Haplotype inference and applications

Olivier DELANEAU

Chaire de bioinformatique

Conservatoire des Arts et Métiers

olivier.delaneau@gmail.com

Cnam
CONSERVATOIRE NATIONAL DES ARTS ET METIERS

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

V. Comparison of accuracy

VI. Application 1 : haplotype association tests

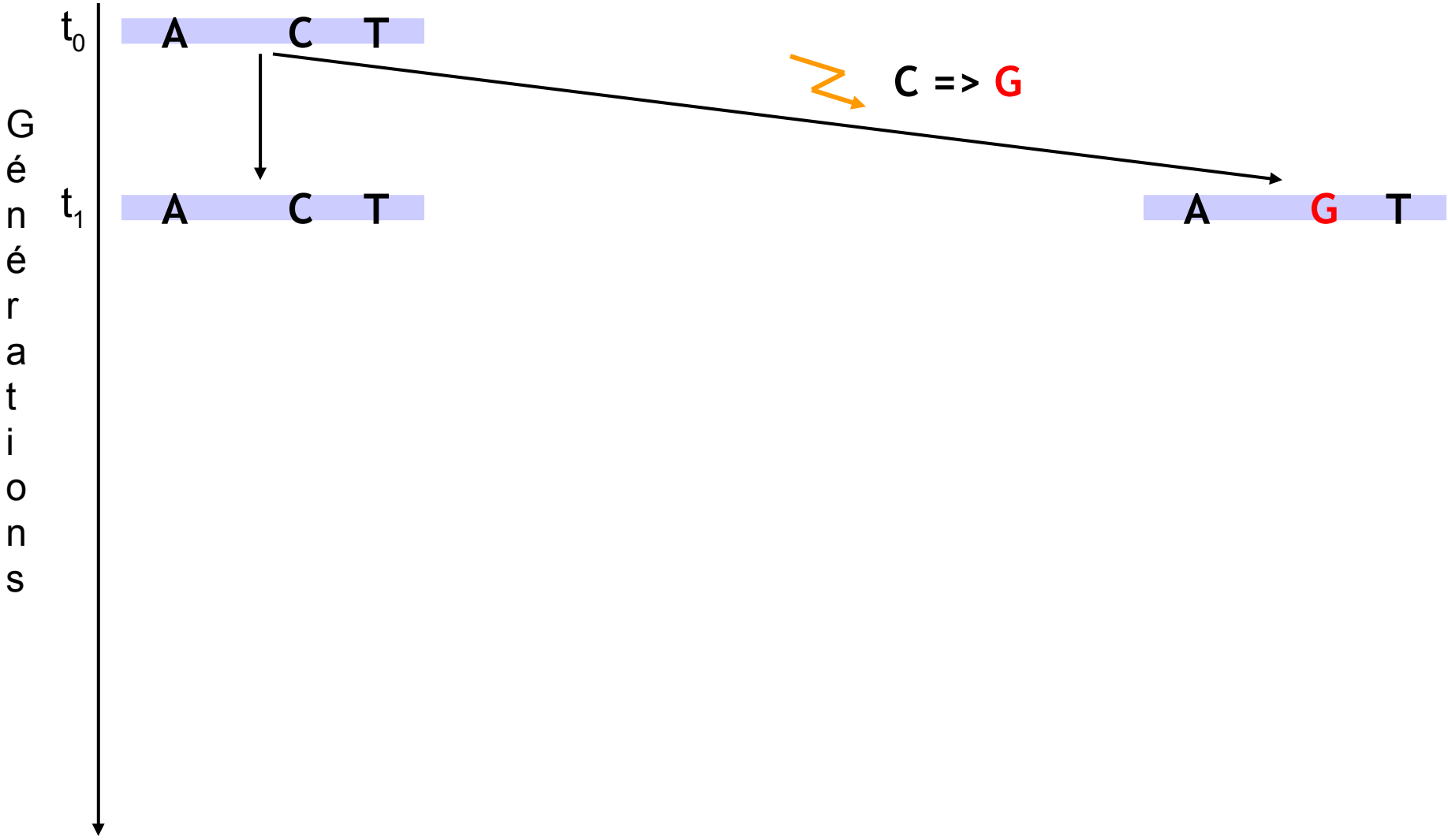
VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

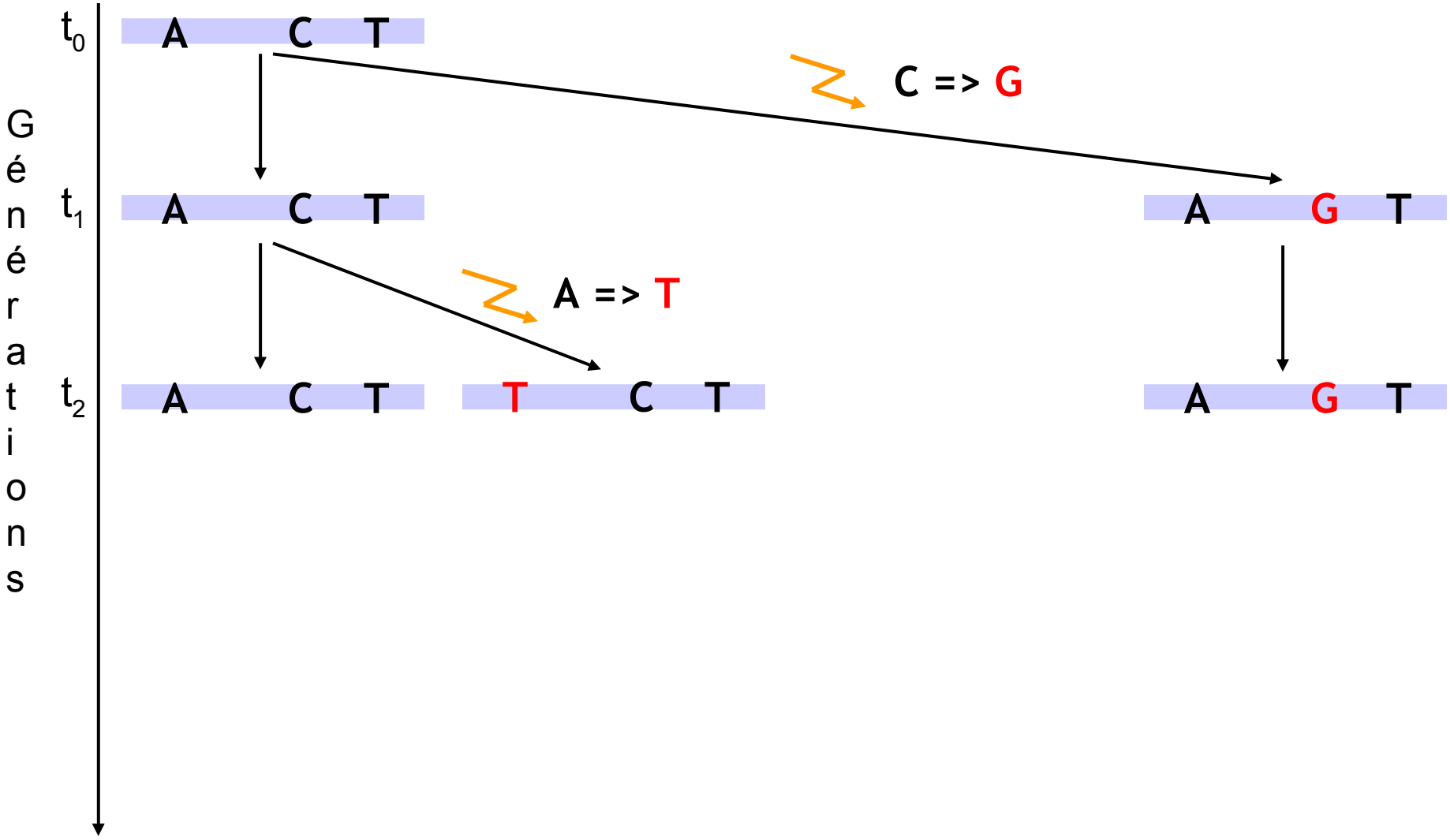
HAPLOTYPES GENESIS

A C T

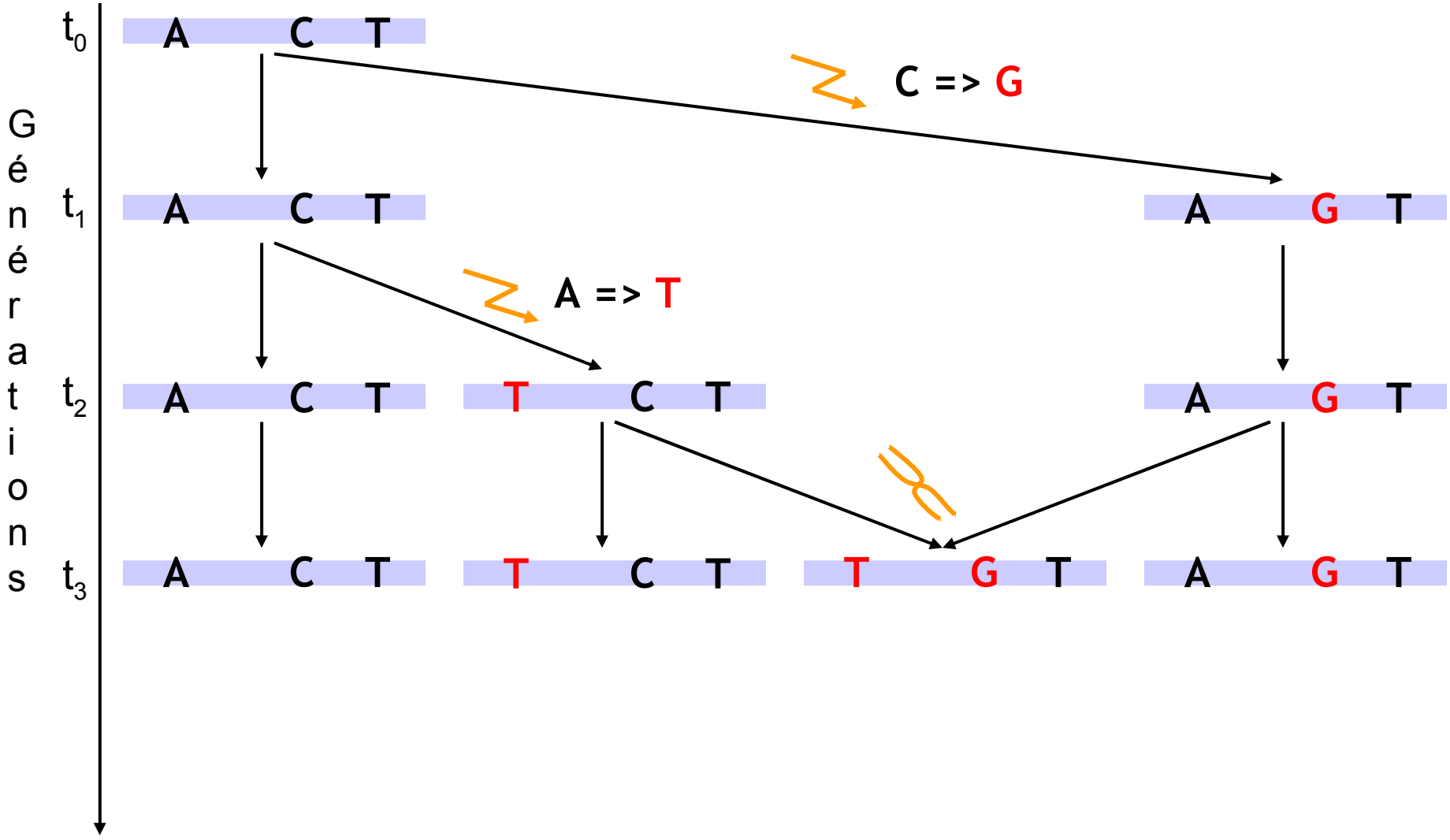
HAPLOTYPES GENESIS



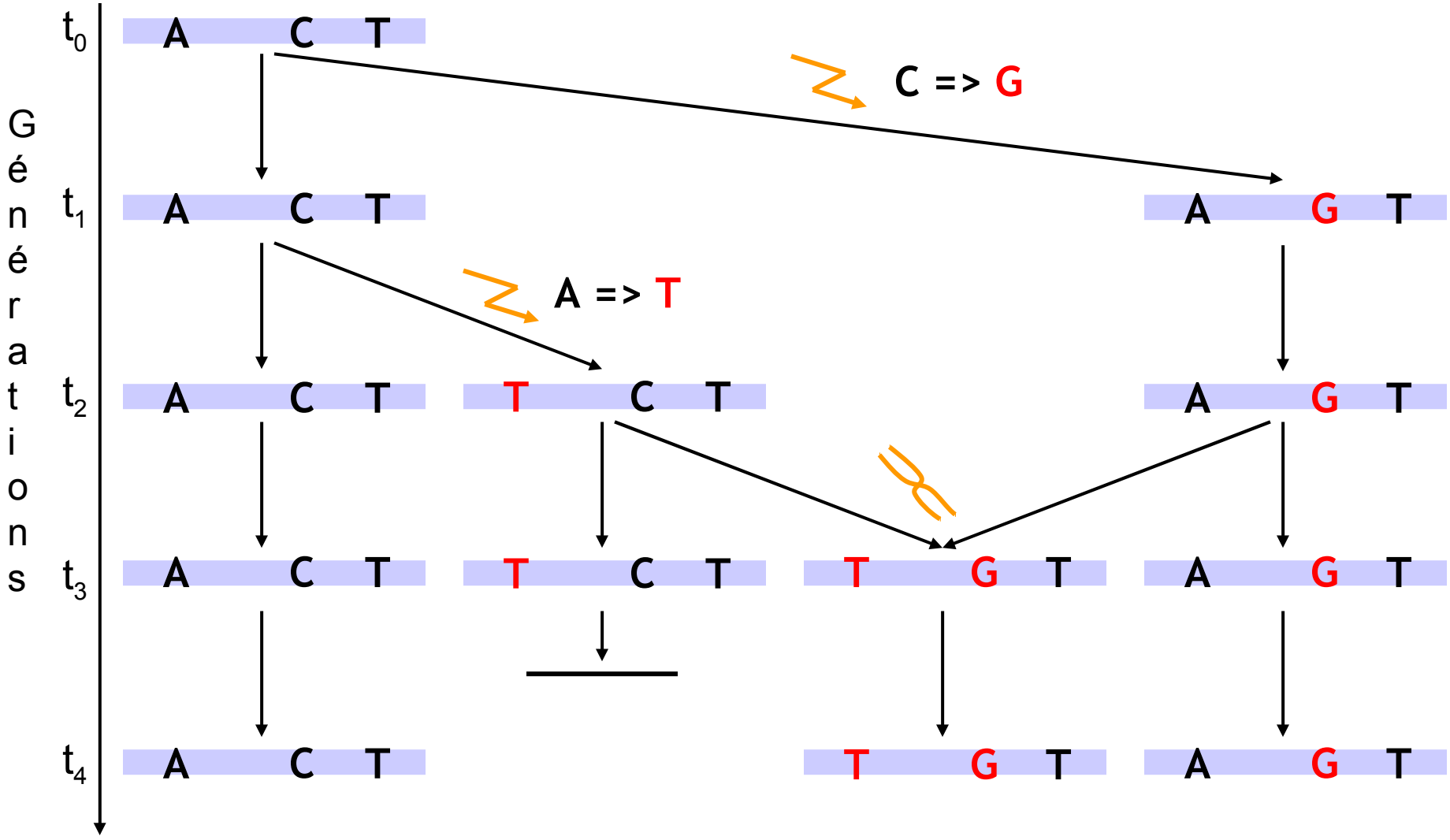
HAPLOTYPES GENESIS



HAPLOTYPES GENESIS

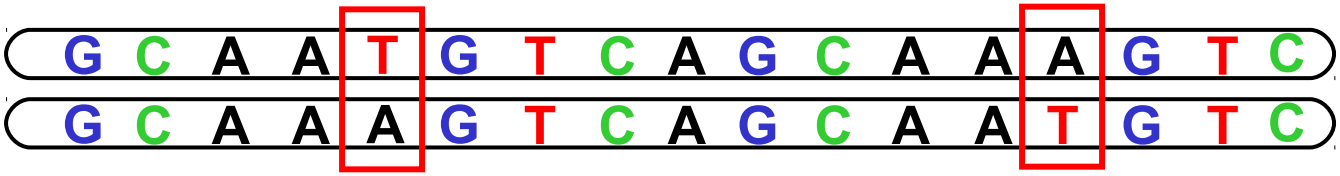


HAPLOTYPES GENESIS



PROBLEMATIC

Pair of chromosomes of an individual :



