Les microsatellites codants,

« *Y'en a pas un sur cent et pourtant ils existent !* »

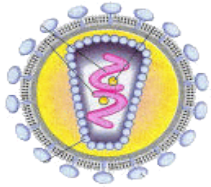
E. Loire, M. Lapierre

F. Praz, D. Higuet, P. Netter
et G. Achaz

Systématique, Evolution, Adaptation (UMR 7138) - UPMC
Atelier de Bio-Informatique - UPMC
Stochastic Models for the Inference of Life Evolution – Collège de France

Whole genome mutation rates

(from Drake *et al.*, 1998)

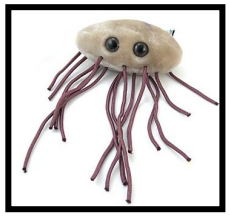


HIV-1 : 3.10^{-5}

***H. sapiens* : 5.10^{-11}**

Substitutions /base /replication

***E. coli* : 5.10^{-10}**



***C. elegans* : 2.10^{-10}**

Is the mutation rate itself subject to selection ?

Across eukaryote genomes

Chromosomal differences

- Autosomes *vs.* sexual chromosome

Local differences

- Hotspot of insertion for transposable elements
- Recombination rate
- GC-biased gene conversion
- …

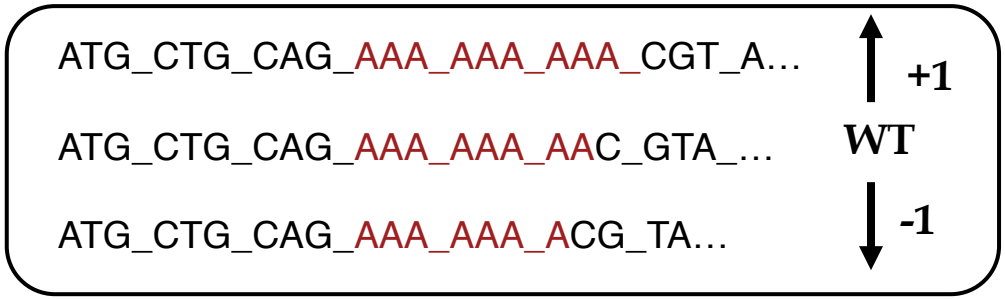
Site differences

- CpG methylation
- Microsatellites (SSRs)
- …

SSRs in coding sequences

SSRs: tandem repepepepepeats of small motifs (few bp)

"High" rate of slippage during replication



Rate of slippage increases exponentially with the number of units

"Long enough" non-3-SSRs confer hypermutability

Mutability of a gene

Let's define for a given gene

Mutability as its rate of STOP mutations

From substitution

~ 5% of all mutations create a STOP

~ 10^{-9} non-sense subst. /kb /replication

From frameshift due to indels in SSRs

~ [10^{-3} , 10^{-6}] /replication

SSRs are likelily the MAJOR source of gene mutability

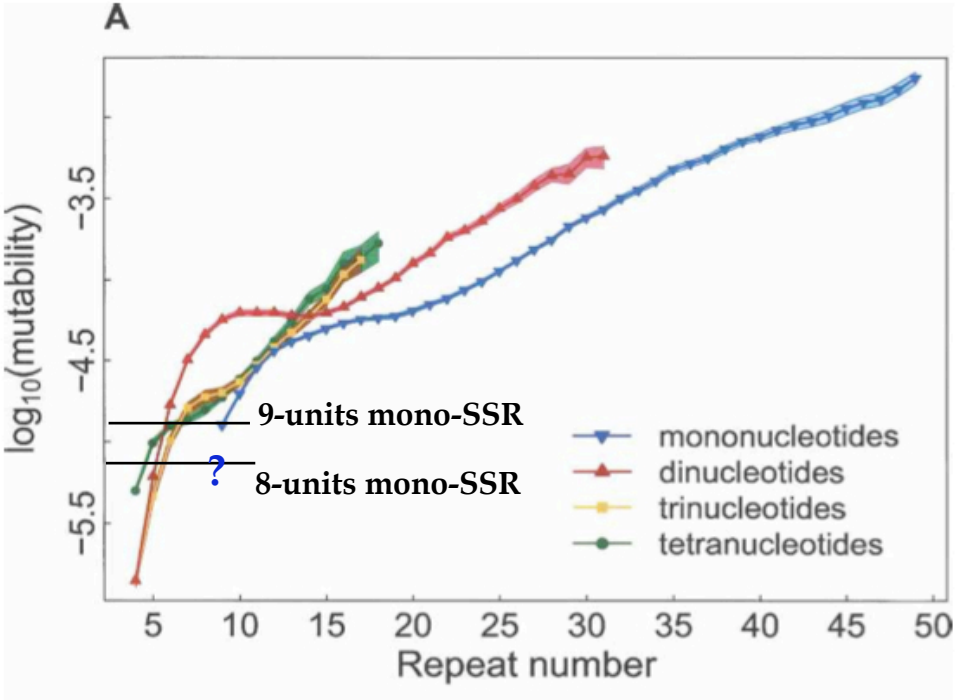
A threshold for SSR instability

For mono-SSR

Experimental observation (tumors)
≥ 8-units (Rose *et al.*, 1998)

Bioinformatics inference
≥ 9-units (Lai *et al.*, 2003)

Human-Chimpanzee SSR mutability



(from Yogeshwar *et al.*, 2007)

