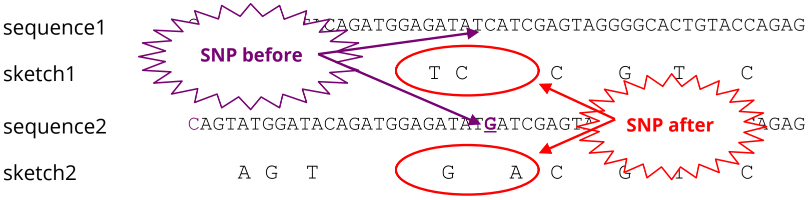
$$f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0,0.05]$$
$$f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05,0.1]$$
$$f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1,0.15]$$
$$f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15,0.2]$$
$$f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2$$

```
sequence1  CAGTATGGATACAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1        A G   T        T  C      C     G   T     C

sequence2  CAGTATGGATACAGATGGAGATATGATCGAGTAGGGGCACTGTACCAGAG

sketch2        A G   T          G    A  C     G   T     C
```

# MSR sketching keeps and amplify SNPs

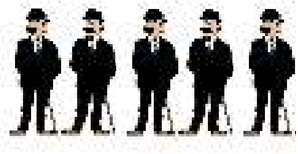
$$f:\{A,C,G,T\}^{10} \rightarrow \{A,C,G,T,\emptyset\}$$

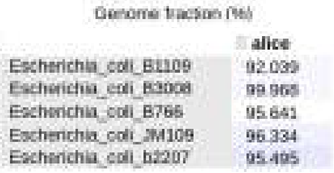
$$\begin{aligned}
&f(10-mer) \rightarrow A \quad if \quad hash(10-mer) \in [0, 0.05]\\
&f(10-mer) \rightarrow C \quad if \quad hash(10-mer) \in [0.05, 0.1]\\
&f(10-mer) \rightarrow G \quad if \quad hash(10-mer) \in [0.1, 0.15]\\
&f(10-mer) \rightarrow T \quad if \quad hash(10-mer) \in [0.15, 0.2]\\
&f(10-mer) \rightarrow \emptyset \quad if \quad hash(10-mer) > 0.2
\end{aligned}$$



# Results: Alice assemblies are complete

- Assembly of the Zymobiomic Gut Microbiome Standard containing 5 strains of *E. coli*





Genome fraction (%)

| | alice |
|---|---|
| Escherichia_coli_B1109 | 92.039 |
| Escherichia_coli_B3008 | 99.969 |
| Escherichia_coli_B766 | 95.641 |
| Escherichia_coli_JM109 | 96.334 |
| Escherichia_coli_b2207 | 95.495 |

## Results: Alice assemblies are fast

| | hifiasm | metaFlye +HairSplitter | Alice-asm |
|---|---|---|---|
| Zymobiomics Gut Microbiome Standard | 20 days | 4 days | 1h20 |
| human genome | 34 days | 25 days | 8h40 |
| human gut microbiome[1] | $\geq$ 60 days | $\geq$ 60 days | 5h00 |

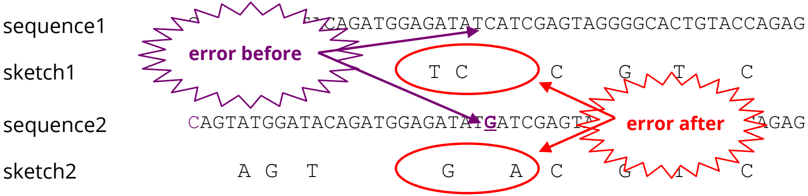
N.B. only assemblers that distinguish strains are shown

[1] Highly accurate metagenome-assembled genomes from human gut microbiota using long-read assembly, binning, and consolidation methods - BiorXiv