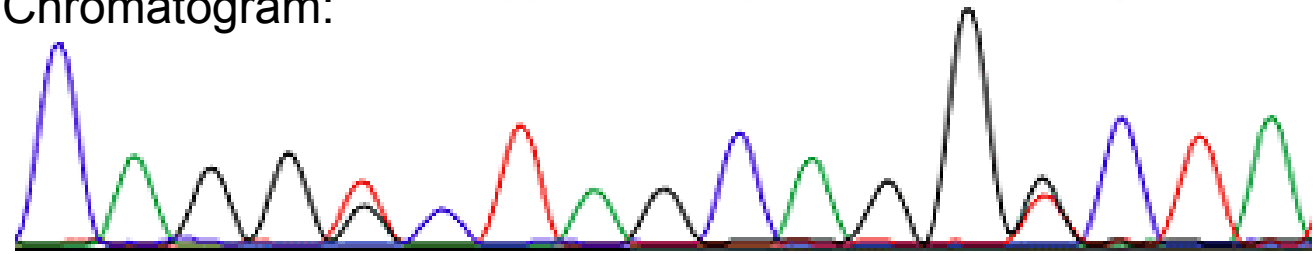
PROBLEMATIC

Pair of chromosomes of an individual :

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

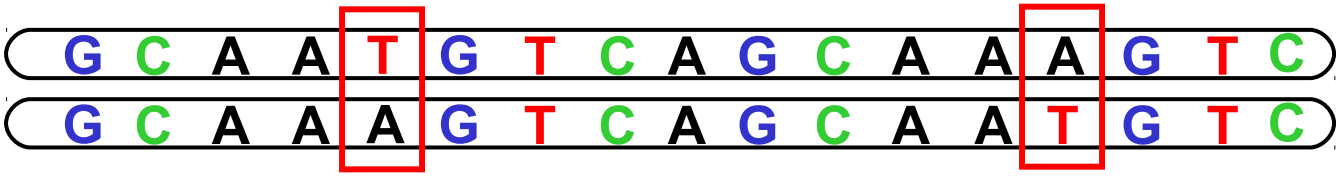
G C A A A G T C A G C A A T G T C

Chromatogram:

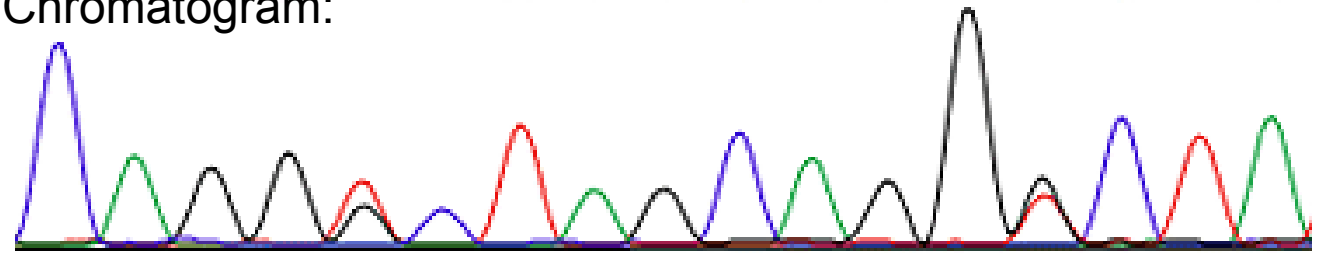


PROBLEMATIC

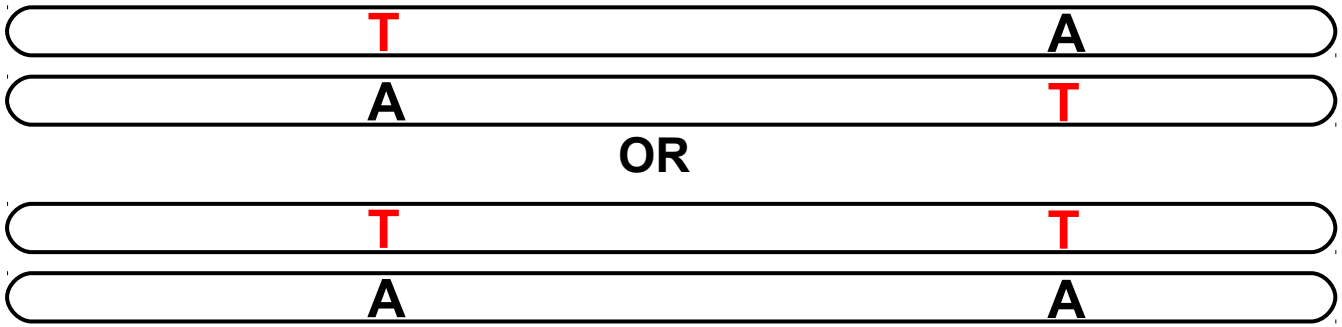
Pair of chromosomes of an individual :



Chromatogram:

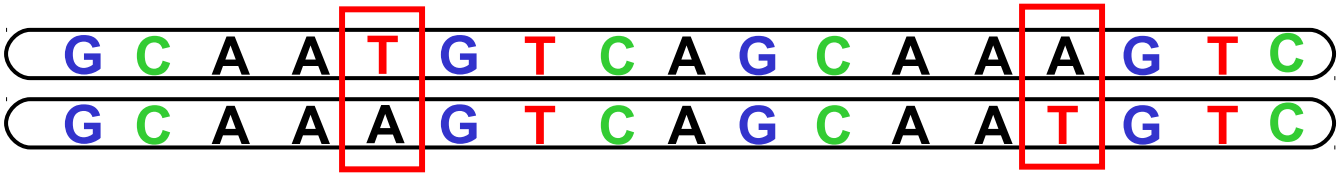


What is the true haplotype' pair ?

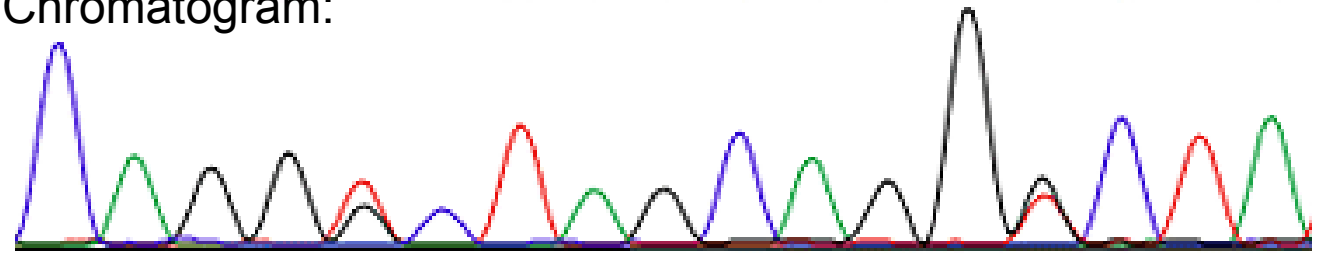


PROBLEMATIC

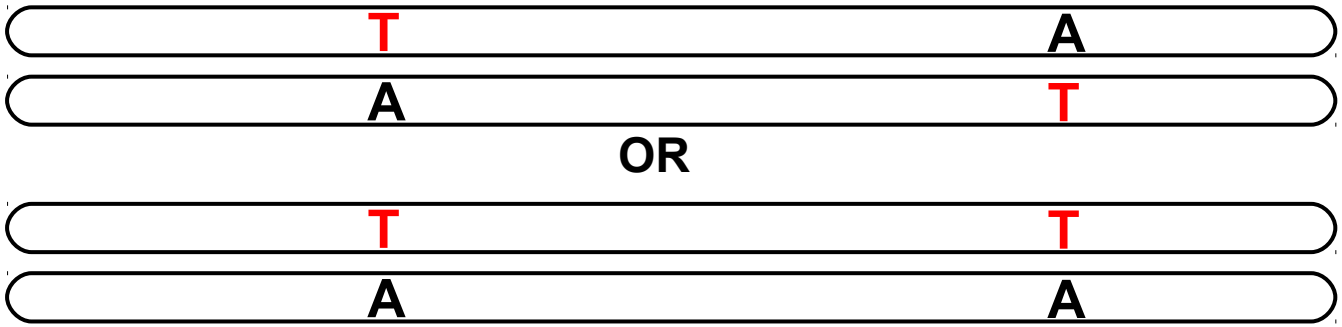
Pair of chromosomes of an individual :



Chromatogram:

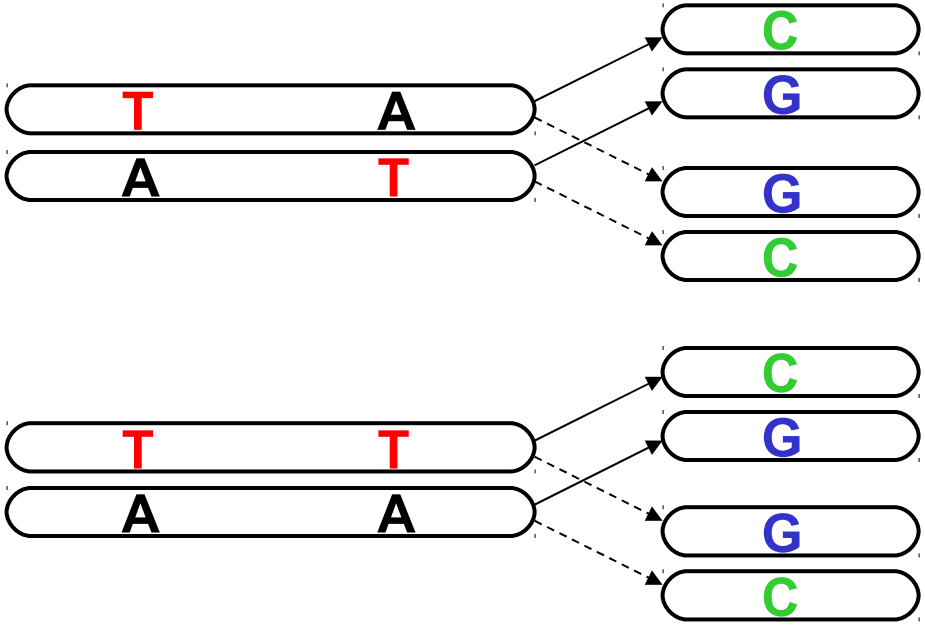


What is the true haplotype' pair ?

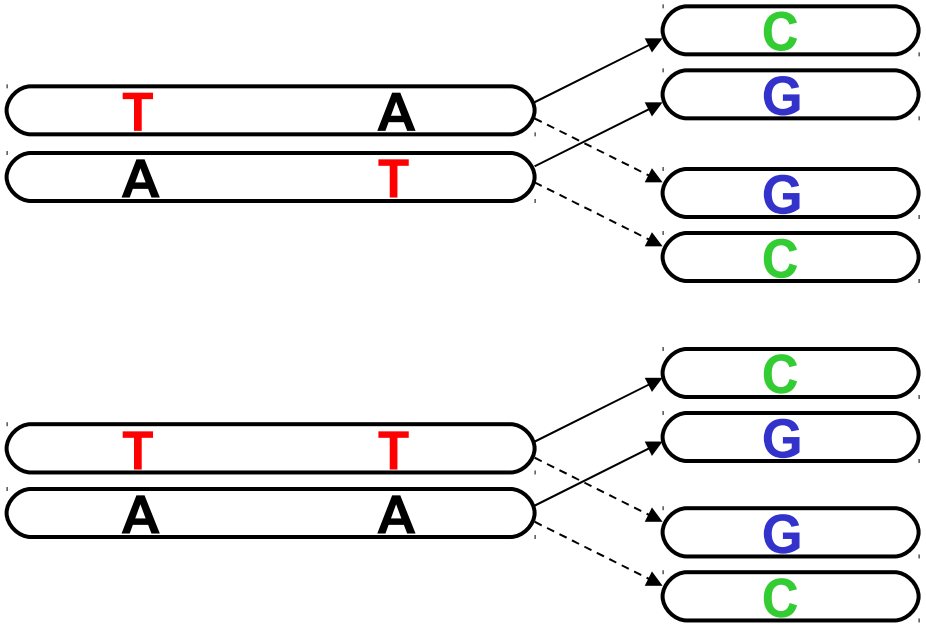


Haplotype inference

NUMBER OF POSSIBLE HAPLOTYPES



NUMBER OF POSSIBLE HAPLOTYPES

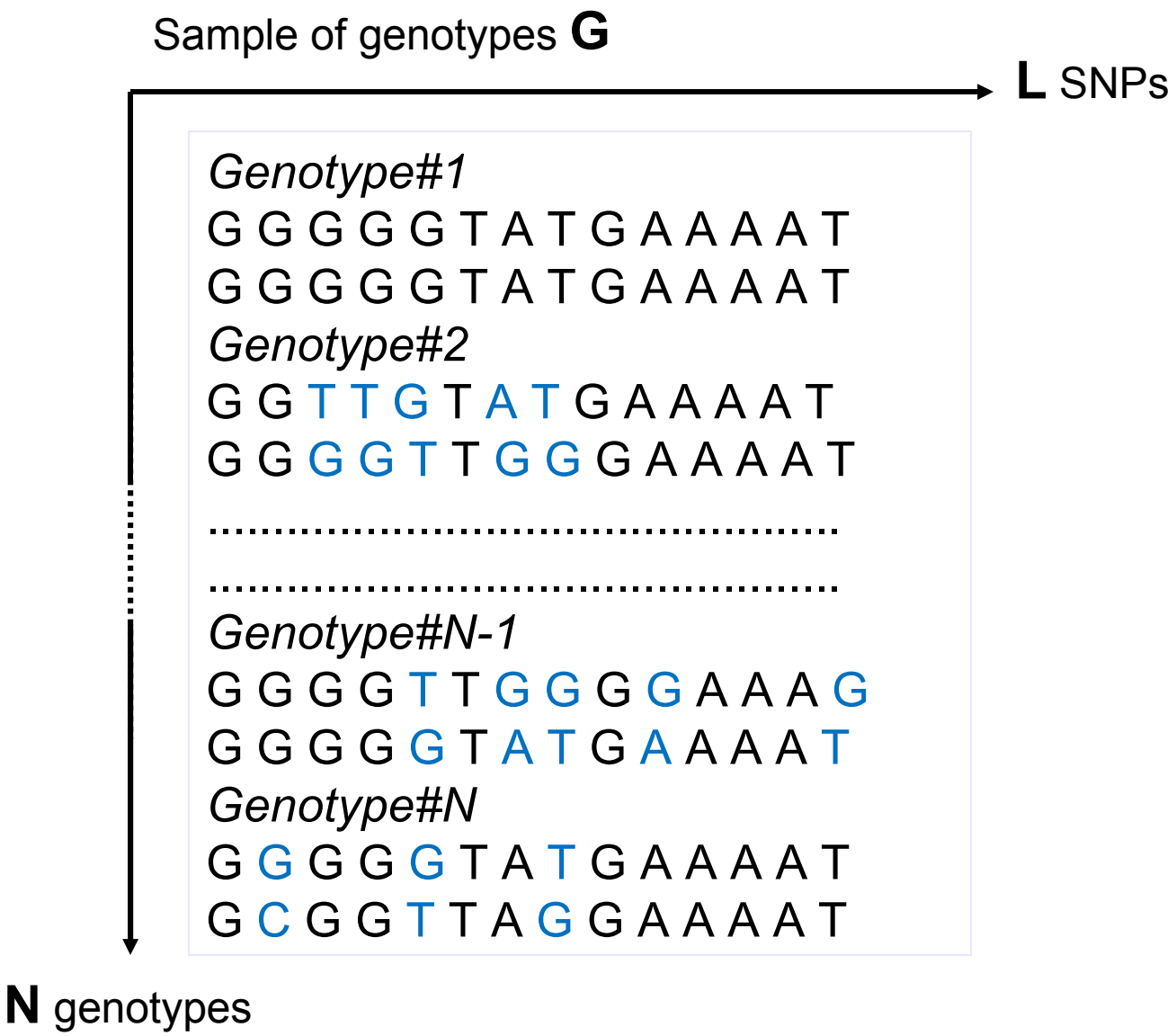


For a genotype with ***s*** heterozygous SNPs:

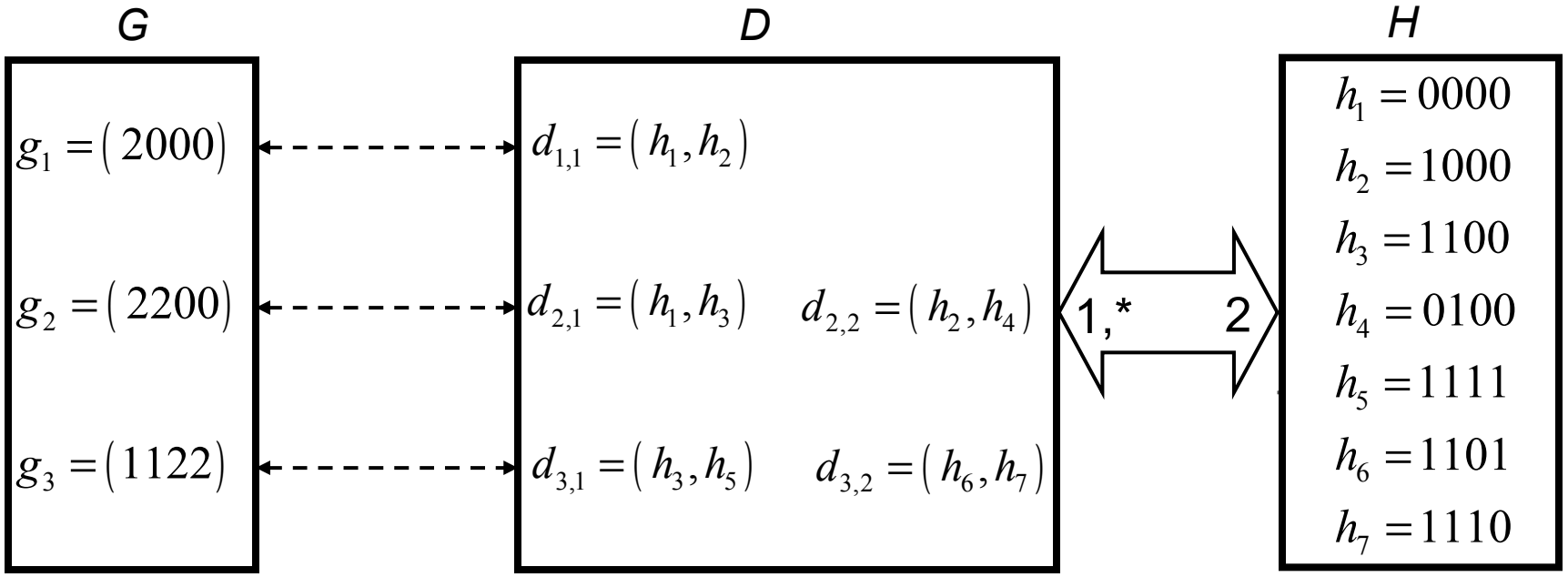
1. 2^{s-1} possible haplotype' pairs.

2. 2^{s} possible haplotypes.

EXAMPLE AND NOTATIONS



HAPLOTYPE SPACE



Genotype :

0 : homozygous wild allele (0 / 0)

1 : homozygous mutant allele (1 / 1)

2 : heterozygous (0 / 1)

Haplotype :

0 : wild allele

1 : mutant allele

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

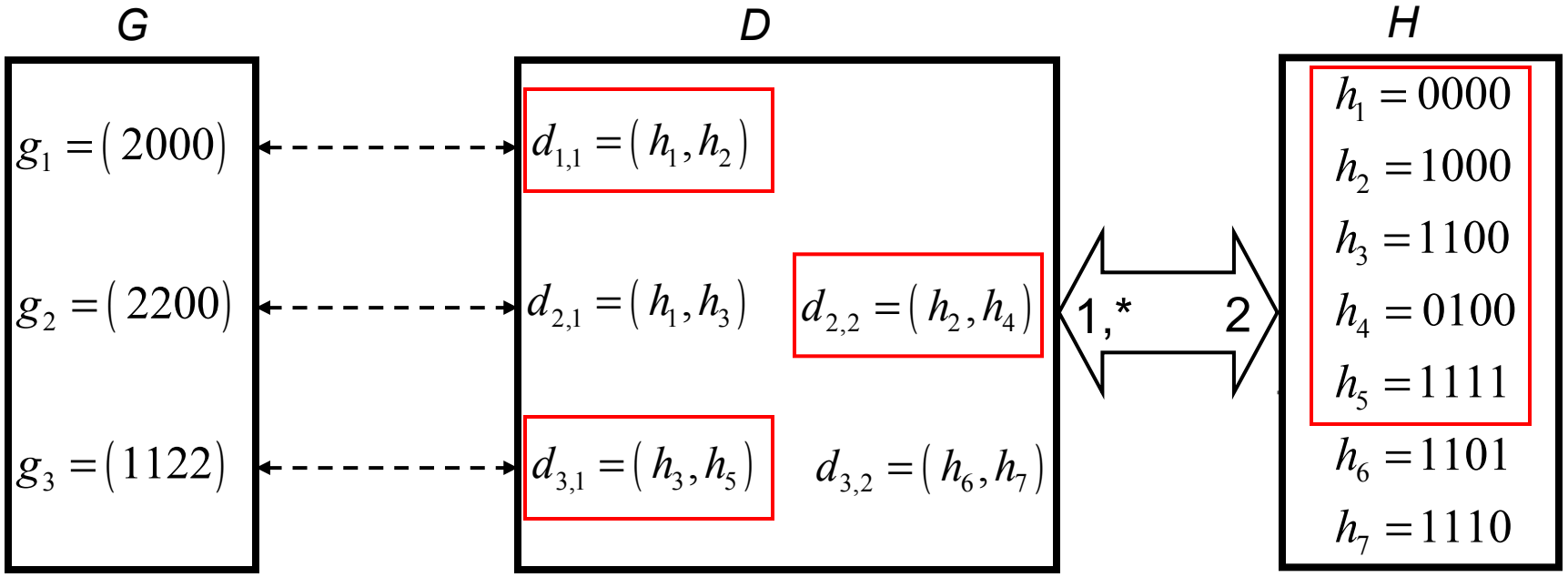
V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

COMBINATORIAL APPROACH



$D_1 = \{d_{1,1}, d_{2,2}, d_{3,1}\}$ $H_1 = \{h_1, h_2, h_3, h_4, h_5\}$

PARSIMONY

Idea :

The individuals share a common evolutionary history

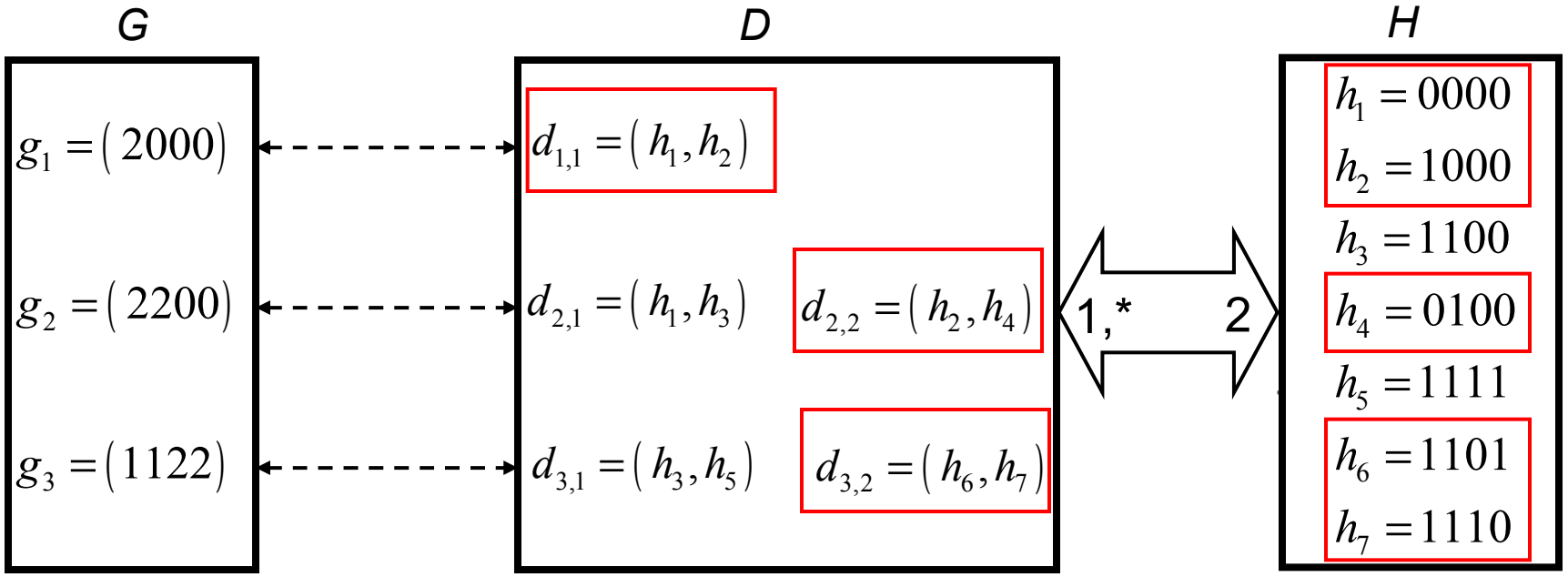
=>The individuals share common haplotypes

Principle :

Minimize the number of distinct haplotypes required to solve all the individuals of the sample

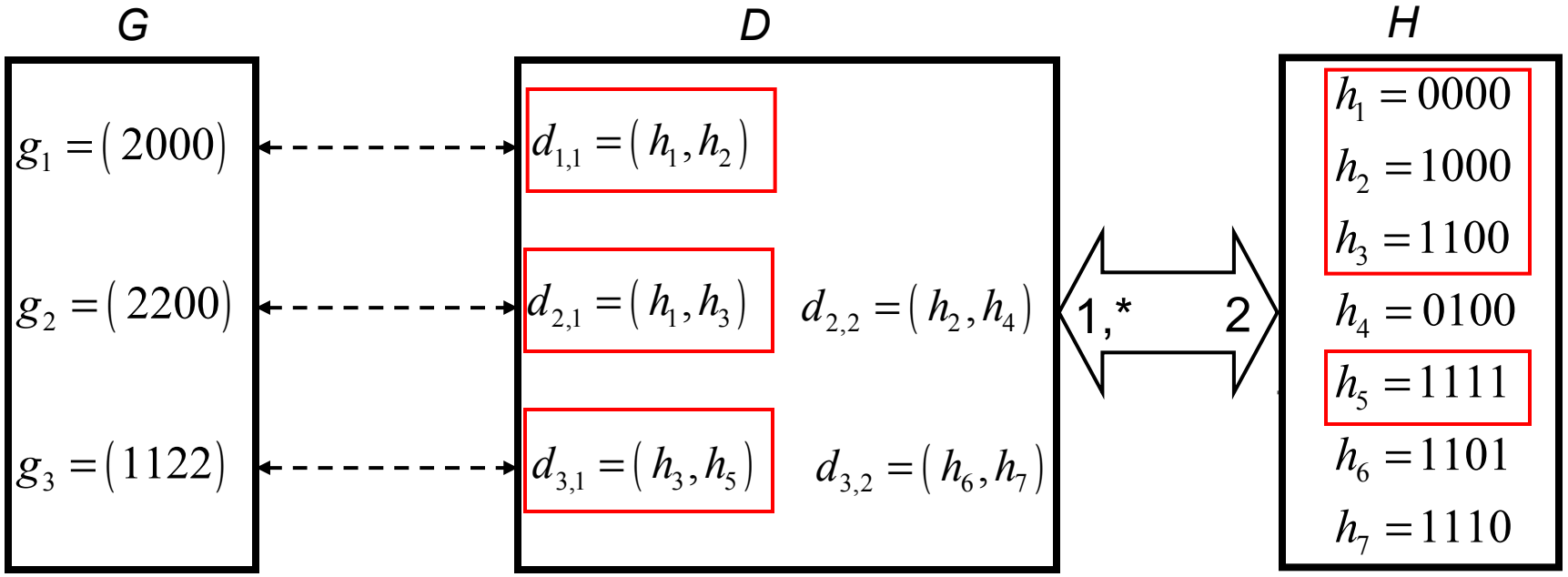
Clark et al, 1990

PARSIMONY



$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1, h_2, h_4, h_6, h_7\}$ **=> 5** haplotypes

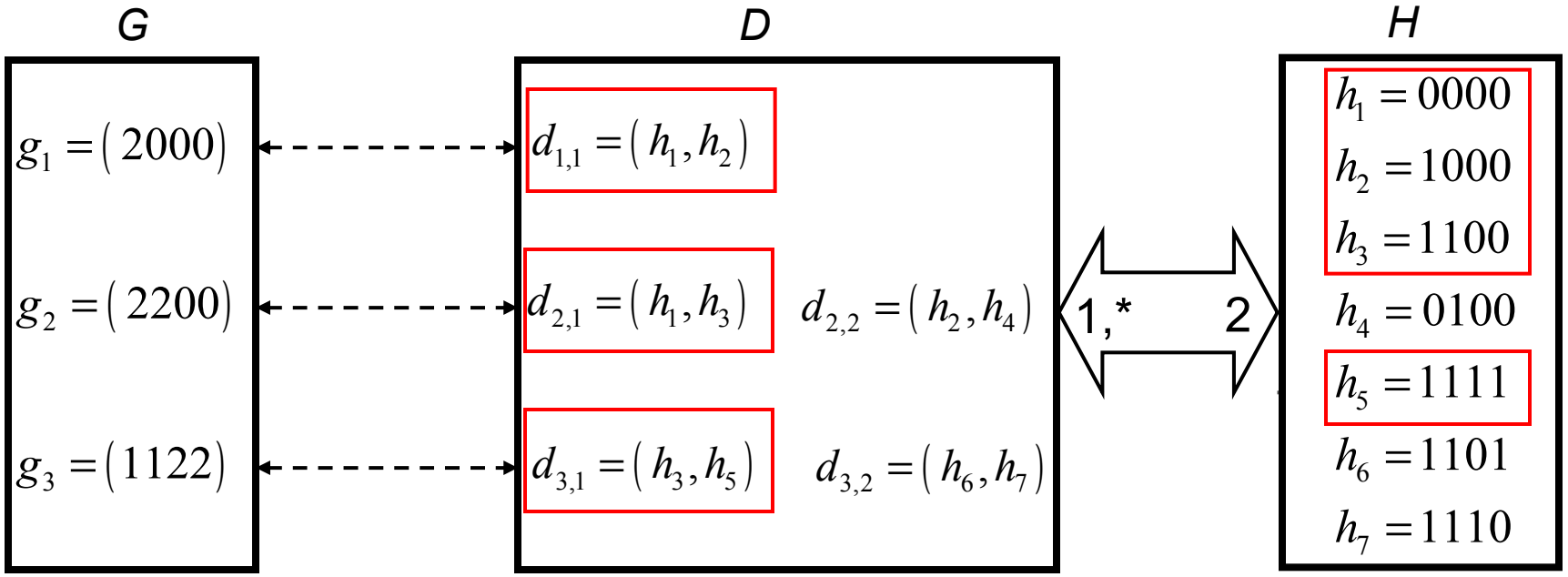
PARSIMONY



$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1, h_2, h_4, h_6, h_7\}$ => **5** haplotypes

