Méthodes de reconstruction d'haplotypes

Olivier DELANEAU et Jean François ZAGURY

Chaire de Bioinformatique - Conservatoire National des Arts et Métiers

Inserm

Institut national
de la santé et de la recherche médicale

PLAN

I. La problématique des haplotypes.

II. Les principales méthodes de reconstruction des haplotypes.

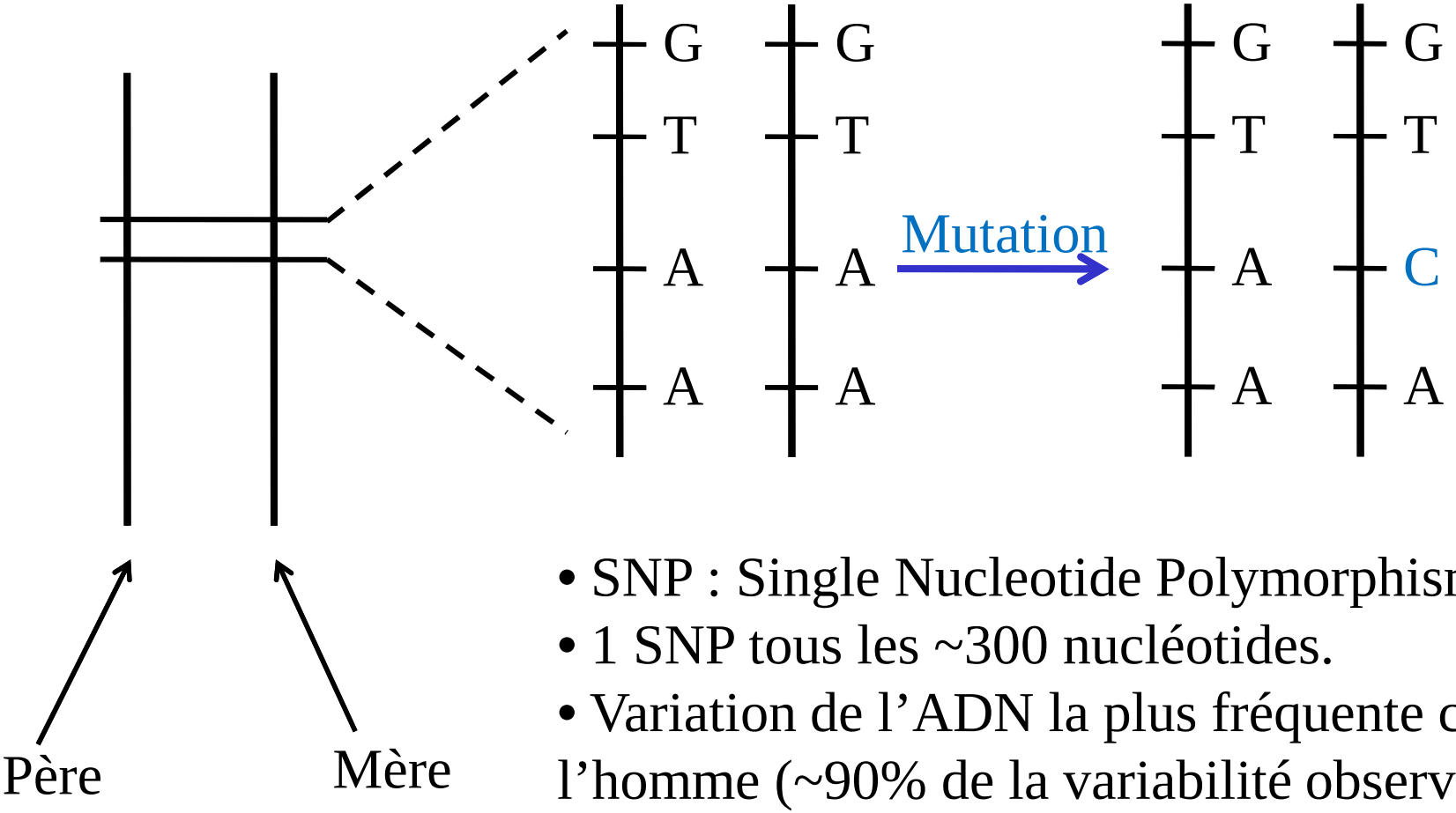
III. Une nouvelle approche : ISHAPE.