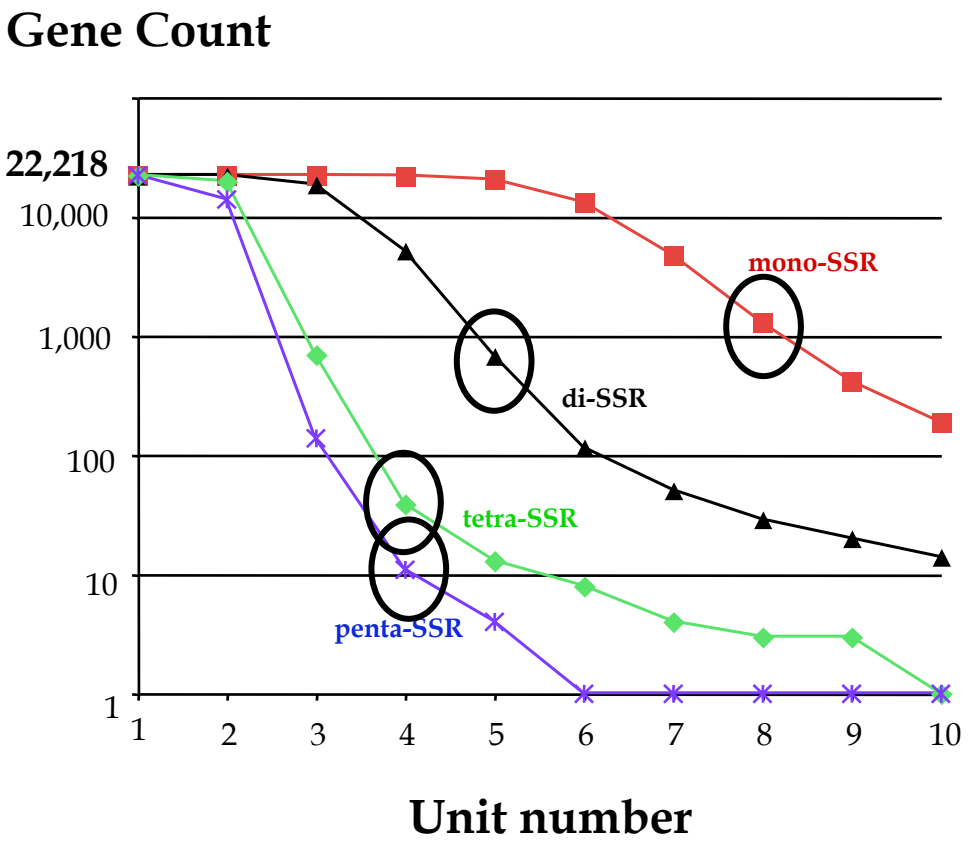
A threshold for mono-, di-, tetra-SSR : 8-, 5-, 4-units

Inferences from "The" Human Genome

Results from Loire et al., MBE, 2009

The Human Genome: Consortium, Nature, 2003; Venter et al., Science, 2003

How many genes with SSRs?



mono-SSR
1,291 genes (5.8%)

di-SSR
678 genes (3.1%)

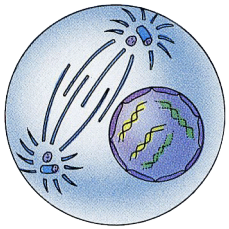
tetra-SSR
39 genes (0.2%)

penta-SSR
11 genes (0.05%)

Total
1,935 genes (8.7%)

Mostly mono- and di-SSRs within genes

GO terms over-representation



Biological Process

cell-cyle* and *DNA maintenance

Molecular Functions

ATPase, GTPase* and *Helicase

Cellular Component

nucleus* and *intracell. non-mbr. bound organelle

A cohesive restricted set of GO-terms

(see Moxon and Wills 1999; Chang et al. 2001; Kashi and King 2006).

Impact of gene structure

The probability of a long mono-SSR is altered by

sequence length
nucleotide composition

Hypothesis

Genes length
and/or composition
explain the results ?

Test

Do we **expect** more SSRs
in the overrepresented
GO. terms ?

Mono-SSRs
a simple substitution model

In a random sequence with independent mono-nucleotides

- of length L
- of composition $\{P_A, P_C, P_G, P_T\}$

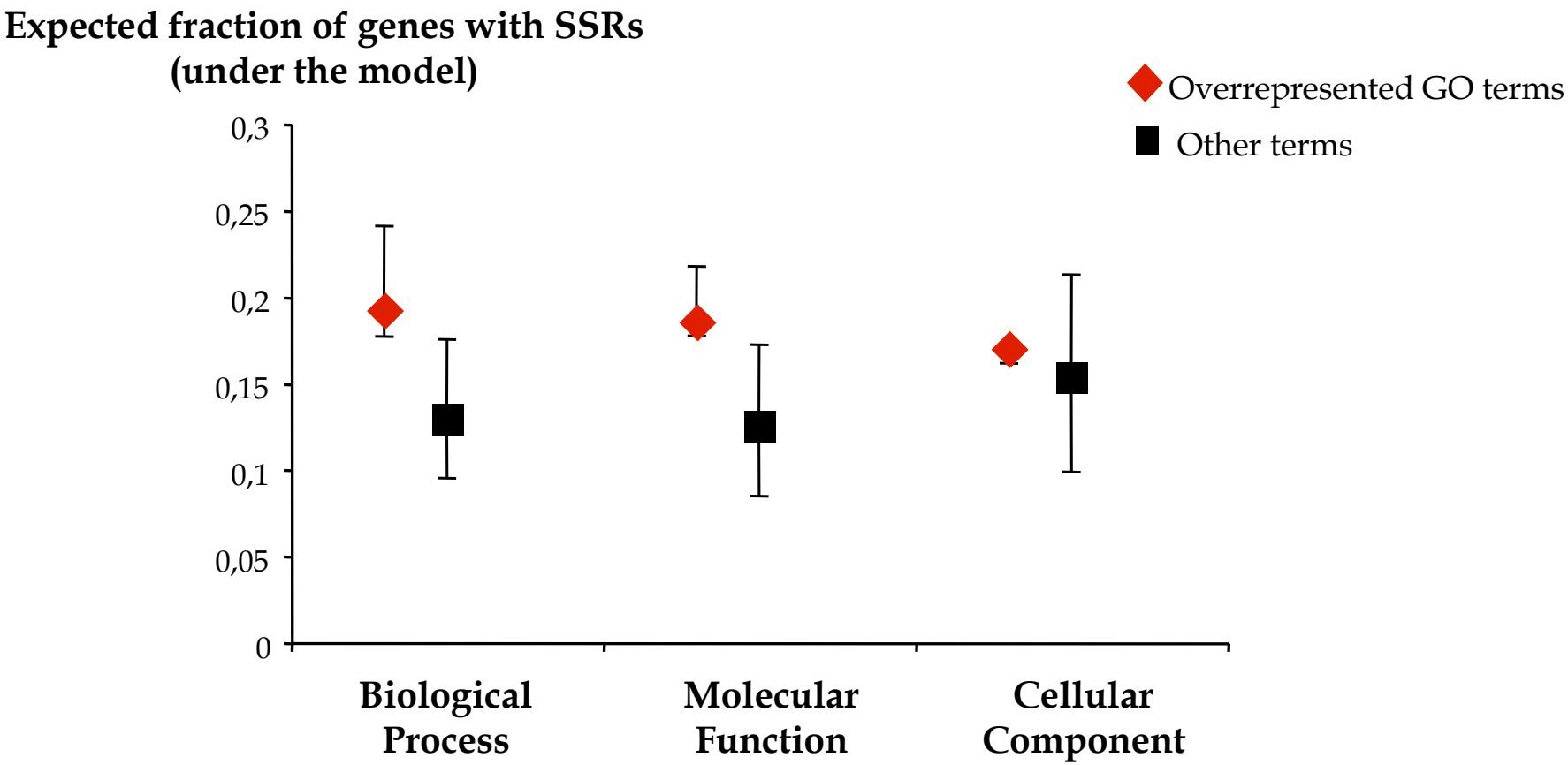
The mean number of runs of nucleotide X of at least size m is:

$$E[\, m^+ \mid L, Px \,] = (L-m+1) \,.\, (1-P_X) \,.\, P_X{}^m$$

The probability of having at least 1 run of this type is:

$$P(\, m^+ \mid L, Px) = 1 - \exp(\, -E[\, m+ \,] \,)$$

Expectations for the functions



We expect more long mono-SSR in the enriched functions

A neutrality test for coding SSRs

The model assumes that all substitutions can occur freely

a neutral model

$m_{1/2}$

Theoretical length $P(m+ \mid L, P_x) = 0.5$

m_{obs}

Length of the longest mono-SSR

Do **50%** of genes have
$m_{obs} > m_{1/2}$?

Gene-by-gene data *vs* theory

	Exons	
mono-SSR	$m_{obs} < m_{1/2}$	$m_{obs} > m_{1/2}$
A	20271	1947
T	20279	1939
G	21660	558
C	21342	876
# expected	11109	11109

(Mono-)SSRs are targeted by purifying selection

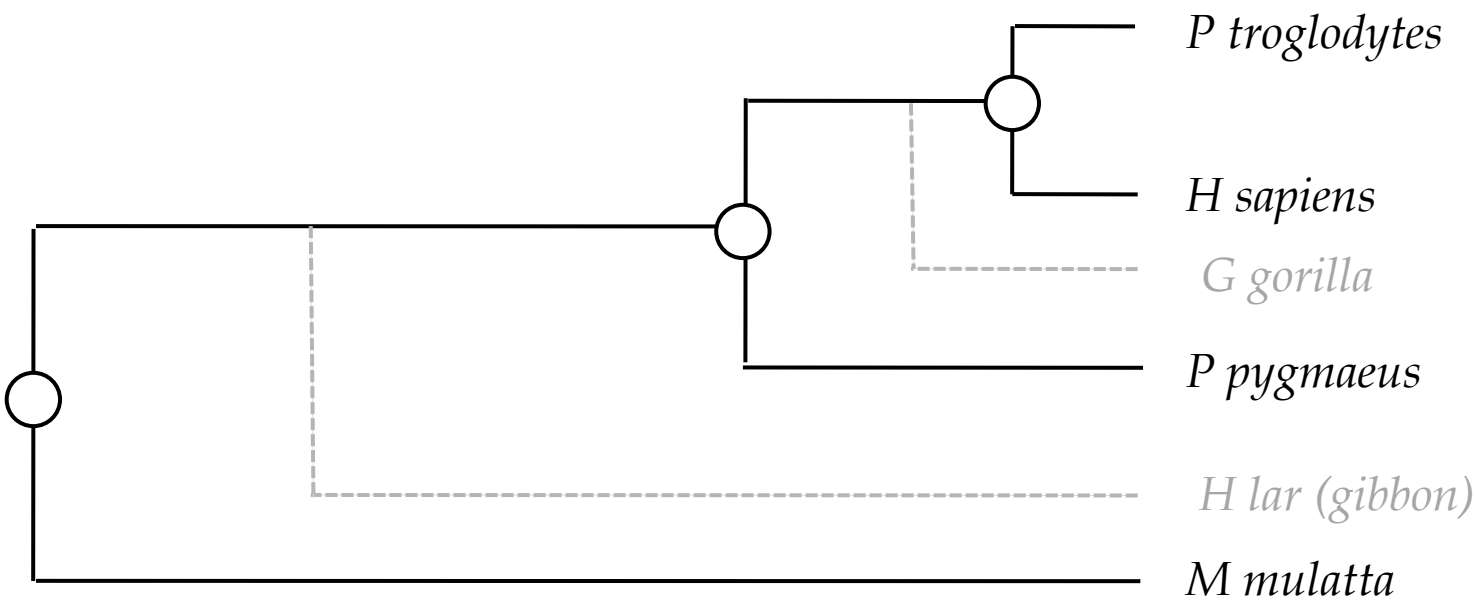
(see also de Wachter 1981; Metzgar et al. 2000; Ackermann and Chao 2006)

Inferences from 4 Primate Genomes

Results from Loire et al., GBE, 2013

Chimpanzee: Consortium, Nature, 2005; Macaca: Zhan et al., Science, 2007
Orang-Utan: Consortium, Nature, 2011

5,015 orthologs from Apes



Multiple alignment (progessive),
Filtering (conserved blocks),
Phylogenetic reconstruction (ML) and
Ancestral states reconstruction (max posterior probabilities)