$D_2 = \{d_{1,1}, d_{2,1}, d_{3,1}\}$ $H_2 = \{h_1, h_2, h_3, h_5\}$ => **4** haplotypes

PARSIMONY



$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1, h_2, h_4, h_6, h_7\}$ => **5** haplotypes

$D_2 = \{d_{1,1}, d_{2,1}, d_{3,1}\}$ $H_2 = \{h_1, h_2, h_3, h_5\}$ => **4** haplotypes

Idea :

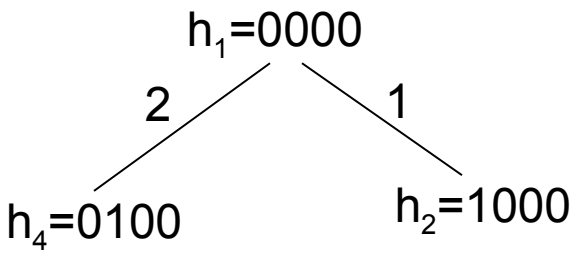
The haplotypes share a common evolutionary history which can be represented by a realistic phylogenetic tree

Principle :

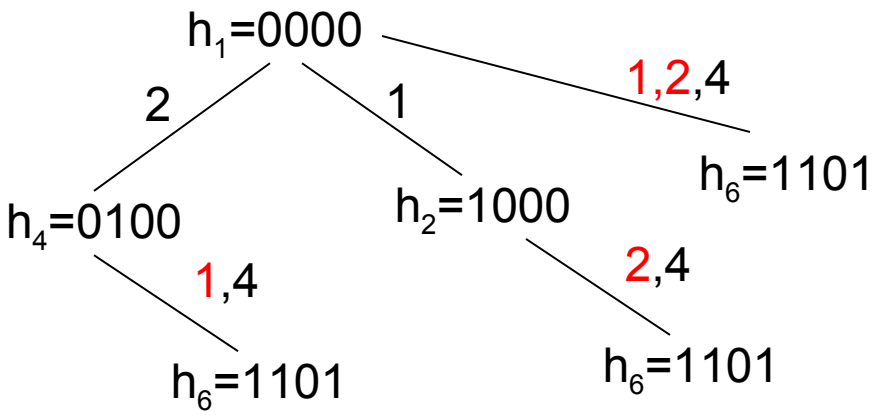
Find a set of haplotypes for G which fits a perfect phylogeny (each node is a haplotype, each branch is a mutation which appears only once)

Gusfield et al, 2002

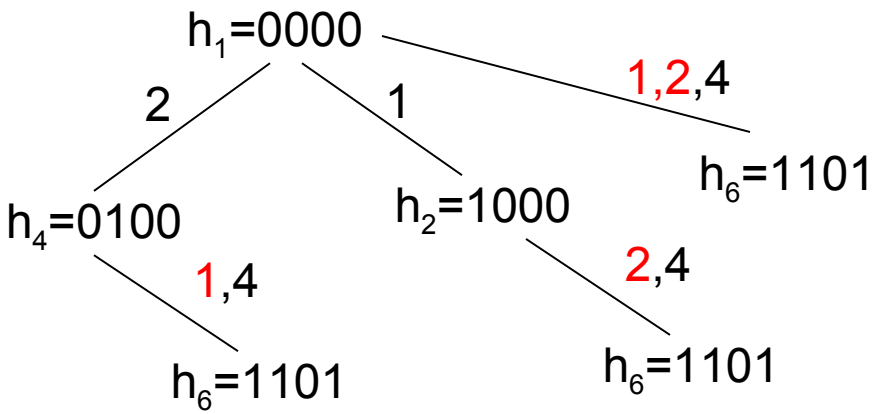
$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1=0000\ h_2=1000\ h_4=0100\ h_6=1101\ h_7=1110\}$



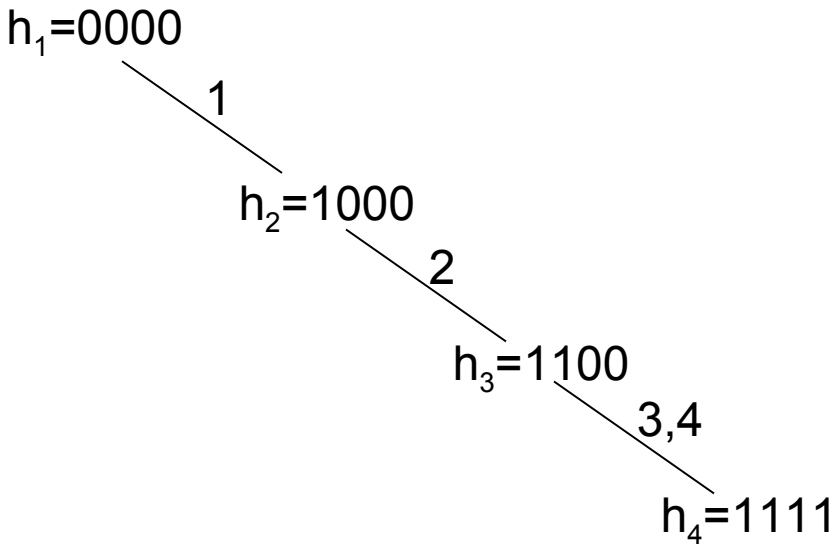
$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1=0000\ h_2=1000\ h_4=0100\ h_6=1101\ h_7=1110\}$



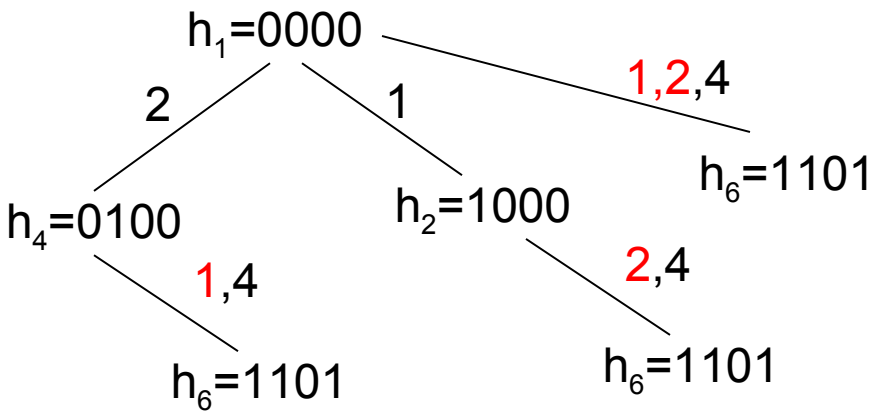
$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1=0000\ h_2=1000\ h_4=0100\ h_6=1101\ h_7=1110\}$



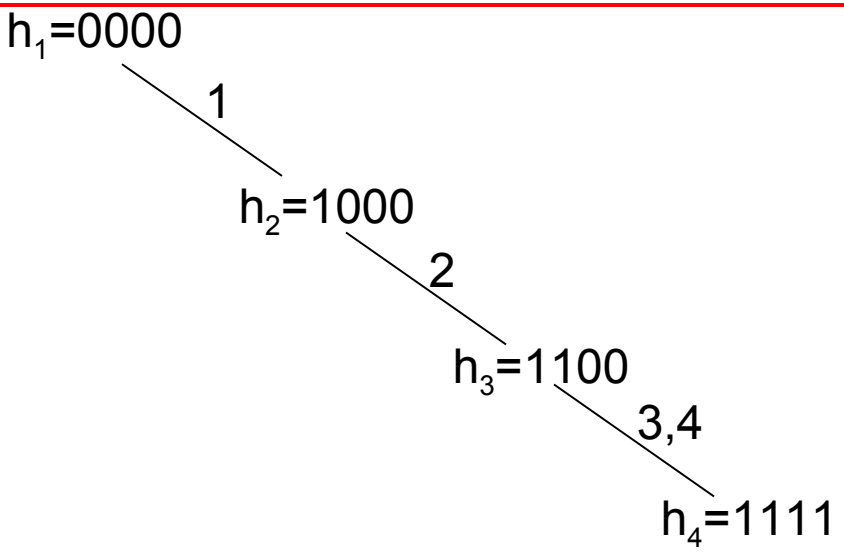
$D_2 = \{d_{1,1}, d_{2,1}, d_{3,1}\}$ $H_2 = \{h_1=0000\ h_2=1000\ h_3=1100\ h_5=1111\}$



$D_1 = \{d_{1,1}, d_{2,2}, d_{3,2}\}$ $H_1 = \{h_1=0000\ h_2=1000\ h_4=0100\ h_6=1101\ h_7=1110\}$



$D_2 = \{d_{1,1}, d_{2,1}, d_{3,1}\}$ $H_2 = \{h_1=0000\ h_2=1000\ h_3=1100\ h_5=1111\}$



CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

HARDY WEINBERG EQUILIBRIUM

H = { h_1, h_2, ..., h_a, ..., h_b, ..., h_m} with frequencies { f_1, f_2, ..., f_a, ..., f_b, ..., f_m}

I. Probability of a pair of haplotypes (h_a,h_b) by Hardy Weinberg :

$$Pr(g_i = (h_a, h_b) \mid H) = 2 \times f_a \times f_b \qquad \text{if } h_a \neq h_b$$

$$Pr(g_i = (h_a, h_b) \mid H) = f_a \times f_b \qquad \text{if } h_a = h_b$$

II. Probability of a genotype by summing over all possible pairs of haplotypes :

$$\Pr(g_i \mid H) = \sum_{(h_a,\, h_b)} \Pr(g_i = (h_a, h_b) \mid H)$$

II. Probability of a genotype by summing over all possible pairs of haplotypes :

$$\Pr(g_i \mid H) = \sum_{(h_a,\, h_b)} \Pr(g_i = (h_a, h_b) \mid H)$$

III. Likelihood of the entire sample **G** by multiplying the N genotype probabilities :

$$\Pr(G \mid H) = \prod_{i=1}^{i=N} \Pr(g_i \mid H)$$

MAXIMUM LIKELIHOOD ESTIMATION

II. Probability of a genotype by summing over all possible pairs of haplotypes :

$$\Pr(g_i \mid H) = \sum_{(h_a,\, h_b)} \Pr(g_i = (h_a, h_b) \mid H)$$

III. Likelihood of the entire sample **G** by multiplying the N genotype probabilities :

$$\Pr(G \mid H) = \prod_{i=1}^{i=N} \Pr(g_i \mid H)$$

IV. Maximum likelihood estimate $\hat{\mathbf{H}}$ by Expectation – Maximization (EM) :

$$\hat{\mathbf{H}} = \max_{H} \Pr(G \mid H)$$

Excoffier et al, 1995

BAYESIAN ESTIMATION

$D = \{ d_1, \ldots, d_i, \ldots, d_N \}$ where d_i is a random variable defined on all possible pairs of haplotypes of g_i.

Construct a Monte Carlo Markov chain with states D^0 , D^1, ..., D^t, ... which converges to the true reconstruction of haplotypes of G.

Gibbs sampler :

0. Start with a random reconstruction D^0

1. Iterate a large number of times the following step to go from state D^t to state D^{t+1} :

> Construct D^{t+1} from D^t by sampling a new pair of haplotypes for a randomly chosen genotype g_i given all the 2N – 2 haplotypes previously sampled D^t_{-i}

Stephens et al, 2001

BAYESIAN ESTIMATION

$D = \{ d_1, \ldots, d_i, \ldots, d_N \}$ where d_i is a random variable defined on all possible pairs of haplotypes of g_i.

Construct a Monte Carlo Markov chain with states D^0 , D^1, ..., D^t, ... which converges to the true reconstruction of haplotypes of G.

Gibbs sampler :

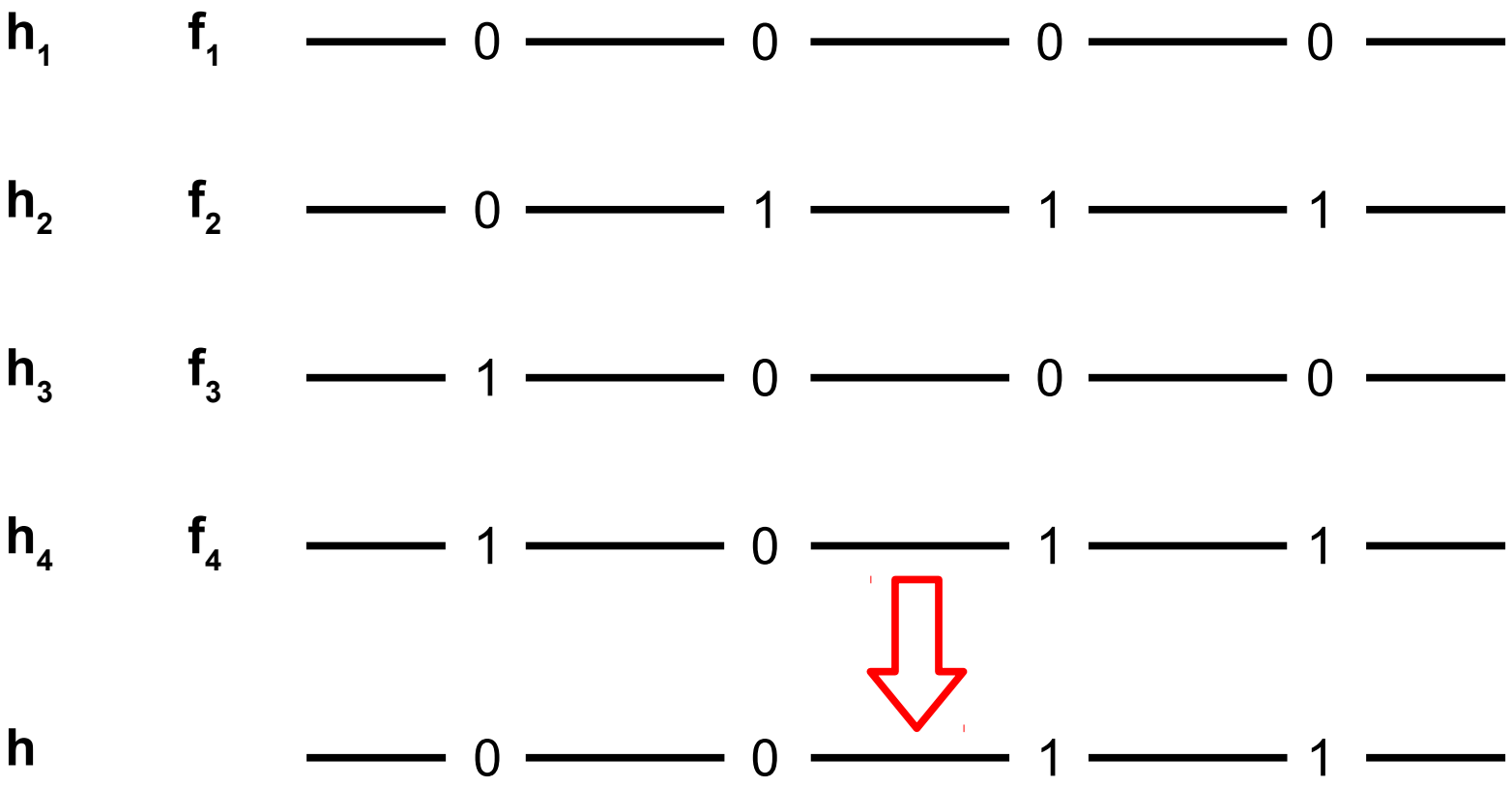
0. Start with a random reconstruction D^0