IV. Conclusion

I. LA PROBLEMATIQUE DES HAPLOTYPES

I.1. LES VARIATIONS GENETIQUES DU GENOME

23 paires de chromosomes



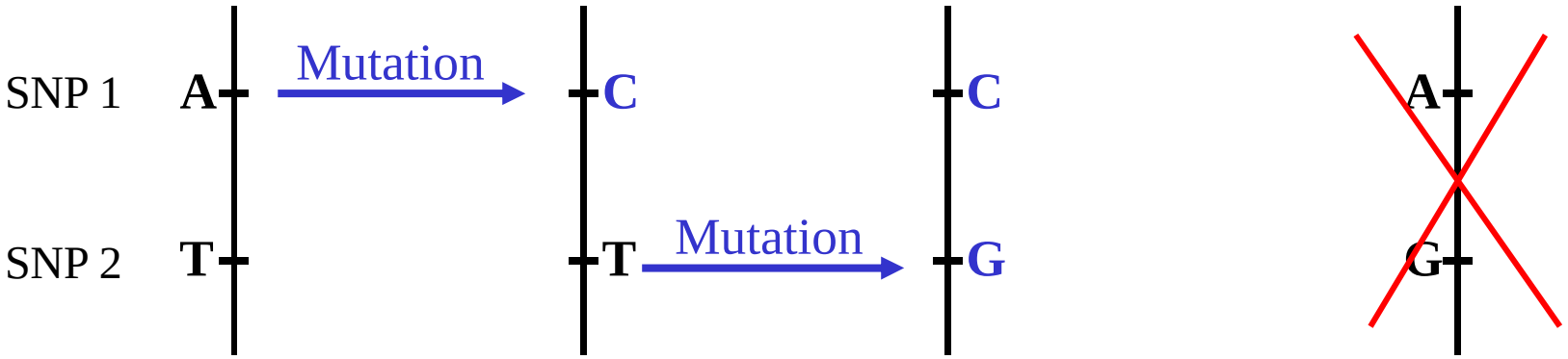
- SNP : Single Nucleotide Polymorphism.
- 1 SNP tous les ~300 nucléotides.
- Variation de l'ADN la plus fréquente chez l'homme (~90% de la variabilité observée).

I.2. DEFINITION DES HAPLOTYPES

1. Combinaison d'allèles sur un même chromosome.

2. Crées au cours de l'évolution au gré :

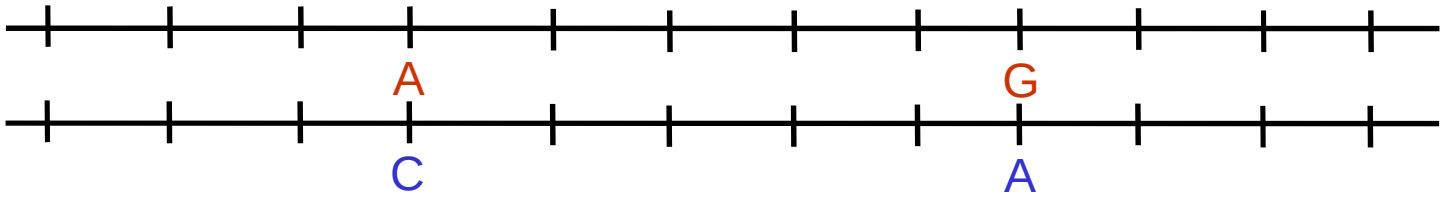
➢ Des mutations,



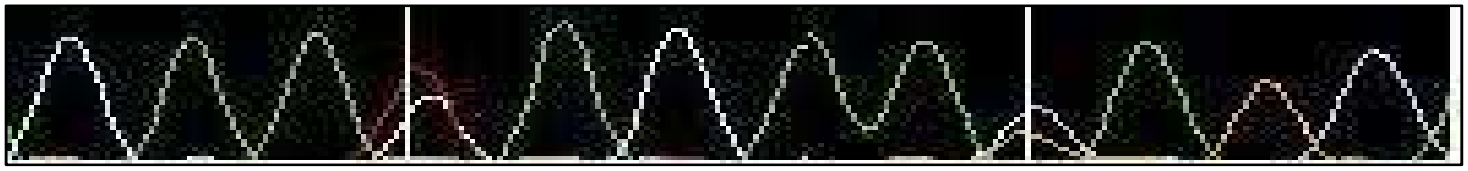
➢ Des recombinaisons, des dérives génétiques et des sélections.

I.3. POURQUOI RECONSTRUIRE LES HAPLOTYPES ?