Definition of an SSR locus

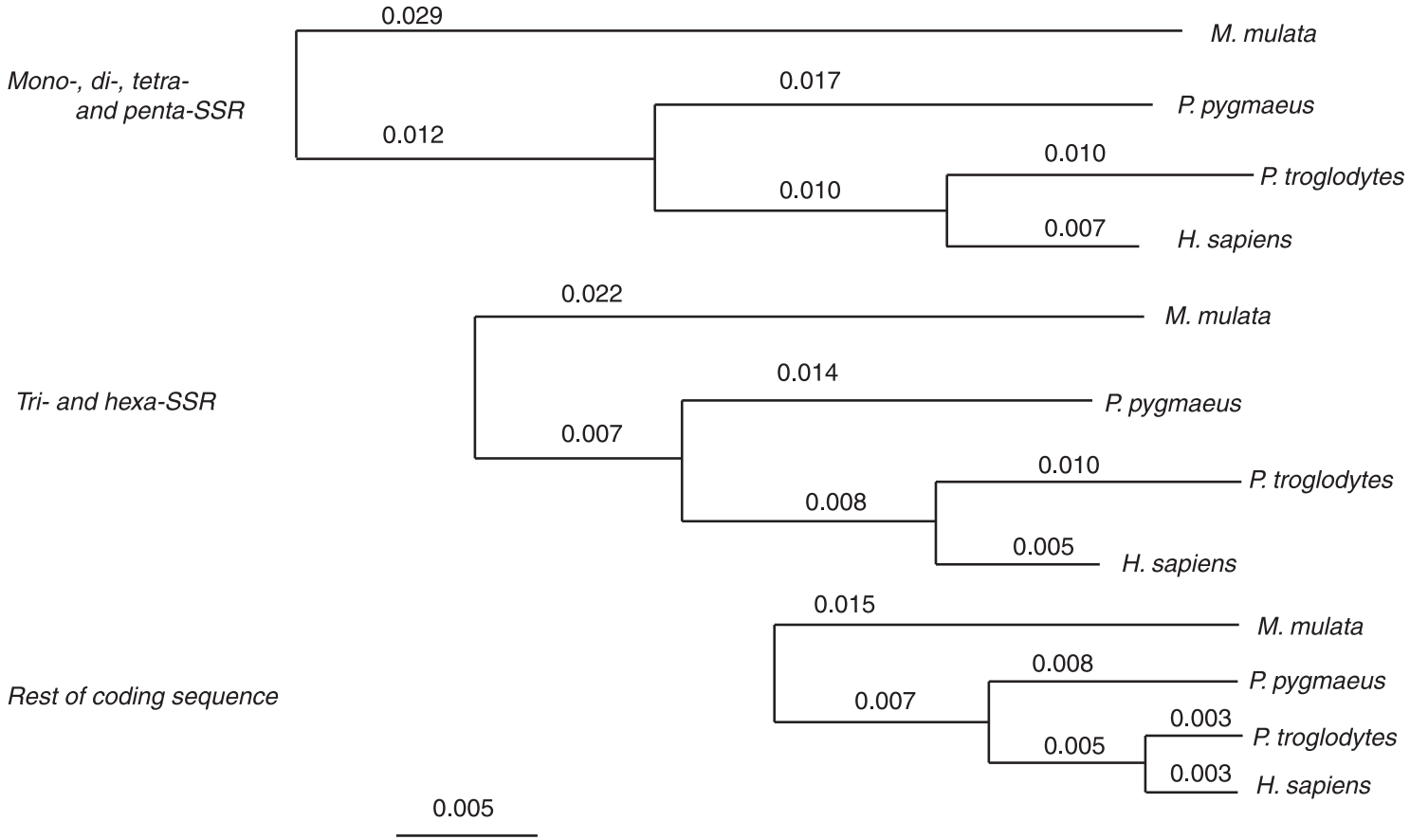
SSR locus = at least one SSR in one species

What kind of mutations ?

Sequence type	# Sites	Indels (% of sites)	Substitutions (% of non-indels)
mono-, di-,tri, penta-SSRs	7,312	130 (1.8%)	557 (7.7%)
tri-, hexa-SSRs	8,499	1,680 (19.8%)	373 (5.5%)
Rest of coding	8,185,286	31,720 (0.4%)	316,408 (3.9%)

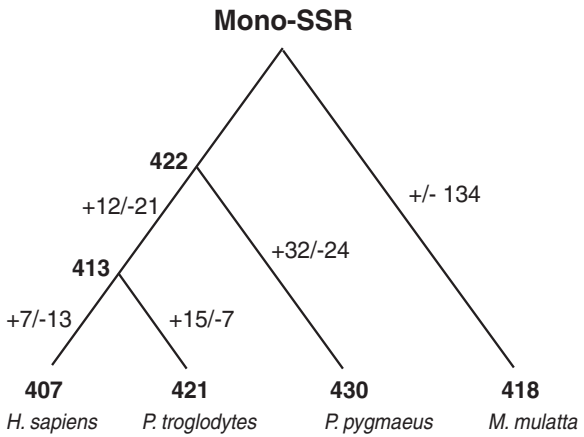
Only tri- and hexa-SSRs expand and contract

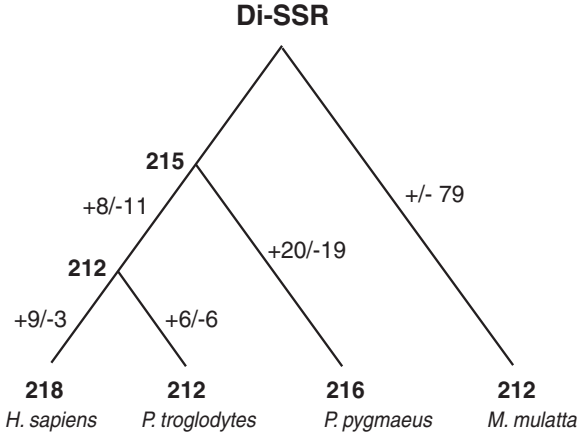
Evolutionary distances

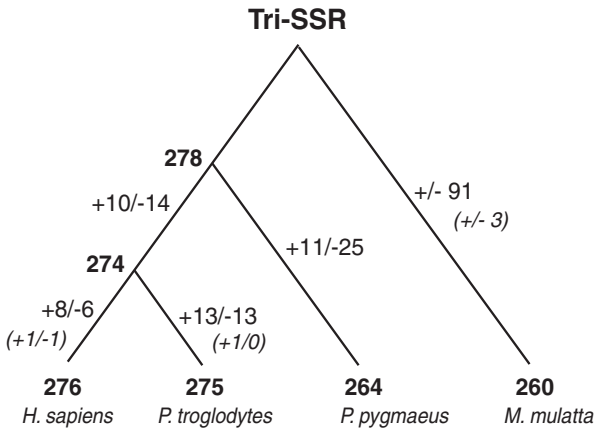


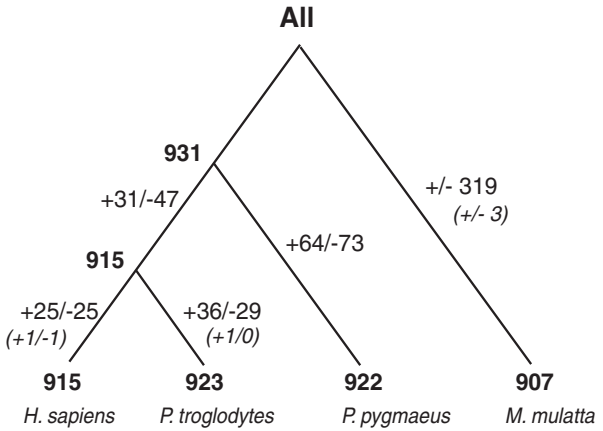
SSRs evolve twice faster than rest of coding sequence