1. Iterate a large number of times the following step to go from state D^t to state D^{t+1} :

Construct D^{t+1} from D^t by sampling a new pair of haplotypes for a randomly chosen genotype g_i given all the 2N – 2 haplotypes previously sampled D^t_{-i}

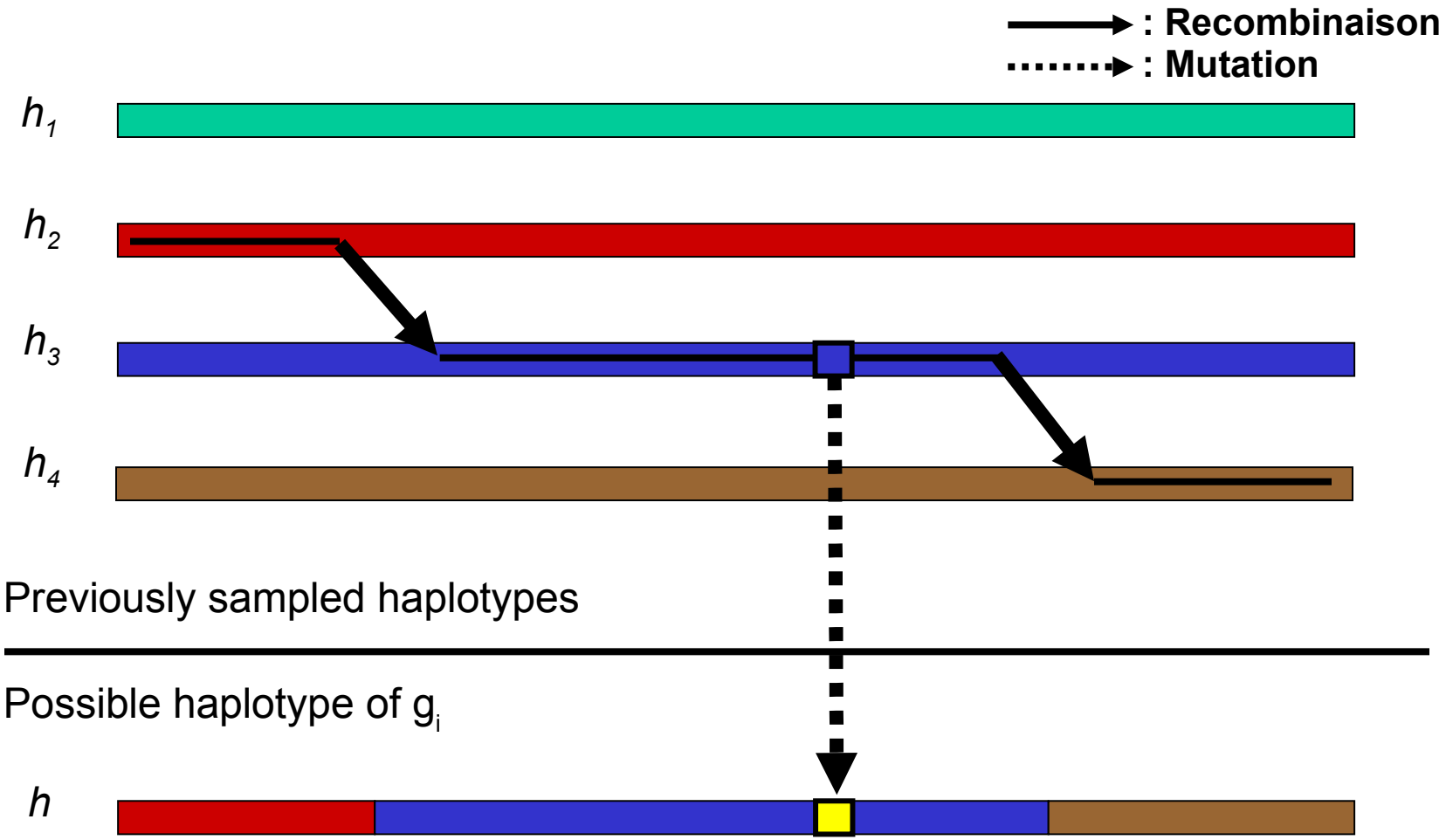
Stephens et al, 2001

HIDDEN MARKOV MODEL OF HAPLOTYPES



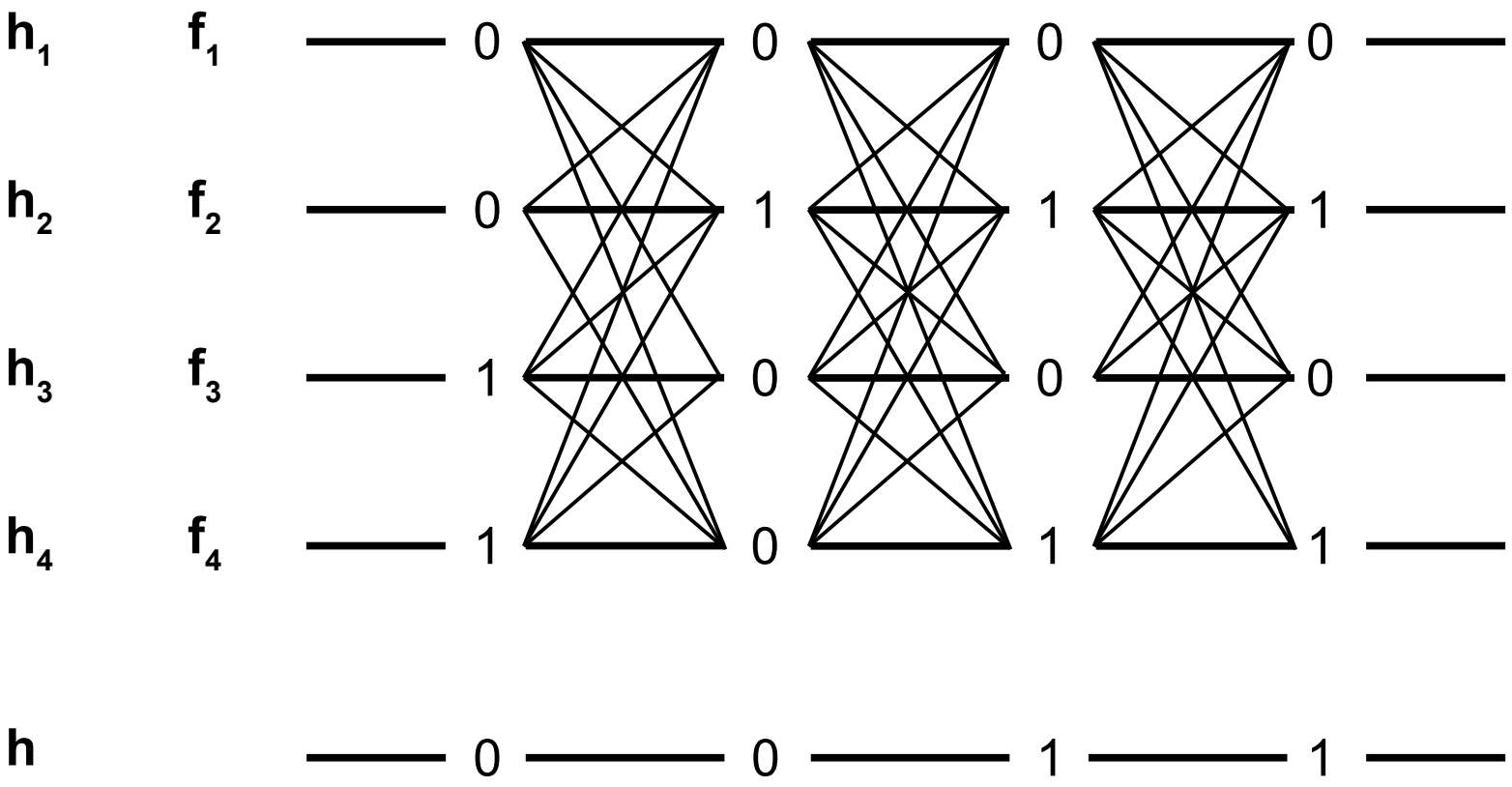
1. How likely ($\mathbf{h}_1$,$\mathbf{h}_2$,$\mathbf{h}_3$,$\mathbf{h}_4$) have produced **h** ?

HIDDEN MARKOV MODEL OF HAPLOTYPES

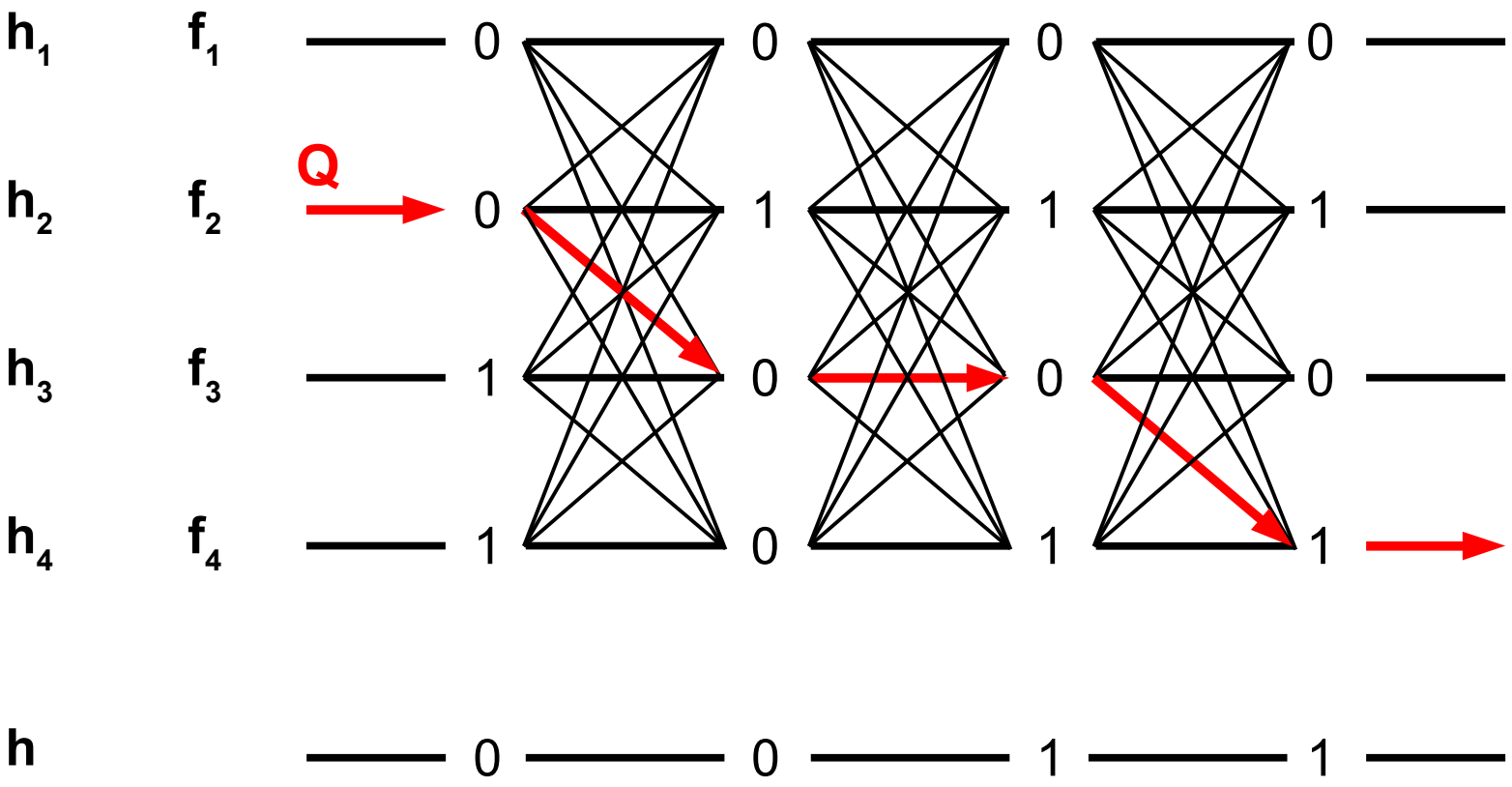


Haplotype **h** is an imperfect mosaic of $(\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3,\mathbf{h}_4)$

HIDDEN MARKOV MODEL OF HAPLOTYPES



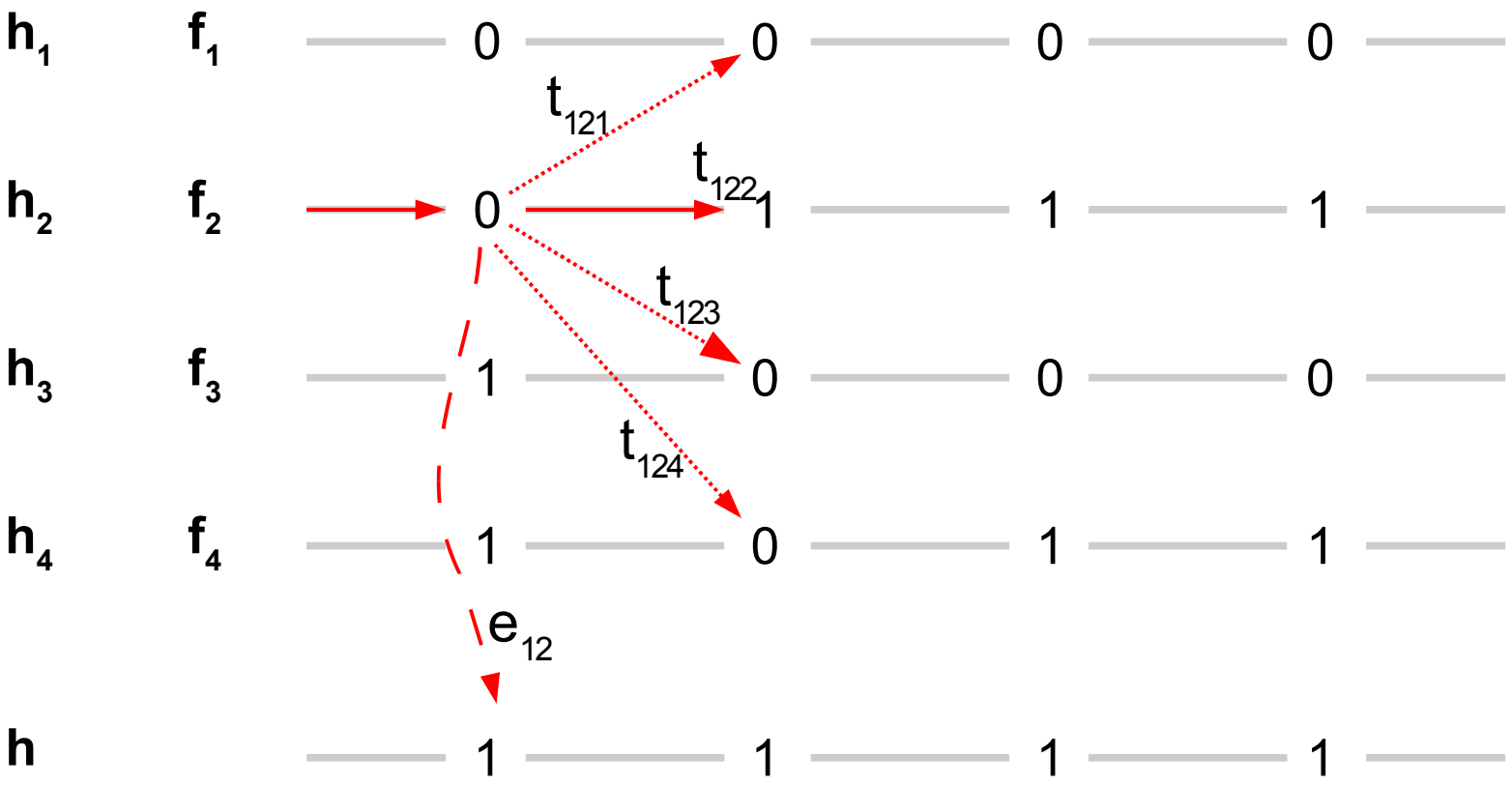
HIDDEN MARKOV MODEL OF HAPLOTYPES



Q is a path in the graph = **Q** is a mosaic of h_2, h_3 and h_4

Haplotype **h** is an imperfect copy of **Q**

HIDDEN MARKOV MODEL OF HAPLOTYPES



K = 4 haplotypes
L = 4 SNPs

HMM λ =
- F : initial probabilities $\mathbf{f_k}$
- T : transition probabilities $\mathbf{t_{jkl}} = Pr(X_{j+1} = h_l | X_j = h_k)$
- E : emission probabilities $\mathbf{e_{jk}} = Pr(o_j | X_j = h_k)$

