Solving the haplotyping problem with linear programming

Tam K.M. Truong[1], **Roland Faure**[1,2], Rumen Andonov[1]

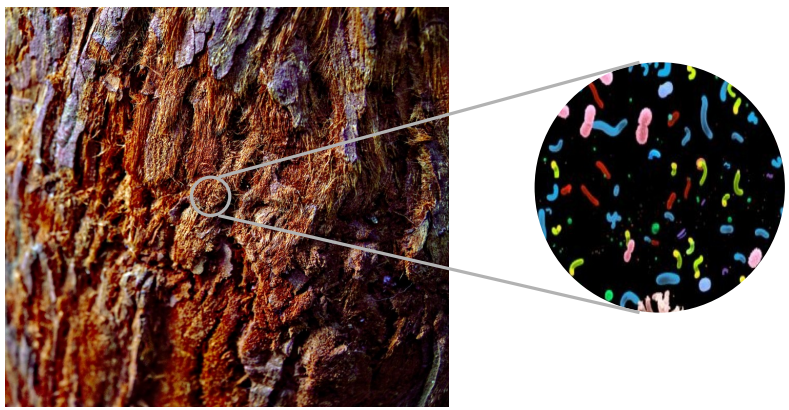
[1]Université de Rennes, IRISA, France

[2]Université libre de Bruxelles (ULB), Belgium

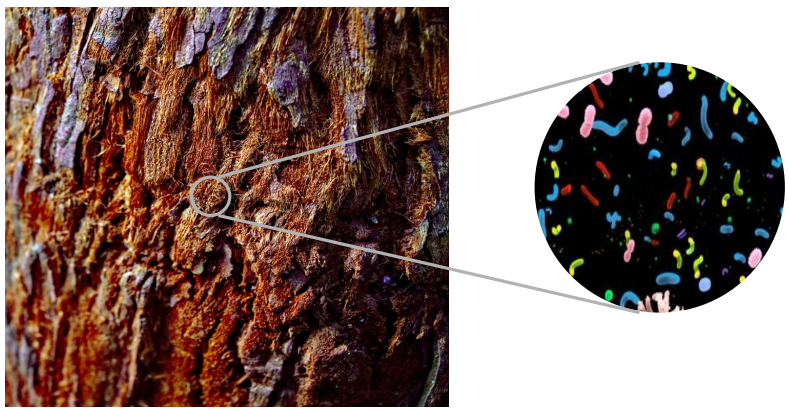
February 2023

The ultimate goal: understand ecosystems

The ultimate goal: understand ecosystems

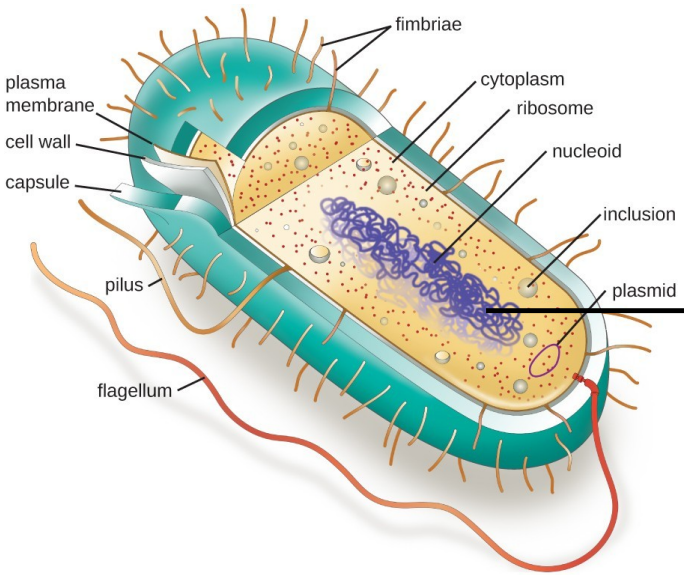


The ultimate goal: understand ecosystems



- Genome sequencing

Biology reminder



AGCAGCAGCGGAGAGGGTT

- DNA is the instruction manual of the bacteria
- Typically 5Mb long

Studying an ecosystem through sequencing

Ecosystem

Studying an ecosystem through sequencing

Ecosystem

DNA extraction

Studying an ecosystem through sequencing

Ecosystem

DNA extraction

Sequencing

Studying an ecosystem through sequencing

Ecosystem

DNA extraction

Sequencing

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

Long reads

Studying an ecosystem through sequencing

Ecosystem

DNA extraction

Sequencing

CGTAGCTAGGAT
GTGCTAATCACGT AAAGCTAATCACTT
TGCGAGCGATCAG TGTCTGAAACCACA
CTCTGGGGTGACA

Long reads

Assembly

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

Genomes

Studying an ecosystem through sequencing

Ecosystem

DNA extraction

Sequencing

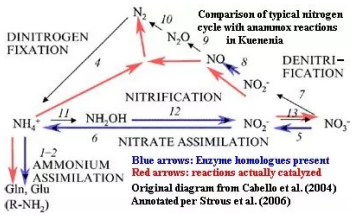
```
CGTAGCTAGGAT
GTGCTAATCACGT  AAAGCTAATCACTT
TGCGAGCGATCAG  TGTCTGAAACCACA
CTCTGGGGTGACA
```

Long reads

Assembly

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

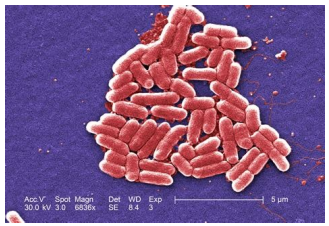
Genomes



Metabolisms & interactions

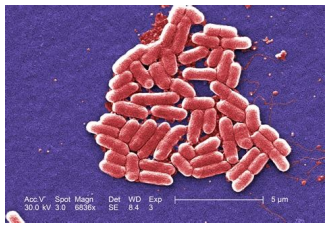
Strains are important

Escherichia coli K12



...ACGCTGAGGCAGCATGTGCCA...