Gains and losses of coding SSRs



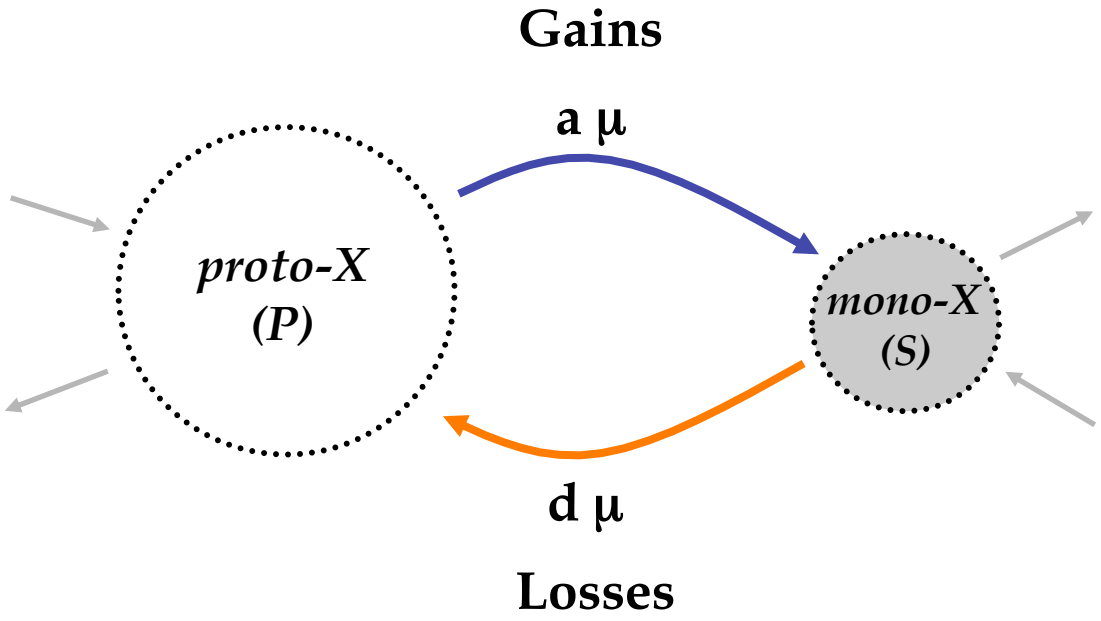






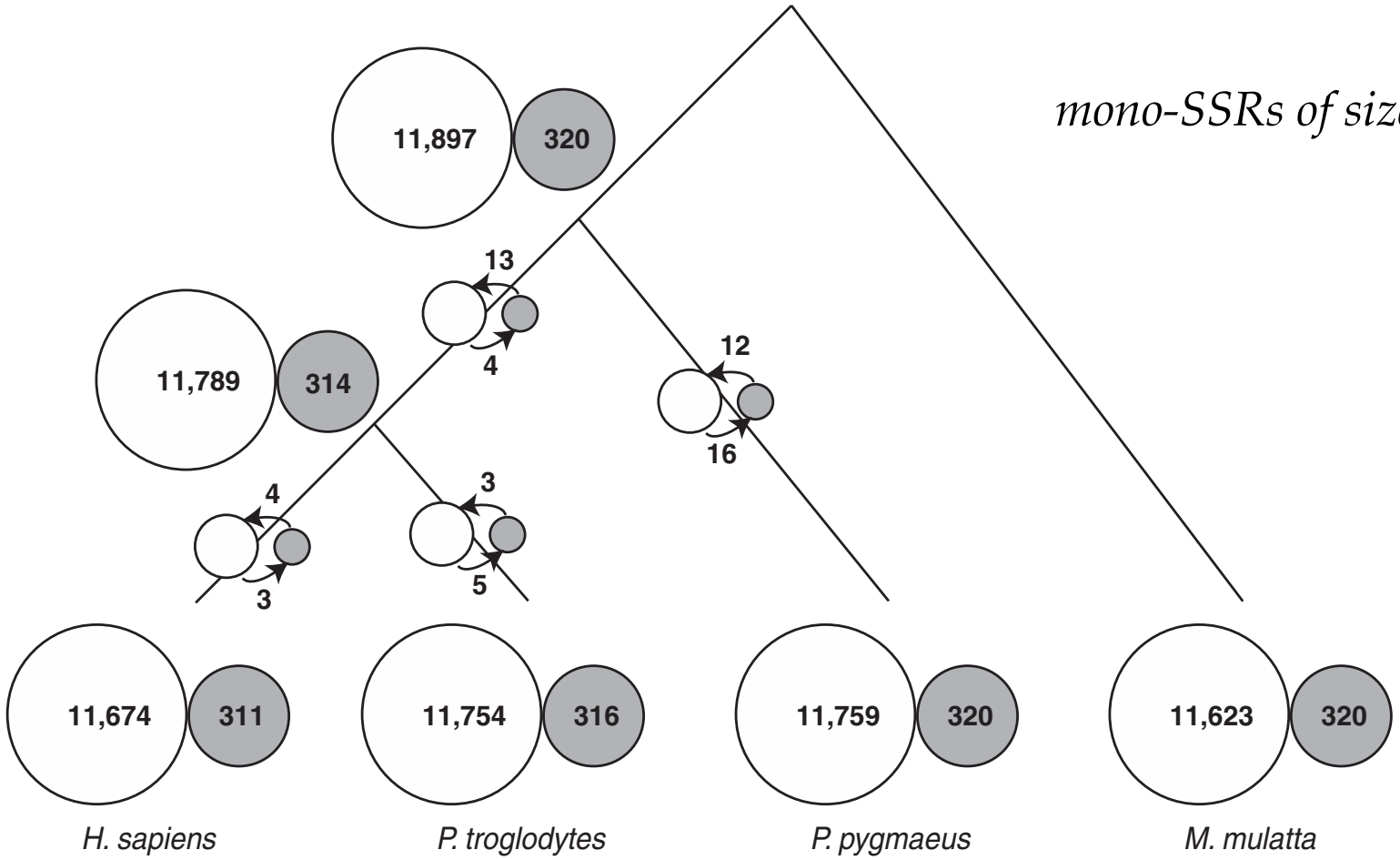
A dynamic equilibrium: Gains ~ Losses

Mono-SSRs
(toward a simple 2-alleles model)

