

# **MICROSATELLITES ET MODELES MUTATIONNELS THEORIQUES**

**Arnaud Estoup**

Laboratoire Modélisation et Biologie Evolutive, CBGP-INRA, 34090

Montpellier, France.

Department of Zoology & Entomology, University of Queensland, Qld, 4072,  
Australia.

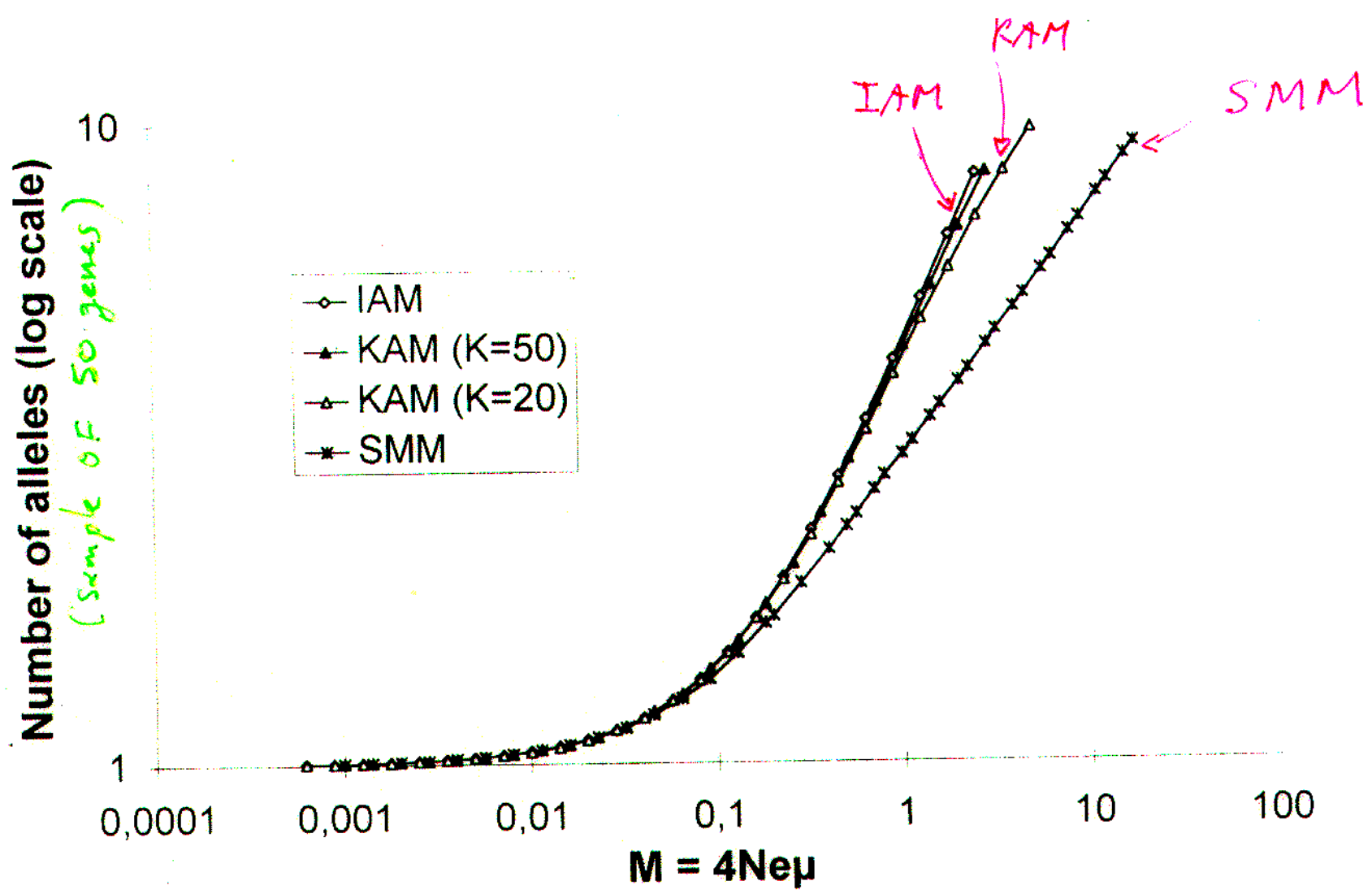
- 1/ Intérêts des modèles théoriques
- 2/ Modèles théoriques simples
- 3/ Tests des modèles théoriques simples
- 4/ Modèles théoriques complexes
- 5/ Homoplasie de taille

## Why choosing a theoretical model?

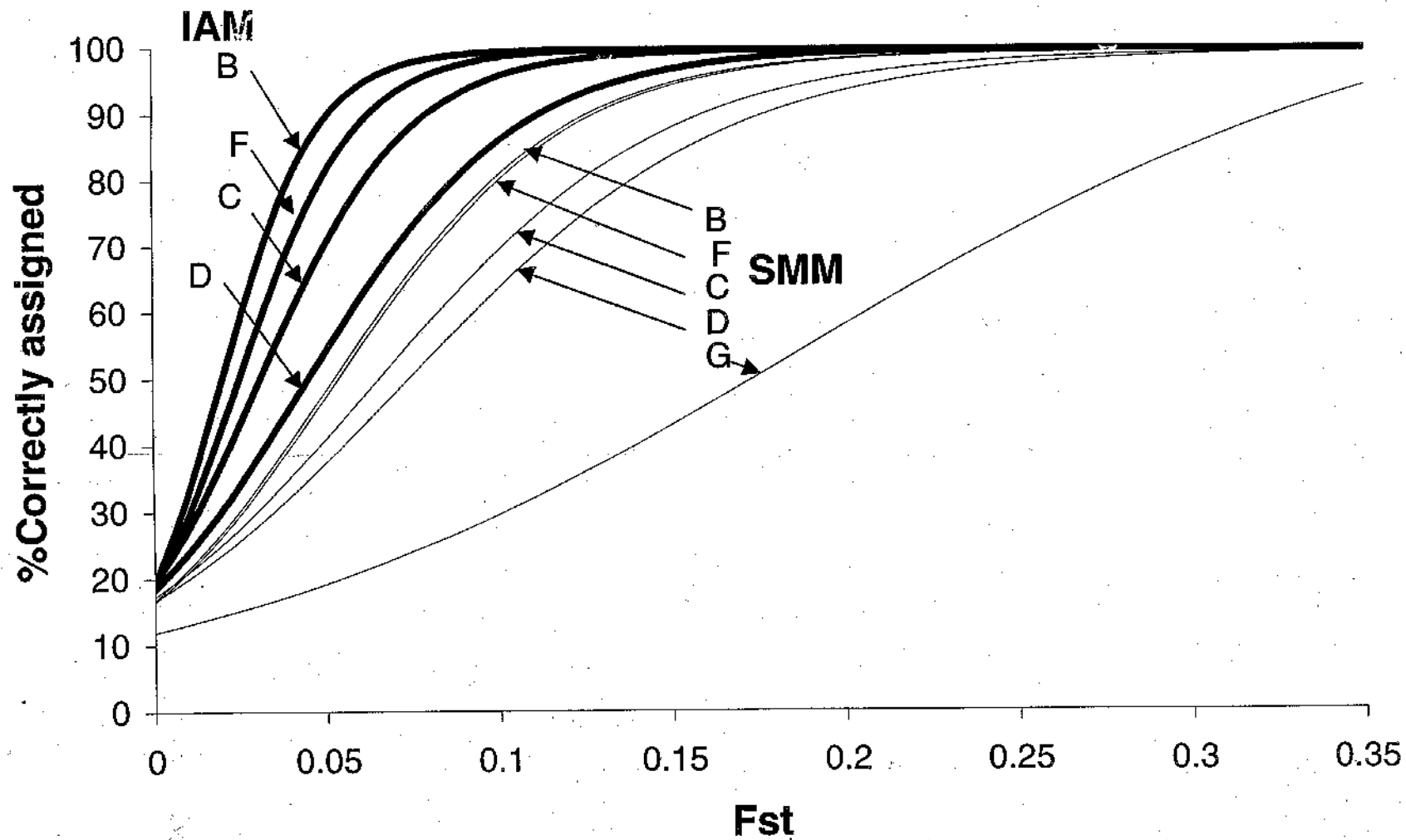
① The estimation of numerous population parameters (e.g. genetic differentiation, number of migrants per generation, etc.) is dependent upon the mutation model assumed for the markers