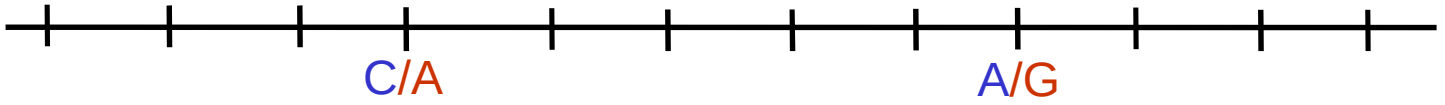
Individu



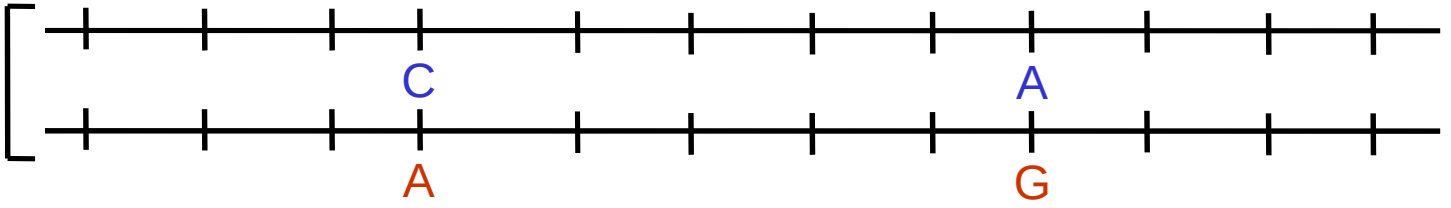
Génotypage



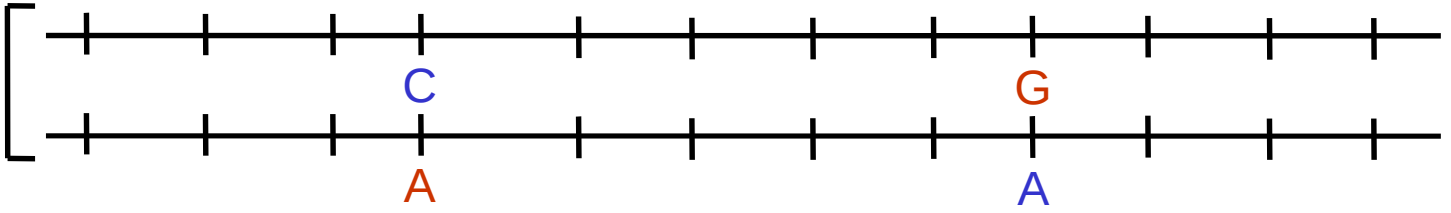
Génotype



2 diplotypes possibles



4 haplotypes possibles