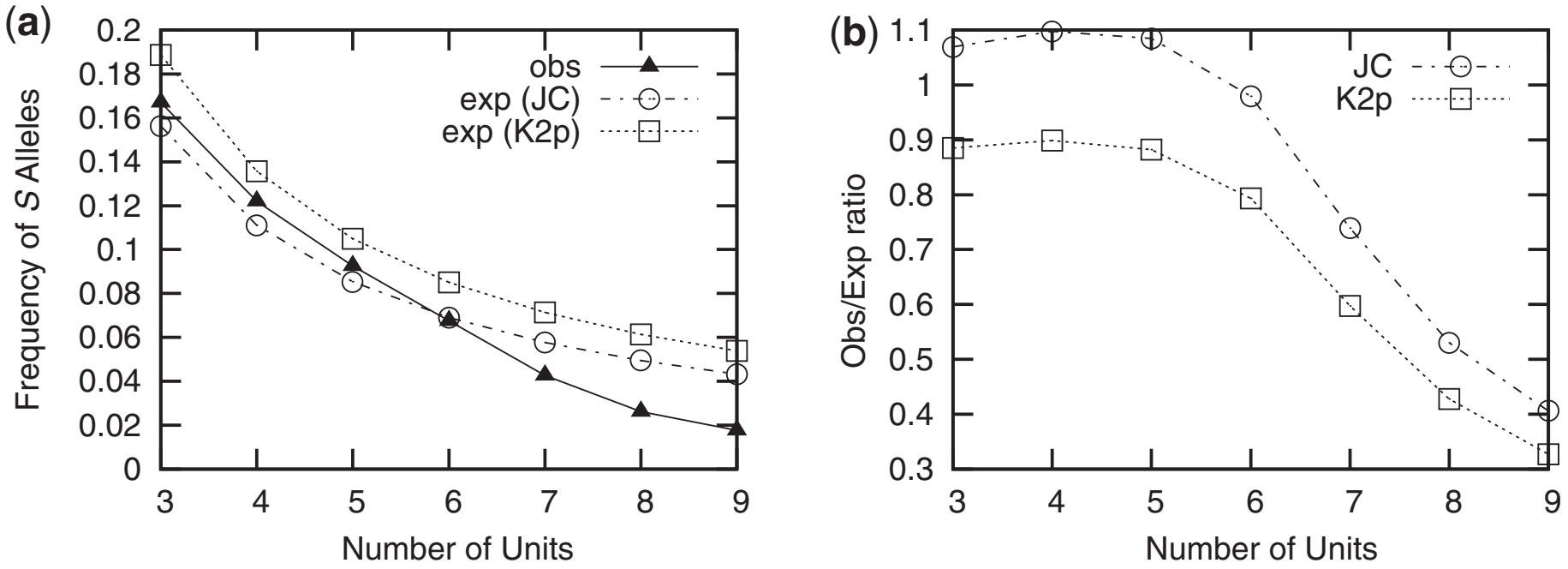


($\mu = 10^{-8}$; a = 1/3 ; d = X)

S/P alleles at equilibrium

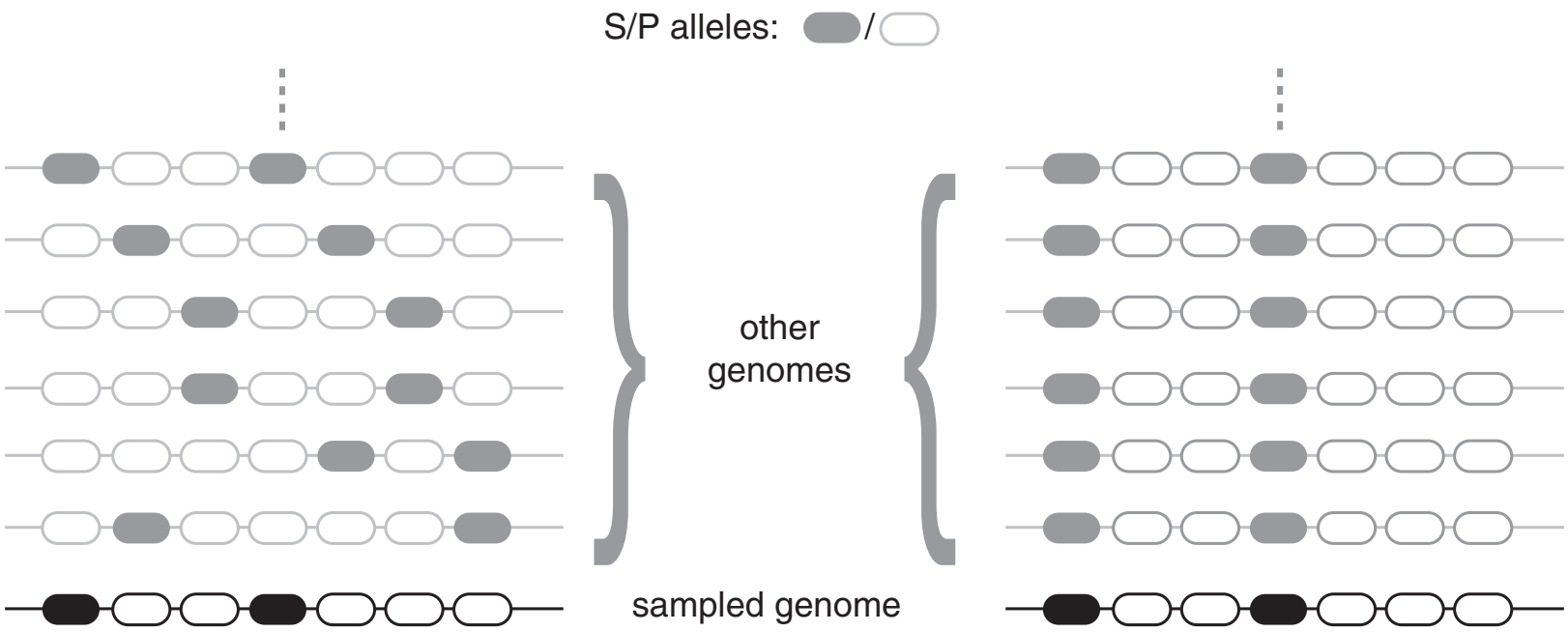


The mutation-only model



It only fits for small (stable) SSRs

Single genome "Population genetics"