② The sensitivity to the mutation model of these population parameter estimates increases with the mutation rate and microsatellites have generally high mutation rates.

- Pop x locus at mutation-drift equilibrium -



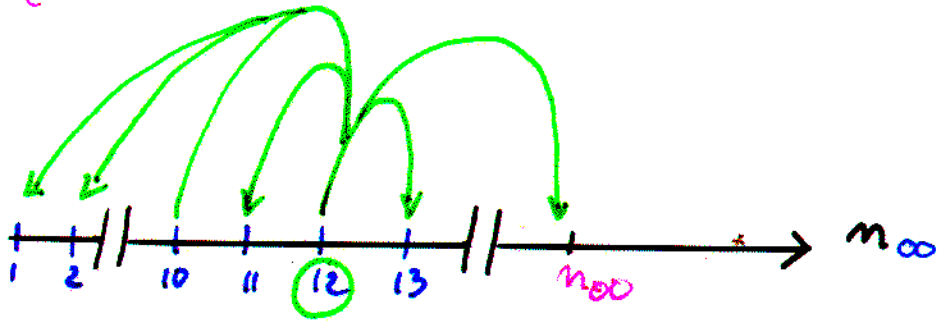
(unp. results)



# Theoretical mutation models

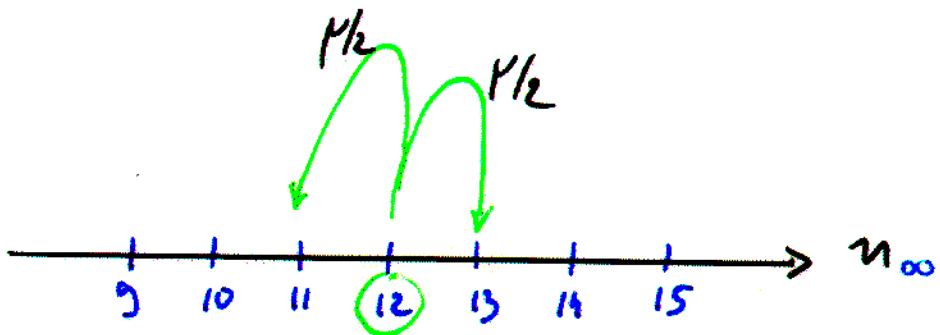
IAM (Infinite allele model)

(CT)<sub>12</sub>



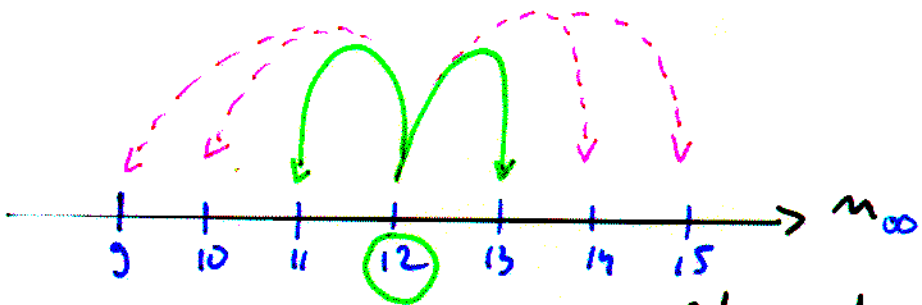
SMM (stepwise mutation model)

(CT)<sub>12</sub>



TPM (Two phase model) → GSM

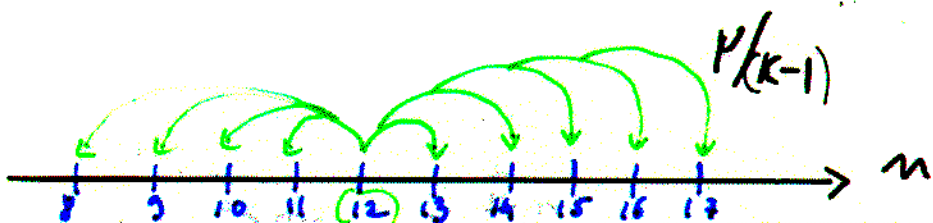
(CT)<sub>12</sub>



Gain/loss of  $X$  repeats -  $X=1$  with proba  $P$   
 $(1-P)^{|i-j|-1}$  -  $X$  follows a geometric law with proba  $1-P$

KAM (K-allele model)

(CT)<sub>12</sub>  
 $K=10$



INFORMATION FROM PEDIGREE STUDIES: (+ Small pool PCR)

MICROSATELLITES

1/ Human pedigrees: → TPM

(e.g.: Weber and Wang 1993)

(85-95% of one repeat change)

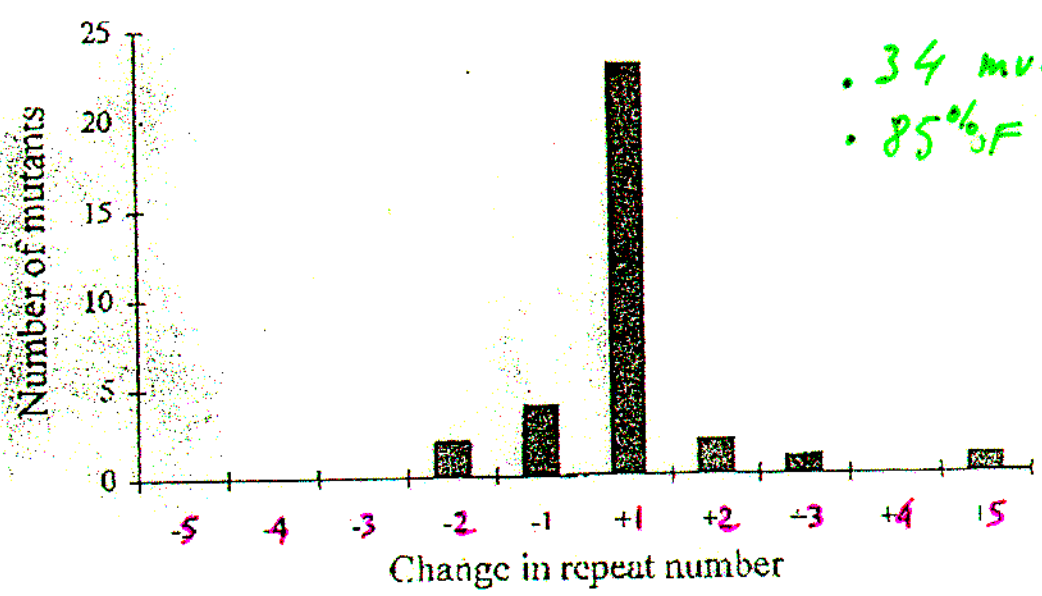
**PB:** low number of mutation events / "multilocus"

conclusions / somatic mutations

2/ Hypervariable tetranucleotide locus in a swallow:

→ TPM

(Primmer et al. 1996)



. 34 mutation events  
 . 85% of one repeat changes

**PB:** Generalization ?

MINISATELLITES

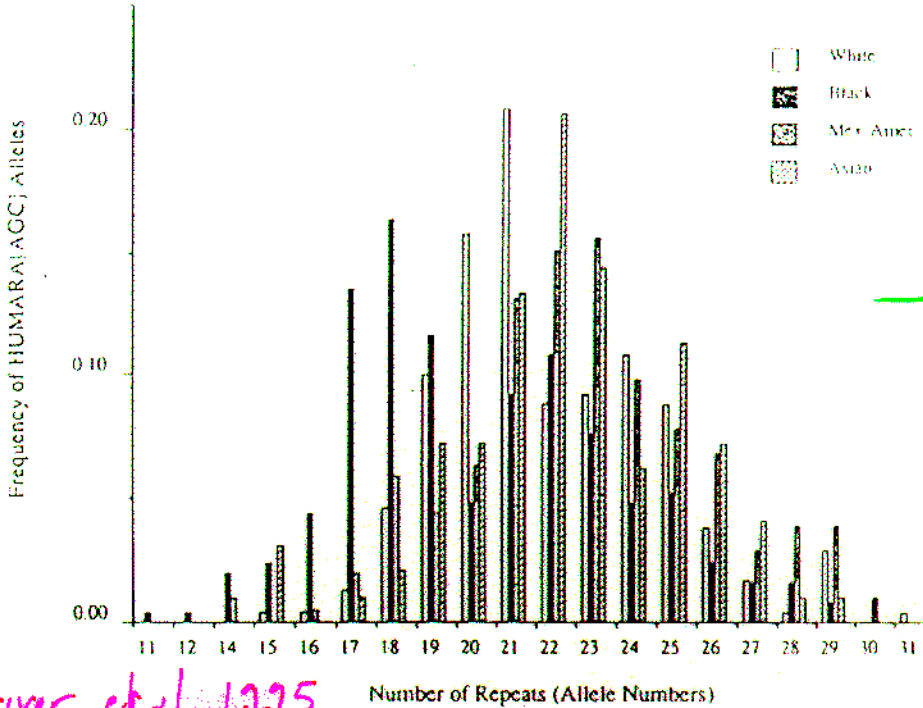
SP-PCR + Human pedigrees: High rate of multiple repeat unit changes ("Small" changes = most frequent)

→ TPM ← IAM  
 KAM

(e.g. Berglund et al. 1991)  
 Jeffrey et al. 1994)

Theoretical mutation models:  
direct information from population data

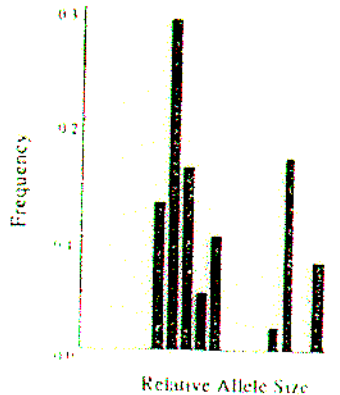
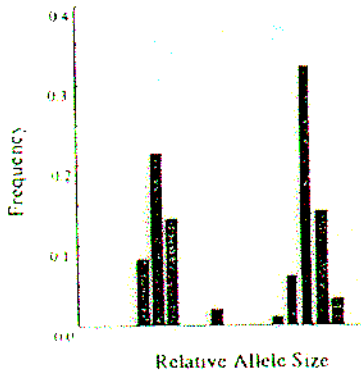
Microsatellite



→ SMM?

1995

Microsatellite



→ ~~SMM?~~  
 • large mutations?

Valdes et al. 1993

# Tests statistiques des modèles mutationnels théoriques : deux approches populationnelles différentes

1/ Comparer la valeur observée de l'hétérozygotie ( $H_o$ ) ou du nombre d'allèle ( $k_o$ ) à celle attendue sous un modèle mutationnel donné ( $H_e$  ou  $K_e$ )

Eq mutation - dérive :

$$k_o \text{ (n gènes)} \rightarrow \ominus_{IAM/SMM/TPM/KAM} \rightarrow H_e \text{ / comparer avec } H_o$$

$$H_o \text{ (n gènes)} \rightarrow \ominus_{IAM/SMM/TPM/KAM} \rightarrow k_e \text{ / comparer avec } k_o$$

2/ Le paramètre population  $\times$  locus considéré est la variance intrapopulation du nombre de répétitions ( $V$ )

Eq mutation - dérive :

$$V = D_{SMM/TPM/KAM} (\ominus = 4N_e\mu)$$

$$V = \text{aucun sens sous IAM}$$

Simulations :  $V_o \rightarrow IC (95\%) \text{ sur } H_o$

## Application of the statistical tests to actual population data

Locus	Method	Model tested	Conclusion	Authors
2 bp repeat unit	1	SMM, IAM	closer to SMM / IAM	Edwards <i>et al.</i> , 1992
2 bp repeat unit	2	SMM	fit SMM	Valdes <i>et al.</i> , 1993
3-5 bp repeat unit	1	SMM, IAM	fit SMM	Shriver <i>et al.</i> , 1993
1 bp and 2 bp repeat unit	1	SMM, IAM	Bad fit to the SMM	Shriver <i>et al.</i> , 1993
15-70 bp repeat unit	1	SMM, IAM	<u>Strong deviation to the IAM</u>	Shriver <i>et al.</i> , 1993
2 bp repeat unit	1	SMM, IAM	SMM not appropriate for all loci studied	Deka <i>et al.</i> , 1995
2 bp repeat unit	2	SMM, TPM	better fit to a TPM (p=0.80-0.95) than to a SMM	Di Rienzo <i>et al.</i> , 1994
2 bp (interrup.) repeat unit	2	SMM	Bad fit to the SMM	Estoup <i>et al.</i> , 1995a
2 bp + Compound	1	SMM, IAM	IAM (SMM) not ruled out for any loci	Estoup <i>et al.</i> , 1995b

MINI

CCL "contradictory" or inconclusive results
 

- low power of tests? (+ basic assumptions)
- large variance among loci
- involvement of more complex mutation processes than those assumed for the theoretical r.n.m.