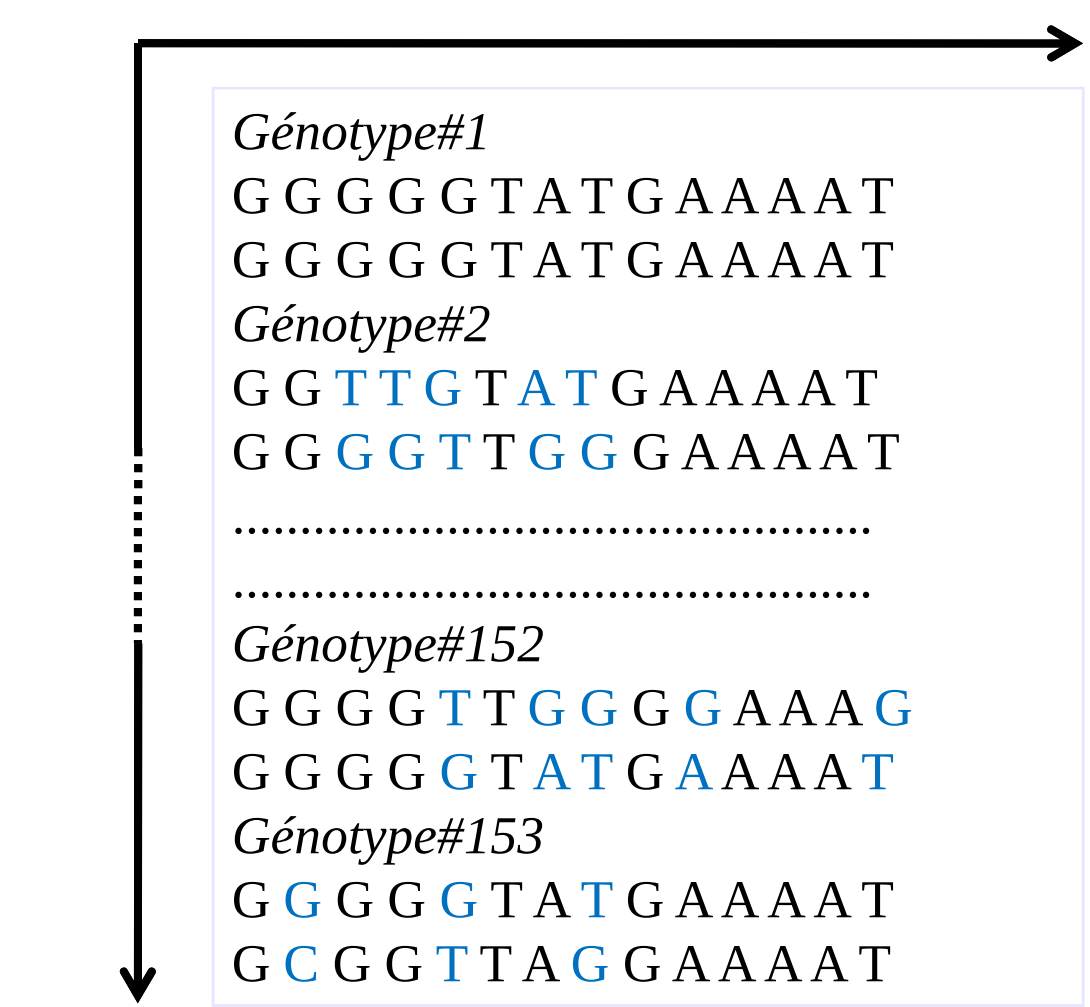
I.4. COMPLEXITE

Un génotype de N SNPs avec S sites hétérozygotes a :

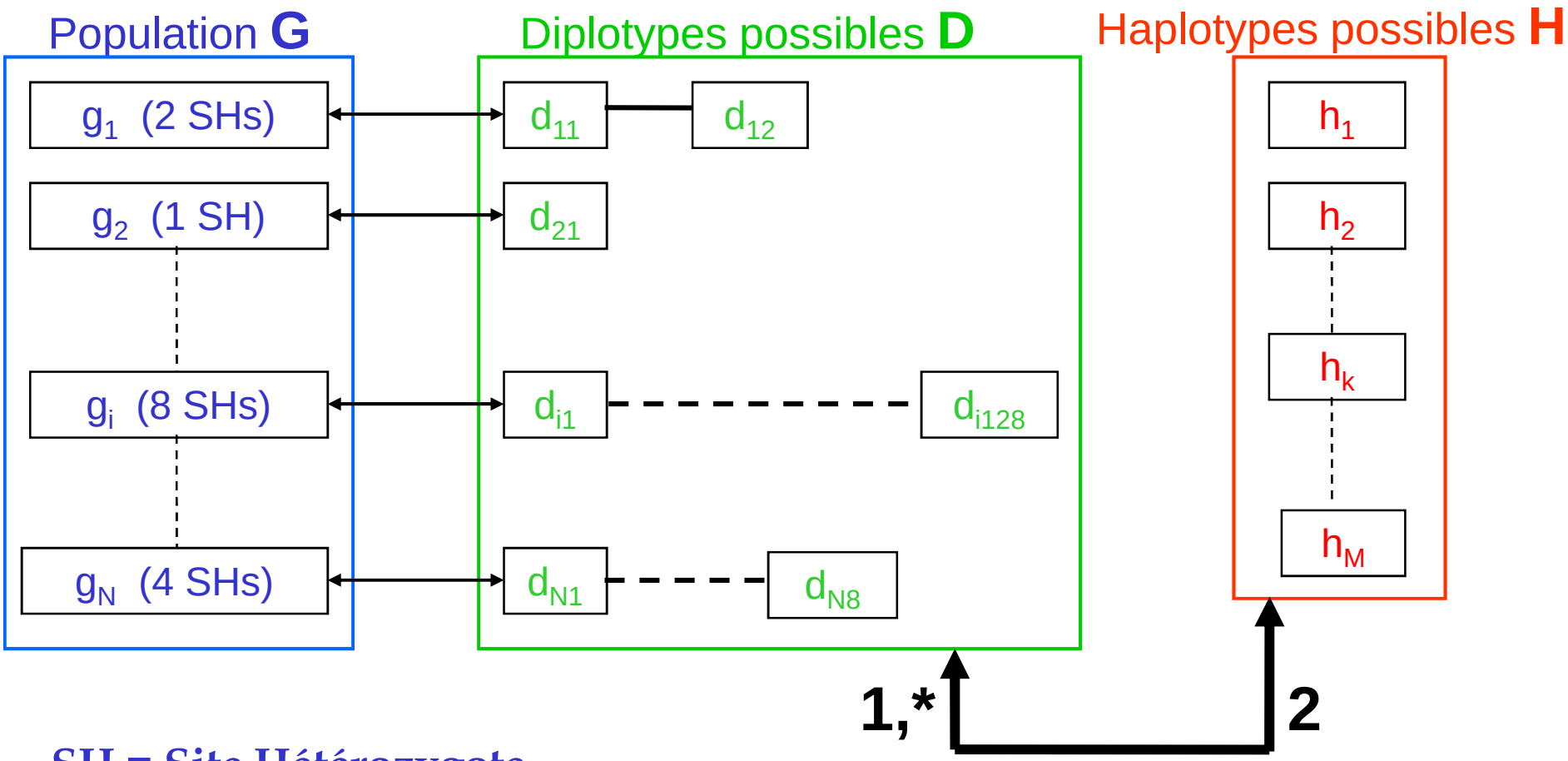
- 2^S haplotypes compatibles possibles,
- 2^{S-1} diplotypes compatibles possibles.