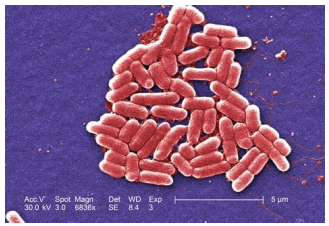
Escherichia coli Sakai



...GCGCTGAGGCAGCATGTGCGA...

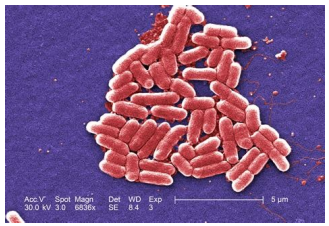
Strains are important

Escherichia coli K12



...ACGCTGAGGCAGCATGTGCCA...

Escherichia coli Sakai



...GCGCTGAGGCAGCATGTGCGA...

- "Knowledge gaps remain: strain diversity at a microscale and between species" R. Caldwell *et al.*, doi.org/10.1016/j.mib.2022.102222, 2022

Problem: differentiating strains

Ecosystem

DNA extraction

Sequencing

Strains get lost here

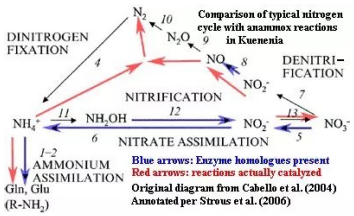
```
CGTAGCTAGGAT
GTGCTAATCACGT  AAAGCTAATCACTT
TGCGAGCGATCAG  TGTCTGAAACCACA
CTCTGGGGTGACA
```

