Haplotype assembly from long reads

Roland Faure[1,2]

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

[2]Université de Rennes, IRISA - France

Genome Assembly Course - December 18th, 2024

"What species do you work on?"

- Kahoot n°1

Using genome sequencing to study stuff

Organism, microbiota...

sample

DNA extraction & preparation

Sequencing

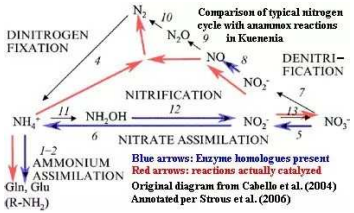
```
CGTAGCTAGGAT
GTGCTAATCACGT  AAAGCTAATCACTT
TGCGAGCGATCAG
CTCTGGGGTGACA
```

Long reads

Assembly

```
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...
```

Genomes or amplicons



Understand things

Using genome sequencing to study stuff

Organism, microbiota...

sample

DNA extraction
& preparation

Sequencing

My Ph.D.

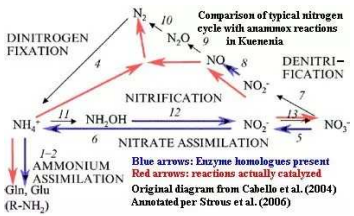
```
CGTAGCTAGGAT
GTGCTAATCACGT  AAAGCTAATCACTT
TGCGAGCGATCAG  TGTCTGAAACCACA
CTCTGGGGTGACA
```

Long reads

Assembly

```
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...
```

Genomes or
amplicons



Understand things

DNA sequencing

Extracted DNA

sequencer

CAGCATCAGTTTTCGAGCACGT

TTACTCAGCAGATCGTCGATCAT

CCCGTAGCTTAGCAGGCATCAG

Reads

DNA sequencing: difficulties

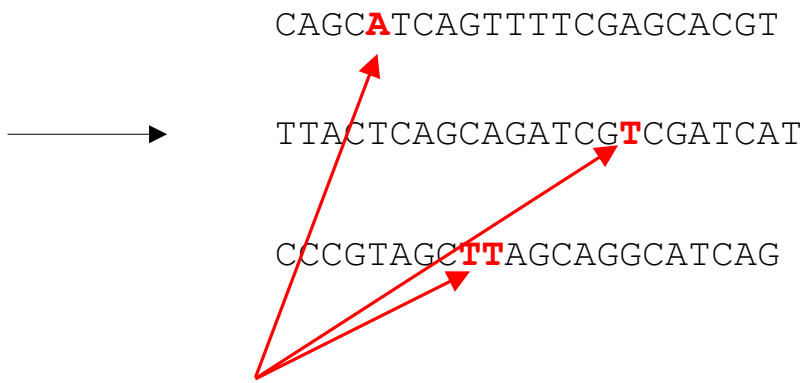
length: 1-20 kbp

CAGCATCAGTTTTCGAGCACGT

TTACTCAGCAGATCGTCGATCAT

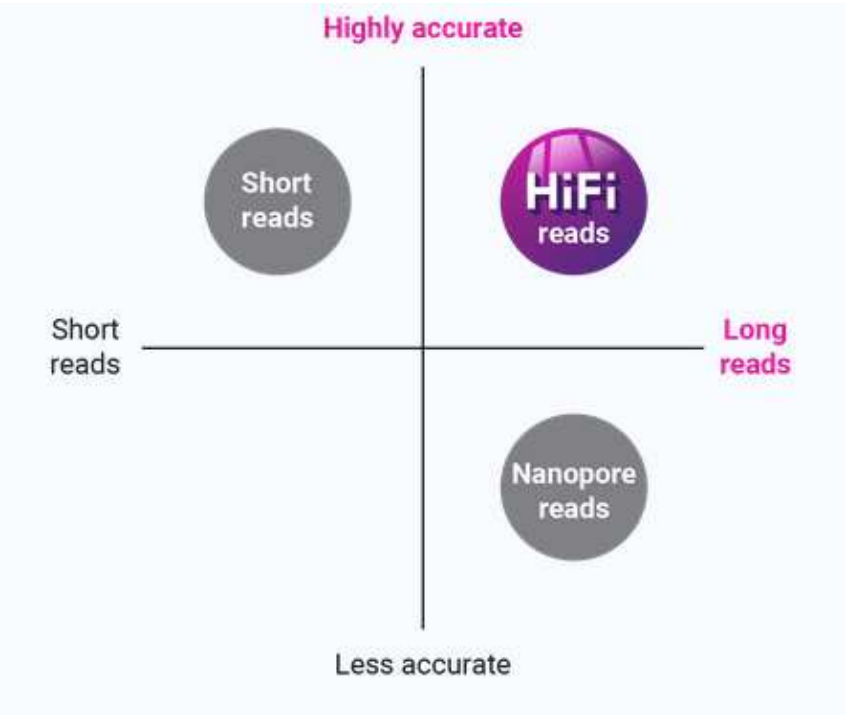
CCCGTAGCTTAGCAGGCATCAG

DNA sequencing: difficulties



sequencing errors: 0.1 – 10 %

The different sequencing technologies



pacb.com

Genome sequencing

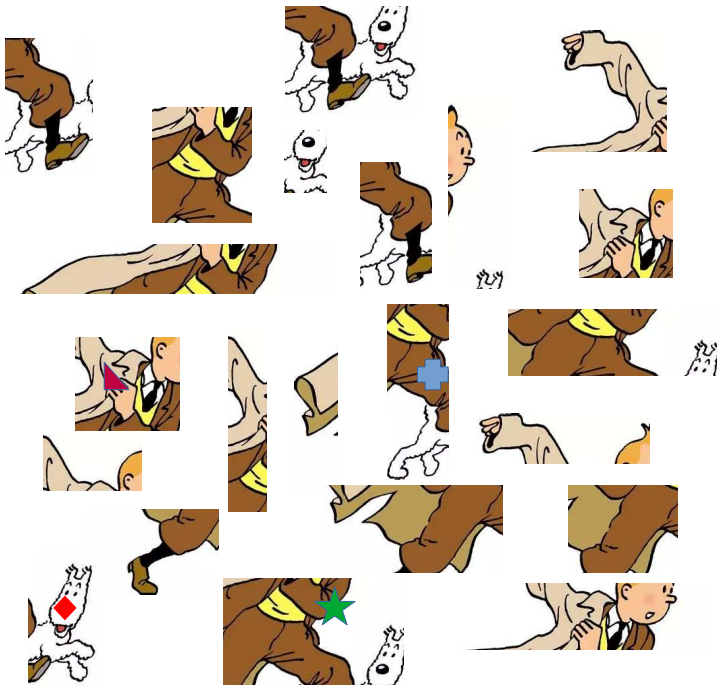
Mysterious genome

Sequencing →

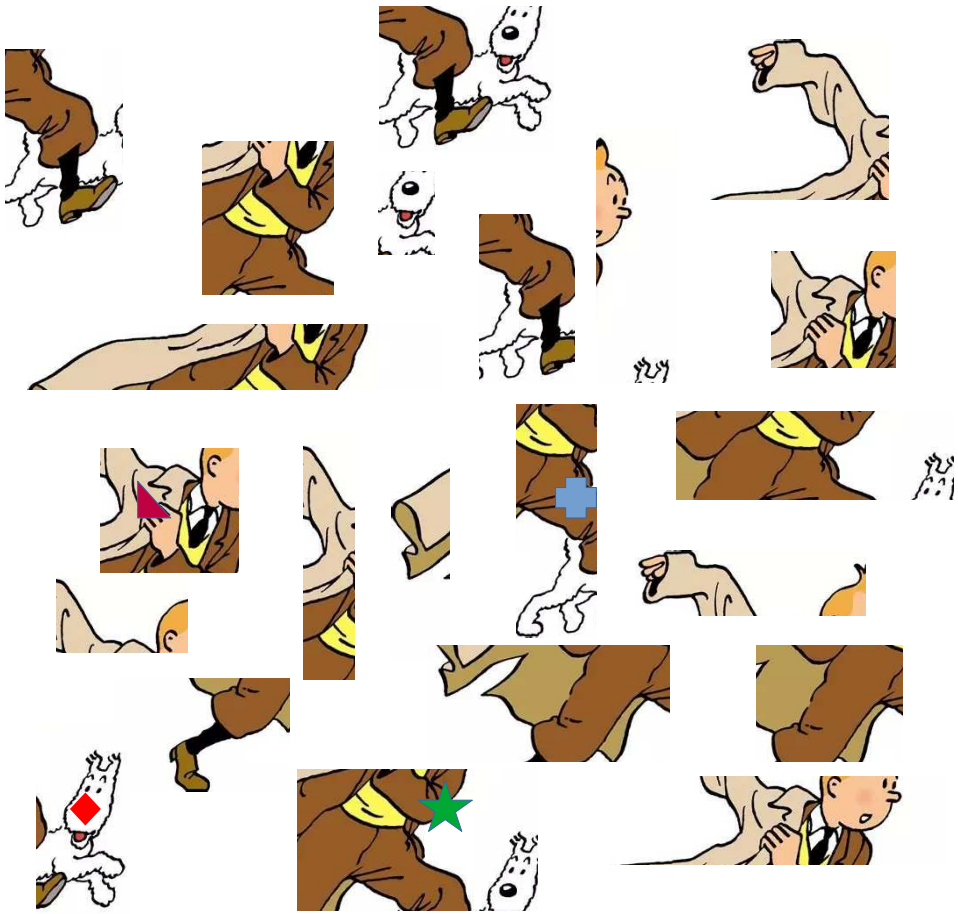
TTCGGCGCTGAGGCAG
CAGCGCTGAGGCAGCAGTGCCA GGCAGCAGTGCCAG
CAGCATTGCCAGGC TCGGC
CGGCGCTGAGGCAGCATTGCC

Mysterious picture

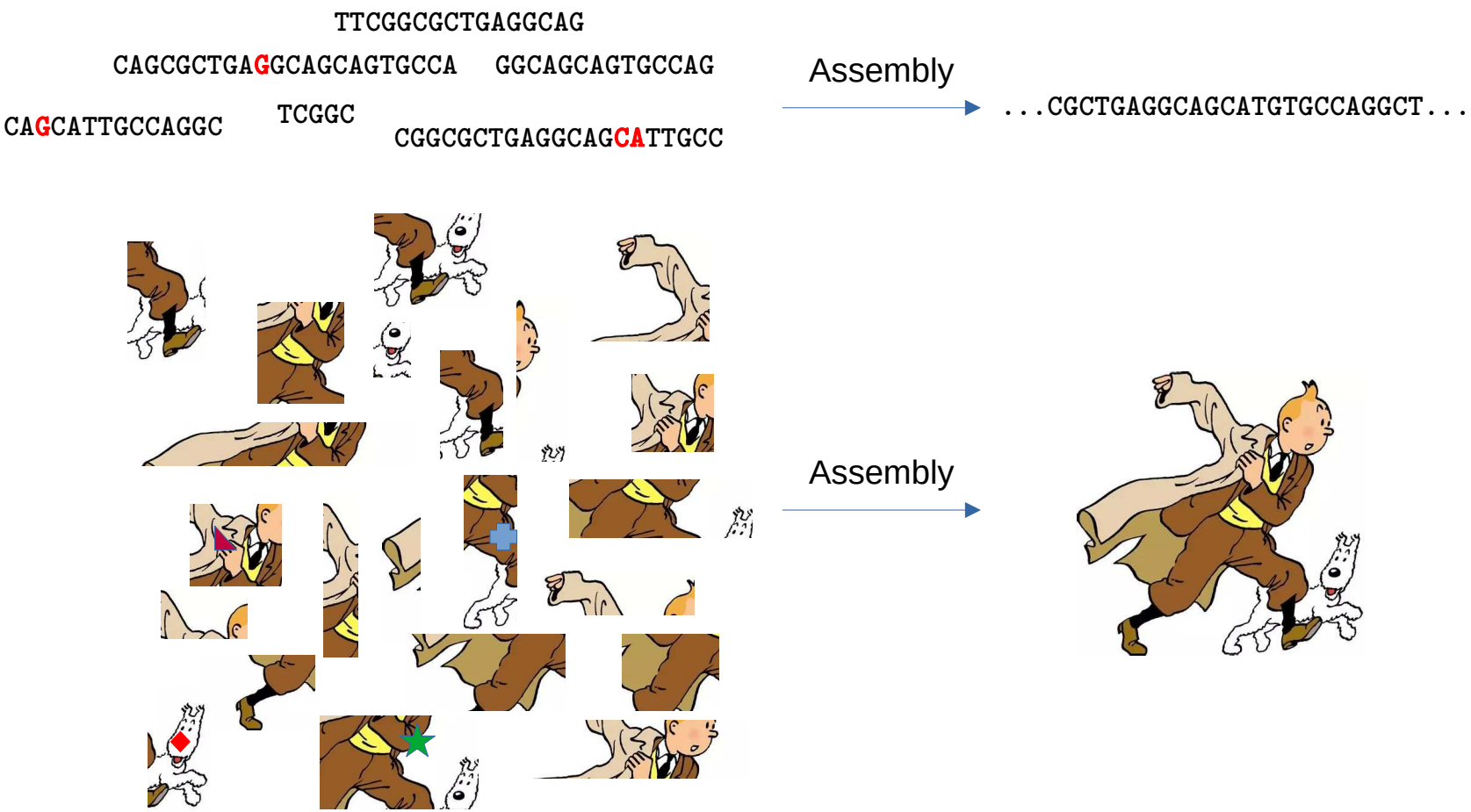
Sequencing →



Imagine you are an assembler: what is this picture? (Kahoot)