Nb de sites hétérozygotes	Nb d'haplotypes possibles	Nb de diplotypes possibles
5	32	16
10	1 024	512
20	1 048 576	524 288

I.5. DONNEES GENOMIQUES



I.6. REPRESENTATION



SH = Site Hétérozygote

II. LES PRINCIPALES METHODES DE RECONSTRUCTION D'HAPLOTYPES

II.1. HISTORIQUE

1990 : Clark AG: **Inference of haplotypes from PCR-amplified samples of diploid populations**. *Molecular biology and evolution.*

1995 : Excoffier L, Slatkin M: **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population**. *Molecular biology and evolution.*

2001 : Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *American journal of human genetics.*

2005 : Stephens M, Scheet P : **Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation**. *Am J Hum Genet*.

2007 : Delaneau O, Coulonges C, Boelle PY, Nelson G, Spadoni JL, Zagury JF : **ISHAPE: new rapid and accurate software for haplotyping.** *BMC Bioinformatics.*

II.3. METHODE STATISTIQUE : EM (1)

Soit f_k la fréquence de h_k, k = 1..M, alors la vraisemblance de P en fonction des f_k s'écrit :

$$L(f_1, ..., f_M) = Pr(G \mid H) = \prod_{i = 1..M} Pr(g_i \mid f_1, ... f_M)$$

Probabilité conditionnelle d'un génotype i à s sites hétérozygotes:

$$Pr(g_i \mid f_1, ..., f_M) = \sum_{j = 1..2^{s-1}} Pr(d_{ij} \mid f_1, ..., f_M)$$

Soit $d_{ij} = (h_l, h_k)$, et si P respecte l'équilibre de Hardy Weinberg :

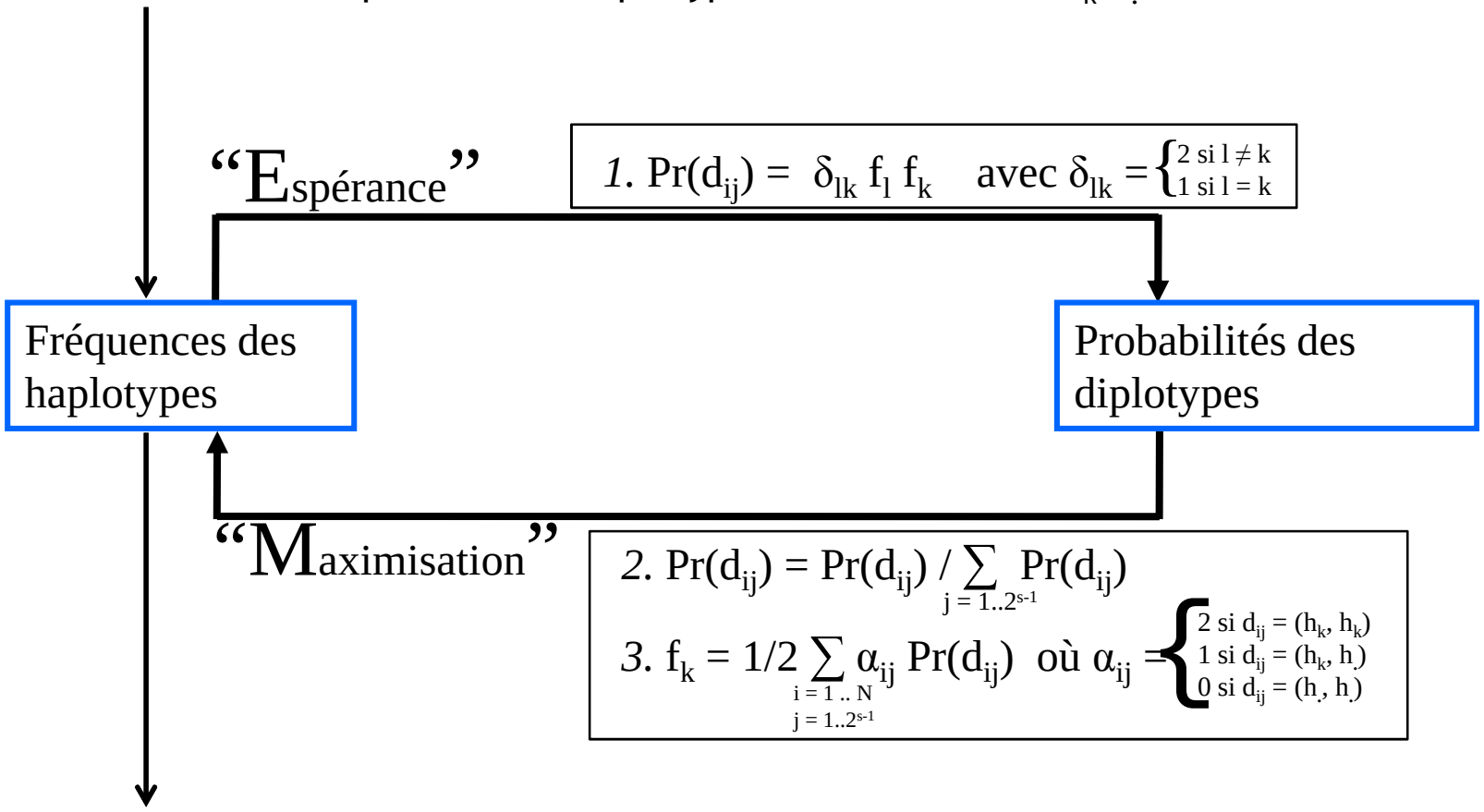
$$Pr(d_{ij}) = \delta_{lk} f_l f_k \quad \text{avec } \delta_{lk} = \begin{cases} 2 \text{ si } l \neq k \\ 1 \text{ si } l = k \end{cases}$$