Long reads

Assembly

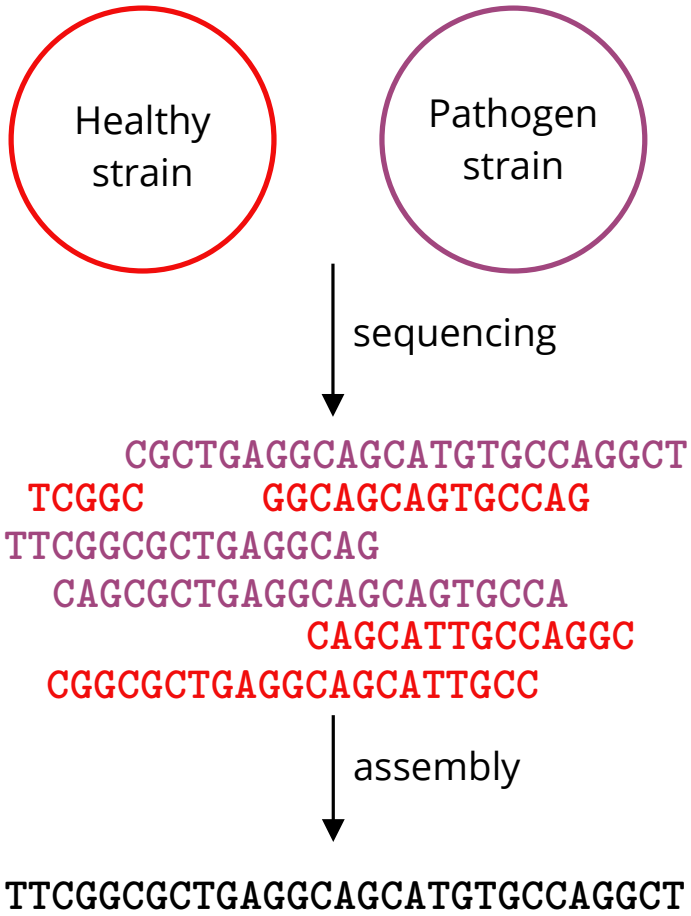
```
...AACGTGCGTCACGTAGTCGAGG...
...CGGCGCTGAGGCAGCAGTGCCA...
```

Genomes

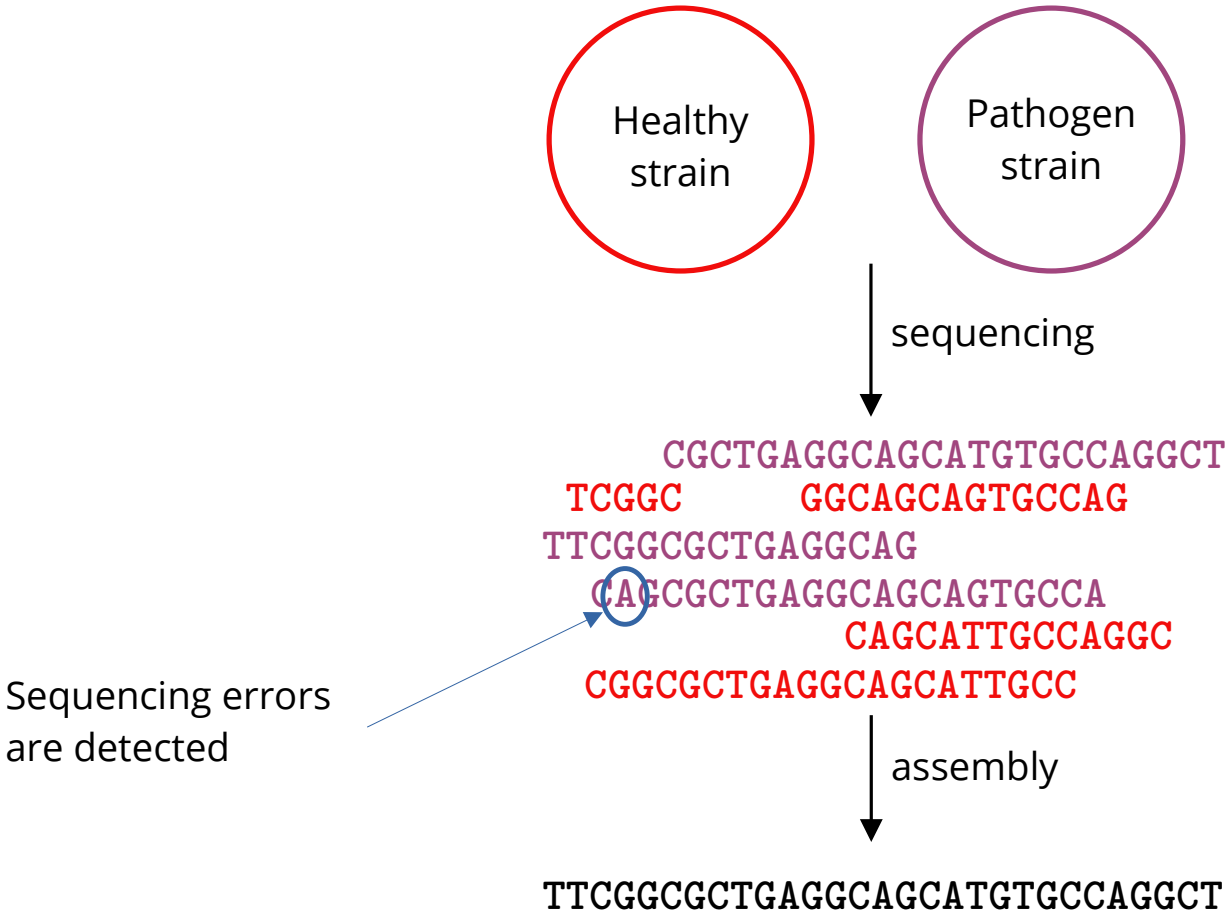


Metabolisms & interactions