Genome assembly



Genome assembly

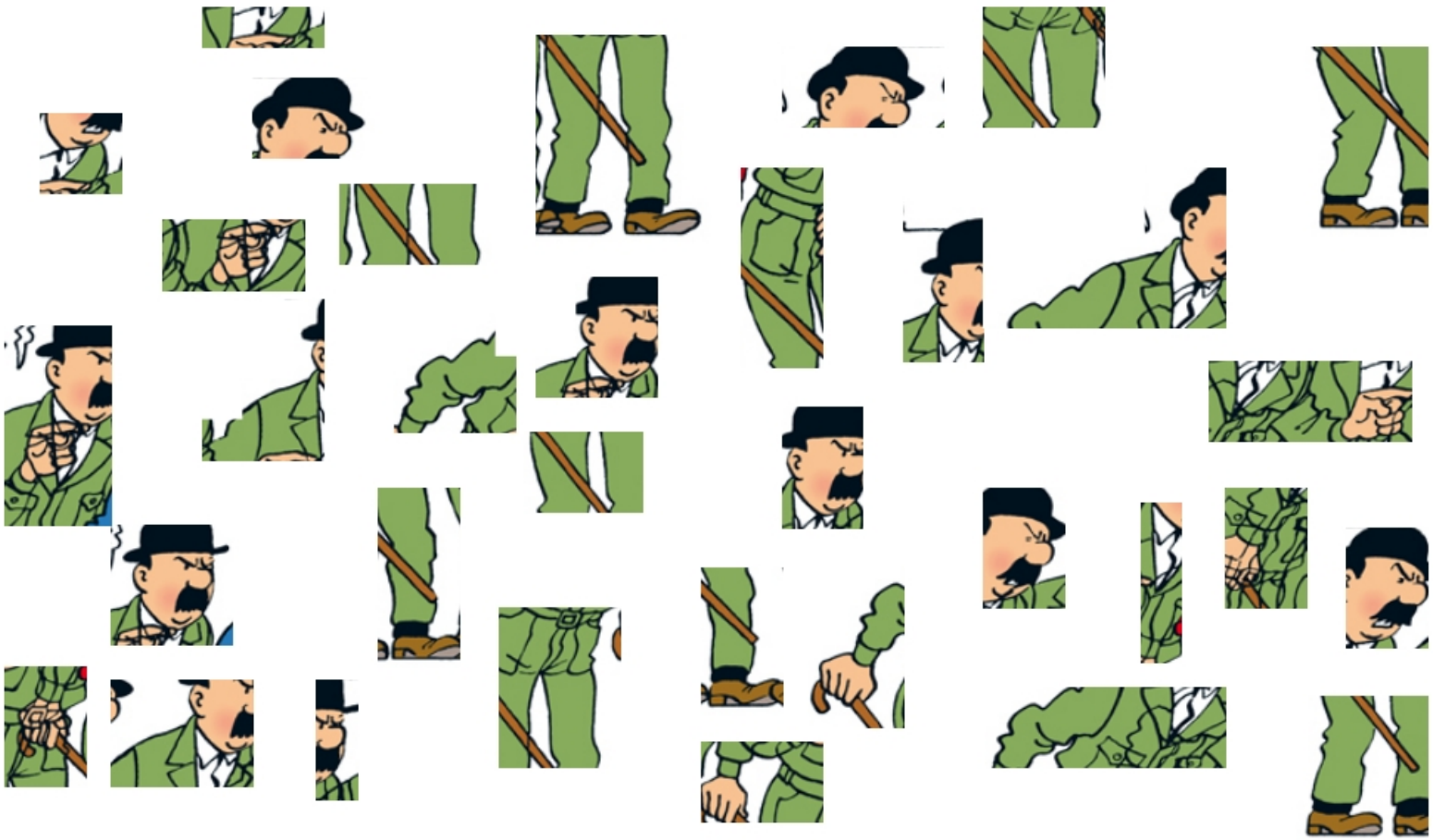
```
CGATGCTGGCTAGCATAGTCGATTTATCT
     CTGGCTAGCTTAGTCGATTTATCTGACAGT
           AGCATAGTCGATTTATCTGACAGTCATAT
                AGTCGATTTATATGACAGTCATATTGCT
                     TTTATCTGACAGTCAGATTGCTACACAC
```

genome assembly: stitching reads
correcting errors

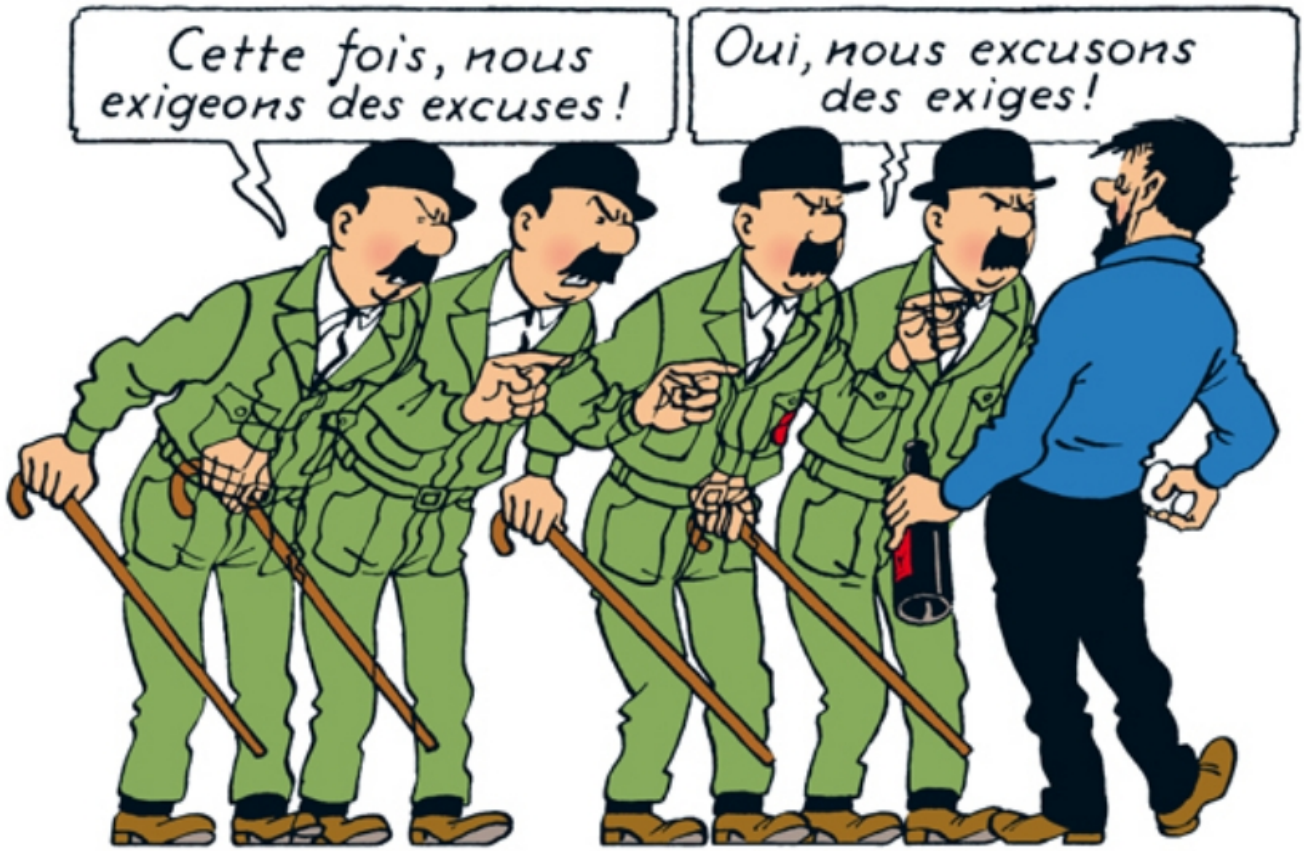
CGATGCTGGCTAGCATAGTCGATTTATCTGACAGTCATATTGCTACACAC

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

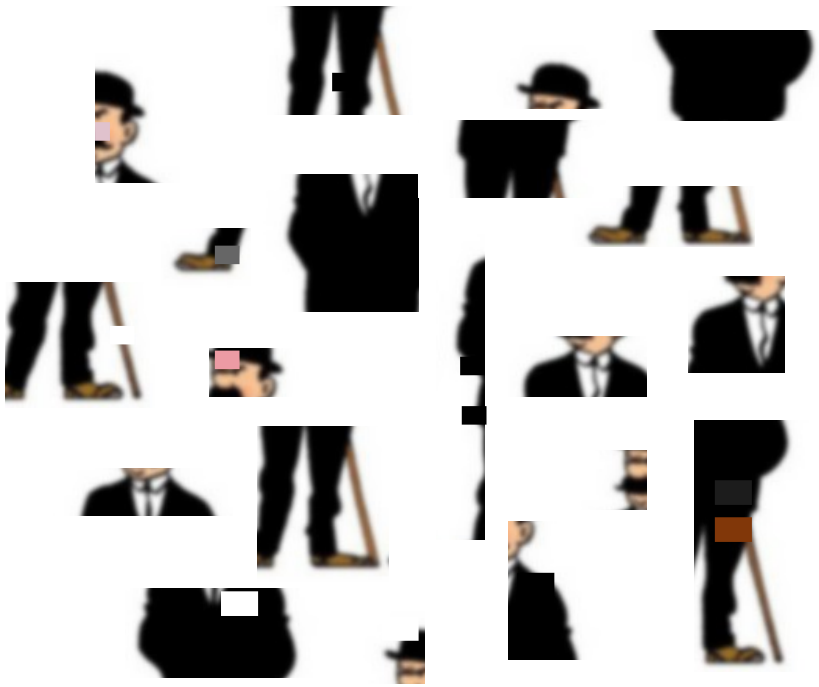
Imagine you are an assembler (Kahoot)



Imagine you are an assembler

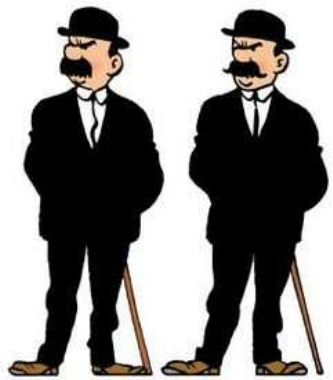


Haplotype assembly

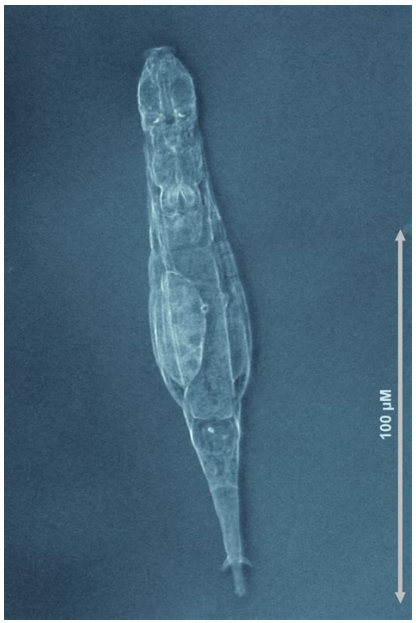


Collapsed assembly

Haplotype assembly



Dupont & Dupond exist in genomes!



Maternal DNA

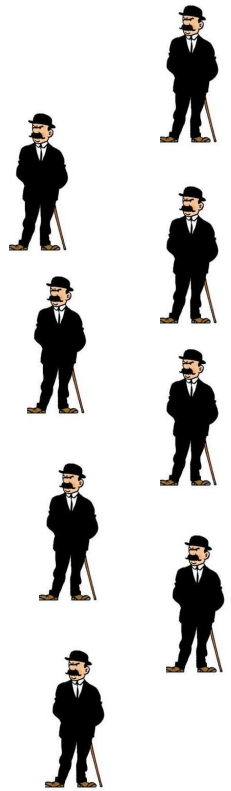
...ACACACCACACACCTCTACGA...

...ACACAC**T**ACACACCTCTACGA...

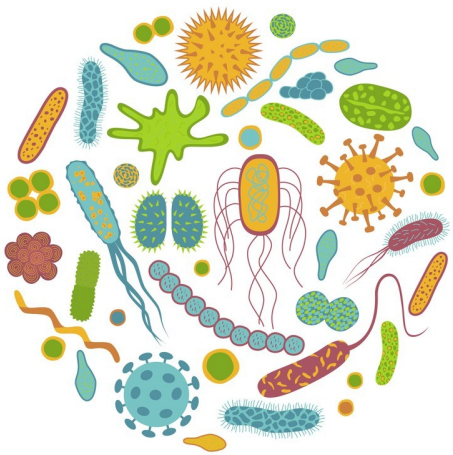
Paternal DNA

Dupont & Dupond exist in genomes!

...ACACACCACACACCT**G**TACGA...
...ACACACCACACACCTCTACGA...
...ACACACCACACACCTCTACGA...
...ACACAC**T**ACACACCTCTACGA...
...ACACACCACACACCTCTACGA...
...ACACACCACACACCTCTACGA...
...ACA**A**ACCACACACCTCTACGA...
...ACACACCACACACCTCTA**A**GA...



Dupont & Dupond exist in genomes!



...ACACACCACACACCTCTACGA...

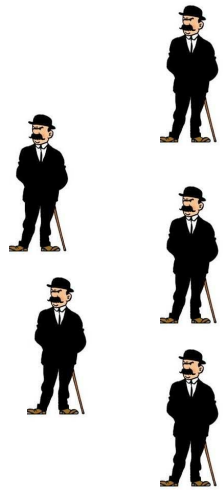
...ACACAC**T**ACACACCTCTACGA...

...ACACACCACACACCTCTACGA...

...ACA**A**ACCACACACCTCTACGA...

...ACACACCACACACCTCTA**A**GA...

...

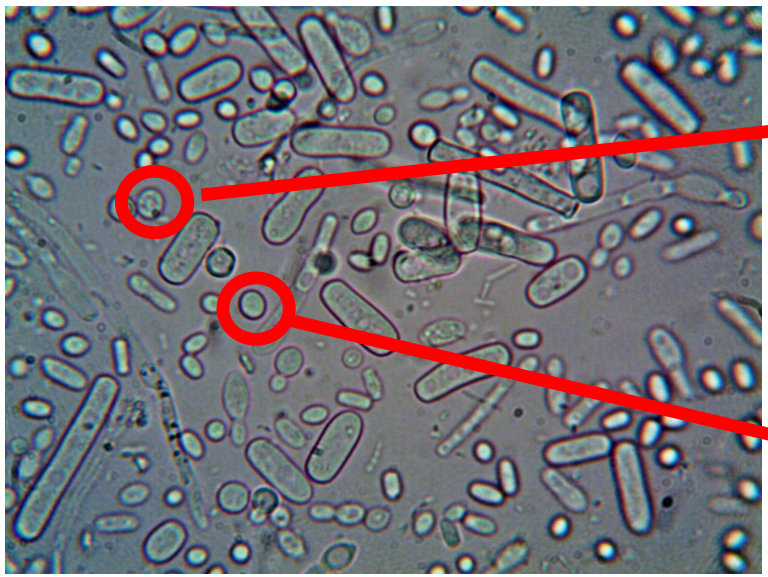


Dupont & Dupond are important in genomes!

- Kahoot

Dupont & Dupond are important in genomes!

- Kahoot



Escherichia coli Sakai

...ACACACCACACACCTCTACGA...

...ACACAC**T**ACACACCTCTACGA...