→ Procédure itérative : algorithme Espérance - Maximisation

Excoffier L, Slatkin M: **Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population**. *Molecular biology and evolution* 1995.

II.3. METHODES STATISTIQUES : EM (2)

On initialise les fréquences des haplotypes aléatoirement : $f_k^{(0)}$.



Quand $f_k^{(t+1)} \approx f_k^{(t)}$ on a atteint un maximum de vraisemblance.

II.4. METHODES STATISTIQUES : PHASE (1)

On associe une variable aléatoire à chacun des haplotypes, et on estime leur distribution a posteriori des génotypes de P; $Pr(H|g_1, ..., g_N)$, avec :

- Un modèle de distribution d'haplotypes : différentes hypothèses sur la constitution en haplotypes d'une population (naïf, coalescence, recombinaison),

- Un echantilloneur de Gibbs (MCMC) pour approximer cette distribution sur les données observées $g_1, ..., g_N$.

Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *American journal of human genetics* 2001

II.2. METHODES STATISTIQUES : PHASE (2)

Pour tout i, on assigne à g_i un d_i' pris aléatoirement parmis les d_{ij} (D').

Soit O; un ordre aléatoire de traitement des g_i .

On itère un grand nombre de fois :

Pour chaque g_i de O :

1. Sélection d'un génotype g_i selon O
2. Pour tout j, calcul de $Pr(d_{ij} | D_{-i}')$: probabilité a posteriori de d_{ij} sachant $D_{-i}' = D' - \{d_i'\}$
3. Echantillonnage sur $Pr(d_{ij} | D_{-i}')$ pour assigner un nouveau d_i' à g_i

Chaque $D'^{(t)}$ constitue un état de la chaine de Markov.

II.2. METHODES STATISTIQUES : PHASE (3)

Modèle naif (Haplotyper)

g_i:	32344	d_{ij}:	32344	32334	32544	32534	33344	33334	33544	33534
	23534		23534	23544	23334	23344	22534	22544	22334	22344

Niu T et al : **Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms.** *The American Journal of Human Genetics* 2002

D_{-i}'
22544
22544
22544
22544
33334
33334
23233
14234

Modèle de coalescence (PHASE v1.0)

g_i:	32444	d_{ij}:	32444	32434	33444	33434
	23434		23434	23444	22434	22444

Stephens M, Smith NJ, Donnelly P: **A new statistical method for haplotype reconstruction from population data**. *American journal of human genetics* 2001

Modèle de recombinaison (PHASE v2.1)