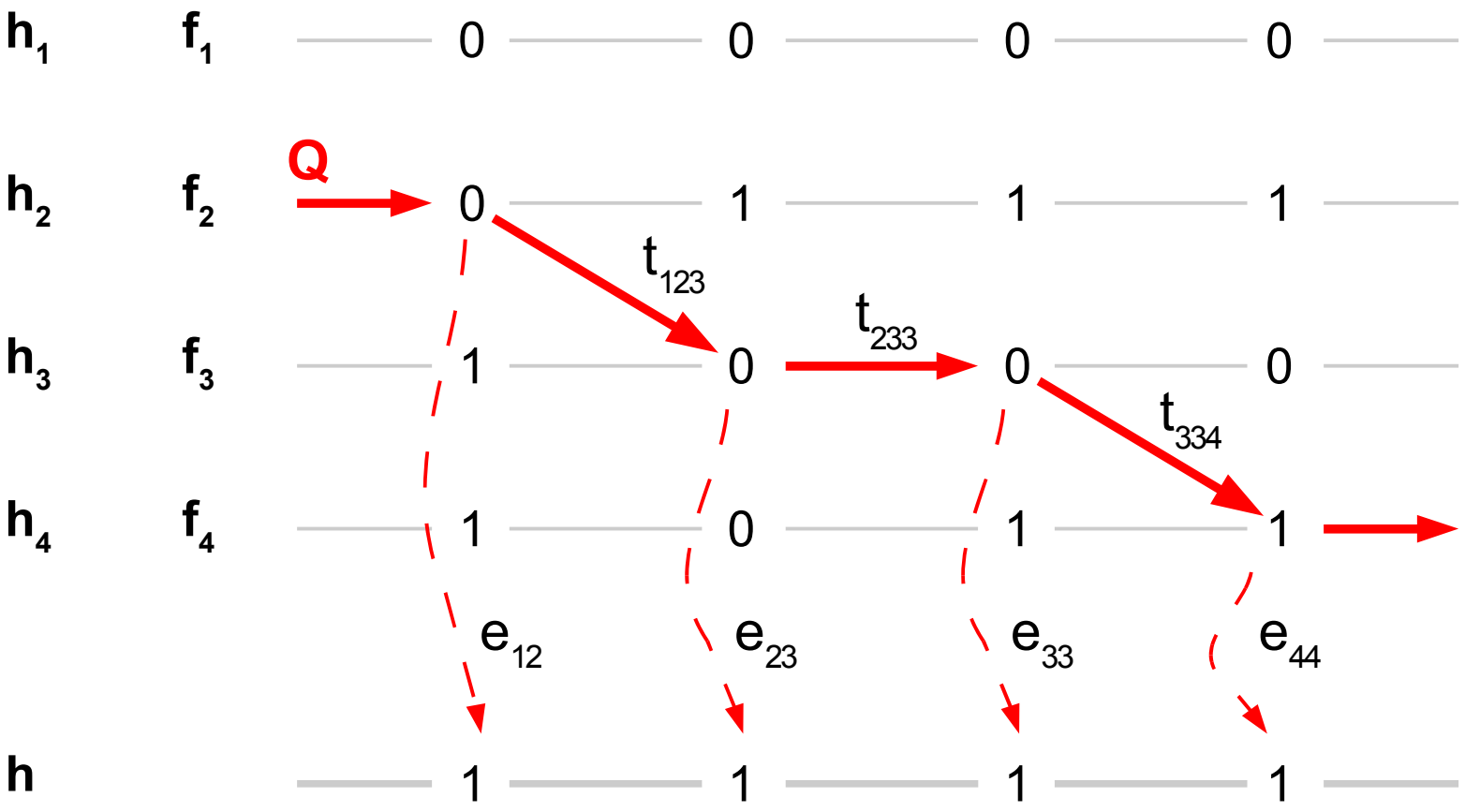
1. How to define the parameters ?

- Prior on the transitions : a function of the genetic distance between SNPs (a large genetic distance implies greater chance of recombination).

- Prior on the emissions : a function of the mutation rate

Li et al, 2003

HIDDEN MARKOV MODEL OF HAPLOTYPES



$$Pr\ (\ Q\ |\ \lambda) = f_2 \times t_{123} \times t_{233} \times t_{334}$$

$$Pr\ (\ h\ |\ Q,\ \lambda) = e_{12} \times e_{23} \times e_{33} \times e_{44}$$

$$Pr(\ h,\ Q\ |\ \lambda\ ) = Pr(\ h\ |\ Q,\ \lambda) \times Pr(\ Q\ |\ \lambda\ )$$

2. How likely ($\mathbf{h}_1$,$\mathbf{h}_2$,$\mathbf{h}_3$,$\mathbf{h}_4$) can produce **h** ?

$$\Pr(h \mid \lambda) = \sum_Q \Pr(h, Q \mid \lambda)$$
$$= \sum_Q \Pr(h \mid Q, \lambda) \times \Pr(Q \mid \lambda)$$

=> Naïve approach $O(K^L)$

=> Forward-backward algorithm $O(LK^2)$

Rabiner, 1989

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

V. Comparison of accuracy

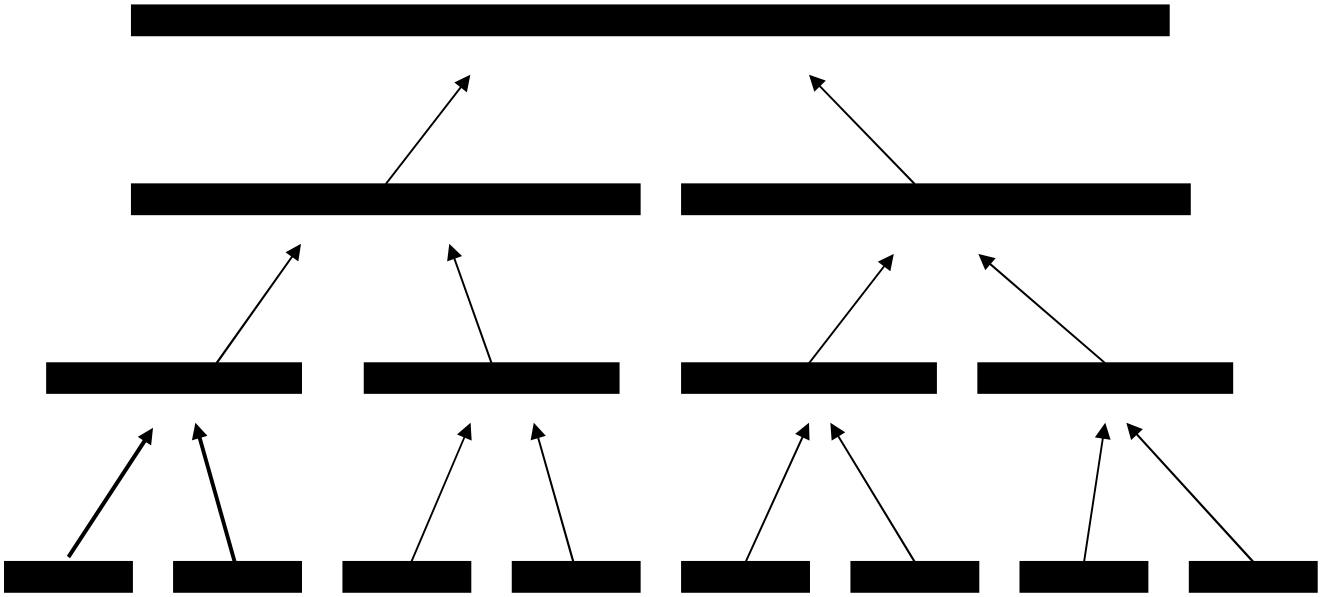
VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

PARTITION LIGATION

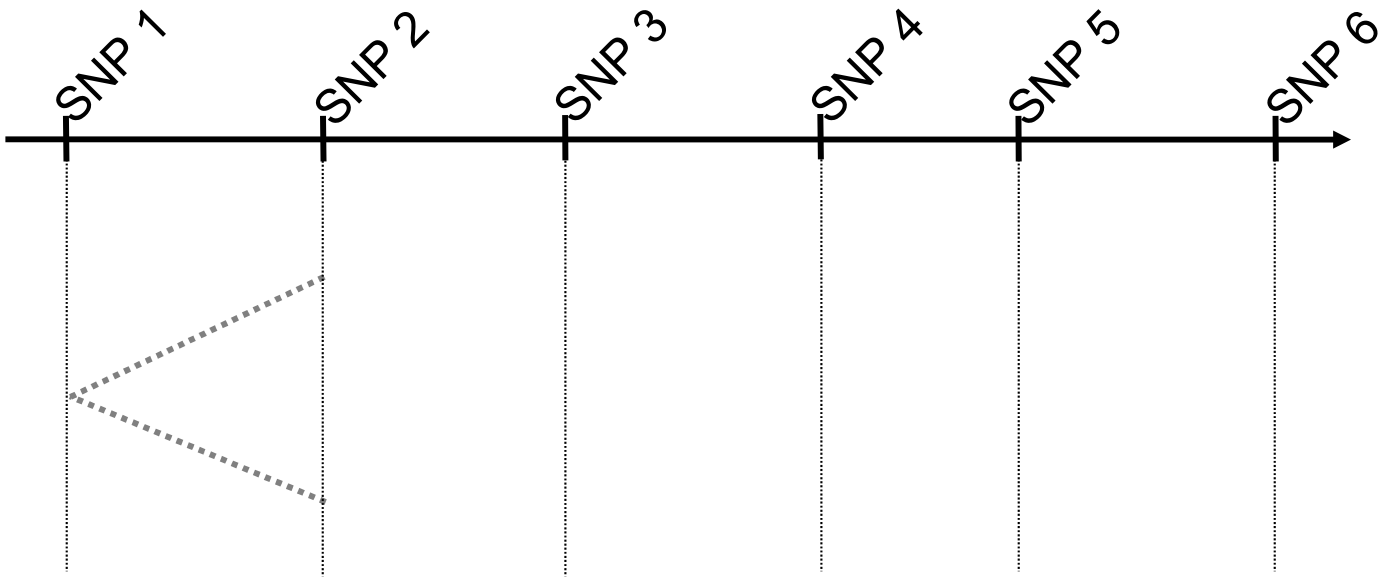
1 segment of 32 heterozygous SNPs = ~2 000 000 000 possible haplotype' pairs



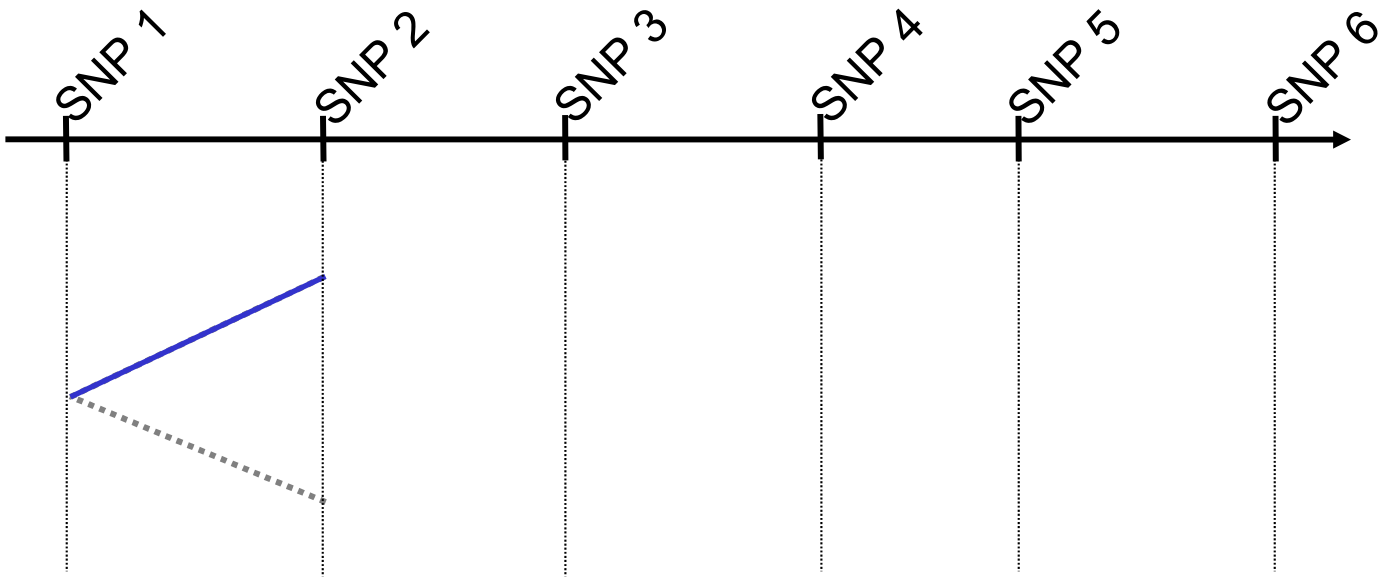
8 segments of 4 heterozygous SNPs = 2^3 x 8 = 64 possible haplotype' pairs