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
     CTGGCTAGCTTAGTCGATTTATCTGACAGT
           AGCATAGTCGATTTATCTGACAGTCATAT
                AGTCGATTTATATGACAGTCATATTGCT
                     TTTATATGACAGTCAGATTGCTACACAC
```

genome assembly: stitching reads
correcting errors

CGATGCTGGCTAGCATAGTCGATTTATCTGACAGTCATATTGCTACACAC

CGATGCTGGCTAGCATAGTCGATTTATATGACAGTCATATTGCTACACAC

- Not so many software!

Haplotype assembly from long reads

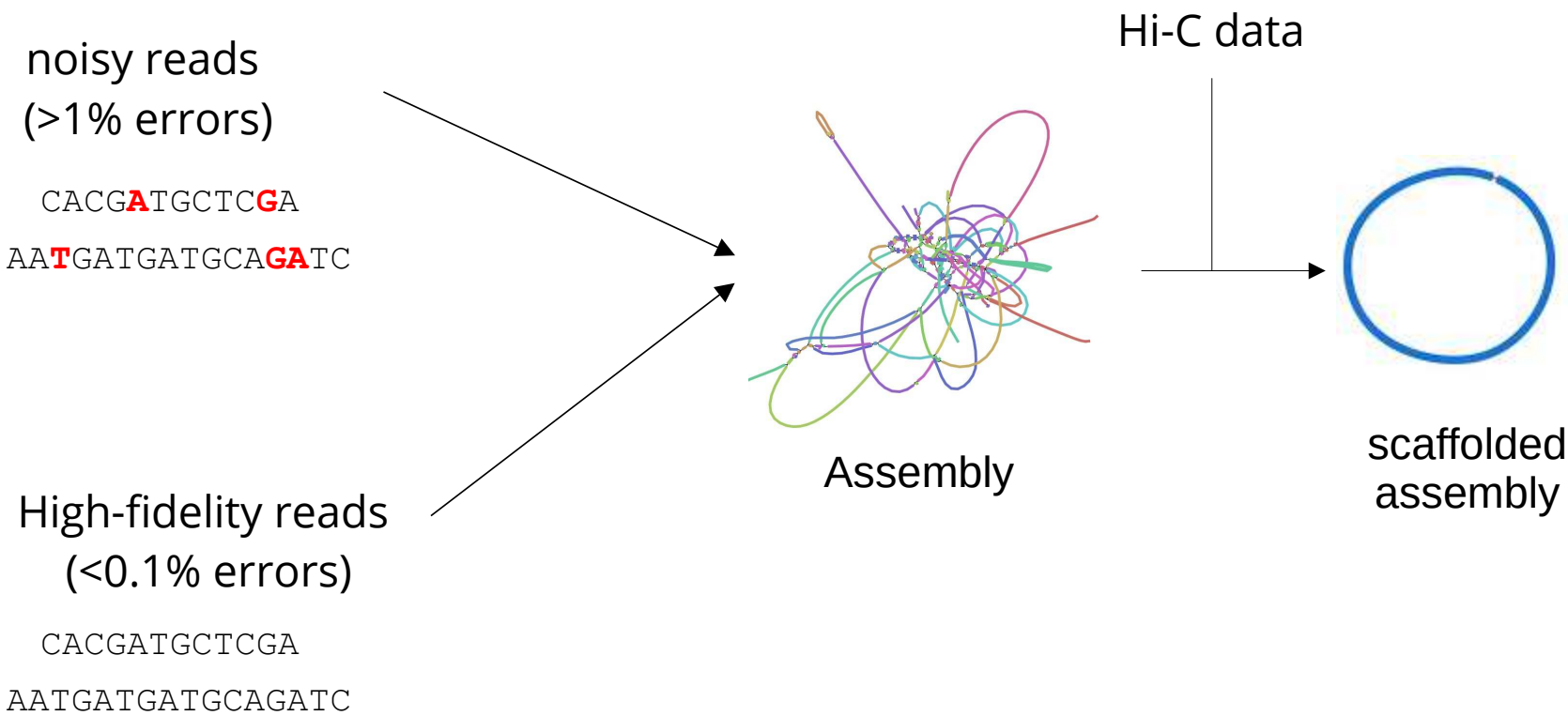
Roland Faure[1,2]

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

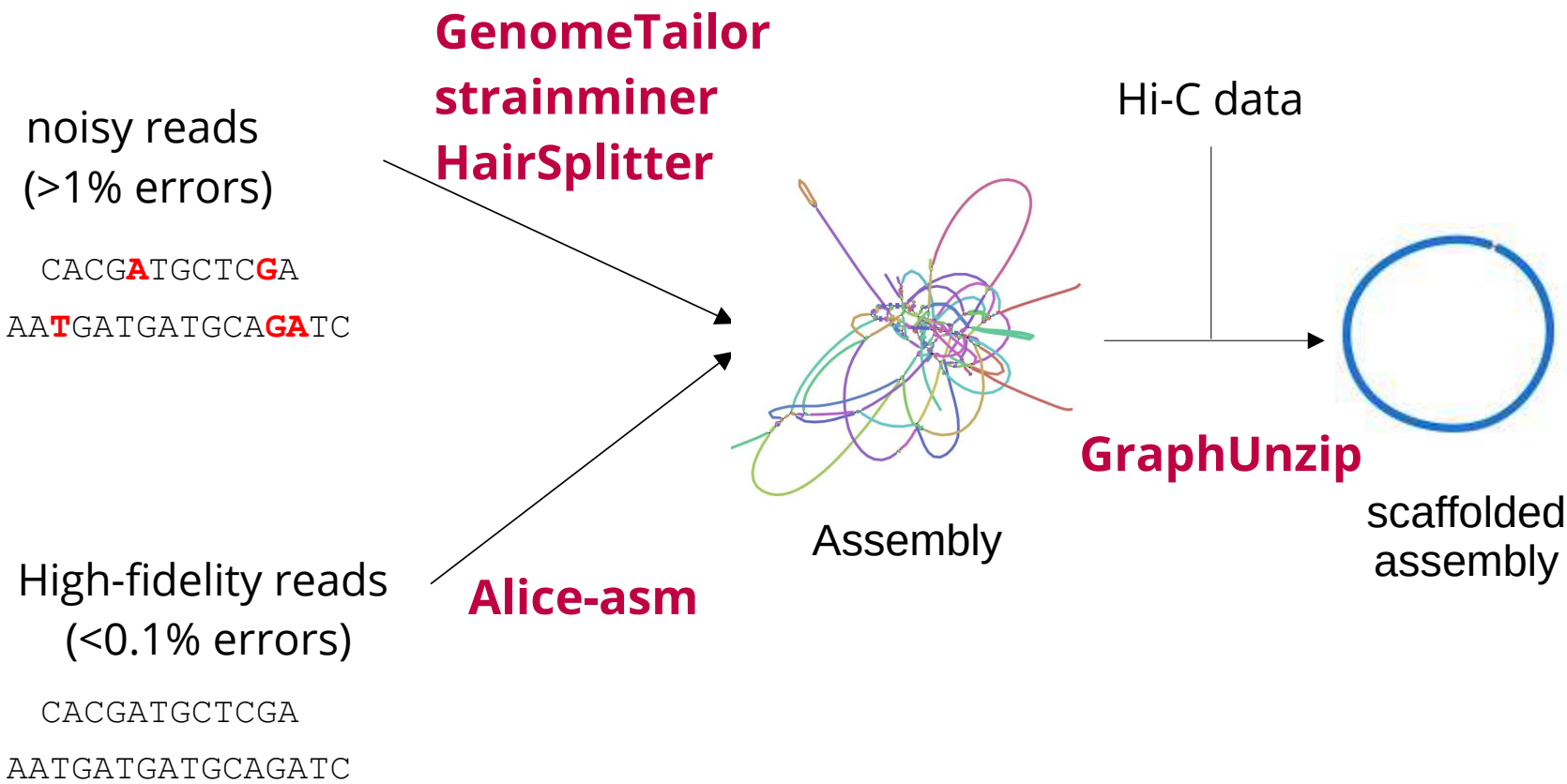
[2]Université de Rennes, IRISA - France

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

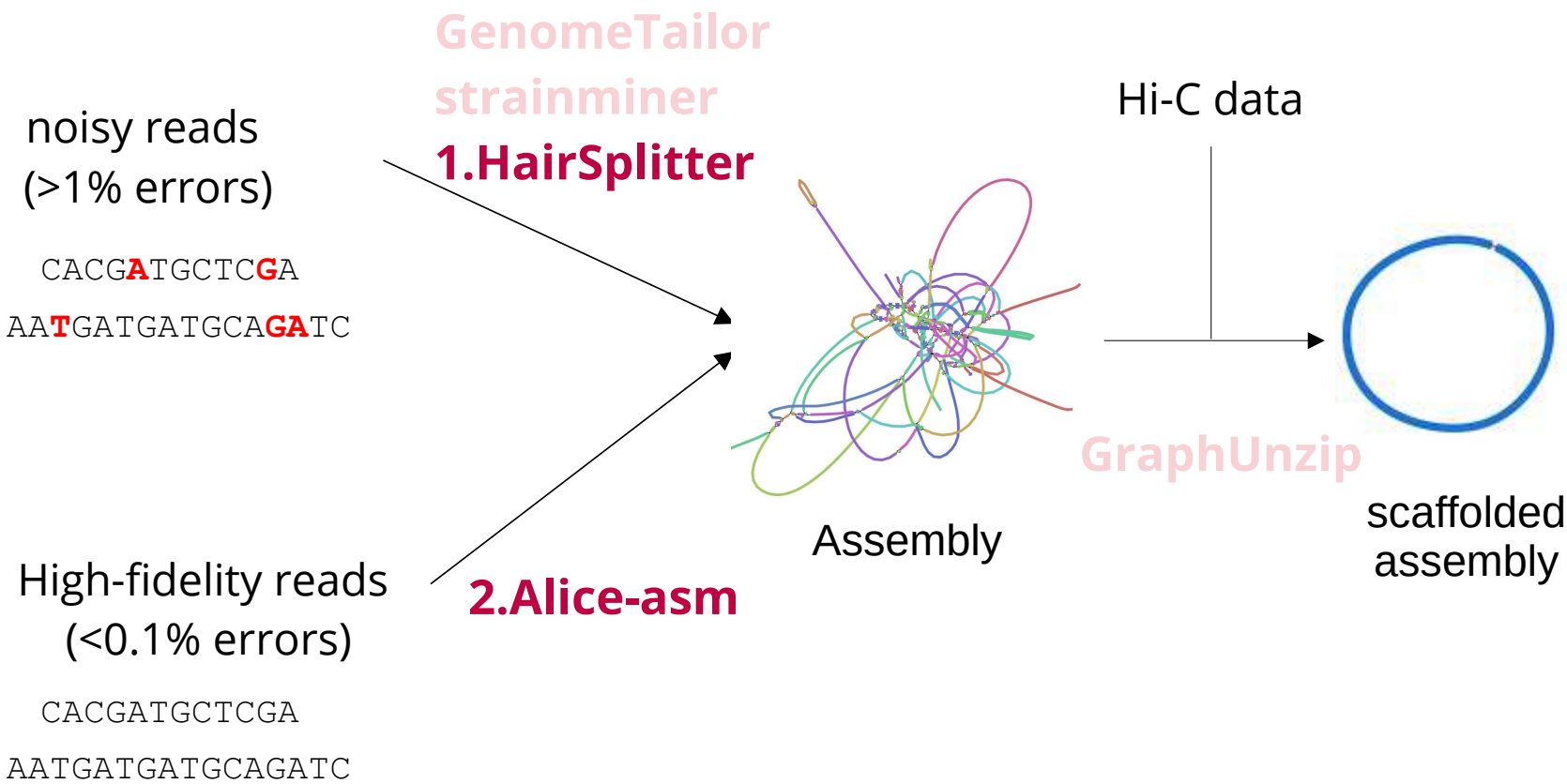
Overview of (meta)genome assembly



Overview of the Ph.D.



Overview of the Ph.D.

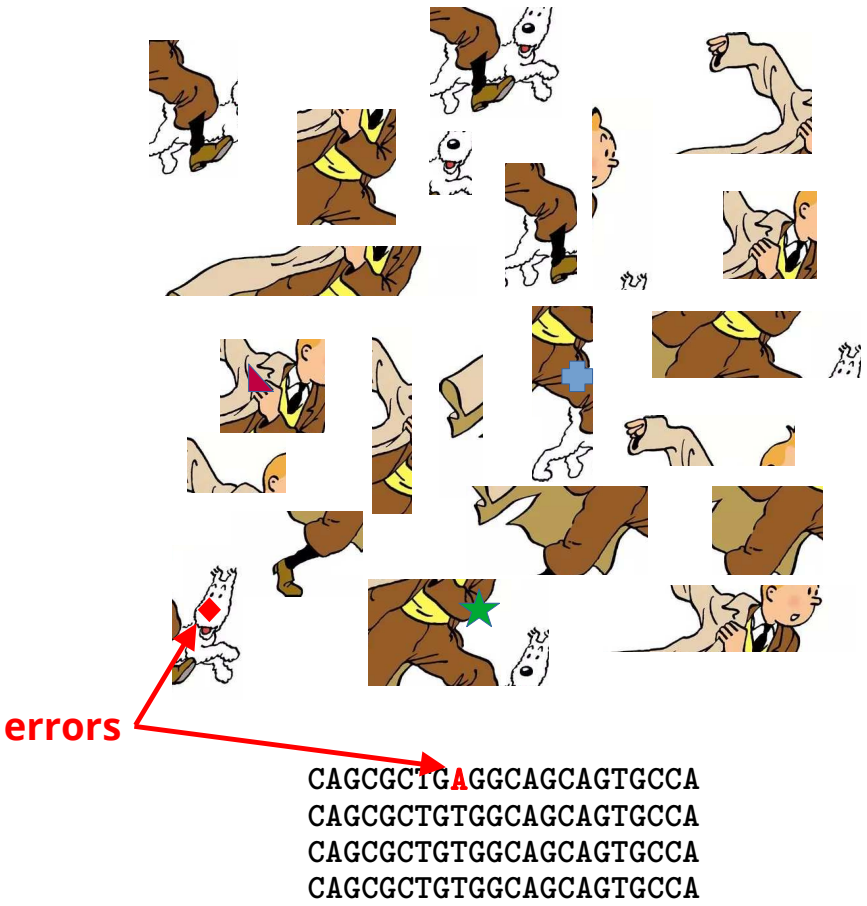


Distiguishing haplotypes with noisy reads - HairSplitter

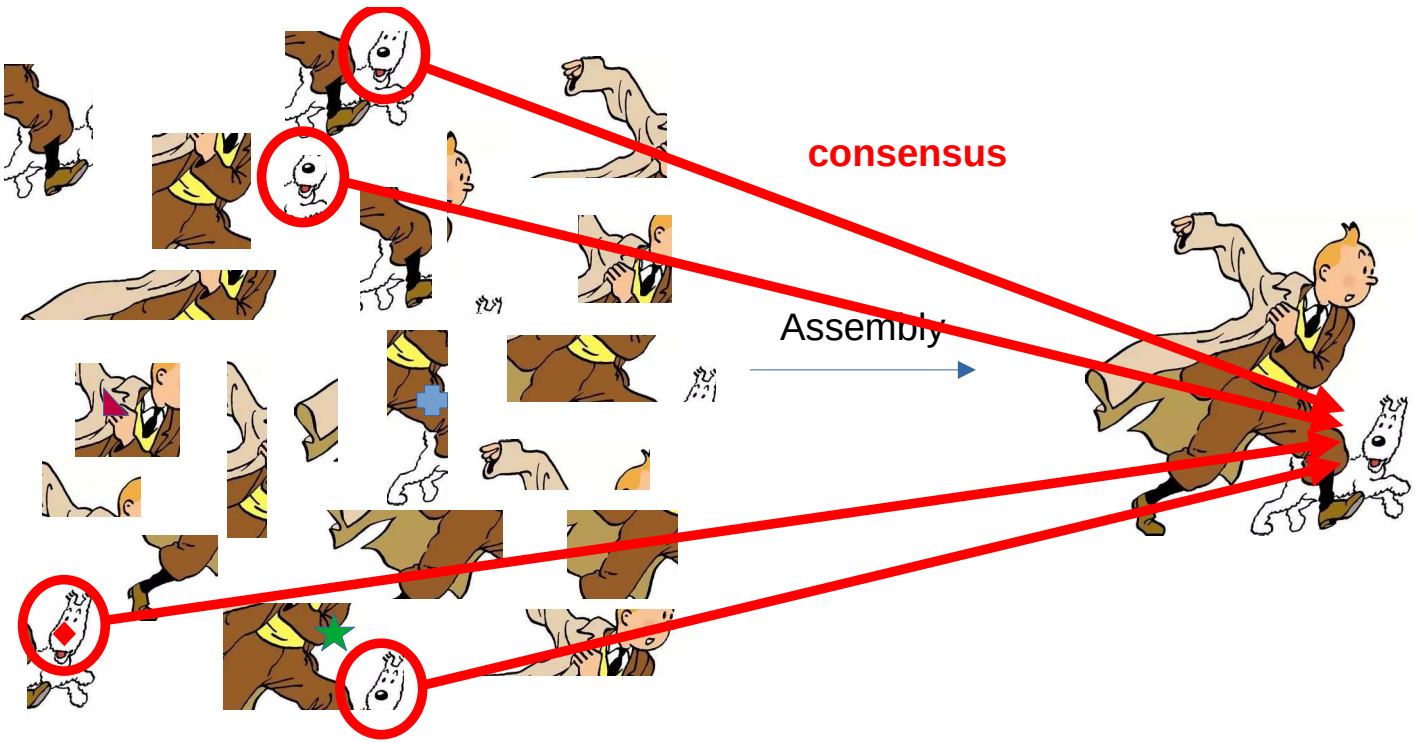
Assemblers correct errors

- I want to correct sequencing errors but I don't know the solution
- Like correcting a text in an unknown language
- Kahoot

Assembling noisy reads: correcting errors by consensus



Assembling noisy reads: correcting errors by consensus

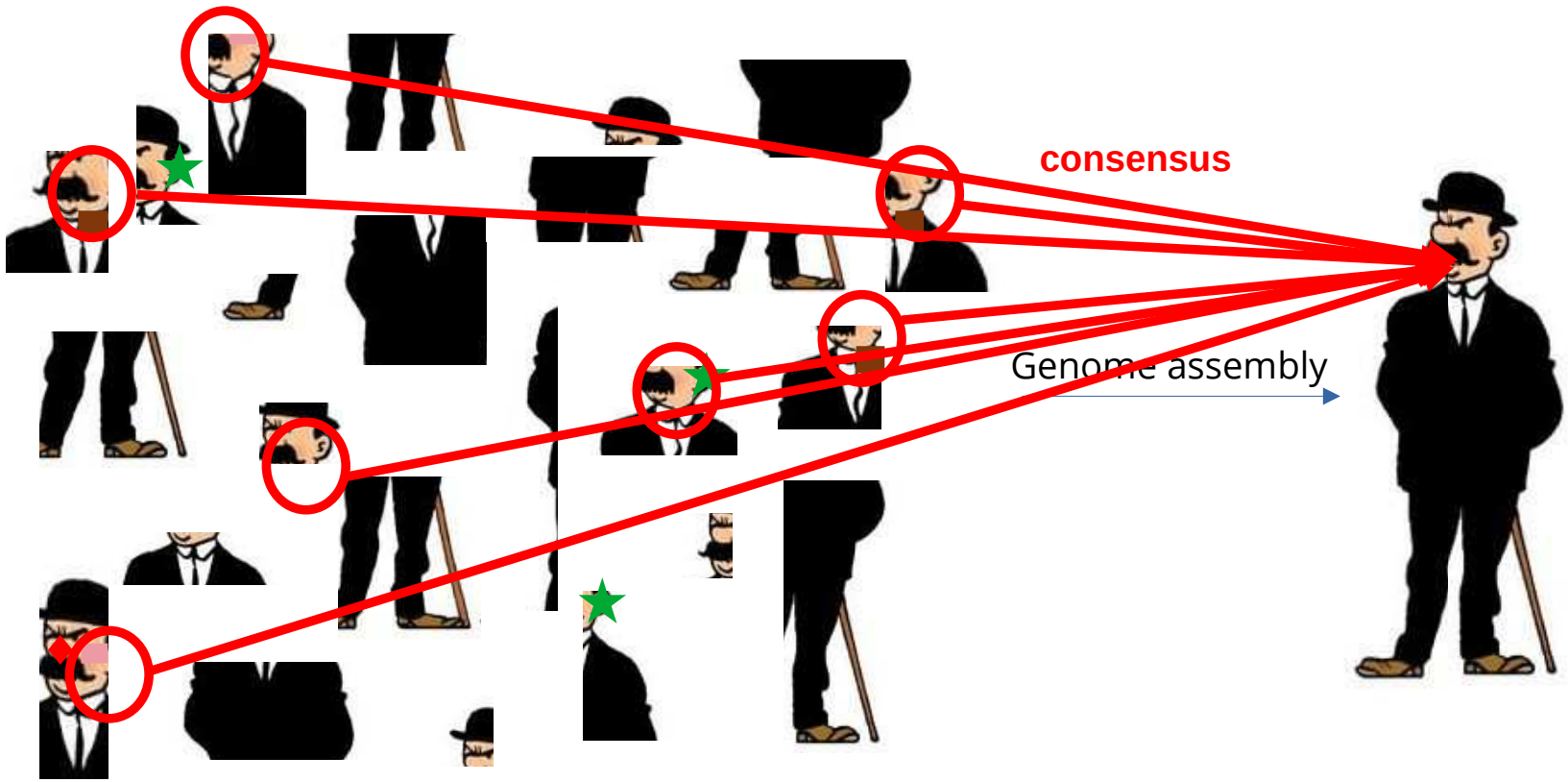


CAGCGCTG**A**GGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA
CAGCGCTGTGGCAGCAGTGCCA

Assembly

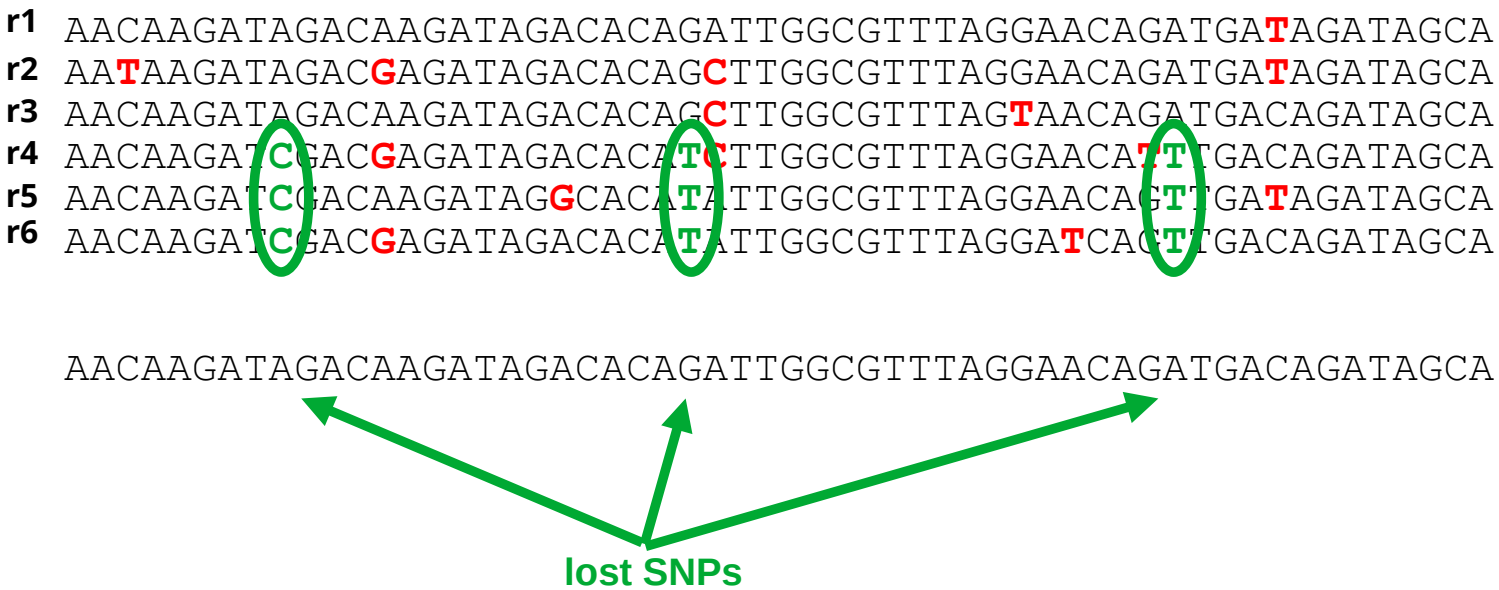
CAGCGCTGTGGCAGCAGTGCCA

Consensus loses the variants



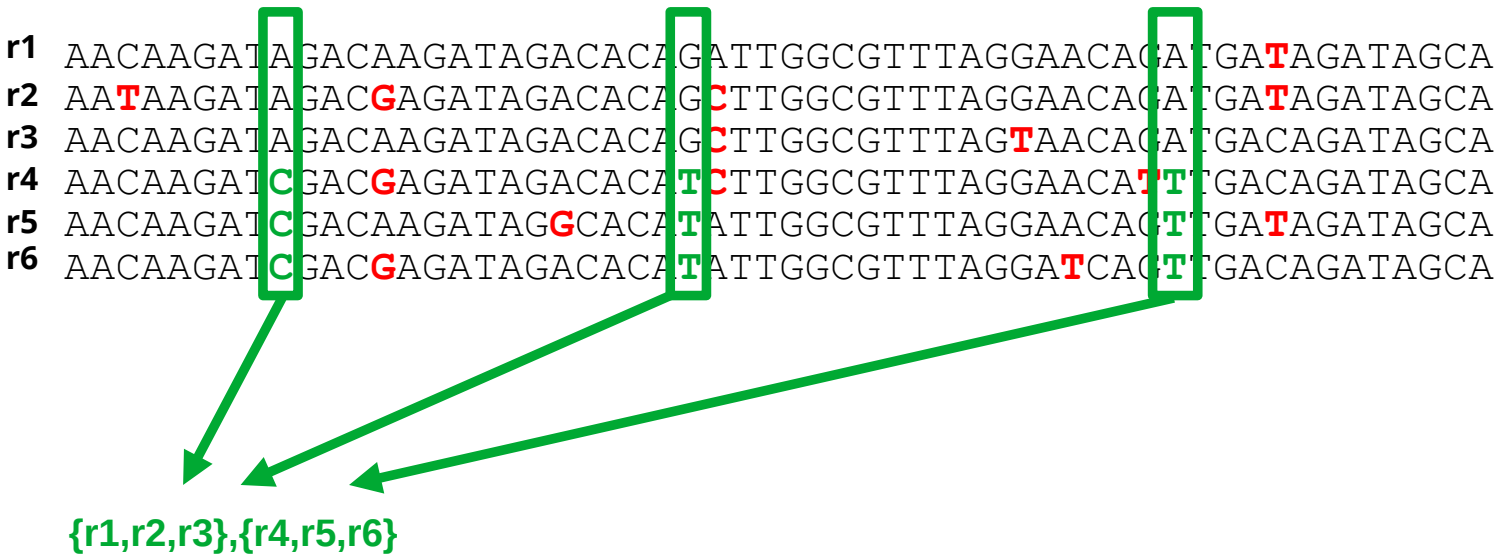
Consensus loses the variants

Consensus loses the variants

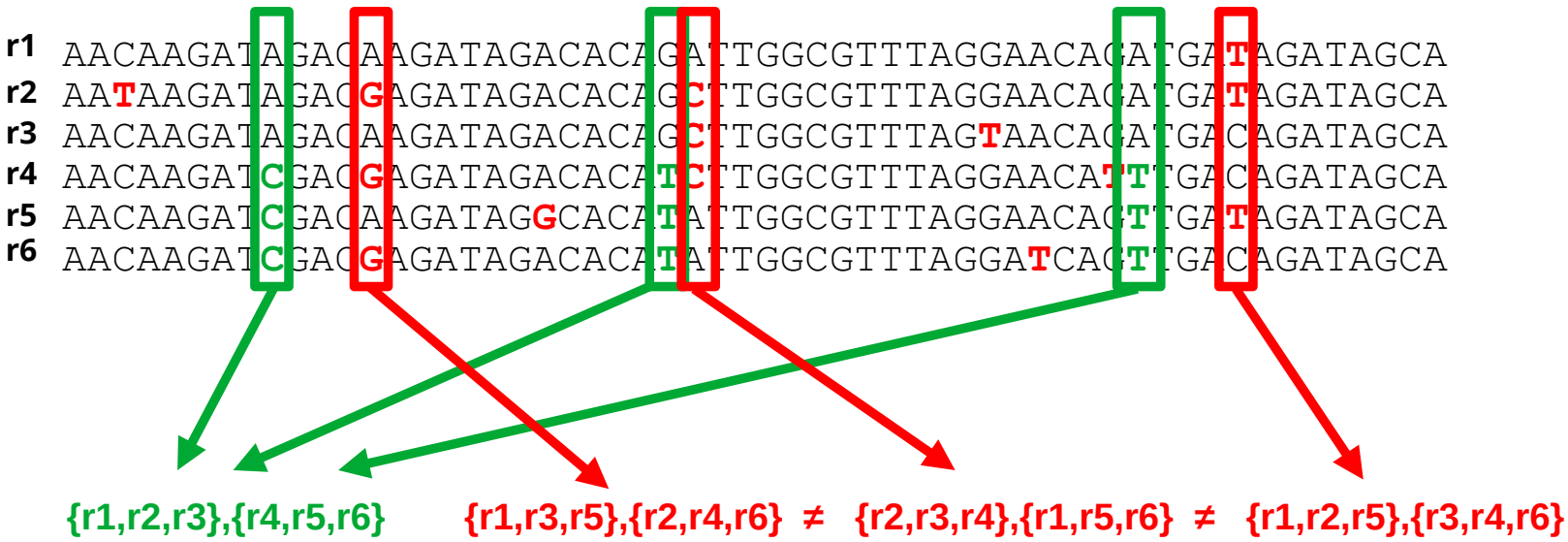


How to distinguish errors and SNPs?