Mutation-selection equilibrium

(textbook population genetics)

Mutation-selection-drift equilibrium

(following Bulmer 1991)

Estimation of selective coefficient

Effective selection coefficient ($2N_e s$)

Model 1, infinite size $N_e = 10\,000$

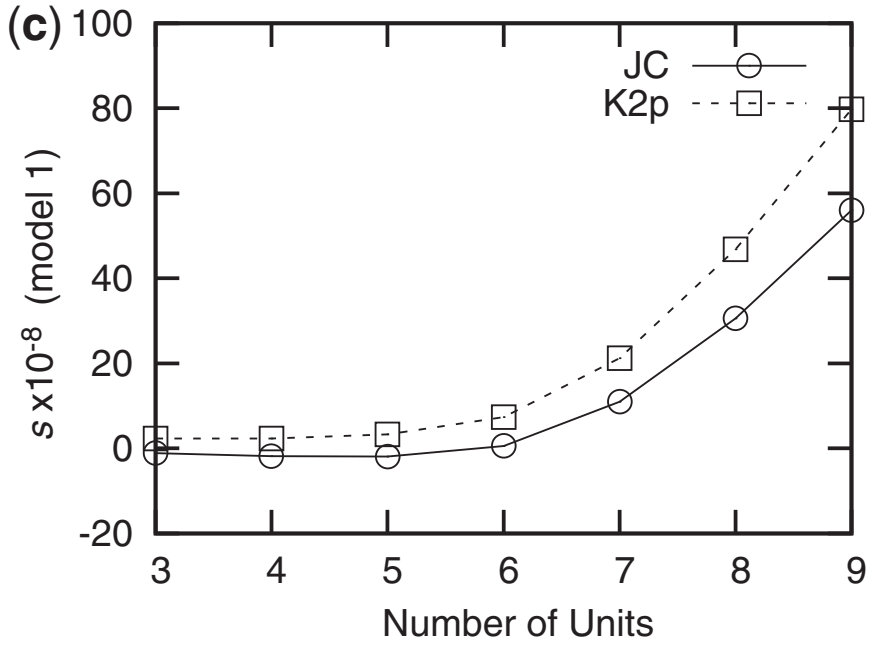
Model 2, finite size direct estimation

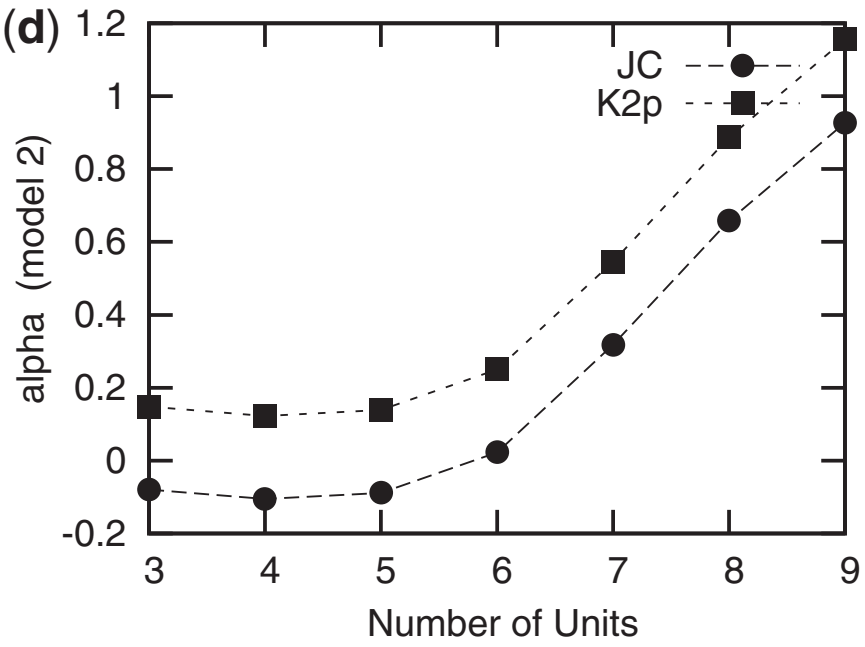
Results for mono-SSRs of 8 units

h	Model 1	Model 2
1	0,0008	0,2
0,1	0,006	2
0	0,03	?

Estimation of selective coefficient

(fix h=0.5)





The longer, the nastier

Comparative genomics

Rate of evolution

Twice faster than the rest of coding sequence

Gains ~ Losses

Coding SSRs are at equilibrium

Selective coefficient

Infinite size model : very small $N_e s << 1$

Finite size model : small $N_e s \sim 1$

Inferences from 1,000 Human Genomes

Results from M. Lapierre (work in progress)