Nui et al, 2002

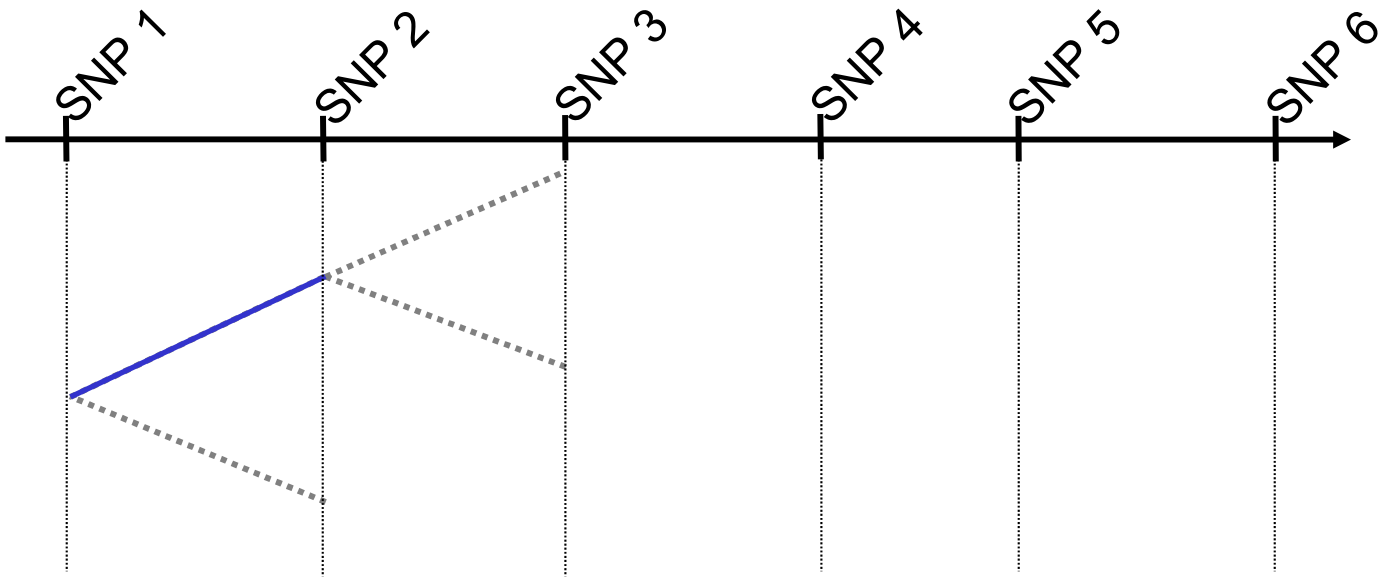
ITERATIVE SNP INCLUSION



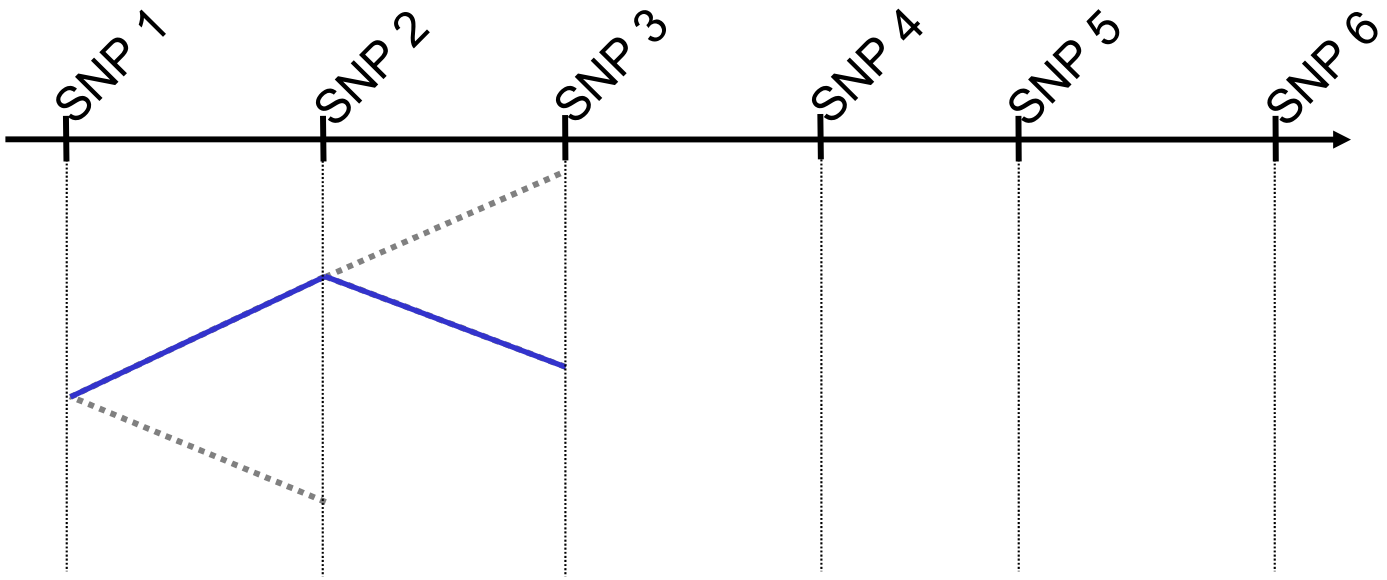
ITERATIVE SNP INCLUSION



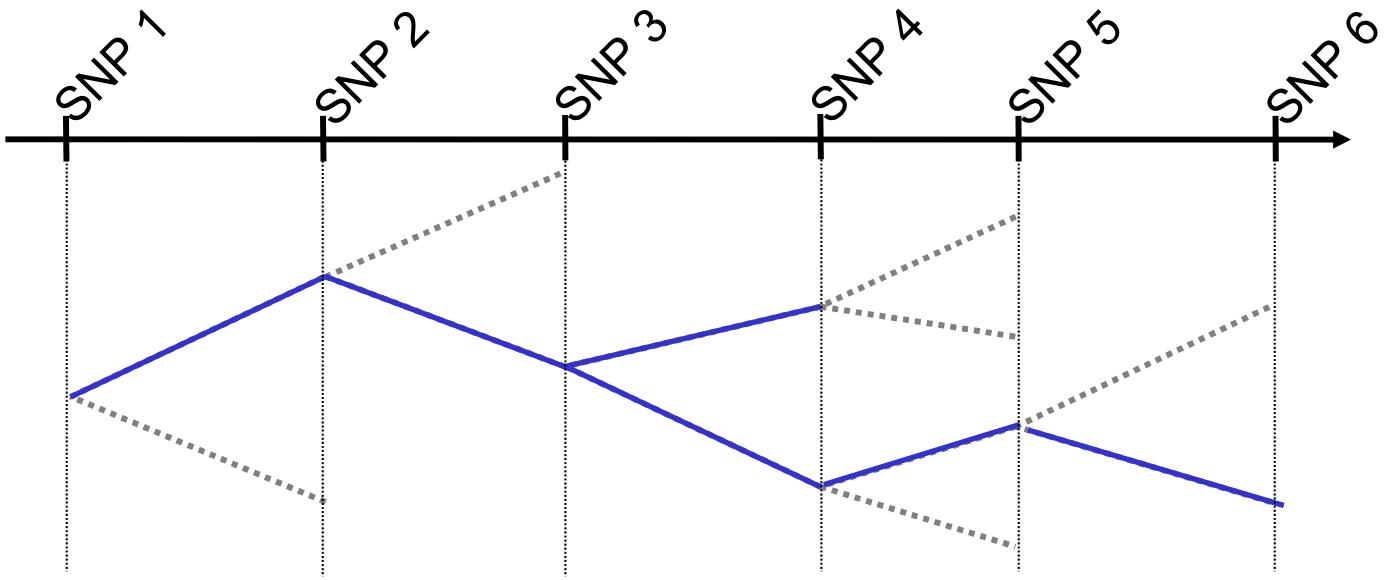
ITERATIVE SNP INCLUSION



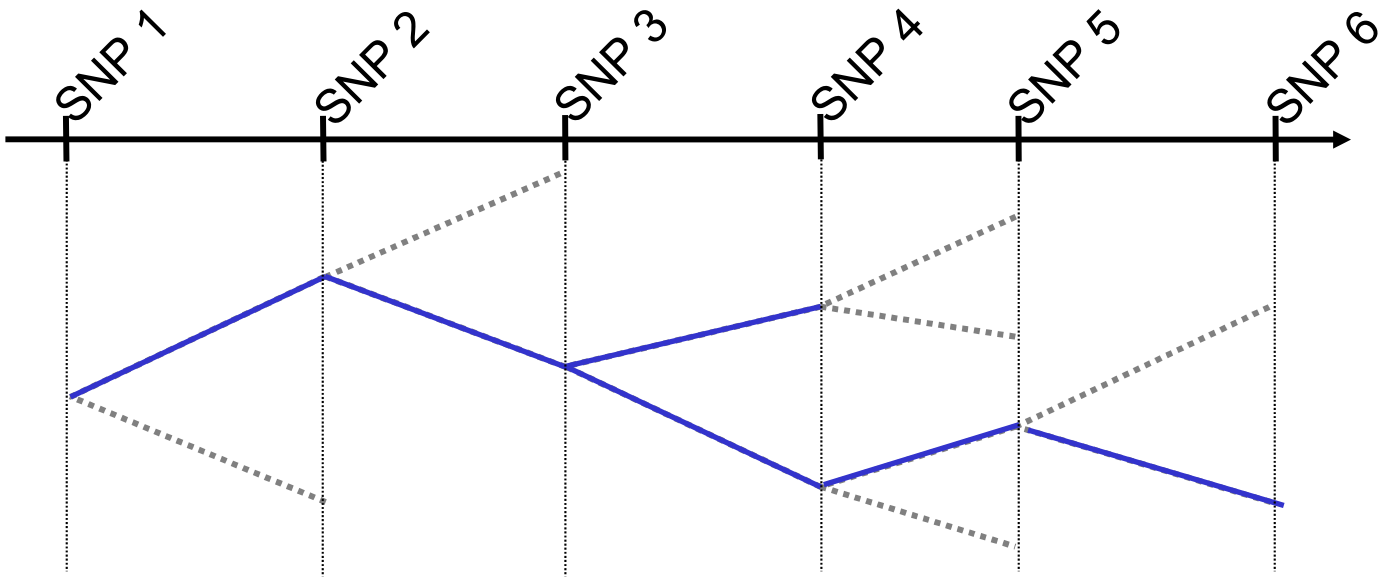
ITERATIVE SNP INCLUSION



ITERATIVE SNP INCLUSION



ITERATIVE SNP INCLUSION



Implementation of the forward – backward algorithm with this tree strategy

Delaneau et al, 2007
Delaneau et al, 2008

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

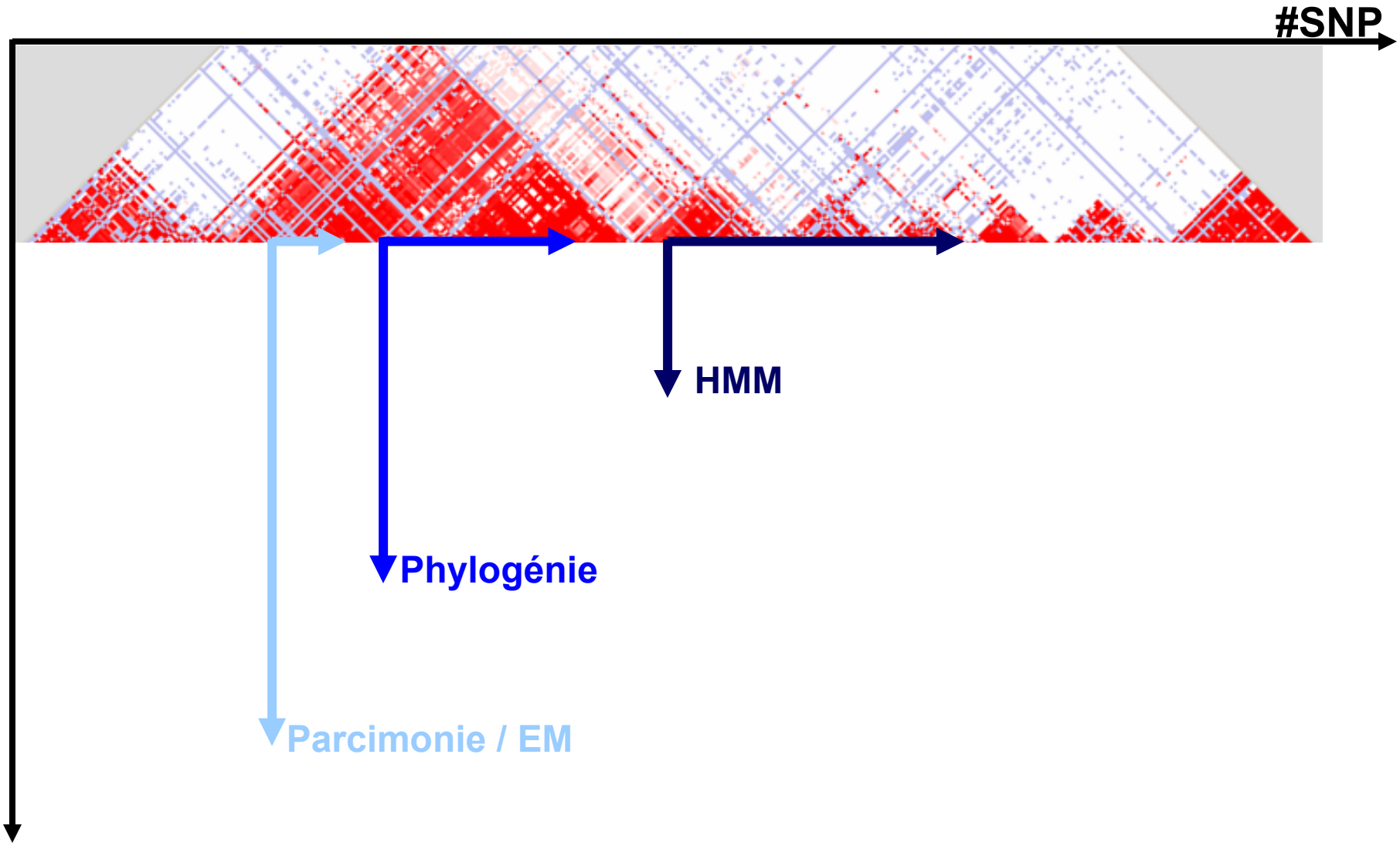
V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

ACCURACY OF HAPLOTYPE INFERENCE



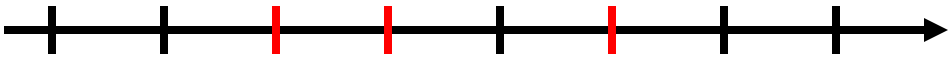
ACCURACY OF HAPLOTYPE INFERENCE

HapMap : 300 x (60 genotypes / 50 SNP)

	Error rate (%)	Running time (sec.)
ShapeIT	6,3	50
Phase v2.1	6,3	1215
EM	12,2	10

HAPLOTYPE INFERENCE STRATEGY

Studied genomic region

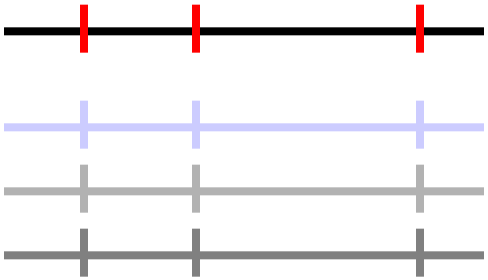


HAPLOTYPE INFERENCE STRATEGY

Studied genomic region

Local haplotype inference

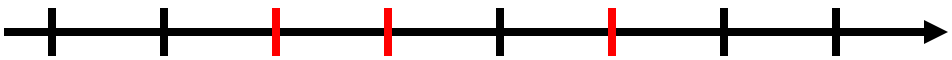
1. Extraction of SNP



2. Haplotype inference

HAPLOTYPE INFERENCE STRATEGY

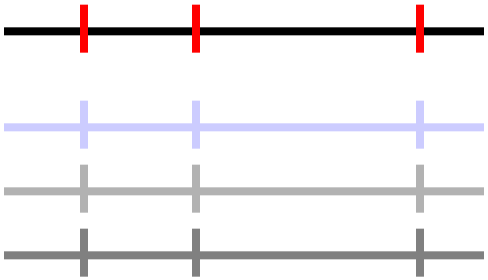
Studied genomic region



Local haplotype inference

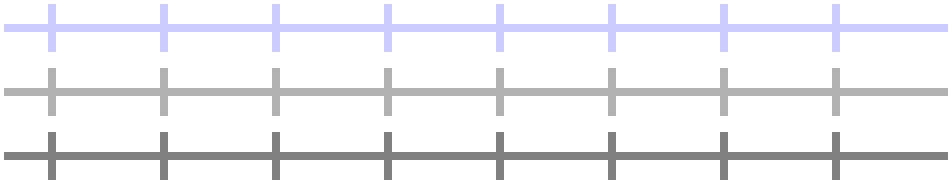
1. Extraction of SNP

2. Haplotype inference

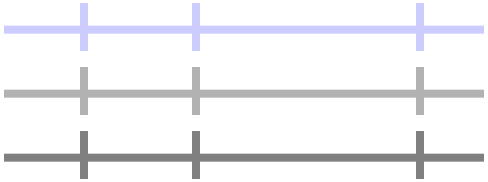


Global haplotype inference

1. Haplotype inference

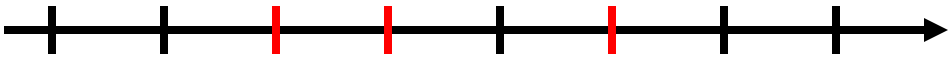


2. Extraction of sub-haplotypes



HAPLOTYPE INFERENCE STRATEGY

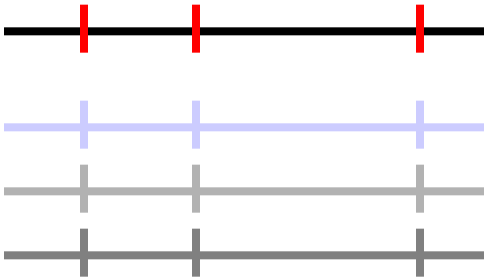
Studied genomic region



Local haplotype inference

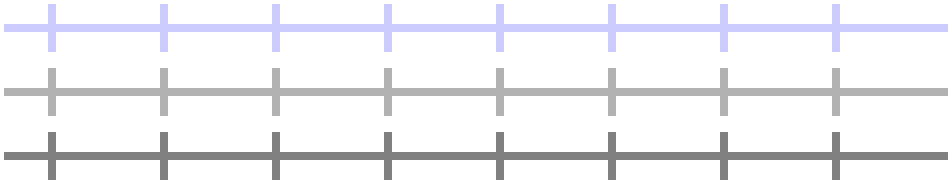
1. Extraction of SNP

2. Haplotype inference

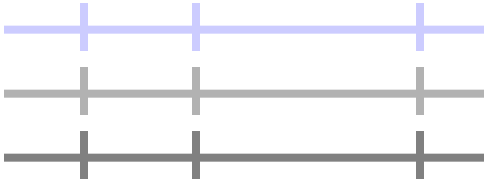


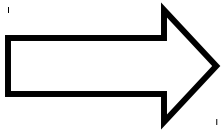
Global haplotype inference

1. Haplotype inference



2. Extraction of sub-haplotypes





Reduce from 1% to 6% the error rate

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

HAPLOTYPE ASSOCIATION TESTS

The simplest approach :

1. Estimate haplotype frequencies in cases and control separately.

2. Find the most likely haplotype pair for each individual.

3. Fill in a contingency table to compare haplotypes in the cases versus the controls.

HAPLOTYPE ASSOCIATION TESTS

The simplest approach :

1. Estimate haplotype frequencies in cases and control separately.

2. Find the most likely haplotype pair for each individual.

3. Fill in a contingency table to compare haplotypes in the cases versus the controls.

Cases		Controls
g_1 : A/A A/A	Versus	g_5 : A/A T/T
g_2 : A/T A/T		g_6 : A/T A/T
g_3 : A/T A/T		g_7 : A/T A/T
g_4 : A/T A/T		g_8 : A/T A/T

Cases		Controls
g_1 : AA/AA		g_5 : AT/AT
g_2 : AA/TT		g_6 : AT/TA
g_3 : AA/TT		g_7 : AT/TA
g_4 : AA/TT		g_8 : AT/TA

! Do not infer haplotypes in two samples separately !