My solution: looking at several positions simultaneously



My solution: looking at several positions simultaneously



Algorithm: 1) looking for variant patterns

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

Algorithm: 1) looking for variant patterns

variant pattern: subset of reads and positions containing alternative bases
size: 2x2

Is this variant too big to be due to sequencing errors?

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

variant pattern: subset of reads and positions containing alternative bases
size: 2x2

- I need a statistical test

Is this variant too big to be due to sequencing errors?

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

variant pattern: subset of reads and positions containing alternative bases
size: 2x2

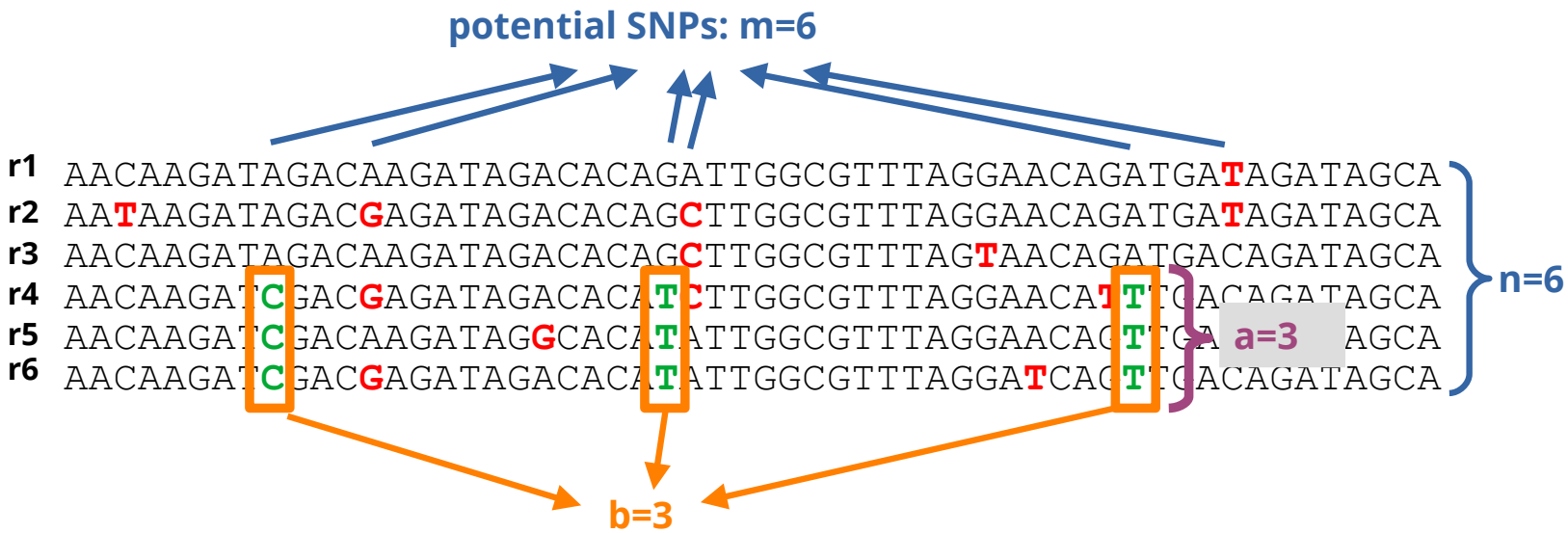
- I need a statistical test
- I need help: Kahoot

Algorithm: 2) Statistical test

- Null hypothesis: there are no haplotypes, just sequencing errors

Algorithm: 2) Statistical test

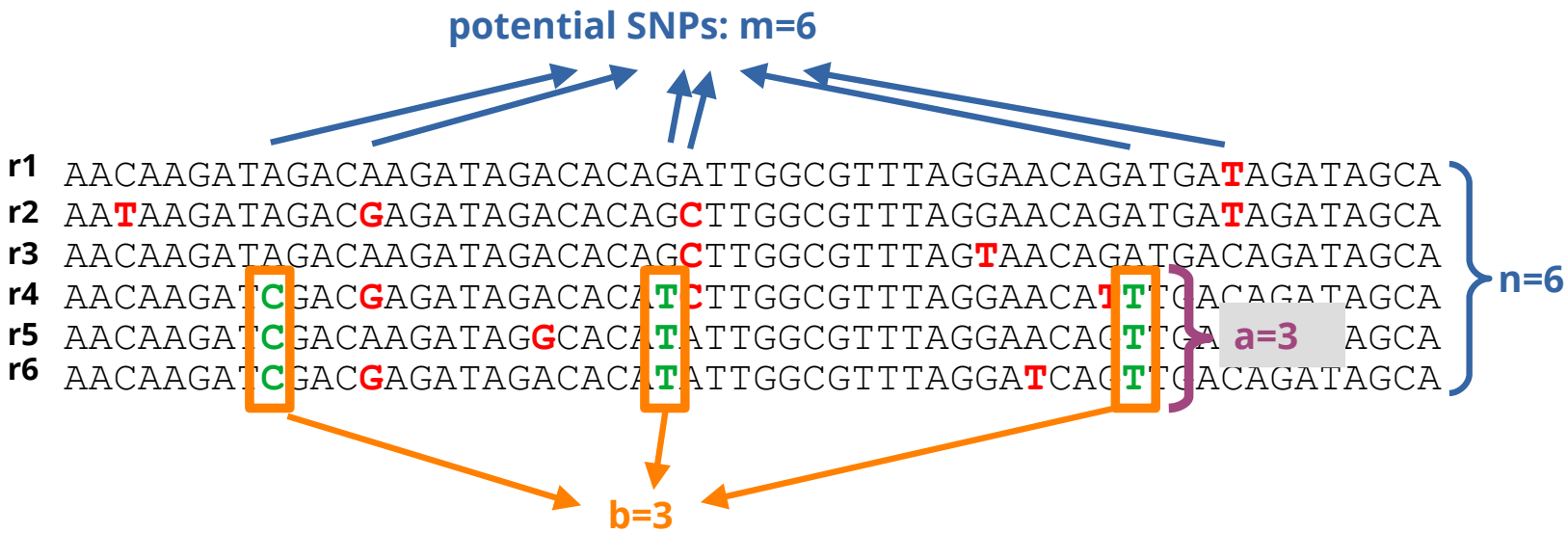
- Null hypothesis: there are no haplotypes, just sequencing errors



P(errors produce pattern of size **ab**) $\leq \binom{n}{a}\binom{m}{b} * \frac{a^{ab}}{n^{ab}} = 0.006$

Algorithm: 2) Statistical test

- Null hypothesis: there are no haplotypes, just sequencing errors



P(errors produce pattern of size **ab**) $\leq \binom{n}{a}\binom{m}{b} * \frac{a^{ab}}{n^{ab}} = 0.006$

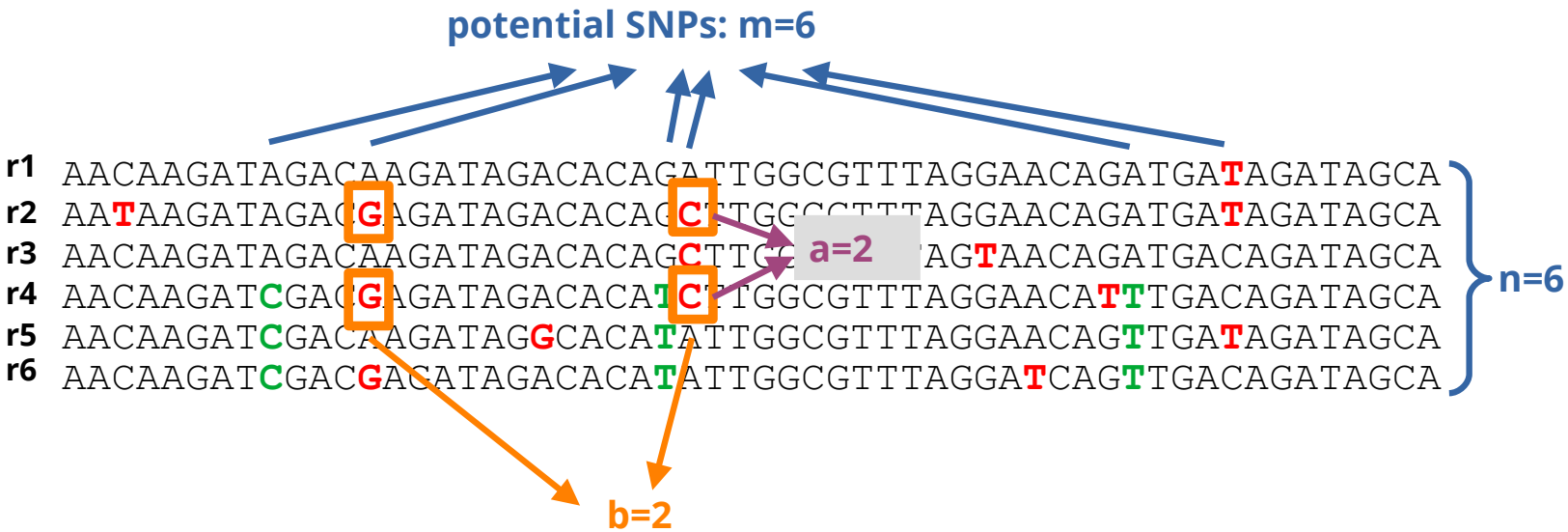
- THIS LOOKS WEIRD

Algorithm: 2) Statistical test

- Null hypothesis: there are no haplotypes, just sequencing errors

Algorithm: 2) Statistical test

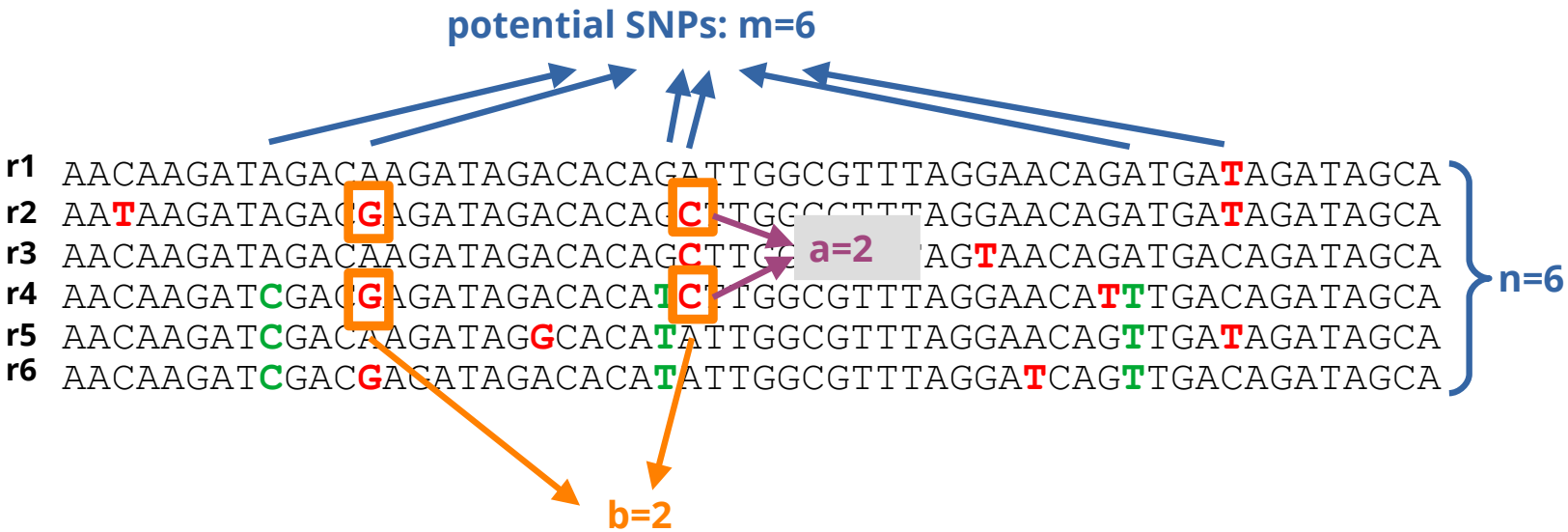