Directional evolution:

(microsatellites)

Pedigree collections:

Multilocus data = Human germline mutations

(Amos et al. 1996)

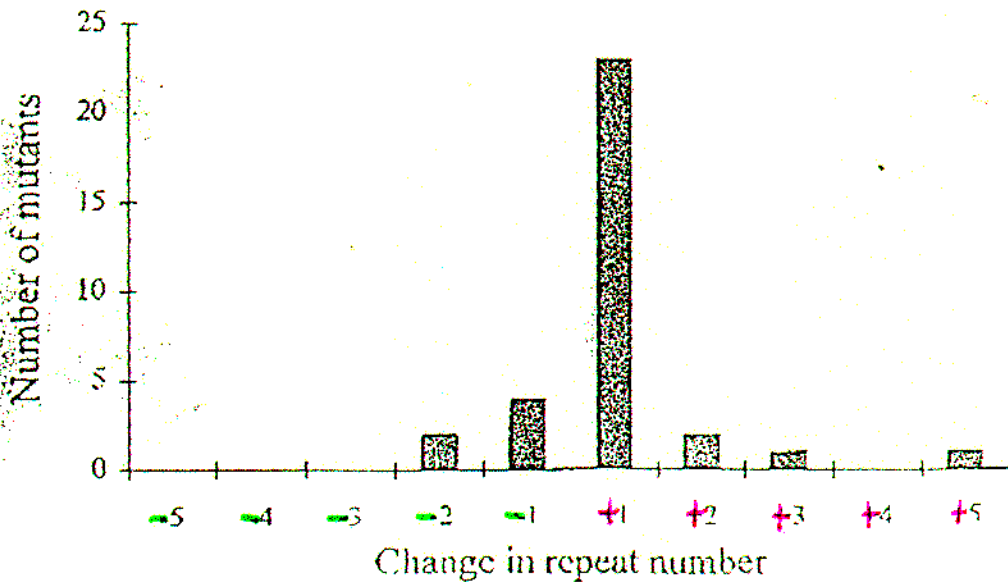
21 gains / 9 losses (unilateral binomial test,  $P=0.021$ )

Monolocus data = germline mutations at a single

hypervariable barn swallow tetranucleotide locus

(Primmer et al. 1996)

26 gains / 7 losses (unilateral binomial test,  $P=7 \times 10^{-4}$ )



Consequences:

• reduce the probability of fixation of microsatellite loci by preventing their evolution to a low mutating state corresponding to very short repeat arrays

• facilitate the emergence of polymorphic microsatellite loci from very short repeat arrays

(Tadchida and Lizuka 1992)

Constraints on allele size:

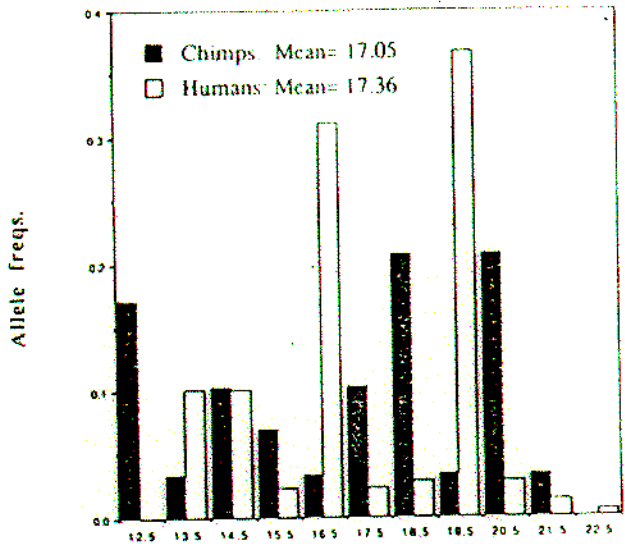
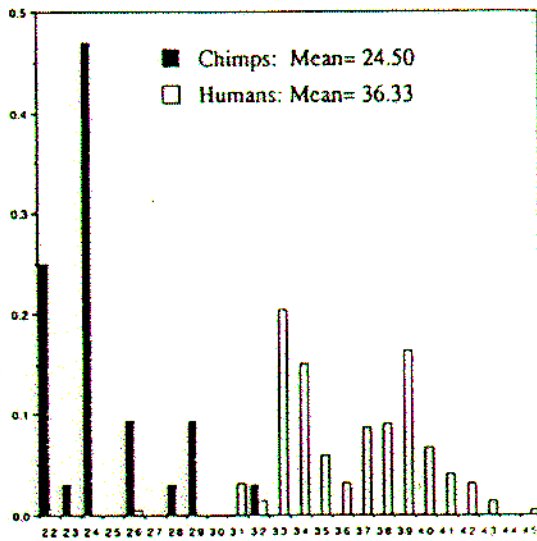
(Microsatellites)

→ Microsatellite loci have a finite size generally shorter than a few tens of repeat units

→ Low allele size variance ratio between non-human primate and human populations

"Mfd 75"

"Mfd 3"



No. of repeats

No. of repeats

(Gerya et al. 1995)

TABLE 1 Variation in allele length at ten microsatellite loci within humans and great apes

Locus	Mean allele length (base pairs)				Variance for non-humans	Variance for humans	F ratio* (d.f.) (non-human/human)
	Human	Chimp	Gorilla	Orang			
ACTC	83.1	78.0	75.9	75.7	21.8	470.88	0.31 (45,263)
D13S133	154.9	144.6	122.4	123.2	132.50	422.76	0.31 (45,265)
D13S137	111.5	105.6	113.5	99.4	93.61	48.53	1.93 (47,267)
D13S227	152.7	152.7	137.6	144.0	54.37	21.41	2.54 (47,259)
D13S193	136.6	134.3	140.6	116.0	123.40	47.37	2.61 (47,273)
O84XCS	83.4	81.5	81.6	87.4	10.19	23.34	0.44 (45,265)
D13S119	132.5	129.0	135.5	124.7	42.46	34.69	1.22 (43,283)
D13S118	193.1	191.5	190.6	190.8	15.17	12.05	1.26 (39,281)
D13S125	147.4	141.4	148.3	141.0	120.57	49.08	2.46 (45,269)
Utsw1523	179.1	175.8	174.9	173.6	18.62	5.23	3.56 (47,259)

\* Critical values for F at P=5% are between 1.44 and 1.49.

(Bowcock et al. 1993)

## The nature of allele size constraints : hypotheses

- 1/ Small alleles tend to increase in size while large alleles tend to decrease in size  
(Gargal et al. 1995)
- 2/ Excess of small gains over losses balanced by rare large mutations  
(Gargal et al. 1995; Weber and Wong 1993; Amos et al. 1996)
- 3/ Counter-selection of large alleles  
(Feldman et al. 1997; Samadi et al. 1998)
- 4/ Counter-selection proportional to the difference in size between the two alleles borne by homologous chromosomes in a diploid individual  
(Samadi et al. 1998)
- 5/ Expansion slow down by accumulation of substitutions (interruptions) within large alleles → lower mutation rate → degeneration into random sequences

**Remark:** if dependent on the absolute size (in bp) of the stretch of repeats constraints on the number of repeated motifs may be stricter when the repeat size increases

constraints: TETRA > TRI > DI

Allele size: DI > TRI > TETRA  
variance

## Variability and repeat size

(microsatellites)

Two-way ANOVA for the natural logarithm of the within-population variance ( $V$ ) of microsatellite loci

(Chakraborty et al. 1997)

Source of data	Component of variance	Mean squares	df	P value
DeKal et al 1995	<u>Locus motif type</u>	13.0	3	<0.001
	Population	0.16	8	0.99
	Interaction	0.07	24	1
	Within/residual	0.76	126	—
Hammond et al 1994	<u>Locus motif type</u>	5.0	1	0.001
	Population	0.16	3	0.75
	Interaction	0.04	3	0.96
	Within/residual	0.38	40	—

$\mu_{DI} > \mu_{TRI} > \mu_{TETRA}$

- Studies based on direct observations of mutations suggest a higher mutation rate for tetranucleotide loci (e.g. Weber and Wang 1993)
- Variances in number of repeats are proportional to the product of the mutation rate and the variance of the allele size change by mutation
- differences in the modalities of mutation for DI / TRI / TETRA ?