Schaid et al, 2002
Zaytkin et al, 2002

HAPLOTYPE ASSOCIATION TESTS

The simplest approach :

1. Estimate haplotype frequencies in cases and control separately.

2. Find the most likely haplotype pair for each individual.

3. Fill in a contingency table to compare haplotypes in the cases versus the controls.

```
g :    AA/TT with probability of 0.7
       AT/AT with probability of 0.3

g :    AA/TT
```

! Does not take into account uncertainty !

Solution 1 : Likelihood ratio test + label permutations

$$L_A = (Pr(G_{cases}|H))$$

$$L_B = (Pr(G_{controls}|H))$$

$$L_C = (Pr(G_{cases+controls}|H))$$

$$2\ln\left(\frac{L_A \times L_B}{L_C}\right) \sim chi2$$

Solution 2 : Regression model

genotype	phenotype	hap pair 1	prob 1	hap pair 2	prob 2
g1	p1	h1/h1	1.0		
g2	p2	h1/h2	0.7	h3/h4	0.3
g3	p3	h3/h4	1.0		

Solution 2 : Regression model

genotype	phenotype	hap pair 1	prob 1	hap pair 2	prob 2
g1	p1	h1/h1	1.0		
g2	p2	h1/h2	0.7	h3/h4	0.3
g3	p3	h3/h4	1.0		

$$\begin{pmatrix} p1 \\ p2 \\ p3 \end{pmatrix} = \begin{array}{c} \begin{matrix} \mu & h1 & h2 & h3 & h4 \end{matrix} \\ \begin{pmatrix} 1.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.35 & 0.35 & 0.15 & 0.15 \\ 1.0 & 0.0 & 0.0 & 0.5 & 0.5 \end{pmatrix} \end{array} \cdot \begin{pmatrix} \mu \\ \beta 1 \\ \beta 2 \\ \beta 3 \\ \beta 4 \end{pmatrix}$$

Solution 2 : Regression model

genotype	phenotype	hap pair 1	prob 1	hap pair 2	prob 2
g1	p1	h1/h1	1.0		
g2	p2	h1/h2	0.7	h3/h4	0.3
g3	p3	h3/h4	1.0		

$$\begin{pmatrix} p1 \\ p2 \\ p3 \end{pmatrix} = \begin{array}{c} \begin{matrix} \mu & h1 & h2 & h3 & h4 \end{matrix} \\ \begin{pmatrix} 1.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 1.0 & 0.35 & 0.35 & 0.15 & 0.15 \\ 1.0 & 0.0 & 0.0 & 0.5 & 0.5 \end{pmatrix} \end{array} \cdot \begin{pmatrix} \mu \\ \beta 1 \\ \beta 2 \\ \beta 3 \\ \beta 4 \end{pmatrix}$$

Testing simultaneously all haplotypes

H0 : β1=0 β2=0 β3=0 β4=0 VS H1 : β1≠0 β2≠0 β3≠0 β4≠0

Testing only haplotype h1

H0 : β1=0 VS H1 : β1≠0

Zaytkin et al, 2002

Strategies :

1. Biological meaning : promoters, exons, etc...

2. Most significant regions obtained by single SNP analysis

3. Systematic screening by sliding window

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

HapMap haplotypes :

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

Genotype :

```
. . . A . . . . . . . C . . . . A . . .
. . . G . . . . . . . A . . . . A . . .
```

HapMap haplotypes :

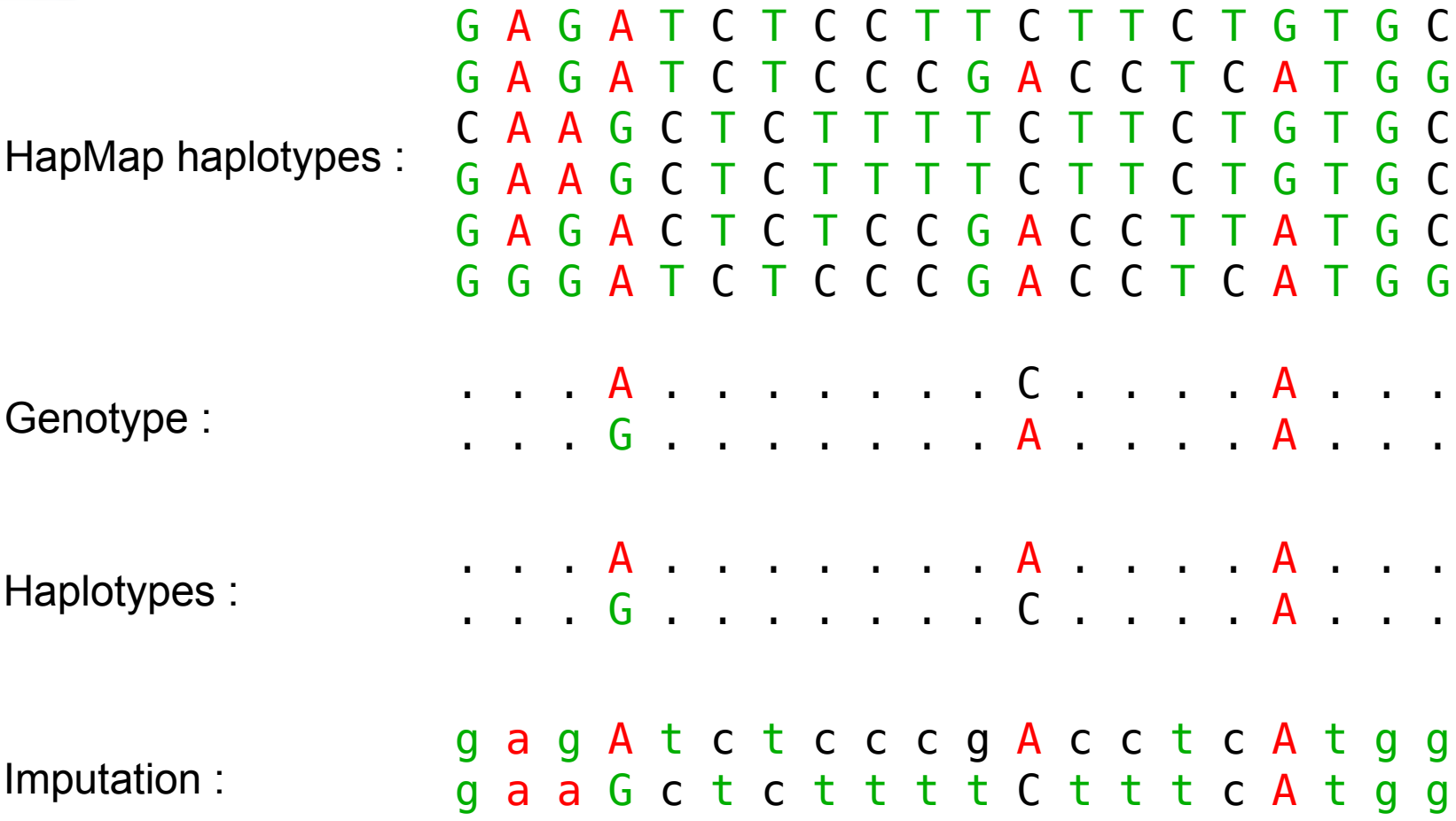
```
G A G A T C T C C T T C T T C T G T G C
G A G A T C T C C C G A C C T C A T G G
C A A G C T C T T T T C T T C T G T G C
G A A G C T C T T T T C T T C T G T G C
G A G A C T C T C C G A C C T T A T G C
G G G A T C T C C C G A C C T C A T G G
```

Genotype :

```
. . . A . . . . . . . C . . . . A . . .
. . . G . . . . . . . A . . . . A . . .
```

Haplotypes :

```
. . . A . . . . . . . A . . . . A . . .
. . . G . . . . . . . C . . . . A . . .
```


HapMap haplotypes :
G A G A T C T C C T T C T T C T G T G C
G A G A T C T C C C G A C C T C A T G G
C A A G C T C T T T T C T T C T G T G C
G A A G C T C T T T T C T T C T G T G C
G A G A C T C T C C G A C C T T A T G C
G G G A T C T C C C G A C C T C A T G G
Genotype :
. . . A . . . . C . . . . A . . .
. . . G . . . . A . . . . A . . .
Haplotypes :
. . . A . . . . A . . . . A . . .
. . . G . . . . C . . . . A . . .
Imputation :
g a g A t c t c c c g A c c t c A t g g
g a a G c t c t t t t C t t t c A t g g

HapMap haplotypes :

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

Genotype :

```
. . . A . . . . . . . C . . . . A . . .
. . . G . . . . . . . A . . . . A . . .
```

Haplotypes :

```
. . . A . . . . . . . A . . . . A . . .
. . . G . . . . . . . C . . . . A . . .
```

Imputation :

```
g a g A t c t c c c g A c c t c A t g g
g a a G c t c t t t t C t t t c A t g g
```

In practice, we have a probability distribution :

c/c	c/t	t/t
0.01	0.18	0.81

Marchini et al, 2007

CONTENTS

I. Introduction

II. Combinatorial haplotype inference

III. Statistical haplotype inference

IV. Algorithmic tricks

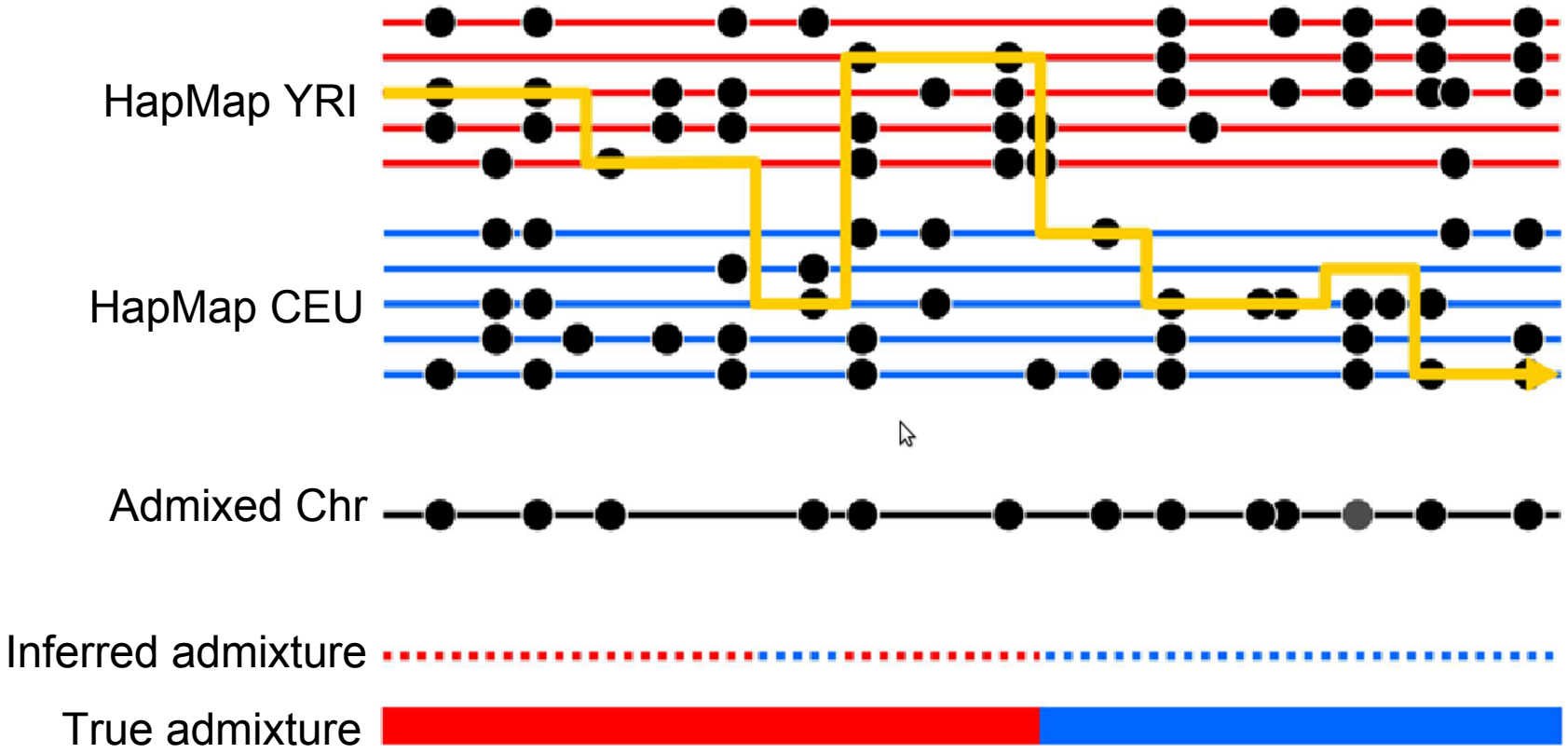
V. Comparison of accuracy

VI. Application 1 : haplotype association tests

VII. Application 2 : genotype imputation

VIII. Application 3 : admixture

ADMIXTURE



Price et al, 2009

CONCLUSION

1. Most accurate haplotype inference methods rely on hidden Markov model.

2. Algorithmic improvements still needed to fit the models on large datasets.

3. Haplotype association tests remain an interesting approach in association studies.

4. Several useful applications rely on haplotype inference: genotype imputation, admixture inference, etc ...

ACKNOWLEGDMENT

Chaire de Bioinformatique du CNAM
Jean-François Zagury
Cédric Coulonges
Sigrid Le Clerc
Sophie Limou
Lieng Taing
Hervé Do
Taoufik Labib
Christiane Morel
Matthieu Montes
Rojo Ratsimandresy
Les nouveaux collaborateurs

CNRS
Christine Sinoquet