- Null hypothesis: there are no haplotypes, just sequencing errors



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

- Null hypothesis: there are no haplotypes, just sequencing errors



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

- This can happen

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

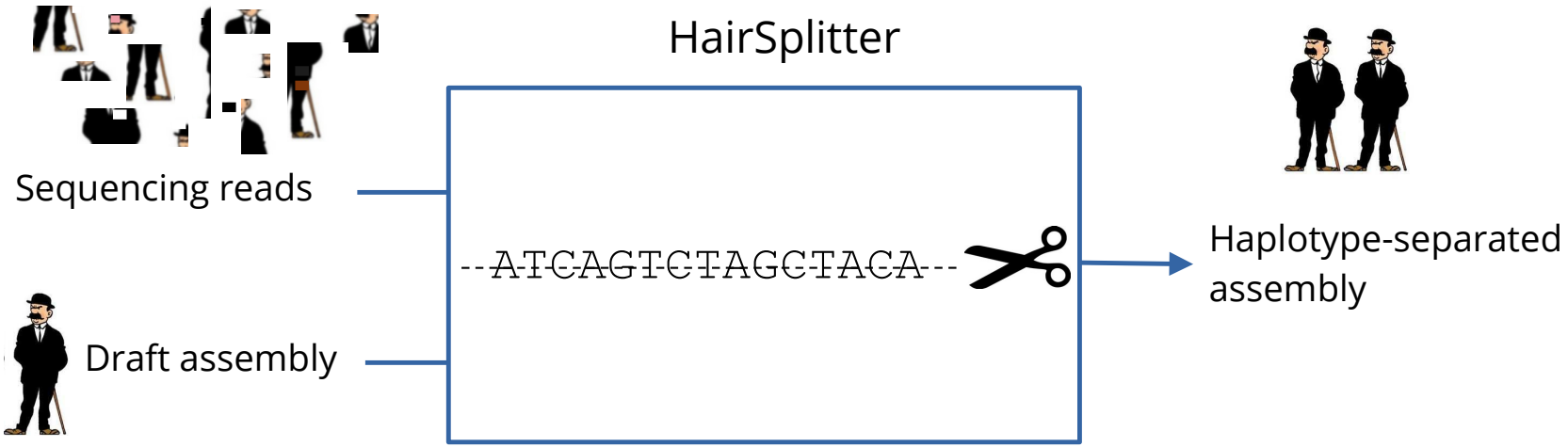
Algorithm: 3) Group reads by haplotype

Passed the test

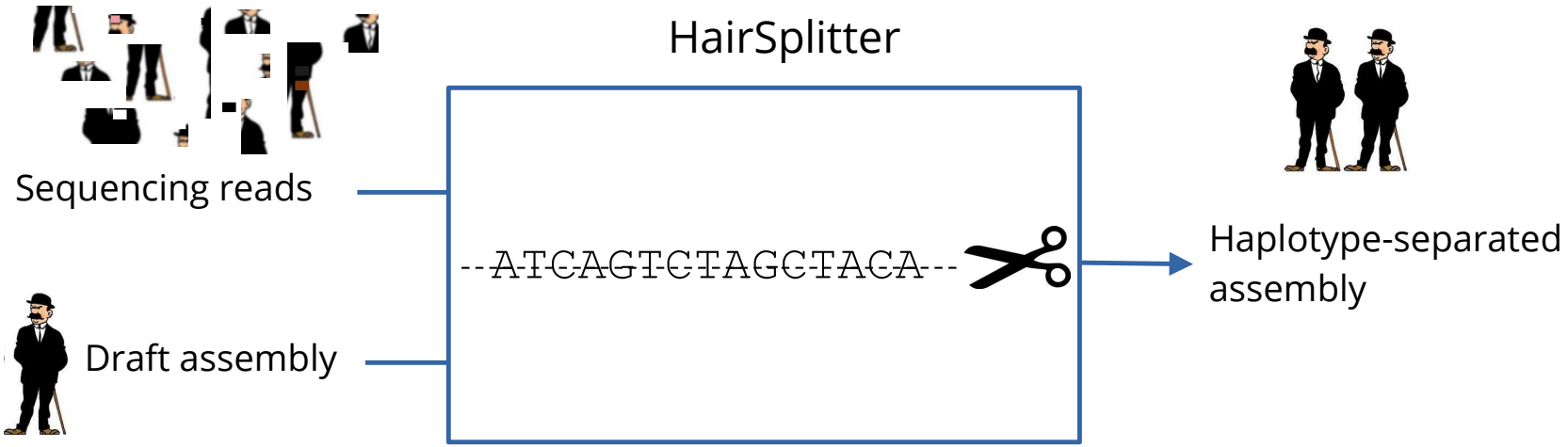
group reads by haplotypes

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

The HairSplitter program



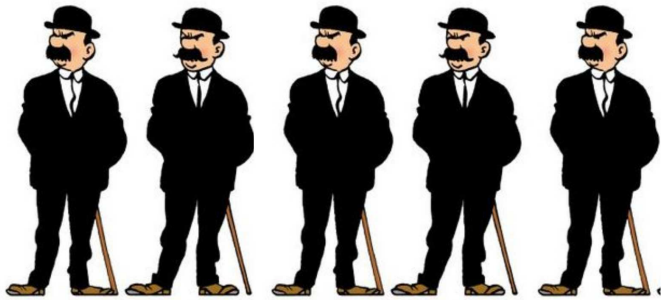
The HairSplitter program



- *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*

Evaluating HairSplitter - results

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



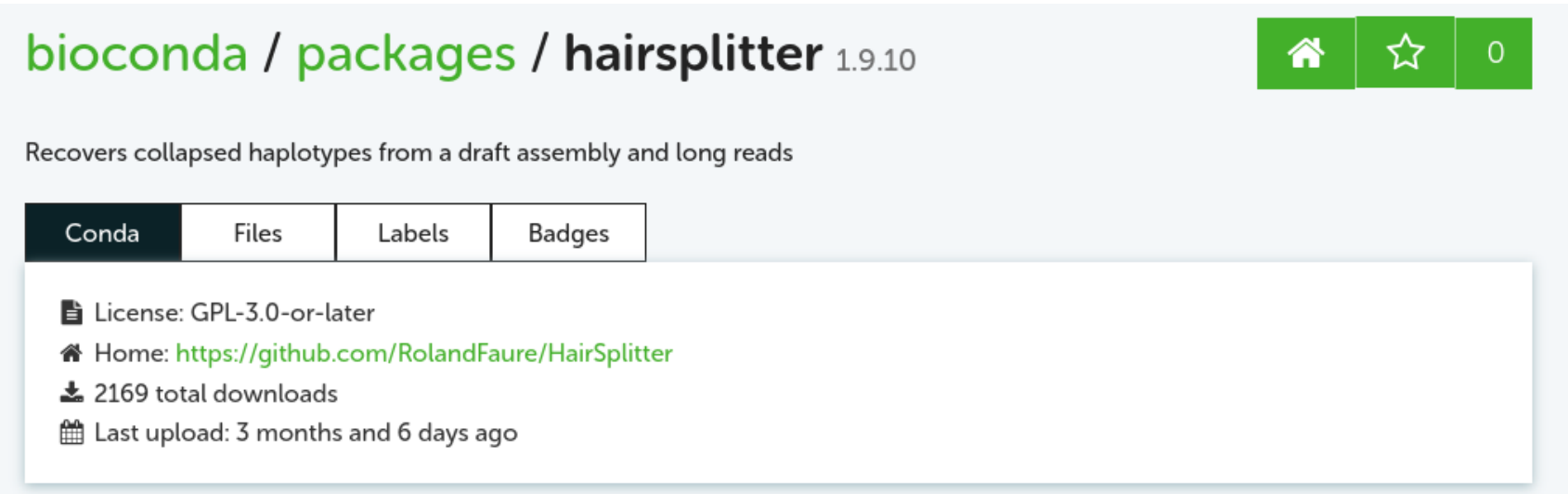
	metaFlye	metaFlye+Strainberry	metaFlye+HairSplitter
Nanopore Q9	0.586	0.749	**0.957**
Nanopore Q20	0.7524	0.9527	**0.961**
PacBio HiFi	0.9589	0.9793	**0.9895**

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

- Improves over the state of the art on complex microbiota

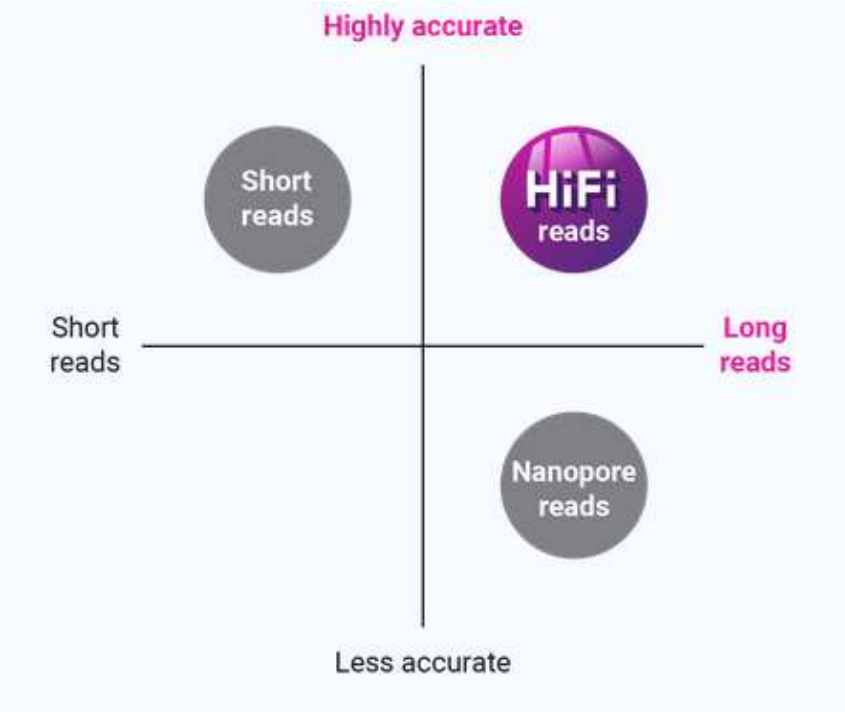
The HairSplitter project

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



Distinguishing haplotypes with high-fidelity reads - Alice

New technology: high-fidelity long reads

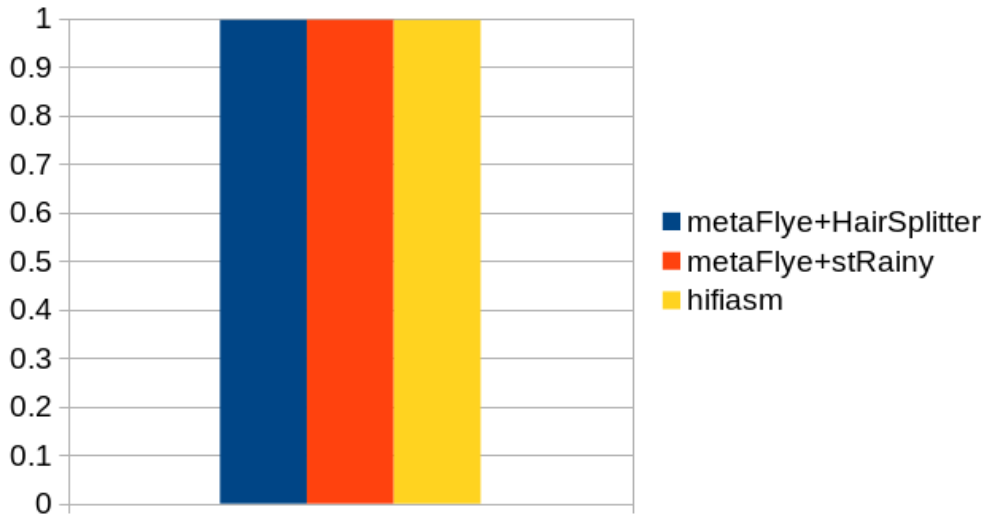


pacb.com

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

Assembly with high-fidelity long reads: easy!

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 - Portik et al.

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	$\geq$ 60 days	$\geq$ 60 days

- Also takes memory, energy and money

Why are assemblers so slow?? (Kahoot)