## Variability and repeat composition

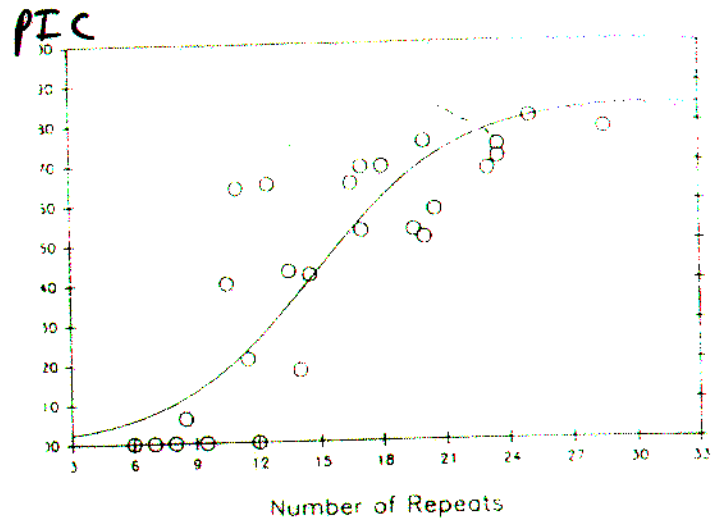
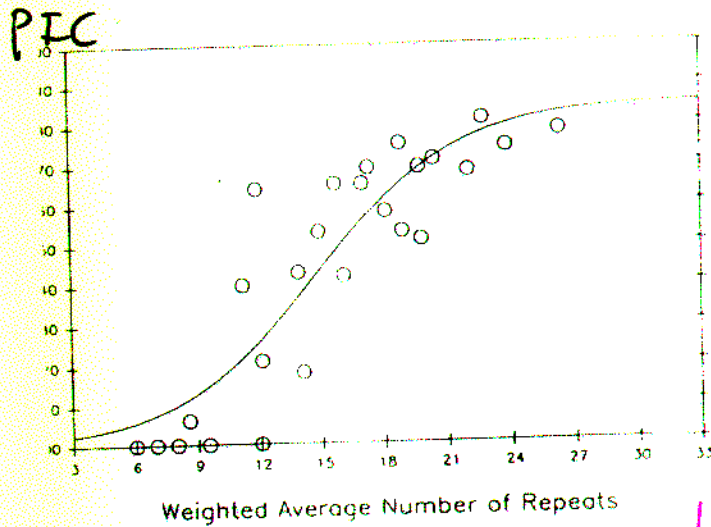
(microsatellites)

- For a given repeat size, the level of microsatellite polymorphism seems to depend upon the composition of the repeated motif (tri- and tetranucleotide motifs) (Gastier et al. 1995 ; Sheffield et al. 1995)
- G/C content appears to be an important factor since the most polymorphic tri- and tetranucleotide motifs are A/T rich (Gastier et al. 1995)

## Variability and tract length / purity

(microsatellites)

The polymorphism of uninterrupted (CA)<sub>n</sub> markers increased with increasing average number of repeats



Weber 1990)

- Stronger correlation between the maximum repeat count and variance than between the mean and variance → the mutation rate does indeed increase with repeat count, but not necessarily smoothly.

(Goldstein and Clark, 1995)

Interrupting bases stabilise the tract of repeats

(microsatellites)

**Locus INRAO11**

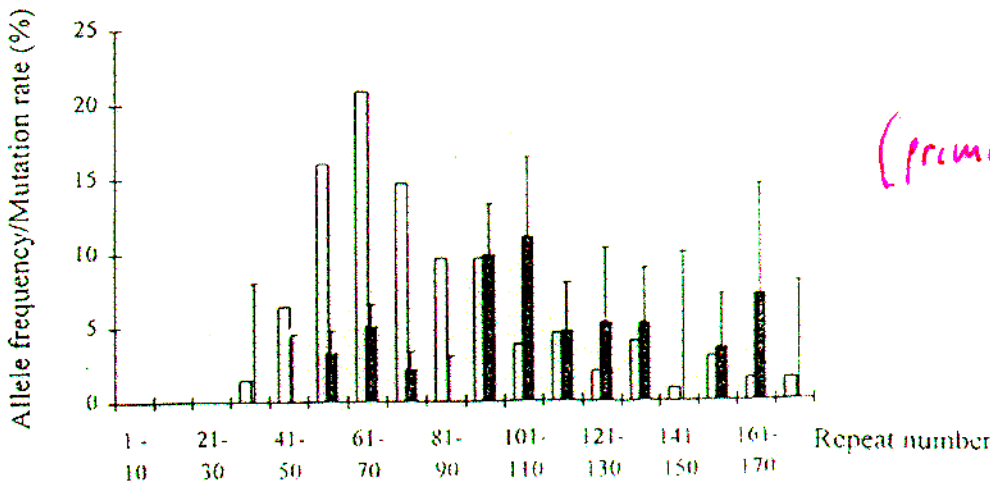
Goat — (CA)<sub>27</sub> — PIC = 0.92

Cattle — (CA)<sub>8</sub>TA(CA)<sub>9</sub> — PIC = 0.41

# Differences in mutability among alleles at the same locus

(Microsatellites)

## Single hypervariable barn swallow tetranucleotide repeat locus



(premmier et al. 1996)

## Trinucleotide microsatellites associated with some human diseases, e.g.

the Spinocerebellar ataxia type 1

(Chong et al. 1993)