# The dark side: MSR sketching keeps errors

$$f:\{A,C,G,T\}^{10}\rightarrow\{A,C,G,T,\emptyset\}$$

$$\begin{aligned}
&f(10-mer)\rightarrow A \quad if \quad hash(10-mer)\in[0,0.05]\\
&f(10-mer)\rightarrow C \quad if \quad hash(10-mer)\in[0.05,0.1]\\
&f(10-mer)\rightarrow G \quad if \quad hash(10-mer)\in[0.1,0.15]\\
&f(10-mer)\rightarrow T \quad if \quad hash(10-mer)\in[0.15,0.2]\\
&f(10-mer)\rightarrow \emptyset \quad if \quad hash(10-mer)>0.2
\end{aligned}$$



sequence1 CAGATGGAGATATCATCGAGTAGGGGCACTGTACCAGAG

sketch1 T C C G T C

sequence2 CAGTATGGATACAGATGGAGATATGATCGAGTA AGAG

sketch2 A G T G A C G T C

error before

error after

# MSR sketching: conclusion & perspectives

- mapping-friendly and keeps SNPs: perfectly adapted to haplotype assembly with high-fidelity reads

# MSR sketching: conclusion & perspectives

- mapping-friendly and keeps SNPs: perfectly adapted to haplotype assembly with high-fidelity reads
- Still a lot to explore on MSR sketching: changing the function, changing the use case...

# MSR sketching: conclusion & perspectives

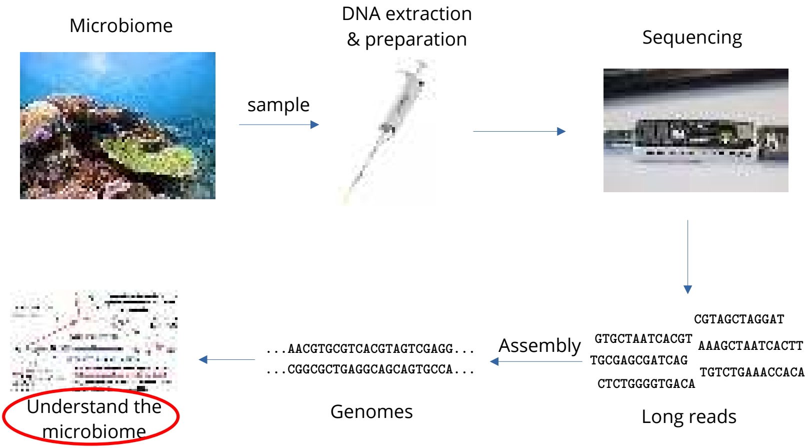
- mapping-friendly and keeps SNPs: perfectly adapted to haplotype assembly with high-fidelity reads
- Still a lot to explore on MSR sketching: changing the function, changing the use case...
- Tune to what extent we want to keep variation

# Conclusion

# Conclusion: achievements

- **Noisy reads**: assemble a mix of haplotypes of unprecedented complexity
- **High-fidelity reads**: assemble very fast while keeping haplotypes with MSR sketching
- **Hi-C data**: improved the scaffolding of haploid and multiploid assemblies

# Why is this thesis useful?



Microbiome

sample

DNA extraction
& preparation

Sequencing

CGTAGCTAGGAT
GTGCTAATCACGT
AAAGCTAATCACTT
TGCGAGCGATCAG
TGTCTGAAACCACA
CTCTGGGGTGACA

Long reads

Assembly

...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...

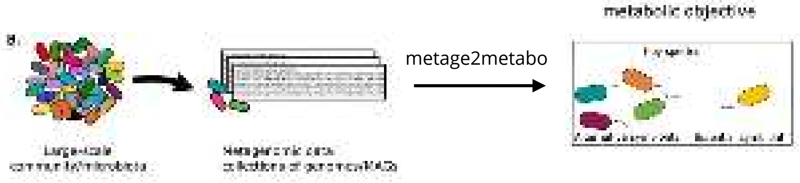
Genomes

Understand the
microbiome

## Why is this thesis useful?

# Example of an application: metage2metabo[1]

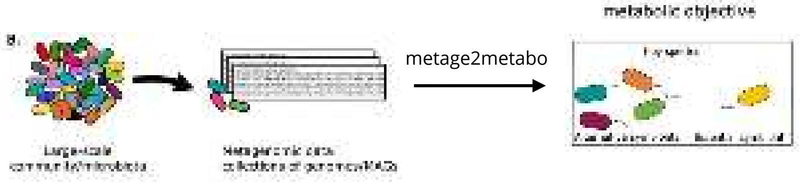
Arnaud Belcour


a.
Large-scale community/microbiota
Metagenomic data collections of genomes/MAGs
metage2metabo
metabolic objective
Key species