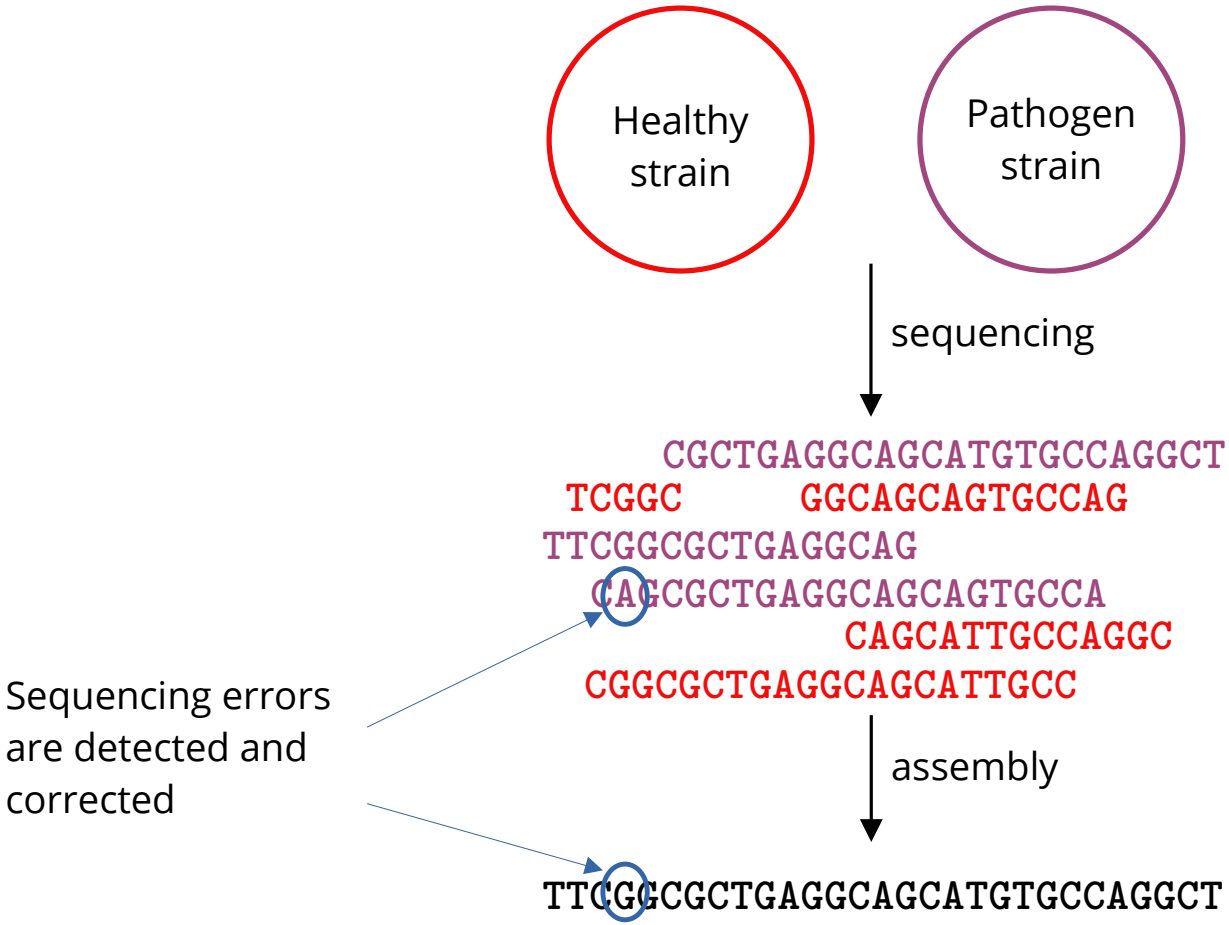
Problem: differentiating strains



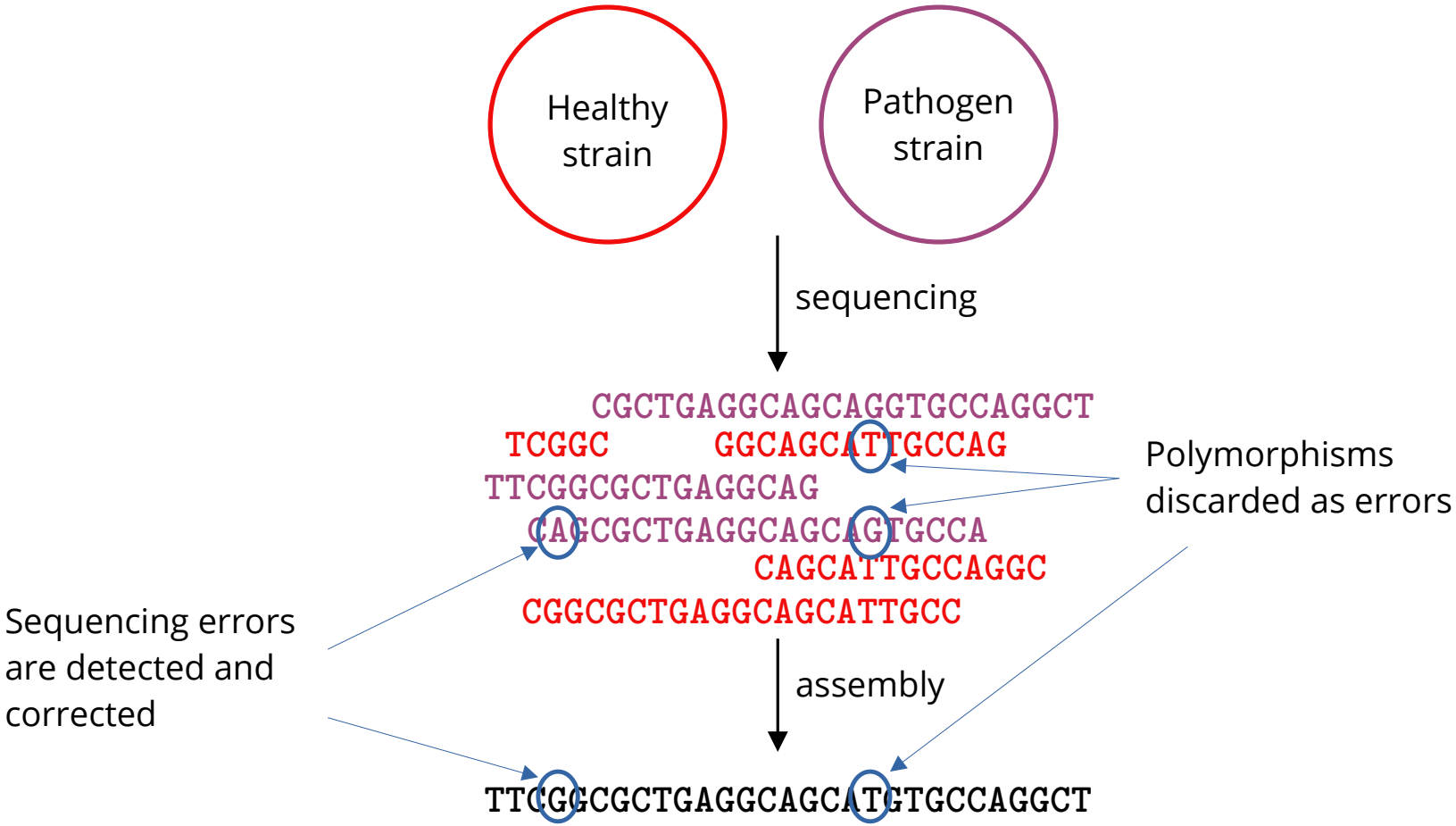
Problem: differentiating strains



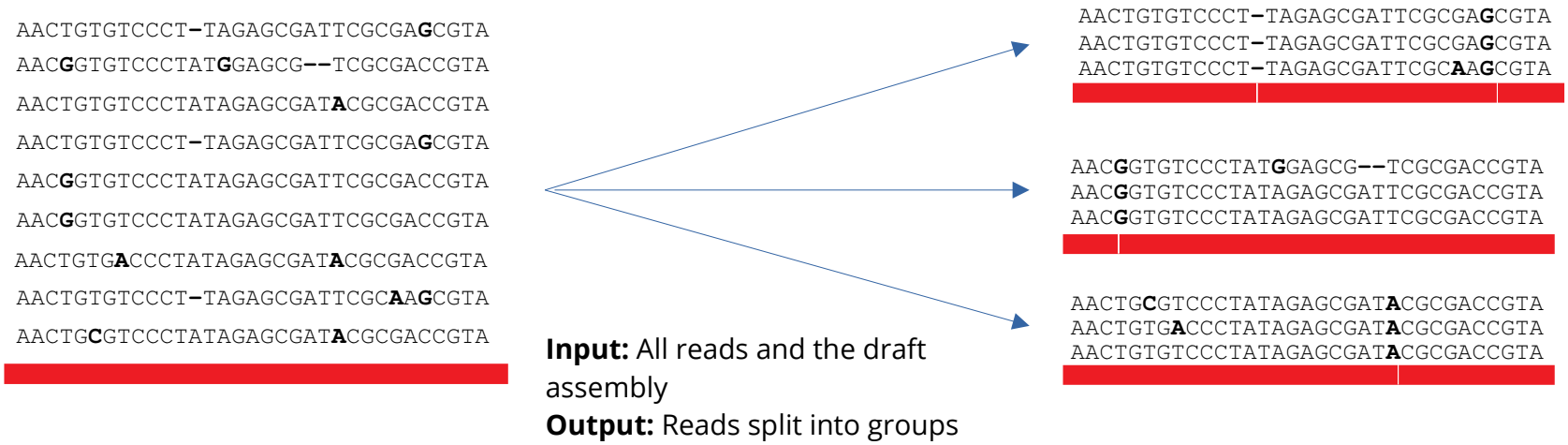