(CAG)<sub>n</sub>

instable

(CAG)<sub>n1</sub>CAT(CAG)<sub>n2</sub>

stable

with n1+n2 ≥ n

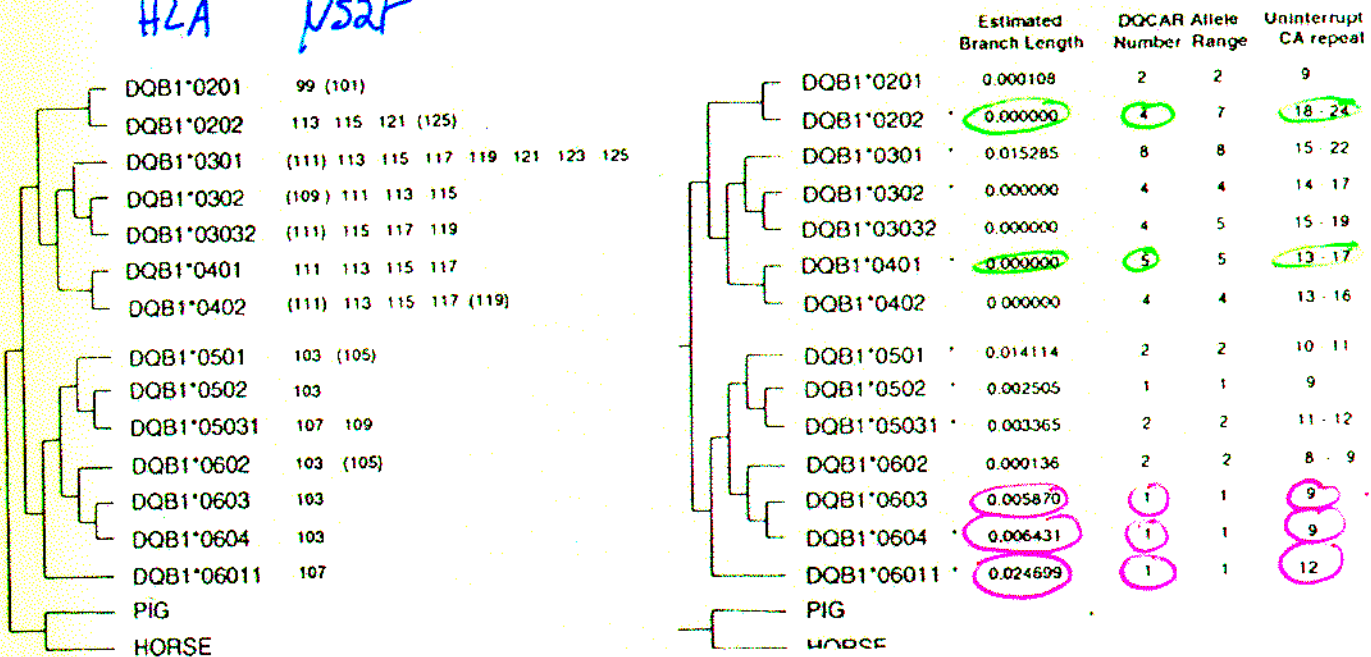
## Phylogenetic approach - a dinucleotide microsatellite locus tightly linked

to a HLA locus

(Jin et al. 1996)

HLA

µsat



# ESTIMATION SIMULTANEE DES PARAMETRES D'UN MODELE MUTATIONNEL COMPLEXE A PARTIR DE DONNEES POPULATIONELLES

Fu & Chakraborty 1998, Genetics 150, 487-497

## Modèle mutationnel Complexe

Distribution  $D_{ij}$  = probabilité qu'une mutation cause un changement de taille de  $i$  à  $j$

$$D_{ij} \text{ tel que } \begin{cases} \alpha(1-P)P^{j-i-1}, & \text{pour } j > i \\ (1-\alpha)(1-P)P^{j-i-1}, & \text{pour } j < i \end{cases}$$

$\alpha$  = probabilité qu'une mutation augmente la taille de l'allèle

Géométric distribution  $(1-P)P^{\text{abs}(i-j)-1}$

Vecteur à quatre paramètres  $\Gamma(\Theta=4Ne\mu, P, \alpha, A)$

$A$  = taille du MRCA

## Méthode des moindres carrés

(ou méthode par maximum de vraisemblance)

Estimation du  $X^2$  minimum du Vecteur  $\Gamma$

$$X^2 = \sum_{i=1}^n \frac{(f_i - e_i(\Gamma))^2}{e_i(\Gamma)}$$

$f_i, i=1, \dots$  nombre d'allèles de taille  $i$  dans un échantillon de  $n$  chromosomes

$$e_i(\Gamma) = E(f_i | \Gamma)$$

Simulations basées sur le processus de coalescence  $\rightarrow$  calcul de  $X^2$  pour  $\Theta = 0.2$  (0.2) 10,  $P = 0.01$  (0.01) 0.10 (0.05) 0.90,  $\alpha = 0.5$  (0.05) 1.00

## Application: Sept locus microsatellites chez l'Homme (8 pop)

- Chaque mutation à une légère tendance à causer une augmentation de taille ( $\alpha > 0.5$ )
- Modèle multi-step plutôt que mono-step ( $P > 0$ )

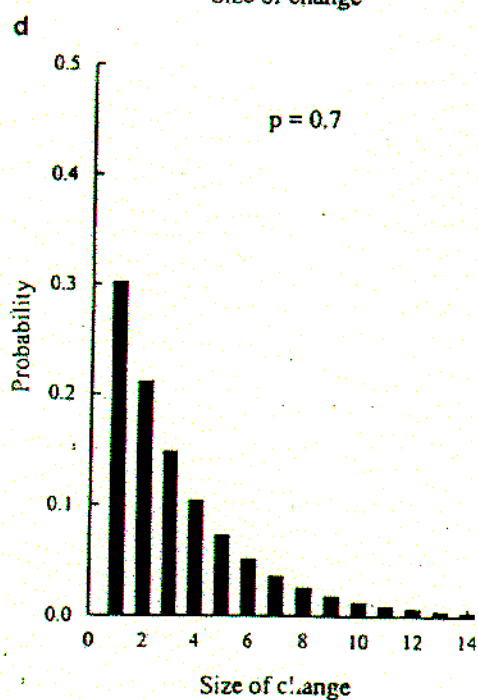
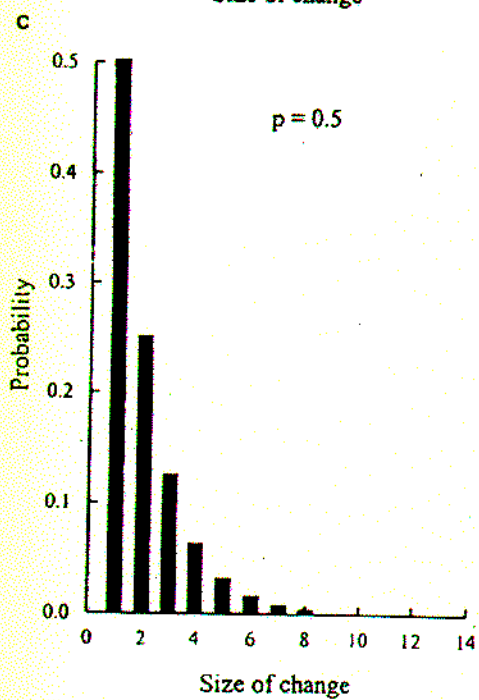
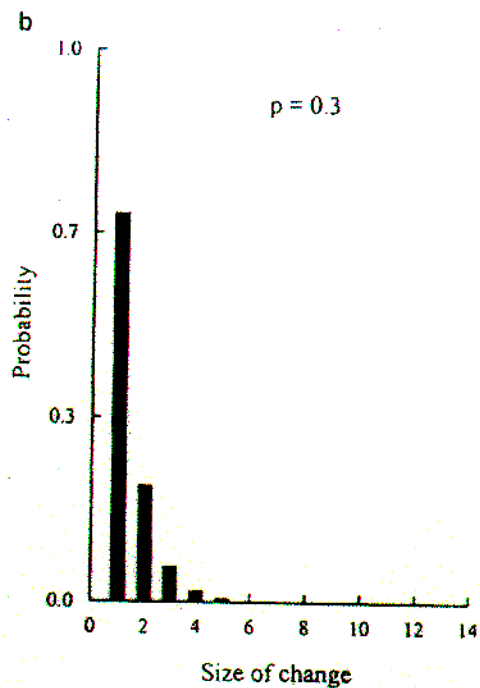
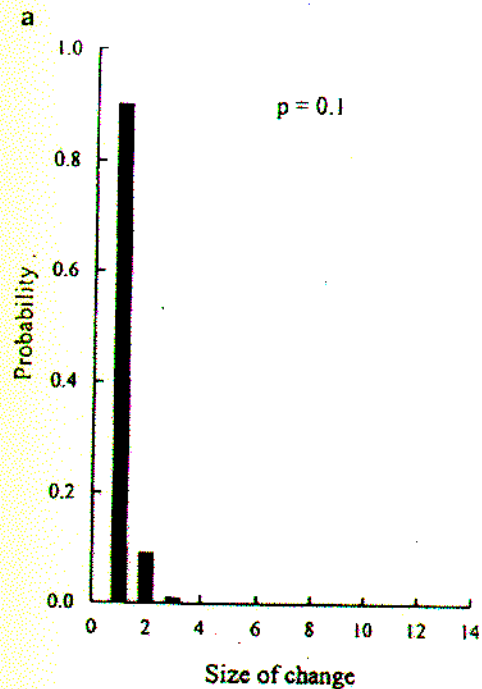