Why are assemblers so slow??

Read 1 ATGCATCGAGTAGGGGCACTGTACC
Read 2 GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT
Read 3 CAGATGGAGAATGCATCGAGTAGG

Read 3 CAGATGGAGAATGCATCGAGTAGG
Read 1 ATGCATCGAGTAGGGGCACTGTACC
Read 2 GAGTAGGGGCACTGTACCAGAGCCAGTAGCAT

CAGATGGAGAATGCATCGAGTAGGGGCACTGTACCAGAGCCAGTAGCAT

The reads and the genome are too long -> let's compress!

- No compression: CCCGGTTTAA
- Lossless compression: 3C 2G 3T 2A
- Lossy compression: CGTA

The reads and the genome are too long -> let's compress!

- No compression: CCCGGTTTAA
- Lossless compression: 3C 2G 3T 2A
- Lossy compression: CGTA

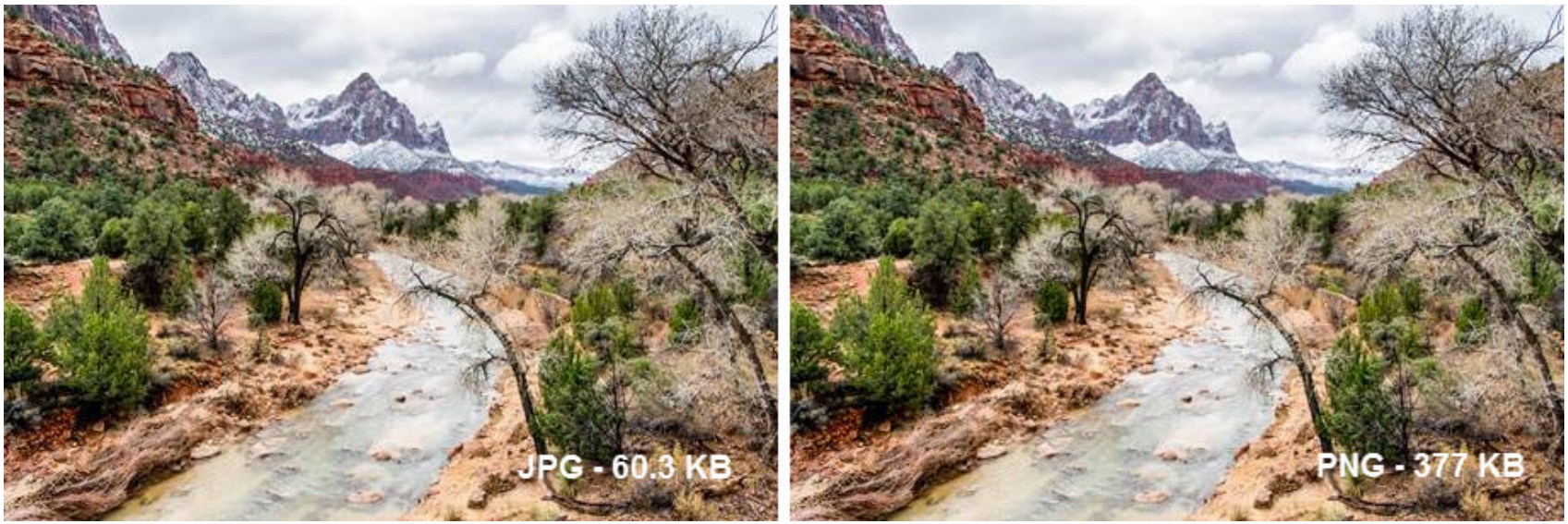
- We want **lossy compression**

The reads and the genome are too long -> let's compress!

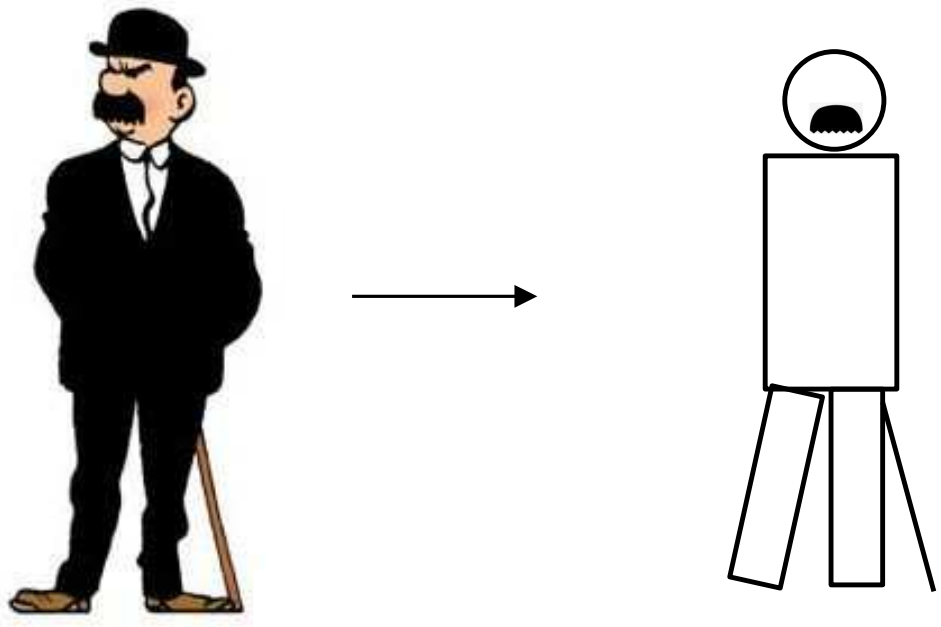
- No compression: CCCGGTTTAA
- Lossless compression: 3C 2G 3T 2A
- Lossy compression: CGTA

- We want **lossy compression**
- You all use lossy compression formats (kahoot)

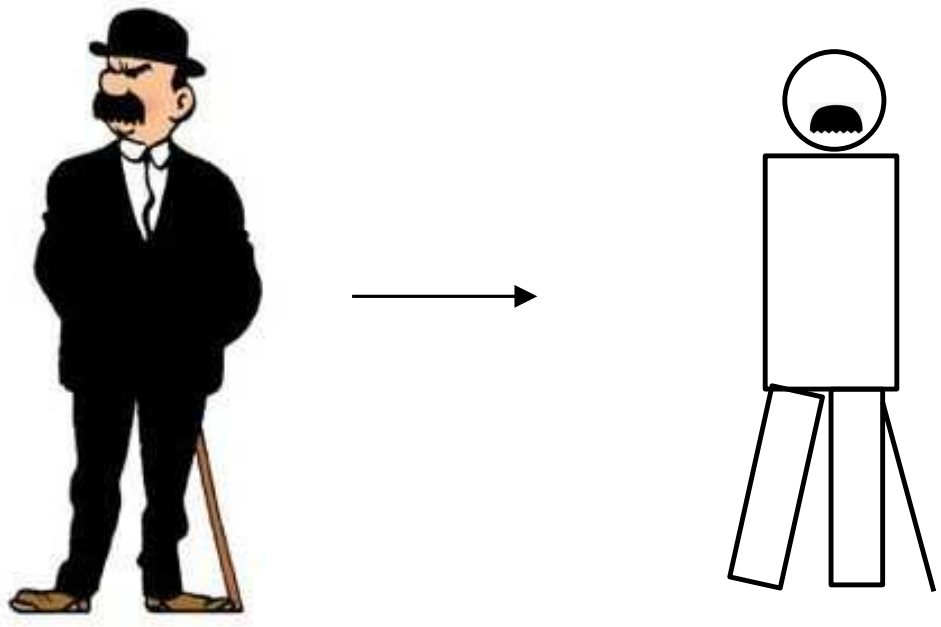
Lossy compression can be extremely good



Lossy data representation = **sketching**



Lossy data representation = **sketching**



...CGACGTATGCATCATGCAG...

?

My contribution: MSR sketching

Sketching rule: "take all letters after an A"

sequence CAGATGGAGAATGCATCGAGTAGGGGCACTGTACCAGAG

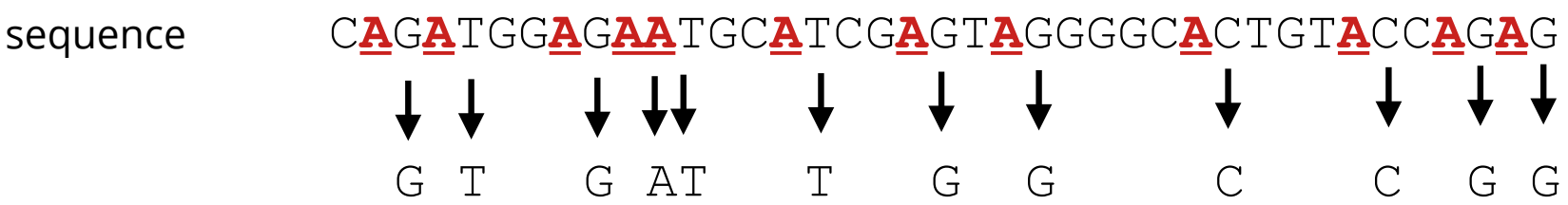
MSR sketching

Sketching rule: "take all letters after an A"

sequence CAGATGGAGAATGCATCGAGTAGGGGCACTGTACCAGAG

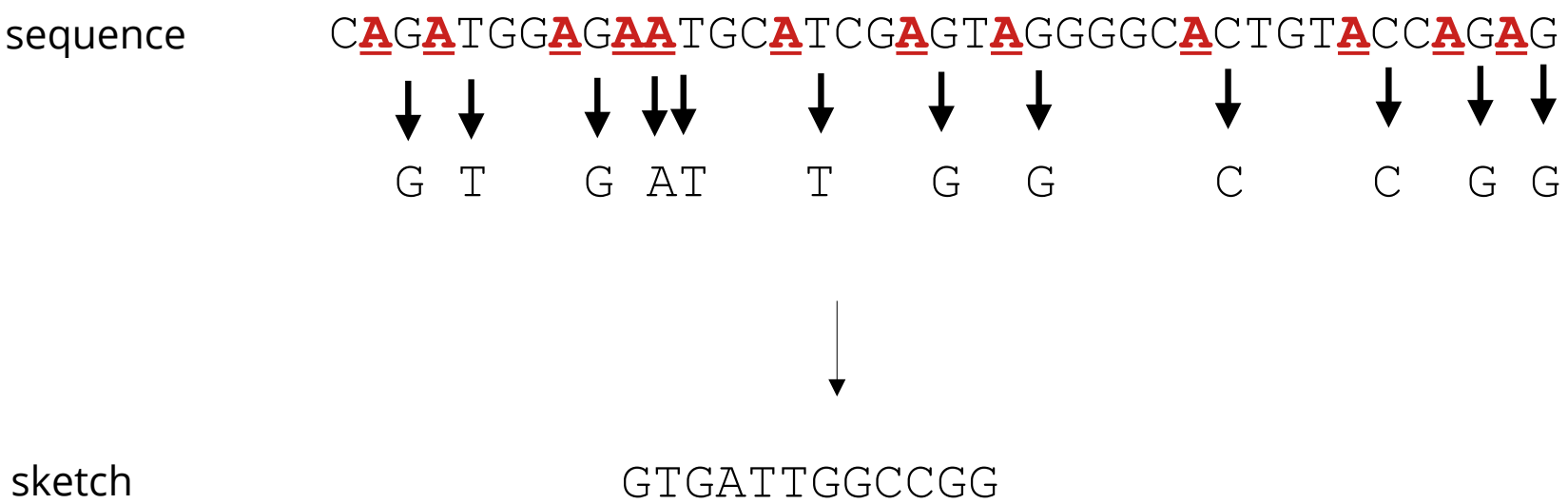
MSR sketching

Sketching rule: "take all letters after an A"



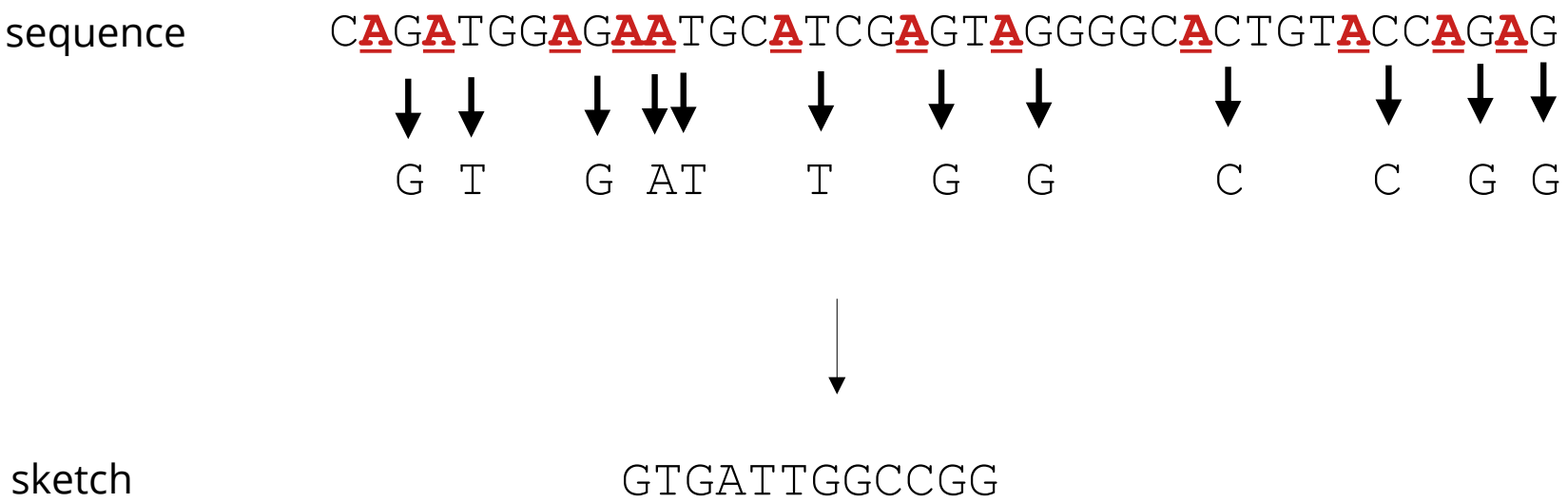
MSR sketching

Sketching rule: "take all letters after an A"