Problem: differentiating strains



Problem: differentiating strains



Goal



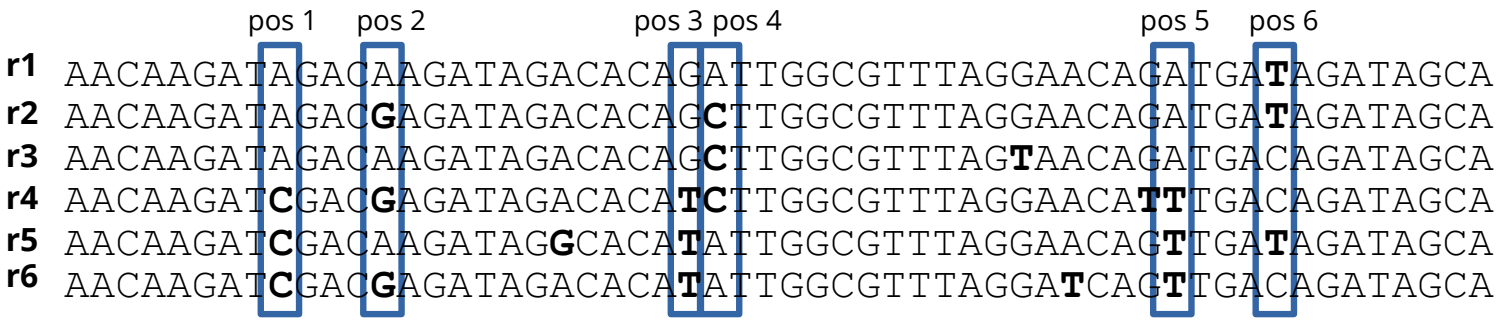
Input: All reads and the draft assembly
Output: Reads split into groups