FIGURE 1.—Numerical examples of geometric distribution  $(1 - P)P^{i-1}$ .

Estimated values ( $\times 100$ ) of parameters  $\theta$ ,  $P$ ,  $\alpha$ , and  $A$ 

Locus	Parameter	Populations								
		SA	DG	PH	NG	KA	GR	CP	SO	CH
FLT1	$\theta$	60	20	20	140	180	40	40	180	180
	$P$	70	65	65	60	50	70	70	60	60
	$\alpha$	95	100	100	90	100	80	80	90	10
	$A$	168	168	168	168	168	168	168	168	176
D13S118	$\theta$	60	320	120	60	220	260	180	140	200
	$P$	55	1	50	55	9	10	25	45	2
	$\alpha$	100	80	95	60	100	45	35	65	35
	$A$	190	190	190	190	188	194	194	190	186
D13S121	$\theta$	60	60	40	60	140	120	120	240	300
	$P$	45	55	60	70	30	50	50	40	35
	$\alpha$	100	90	90	95	75	75	75	60	30
	$A$	166	166	166	166	166	166	166	166	170
D13S71	$\theta$	220	40	40	80	140	240	260	140	20
	$P$	25	50	2	50	25	6	2	20	2
	$\alpha$	45	65	20	55	10	35	15	35	90
	$A$	73	75	75	73	77	75	77	75	71
D13S122	$\theta$	880	160	100	300	140	240	240	300	100
	$P$	3	30	60	30	65	50	50	25	2
	$\alpha$	25	10	80	5	40	45	55	15	40
	$A$	105	103	95	109	95	95	95	103	79
D13S193	$\theta$	240	240	260	140	260	220	220	220	140
	$P$	55	40	55	65	40	55	55	45	40
	$\alpha$	100	100	10	75	95	15	15	80	80
	$A$	131	131	147	133	131	147	147	131	133
D13S124	$\theta$	120	20	20	60	80	100	100	120	300
	$P$	10	2	5	1	10	1	15	15	4
	$\alpha$	25	25	15	0	30	95	95	60	100
	$A$	185	187	187	187	187	185	185	185	179

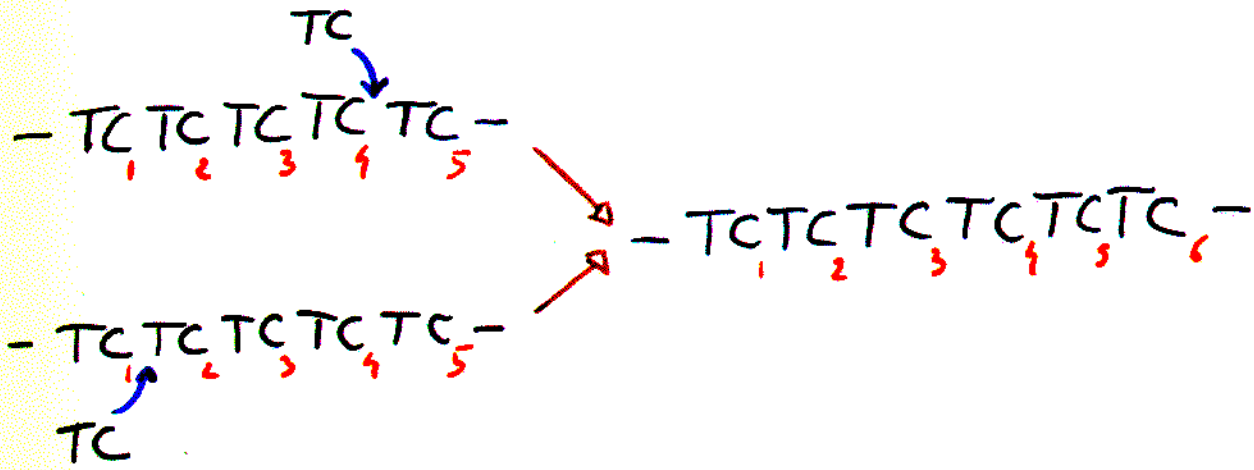
## Homoplasie de taille

- **Définition** : Deux électromorphes de tailles identiques peuvent avoir des séquences ancestrales différentes = identité par état et non par ascendance

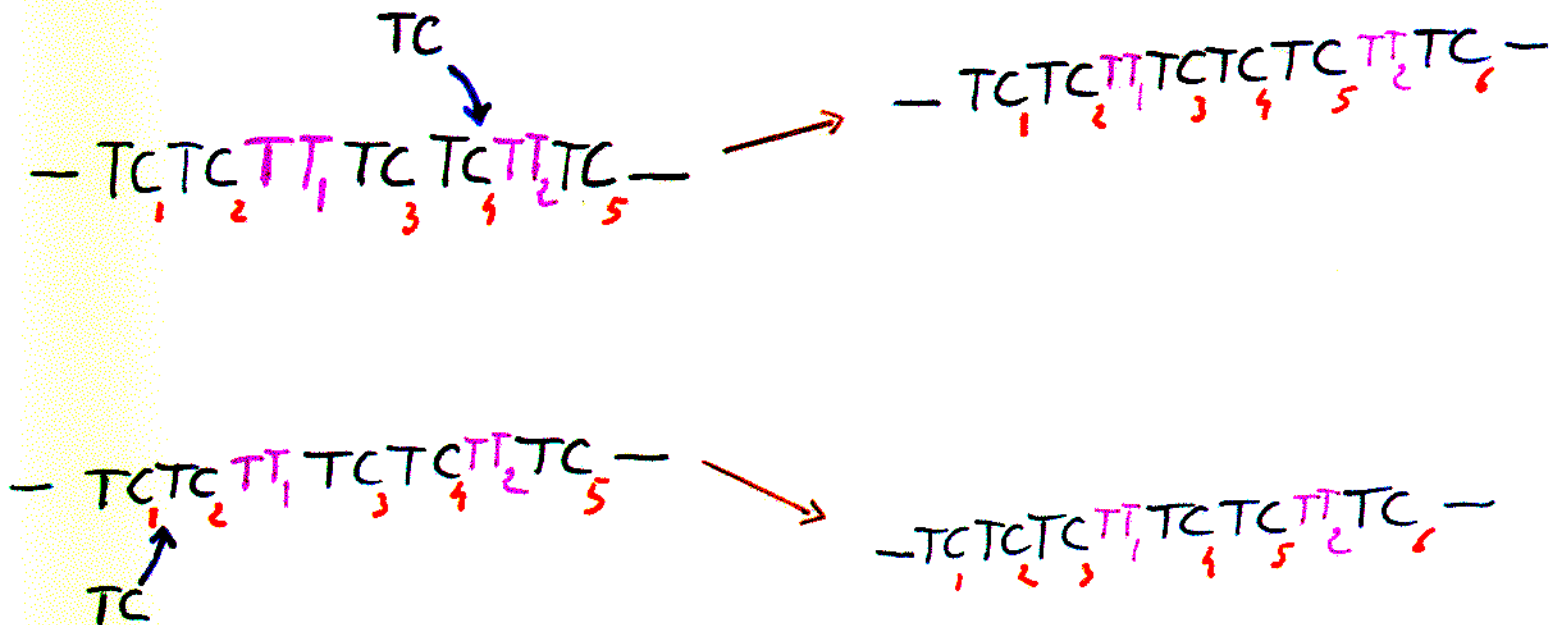
Ex : (TC)10 → (TC)11            (TC)12 → (TC)11