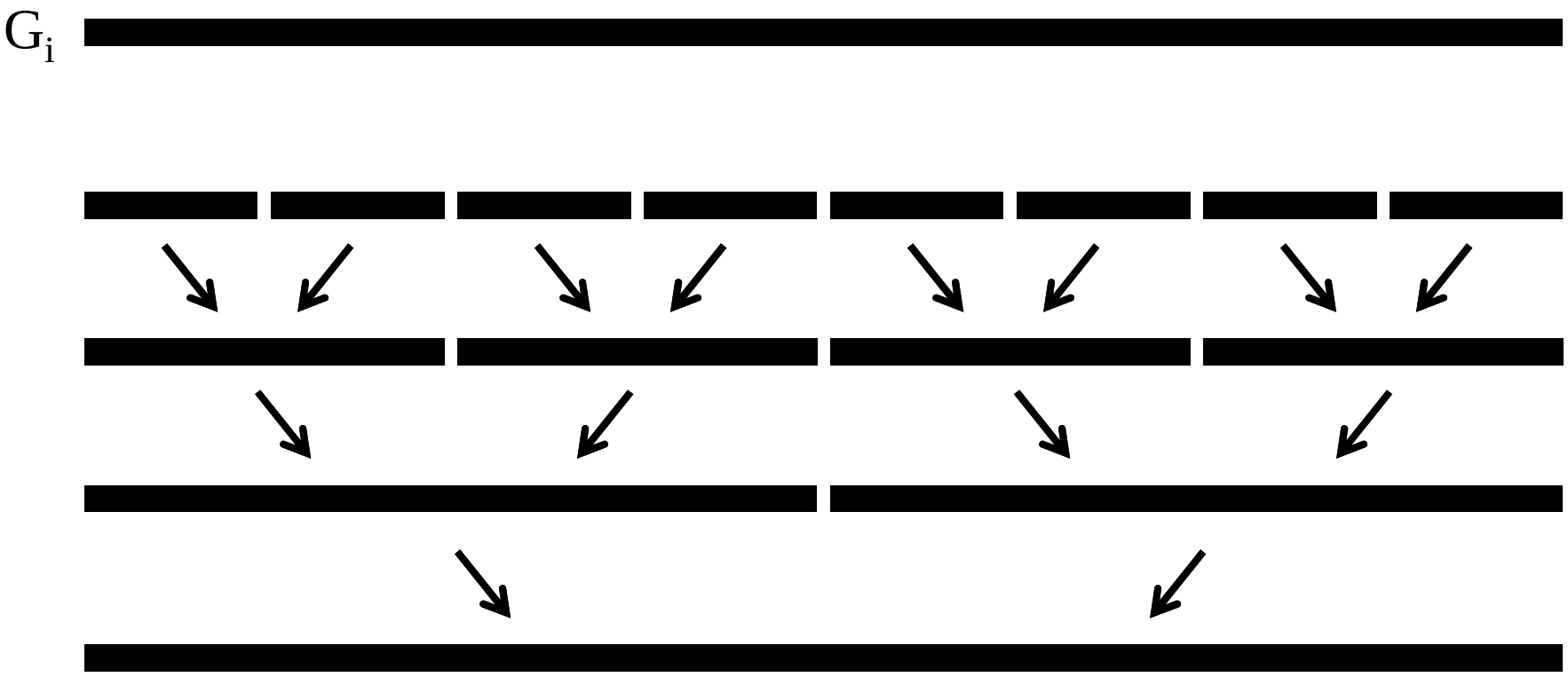
g_i:	22333	d_{ij}:	22333	22334
	22234		22234	22233

Stephens M, Scheet P : **Accounting for Decay of Linkage Disequilibrium in Haplotype Inference and Missing-Data Imputation**. *American Journal of Human Genetics 2005*

II.2. METHODES STATISTIQUES : PL

Partition – Ligation permet :

- de briser l'aspect exponentiel du problème,
- de traiter plus de SNPs.



Niu T et al : **Bayesian Haplotype Inference for Multiple Linked Single-Nucleotide Polymorphisms.** *The American Journal of Human Genetics* 2002

II.3. UNE NOUVELLE APPROCHE : ISHAPE

II.3. ISHAPE (1)

Constat sur PHASE :