MSR sketching

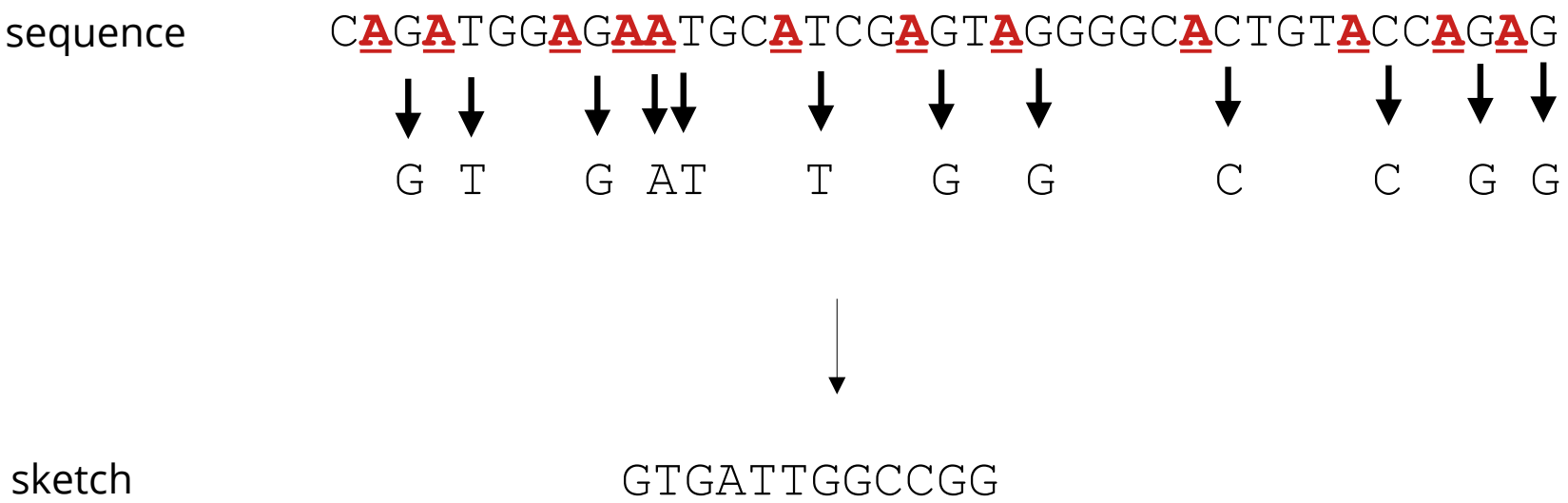
Sketching rule: "take all letters after an A"



- The sketch is 4 times smaller than the original sequence!

MSR sketching

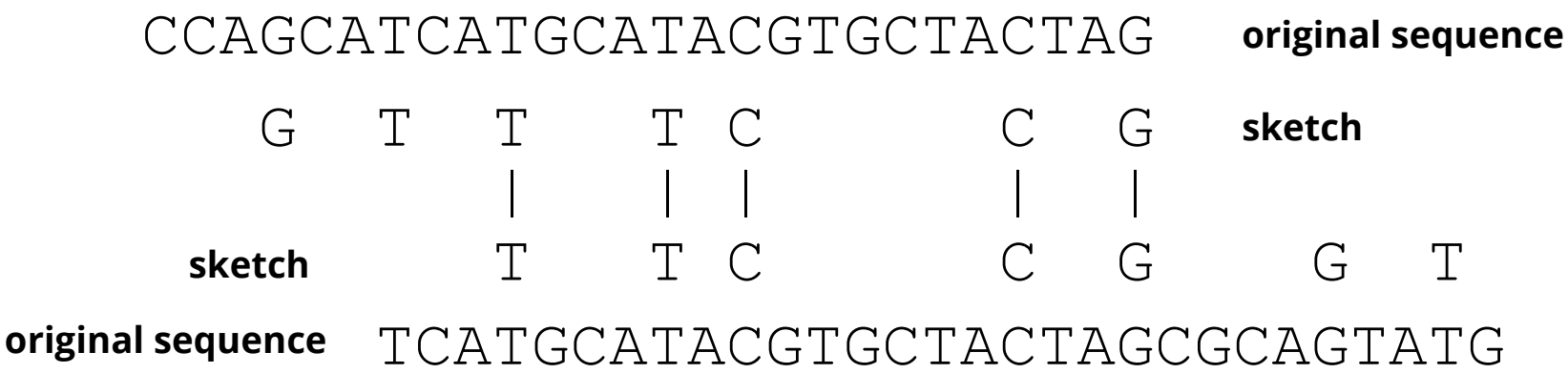
Sketching rule: "take all letters after an A"



- The sketch is 4 times smaller than the original sequence!
- I can choose another rule if I want

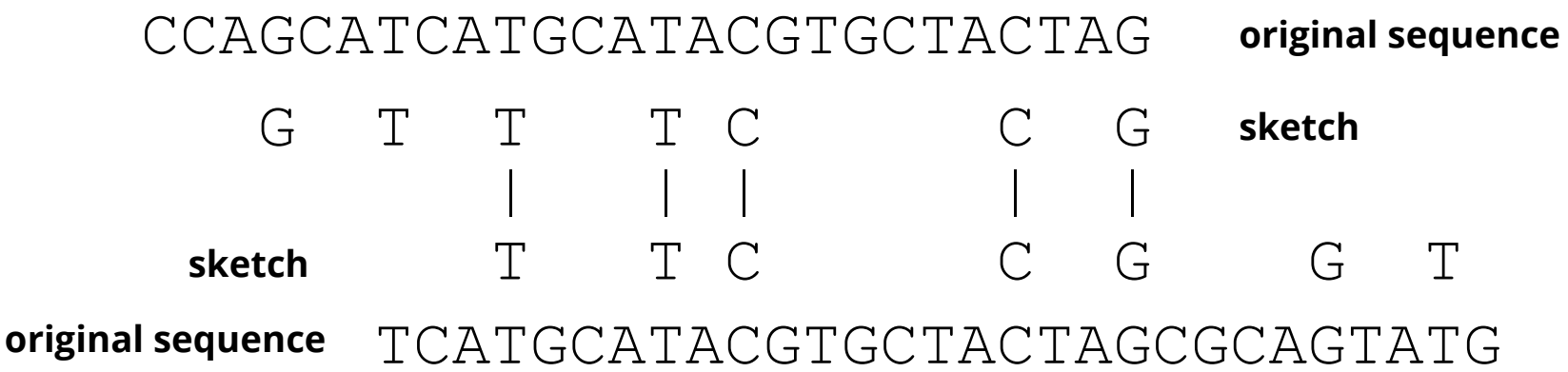
MSR=**Mapping-friendly** Sequence Reductions

- If two reads align, their sketchs align too



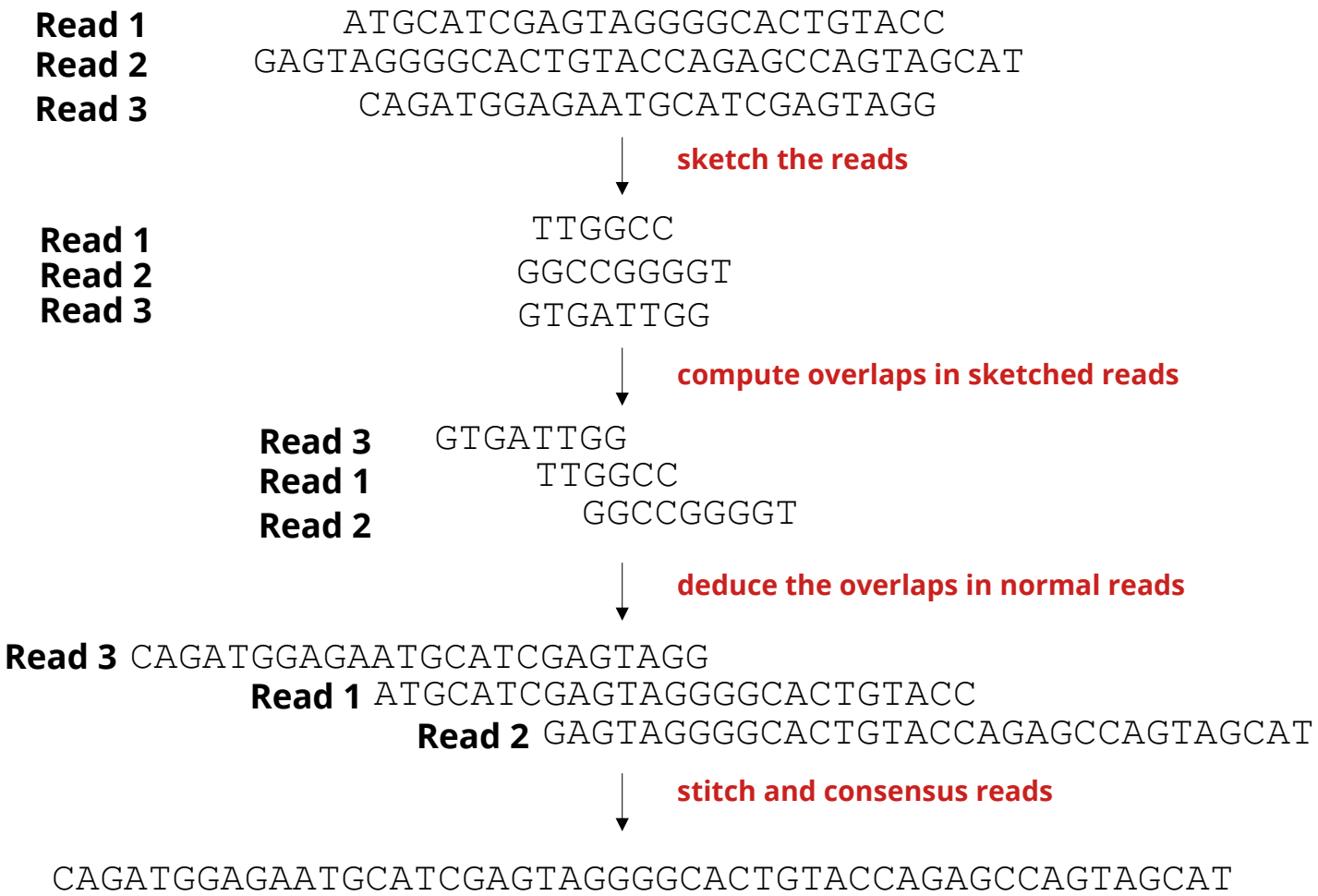
MSR=**Mapping-friendly** Sequence Reductions

- If two reads align, their sketchs align too

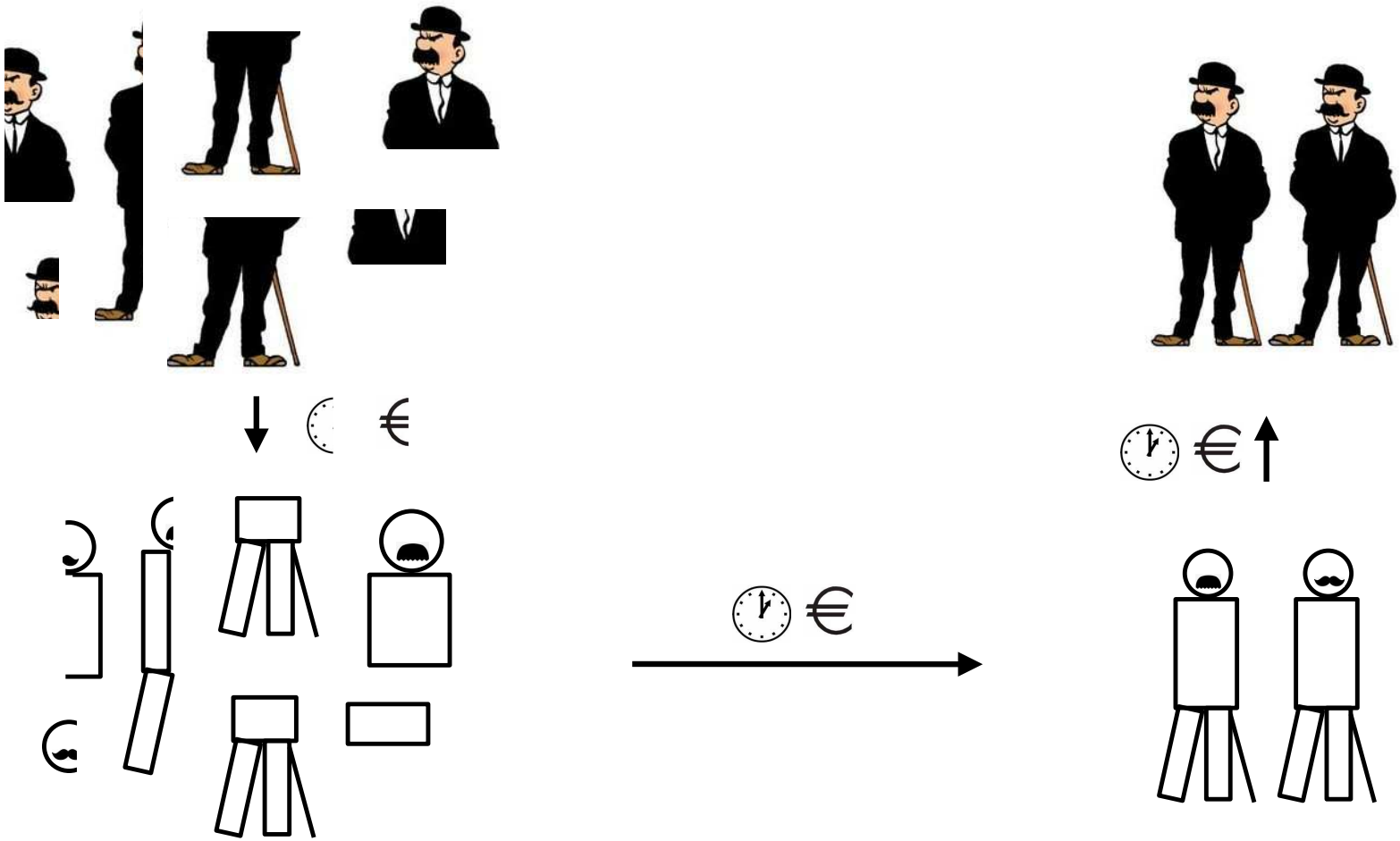


- Let's try to compute the overlaps again: Kahoot

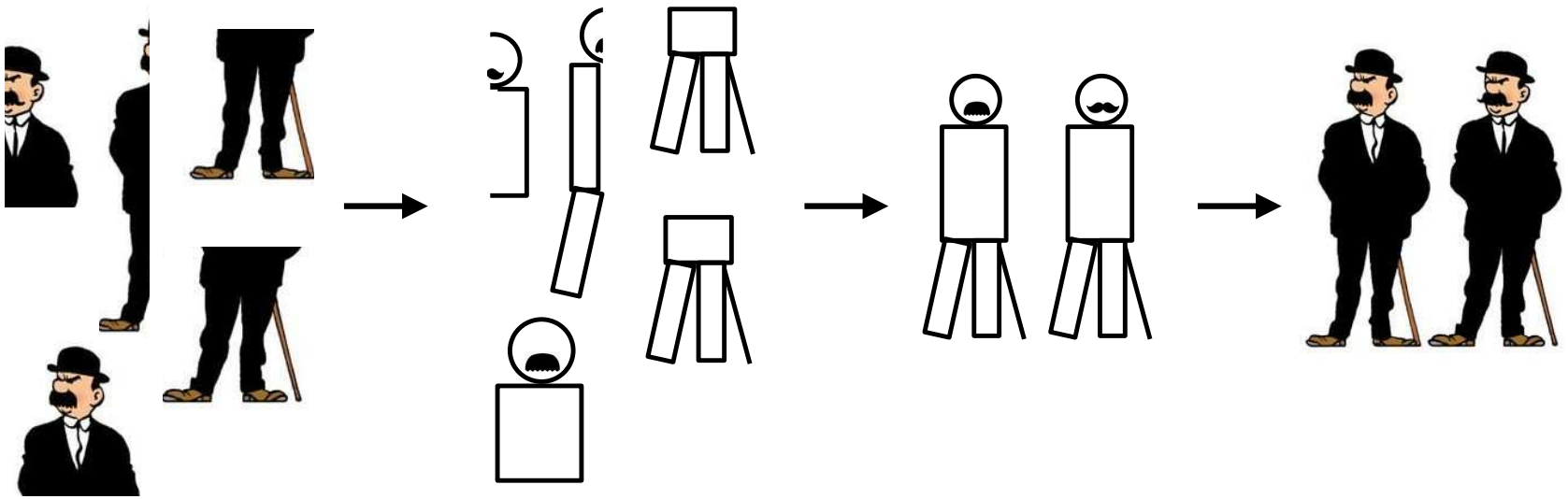
The new assembly strategy

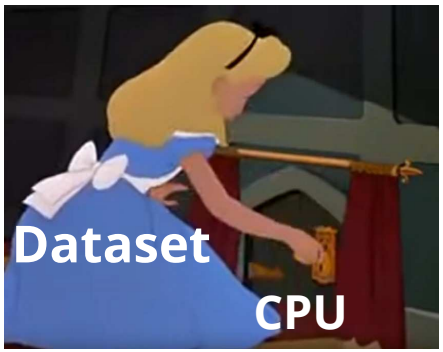


Solution for fast assembly: sketching the reads



Assembling using MSR sketches: the Alice assembler



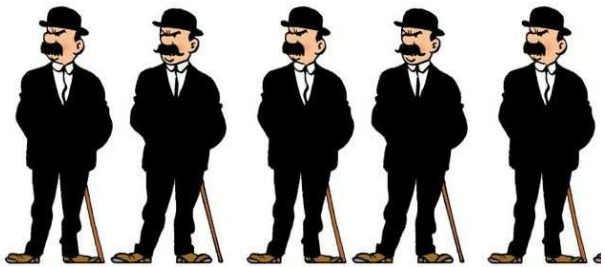



Drink-me potion


Eat-me cake

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	99.968
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

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

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	$\geq$ 60 days	$\geq$ 60 days	5h00

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

Was this thesis really useful?

Organism, microbiota...

sample

DNA extraction
& preparation

Sequencing

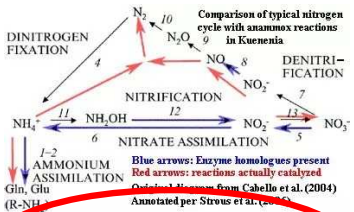
```
                                  CGTAGCTAGGAT
GTGCTAATCACGT                     AAAGCTAATCACTT
TGCGAGCGATCAG                     TGTCTGAAACCACA
CTCTGGGGTGACA
```

Long reads

Assembly