- **Facteurs influençant le taux d'homoplasie de taille**
  - **Modèle mutationnel**
  - **Taux de mutation**
  - **Temps de divergence**
- **Prédiction d'un fort taux d'homoplasie aux locus microsatellites**
  - **Modèle mutationnel = SMM, TPM (KAM), GSM**
  - **Taux de mutation élevé ( $\mu = 5 \times 10^{-4}$  en moyenne)**
  - **Contraintes de taille**

# microsatellite pur / non-interrompu



# microsatellite interrompu



**Table 4:** Core sequences of electromorphs of the locus A113 in different subspecies of *A. mellifera* and in *A. cerana* and *A. dorsata*. The interruptions which beacons the stretches of repeats are in bold characters. The number of copies of each electromorph sequenced in the different species, subspecies and populations is also indicated. A, U and V letters correspond to the honey bee populations of the lineage M from Avignon, Valhousse and Umeå, respectively.

Species	Lineage	Subspecies	Electromorph size in pb	Core sequence (strand orientation 5' - 3')	number of genes (copies) sequenced	
<i>A. mellifera</i>	M	<i>mellifera</i>	202	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>5</sub> GTTTCG(TC) <sub>2</sub>	4A+IU	
			208	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>14</sub> GTTTCG(TC) <sub>2</sub>	1A	
			214	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>2</sub> TT(TC) <sub>9</sub> GTTTCG(TC) <sub>2</sub>	1A+IU	
			220	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>8</sub> TT(TC) <sub>5</sub> GTTTCG(TC) <sub>2</sub>	6A+4V+3U	
			222	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>21</sub> GTTTCG(TC) <sub>2</sub>	1A+IU	
			224	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>22</sub> GTTTCG(TC) <sub>2</sub>	1A+IU	
			226	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>18</sub> GTTTCG(TC) <sub>2</sub>	1A	
			228	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>19</sub> GTTTCG(TC) <sub>2</sub>	1A	
			230	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>25</sub> GTTTCG(TC) <sub>2</sub>	IV	
			234	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>27</sub> GTTTCG(TC) <sub>2</sub>	1A	
	236	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>28</sub> GTTTCG(TC) <sub>2</sub>	1A			
	C	<i>ligustica</i>	214	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>2</sub> TT(TC) <sub>9</sub> GTTTCG(TC) <sub>2</sub>	2	
			220	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>8</sub> TT(TC) <sub>5</sub> GTTTCG(TC) <sub>2</sub>	3	
			224	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>7</sub> TT(TC) <sub>5</sub> GTTTCG(TC) <sub>2</sub>	1	
		<i>carnica</i>	214	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>2</sub> TT(TC) <sub>9</sub> GTTTCG(TC) <sub>2</sub>	2	
			220	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>8</sub> TT(TC) <sub>5</sub> GTTTCG(TC) <sub>2</sub>	3	
			226	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>18</sub> GTTTCG(TC) <sub>2</sub>	1	
			228	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>24</sub> GTTTCG(TC) <sub>2</sub>	1	
			230	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>20</sub> GTTTCG(TC) <sub>2</sub>	1	
			236	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>23</sub> GTTTCG(TC) <sub>2</sub>	1	
		<i>cecropia</i>	214	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>2</sub> TT(TC) <sub>9</sub> GTTTCG(TC) <sub>2</sub>	3	
			220	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>4</sub> TT(TC) <sub>15</sub> GTTTCG(TC) <sub>2</sub>	1	
		A	<i>scutellata</i>	214	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>11</sub> GTTTCG(TC) <sub>2</sub>	2
				218	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>13</sub> GTTTCG(TC) <sub>2</sub>	3
				220	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>14</sub> GTTTCG(TC) <sub>2</sub>	1
	<i>capensis</i>		208	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>8</sub> GTTTCG(TC) <sub>2</sub>	2	
			214	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>11</sub> GTTTCG(TC) <sub>2</sub>	1	
224	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>11</sub> GTTTCG(TC) <sub>2</sub>	1				
224	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>11</sub> GTTTCG(TC) <sub>2</sub>	1				
224	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>5</sub> TT(TC) <sub>11</sub> GTTTCG(TC) <sub>2</sub>	1				
<i>A. cerana</i>			184	(TC) <sub>2</sub> C(TC) <sub>2</sub> TT(TC) <sub>3</sub> GTTTCG(TC) <sub>2</sub>	2	
<i>A. dorsata</i>			184	(TC) <sub>2</sub> C(TC) <sub>5</sub> O(TC) <sub>2</sub> G(TC) <sub>3</sub>	1	
			186	(TC) <sub>2</sub> C(TC) <sub>4</sub> TTTCG(TC) <sub>2</sub> O(TC) <sub>3</sub>	1	
<b>Total</b>			<b>14 electromorphs</b>		<b>63 copies</b>	

## 'Gestion' de l'homoplasie de taille

- Gestion inutile pour certaines études = peu de mutations entre les taxons / individus analysés (ex : assignation de parentés, populations très peu divergentes)  
*len nbre de génération*
- Gestion totale si le locus mute strictement en SMM cf statistiques disponibles
- Gestion très imparfaite si :
  - TPM ( $p$ ,  $Vg$ ) ou KAM ( $k$ )
  - Contraintes de taille ( $k$ )

# RESUME: MODELES MUTATIONNELS THEORIQUES

1/ Essentiels à partir d'une certaine échelle évolutive / tx de mutation

2/ Modèle complexe :

- Multistep ( $P > 0$ )
- Biais ( $\alpha > 0.5$ )
- Contraintes de tailles (?)
- $\mu$  augmente avec le nombre de répétitions (?)
- $\mu$  varie avec la structure moléculaire (?)

3/ Paramètres du Modèle différents selon les locus

→ approche locus par locus ou homogénéisation des locus

(SMM)

→ données suffisantes pour un paramétrage précis ?

4/ Homoplasie: gestion via les modèles mutationnels

A UN NIVEAU MOLECULAIRE  
(+  $\mu_{sat}$  interrompus  
composés  
→ Sequences / SSCP)