Differentiating errors from signal

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

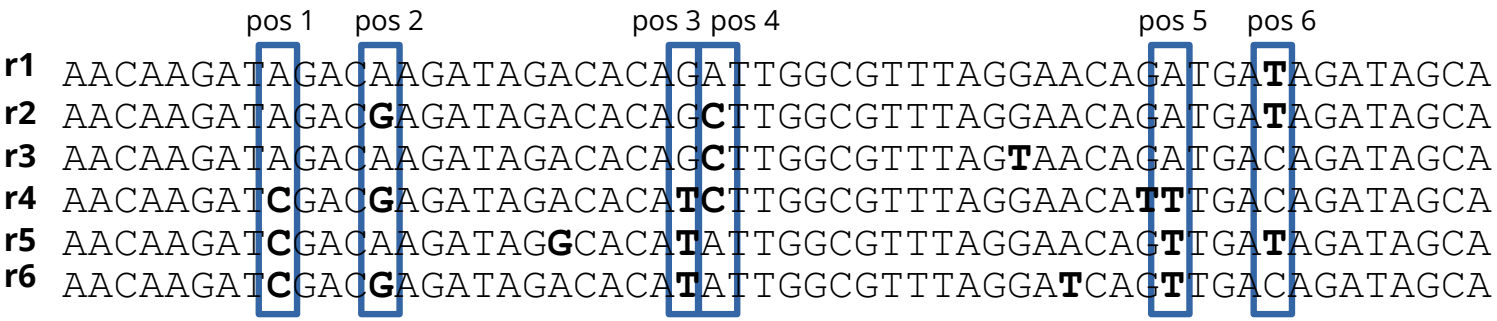
Differentiating errors from signal

The model



Positions with high diversity are selected

The model

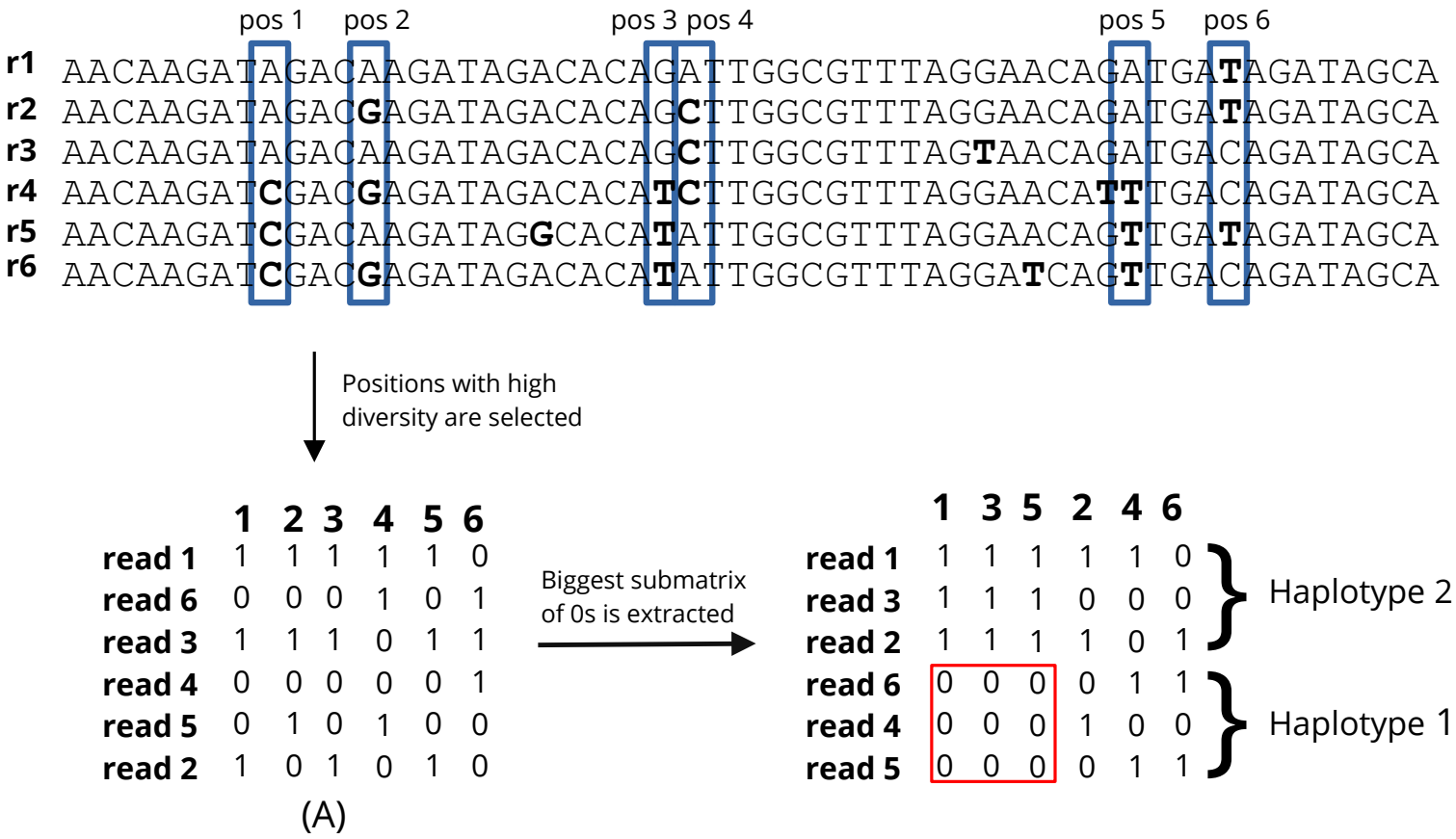


Positions with high diversity are selected

	1	2	3	4	5	6
read 1	1	1	1	1	1	0
read 6	0	0	0	1	0	1
read 3	1	1	1	0	1	1
read 4	0	0	0	0	0	1
read 5	0	1	0	1	0	0
read 2	1	0	1	0	1	0

(A)

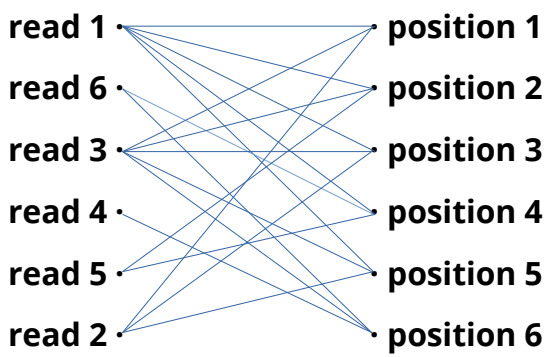
The model


pos 1
pos 2
pos 3 pos 4
pos 5
pos 6
r1 AACAAGATAGACAAGATAGACACAGATTGGCGTTTAGGAACAGATGATAGATAGCA
r2 AACAAGATAGACGAGATAGACACAGCTTGGCGTTTAGGAACAGATGATAGATAGCA
r3 AACAAGATAGACAAGATAGACACAGCTTGGCGTTTAGTAACAGATGACAGATAGCA
r4 AACAAGATCGACGAGATAGACACATCTTGGCGTTTAGGAACATTTGACAGATAGCA
r5 AACAAGATCGACAAGATAGGCACATATTGGCGTTTAGGAACAGTTGATAGATAGCA
r6 AACAAGATCGACGAGATAGACACATATTGGCGTTTAGGATCAGTTGACAGATAGCA
Positions with high diversity are selected
1 2 3 4 5 6
read 1 1 1 1 1 1 0
read 6 0 0 0 1 0 1
read 3 1 1 1 0 1 1
read 4 0 0 0 0 0 1
read 5 0 1 0 1 0 0
read 2 1 0 1 0 1 0
(A)
Biggest submatrix of 0s is extracted
1 3 5 2 4 6
read 1 1 1 1 1 1 0
read 3 1 1 1 0 0 0
read 2 1 1 1 1 0 1
read 6 0 0 0 0 1 1
read 4 0 0 0 1 0 0
read 5 0 0 0 0 1 1
Haplotype 2
Haplotype 1

From the matrix to the graph

	1	**2**	**3**	**4**	**5**	**6**
read 1	1	1	1	1	1	0
read 6	0	0	0	1	0	1
read 3	1	1	1	0	1	1
read 4	0	0	0	0	0	1
read 5	0	1	0	1	0	0
read 2	1	0	1	0	1	0

(A)



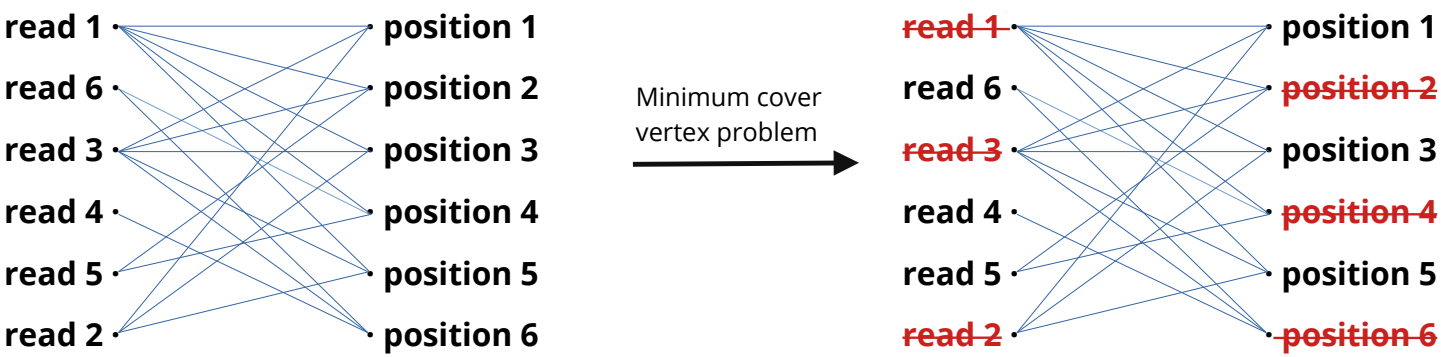
From the matrix to the graph

	1	2	3	4	5	6
read 1	1	1	1	1	1	0
read 6	0	0	0	1	0	1
read 3	1	1	1	0	1	1
read 4	0	0	0	0	0	1
read 5	0	1	0	1	0	0
read 2	1	0	1	0	1	0