```
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...
```

Genomes or
amplicons



Understand things

What is the future of assembly?

Organism, microbiota...

sample

DNA extraction
& preparation

Sequencing

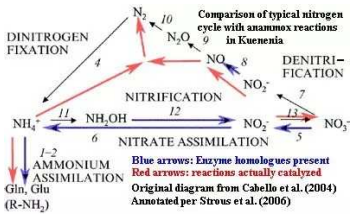
```
                                  CGTAGCTAGGAT
GTGCTAATCACGT                     AAAGCTAATCACTT
TGCGAGCGATCAG                     TGTCTGAAACCACA
CTCTGGGGTGACA
```

Long reads

Assembly

```
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...
```

Genomes or
amplicons

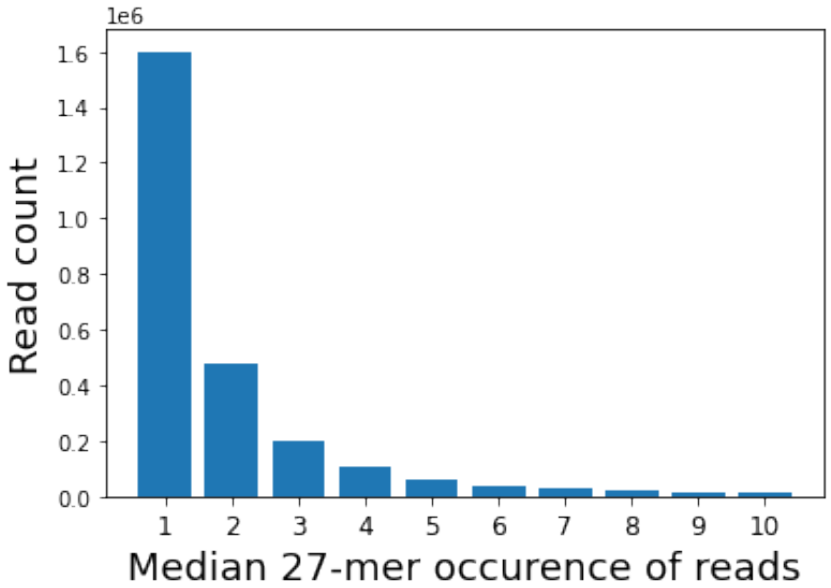


Understand things

All DNA is not captured by sequencing

Nicolas Maurice

- Example of a sequencing of the soil microbiota



Adapted from the work of Nicolas Maurice

- Low-coverage sequencing
- Missing DNA

What is the future of assembly?

Organism, microbiota...

sample

DNA extraction
& preparation

Sequencing

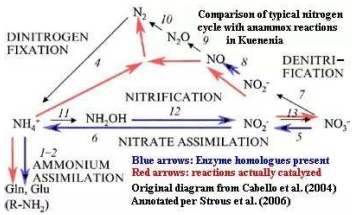
```
                                    CGTAGCTAGGAT
GTGCTAATCACGT                       AAAGCTAATCACTT
TGCGAGCGATCAG                       TGTCTGAAACCACA
CTCTGGGGTGACA
```

Long reads

Assembly

```
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...
```

Genomes or amplicons



Understand things

Conclusion: let's finish this kahoot

- Three general questions
- Double points