[1] Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species - Belcour et al., 2020

# Example of an application: metage2metabo[1]

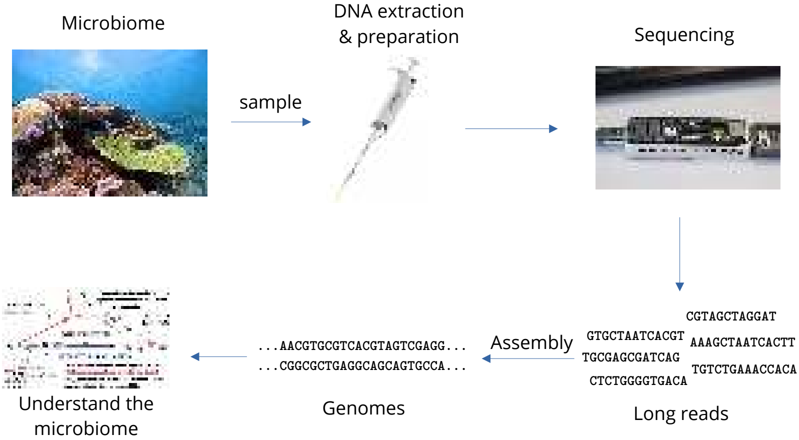
Arnaud Belcour



- Predictions for human health, soil fertility, ecology...

[1] Metage2Metabo, microbiota-scale metabolic complementarity for the identification of key species - Belcour et al., 2020

# What is the future of assembly?



Microbiome

sample

DNA extraction & preparation

Sequencing

CGTAGCTAGGAT
GTGCTAATCACGT
AAAGCTAATCACTT
TGCGAGCGATCAG
TGTCTGAAACCACA
CTCTGGGGTGACA

Long reads

Assembly

...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...

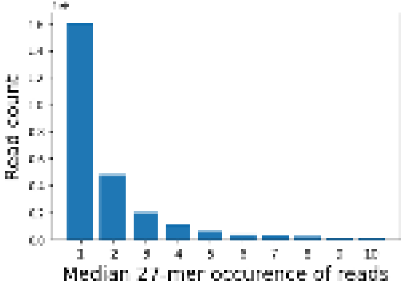
Genomes

Understand the microbiome

# All DNA is not captured by sequencing

Nicolas Maurice

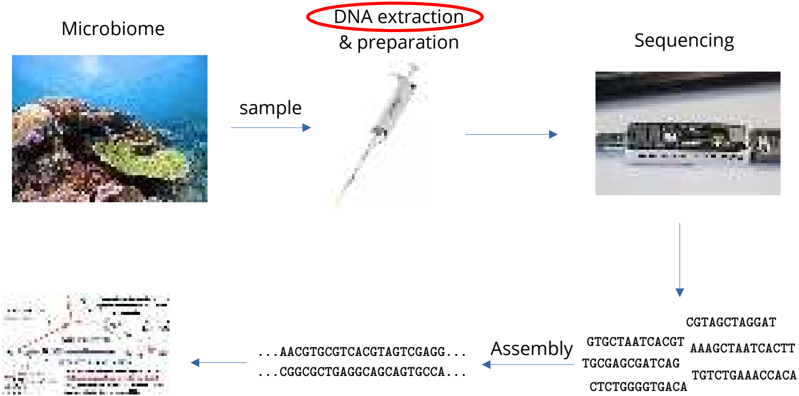
- Example of the sequencing of the soil microbiote



Adapted from the work of Nicolas Maurice

- Low-coverage assembly
- Missing DNA

# What is the future of assembly?



Microbiome

sample

DNA extraction
& preparation

Sequencing

CGTAGCTAGGAT
GTGCTAATCACGT
AAAGCTAATCACTT
TGCGAGCGATCAG
TGTCTGAAACCACA
CTCTGGGGTGACA

Long reads

Assembly

...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...

Genomes

Understand the
microbiome