(A)

Biggest submatrix of 0s is extracted →

	1	3	5	2	4	6
read 1	1	1	1	1	1	0
read 3	1	1	1	0	0	0
read 2	1	1	1	1	0	1
read 6	0	0	0	0	1	1
read 4	0	0	0	1	0	0
read 5	0	0	0	0	1	1



Formulation of the constraints

$$\min \quad \sum_{i \in U \cup V} \deg(i) v_i$$

$$v_i + v_j \geq 1, \ \forall (i,j) \in E$$

$$v_i \in \{0,1\}, \ \forall i \in U \cup V$$

Weighted minimal vertex cover problem

$$\min \quad \sum_{i \in U \cup V} \deg(i) v_i$$

$$1 - v_i \geq c_{ij}, \ \forall i \in U, \forall j \in V$$

$$1 - v_j \geq c_{ij}, \ \forall i \in U, \forall j \in V$$

$$1 - v_i - v_j \leq c_{ij}, \ \forall i \in U, \forall j \in V$$

$$\epsilon \times \sum_{i \in U} \sum_{j \in V} (1 - A_{i,j}) c_{ij} \geq \sum_{i \in U} \sum_{j \in V} A_{i,j} c_{ij}$$

$$v_i \in \{0,1\}, c_{ij} \in \{0,1\} \ \forall i \in U \cup V$$

Weighted minimal vertex cover problem, tolerating an ϵ ratio of 1 in the submatrix of 0s

Formulation of the constraints

$$\min \quad \sum_{i \in U \cup V} \deg(i) v_i$$

$$v_i + v_j \geq 1, \ \forall (i,j) \in E$$

$$v_i \in \{0,1\}, \ \forall i \in U \cup V$$

Weighted minimal vertex cover problem

$$\min \quad \sum_{i \in U \cup V} \deg(i) v_i$$

$$1 - v_i \geq c_{ij}, \ \forall i \in U, \forall j \in V$$

$$1 - v_j \geq c_{ij}, \ \forall i \in U, \forall j \in V$$

$$1 - v_i - v_j \leq c_{ij}, \ \forall i \in U, \forall j \in V$$

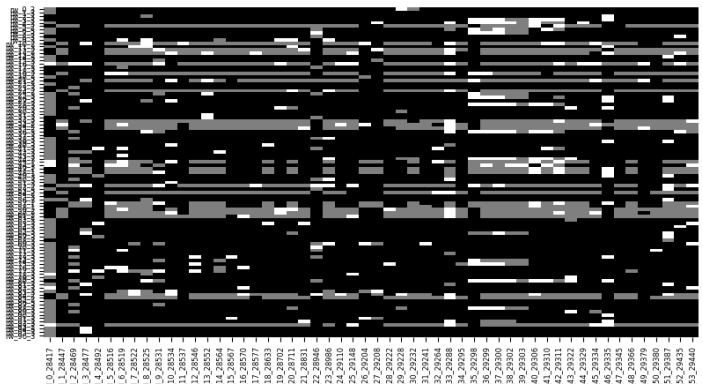
$$\epsilon \times \sum_{i \in U} \sum_{j \in V} (1 - A_{i,j}) c_{ij} \geq \sum_{i \in U} \sum_{j \in V} A_{i,j} c_{ij}$$

$$v_i \in \{0,1\}, c_{ij} \in \{0,1\} \ \forall i \in U \cup V$$

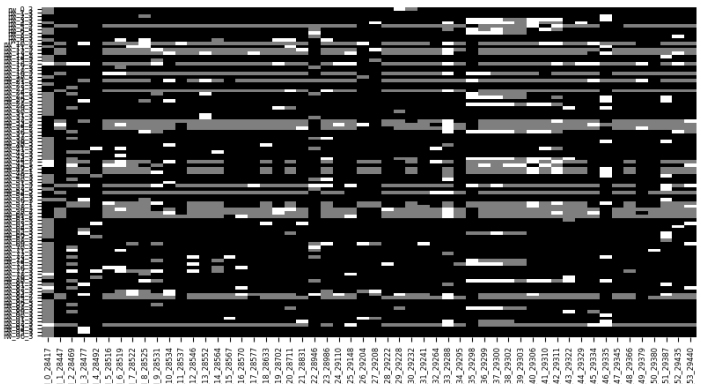
Weighted minimal vertex cover problem, tolerating an ϵ ratio of 1 in the submatrix of 0s