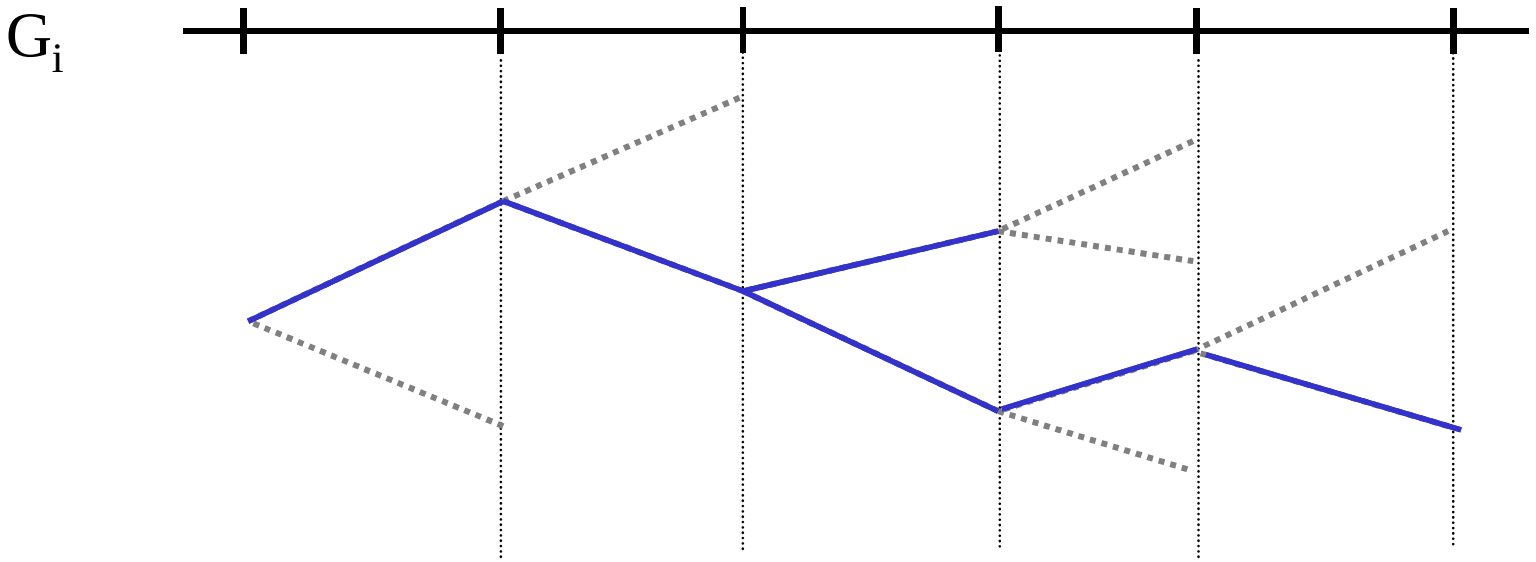
- très performant,
- très lent,
- difficultés à converger correctement.

Une solution consiste à réduire l'espace des diplotypes possibles en résolvant tout d'abord les sites les moins ambiguës.

- Limite les calculs postérieurs,
- Facilite la convergence du Gibbs sampler.

II.3. ISHAPE (2)

IEM (Itérative EM) : algorithme EM très rapide où les haplotypes sont construits progressivement en incluant les SNPS un par un.



II.3. ISHAPE (3)

Bootstrap IEM : On génère X (=500) échantillons bootstrap de **P** dont on estime les fréquences haplotypiques par IEM avec un ordre aléatoire d'inclusion des SNPs.

=> Dessine un espace de diplotypes candidats de taille réduite.

A. Taux de capture sur GH1 (14 SNPs)

Prog.\MD	0%	2%	5%	10%
Ishape	**0.99**	**0.99**	**0.98**	**0.97**
Phase 2.1	0.98	0.97	0.97	0.96
fastPhase	0.88	0.87	0.82	0.76
PLEM	0.91	0.90	0.89	0.86

B. Réduction du nombre de diplotypes sur GH1

\ MD	0%	2%	5%	10%
Nb. diplotypes possibles	9.6	18.7	48.7	244.1
Nb. diplotypes candidats	2.3	3.3	5.4	10.2

Delaneau et Al : **ISHAPE: new rapid and accurate software for haplotyping**. *BMC Bioinformatics 2007*.

II.3. ISHAPE (4)

Echantilloneur de Gibbs : adaptation de PHASE sur un nombre limité de diplotypes candidats.

C. Résultats sur HapMap – CEU (10 à 80 SNPs)

	SNPs contigus			SNPs 5kb		
Prog.	SER	Class.	Temps	SER	Class	Temps
Ishape	**1.10**	**1.83**	34.8	3.60	**1.92**	66.1
Phase 2.1	1.17	2.11	215	**3.57**	2.03	702
Phase 1.0	1.39	2.67	52.1	4.92	3.81	142.5
fastPhase	1.31	2.73	100.3	3.98	2.71	88.8
PLEM	1.56	3.07	**22.1**	5.16	3.71	**19.1**

II.4. CONCLUSION

II.4. Conclusion

1. Nouvelle méthode qui utilise la puissance de l'EM et la précision de PHASE, en s'appuyant sur la réduction de l'espace des solutions possibles.

2. Les résultats obtenus montrent que ce logiciel est jusqu'à 10 fois plus rapide que PHASE et aussi fiable.

3. Utile pour l'exploitation des « Genome-Scan » 500 000 SNPs où ISHAPE peut permettre un important gain de temps et de précision.