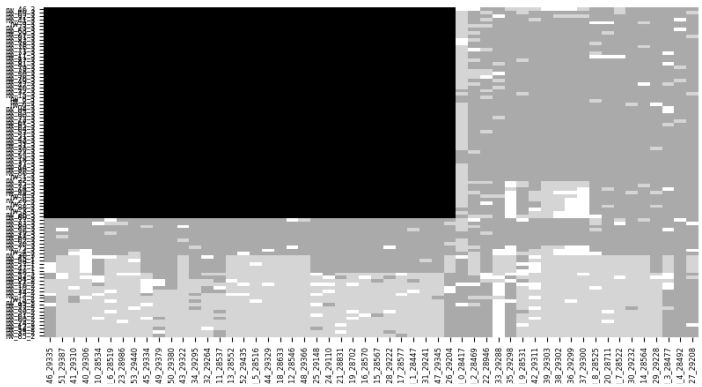
- Run through Gurobi Solver

Result on a mix of 3 *E. coli*

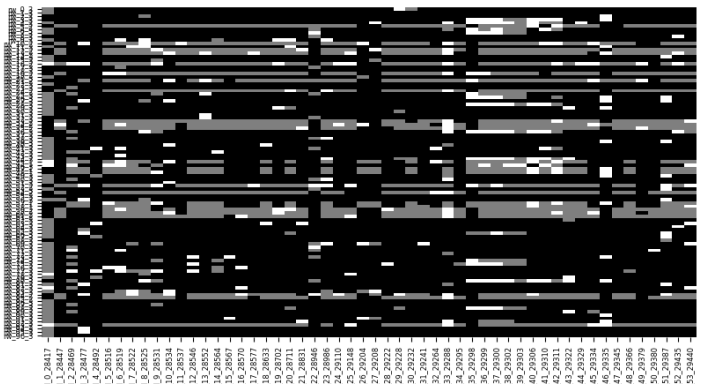


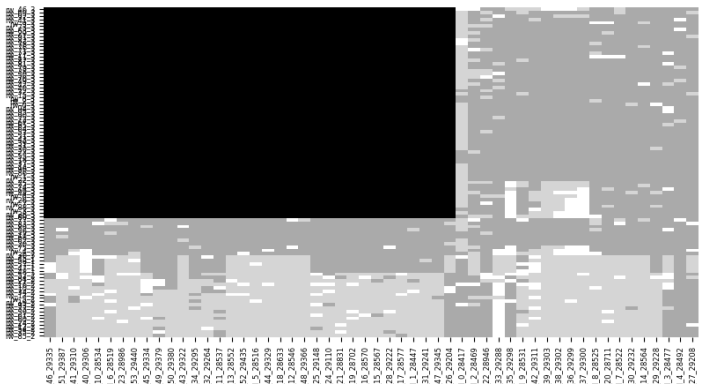
Result on a mix of 3 *E. coli*





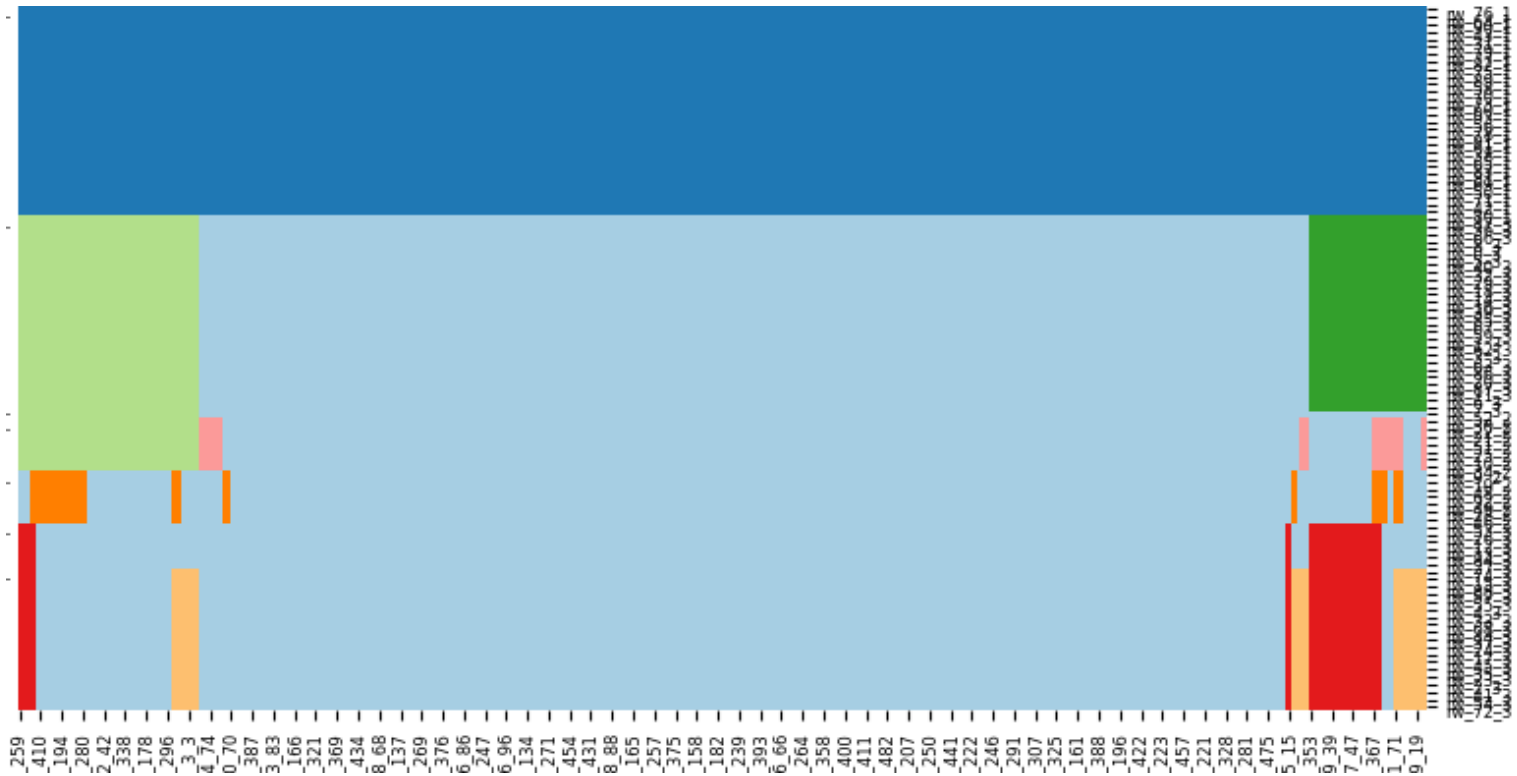
Result on a mix of 3 *E. coli*





- Succesfully separated two haplotypes from the reference
- Some false negative but no false positive

Separating all the haplotypes



- Iteratively apply the method on remaining matrix
- The difficult part is knowing when to stop separating

Conclusion

- This is the first formulation of the haplotyping problem as a linear programming problem
- Not yet competitive with existing heuristics

Acknowledgments

- Tam K.M. Truong (implementation), Rumen Andonov (supervision) and Dominique Lavenier (original idea)