1,092 human genomes: Consortium, Nature, 2010; Consortium, Nature, 2012

1,000 human genomes

1,092 genomes

> ~2,200 haploid genomes

No ascertainment bias

> No need for ad-hoc corrections

Orienting mutation using the Chimpanzee genome

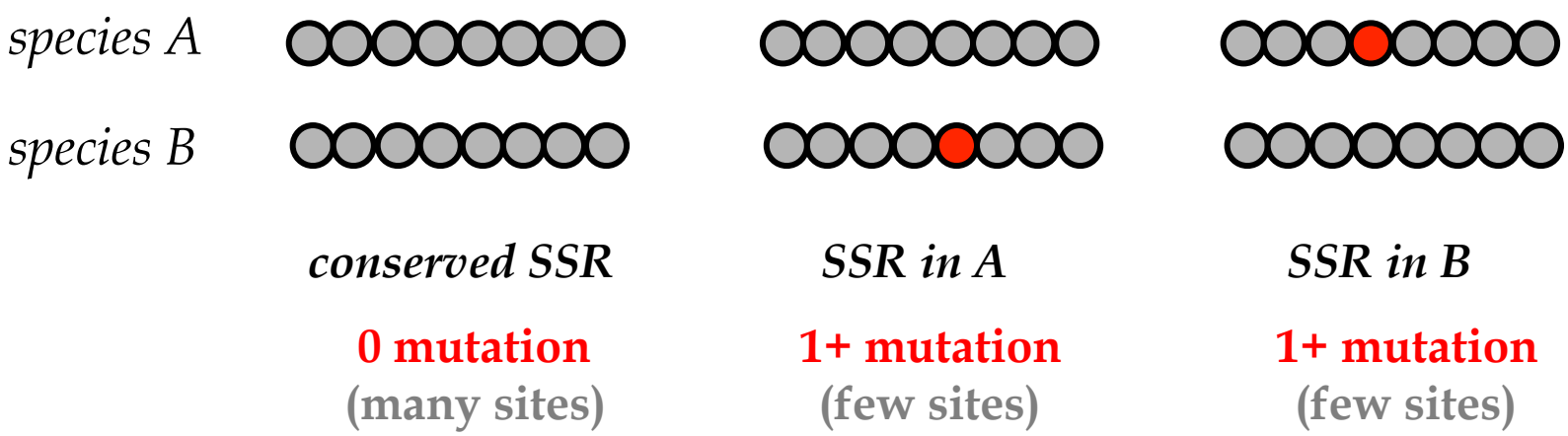
> Ancestral and Derived states

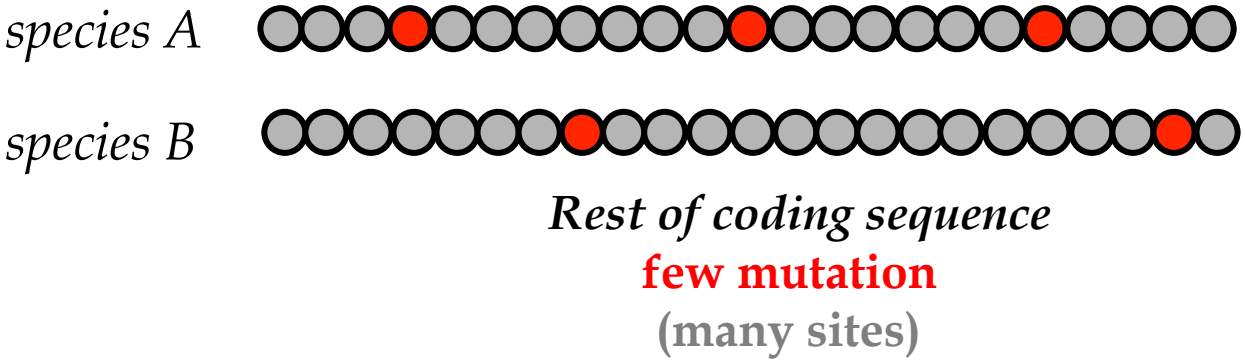
SNPs in coding SSRs

Sequence Type	# SNPs	Density (/bp)
Rest of Coding sequence	179,893	0.5%
mono-, di-, tri-, tetra-, penta-SSRs	324	1.5%
Tri-, hexa-SSRs	109	0.5%

Why some SSRs have more mutations ?

Counting mutations back and forth

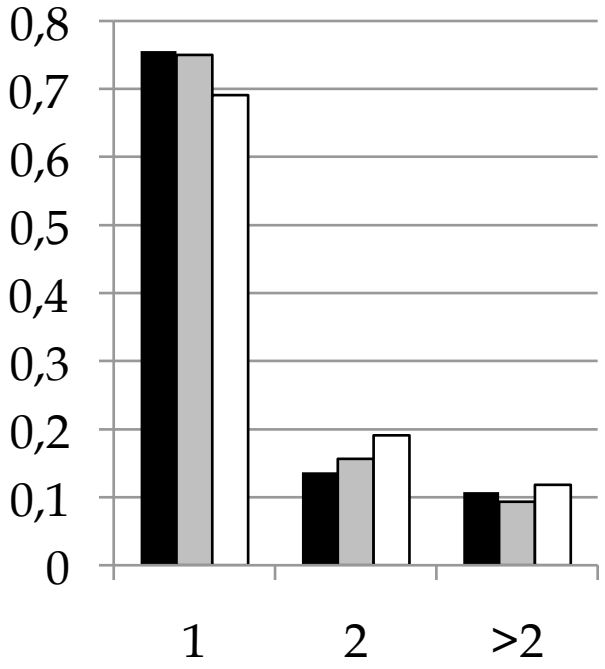


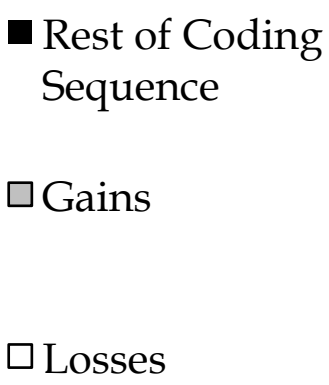


Distances in SSRs are overestimated by a factor 2 !

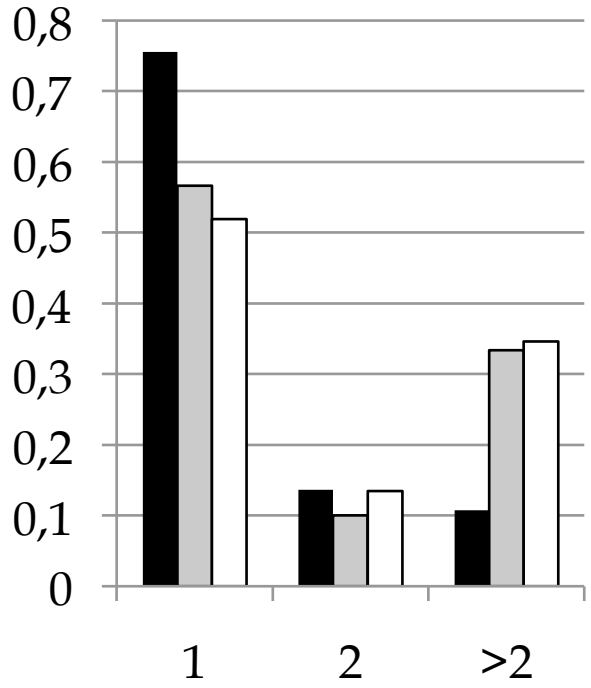
Frequencies of SNPs

Mono-, di-, tetra-, penta-SSRs





Tri-, hexa-SSRs



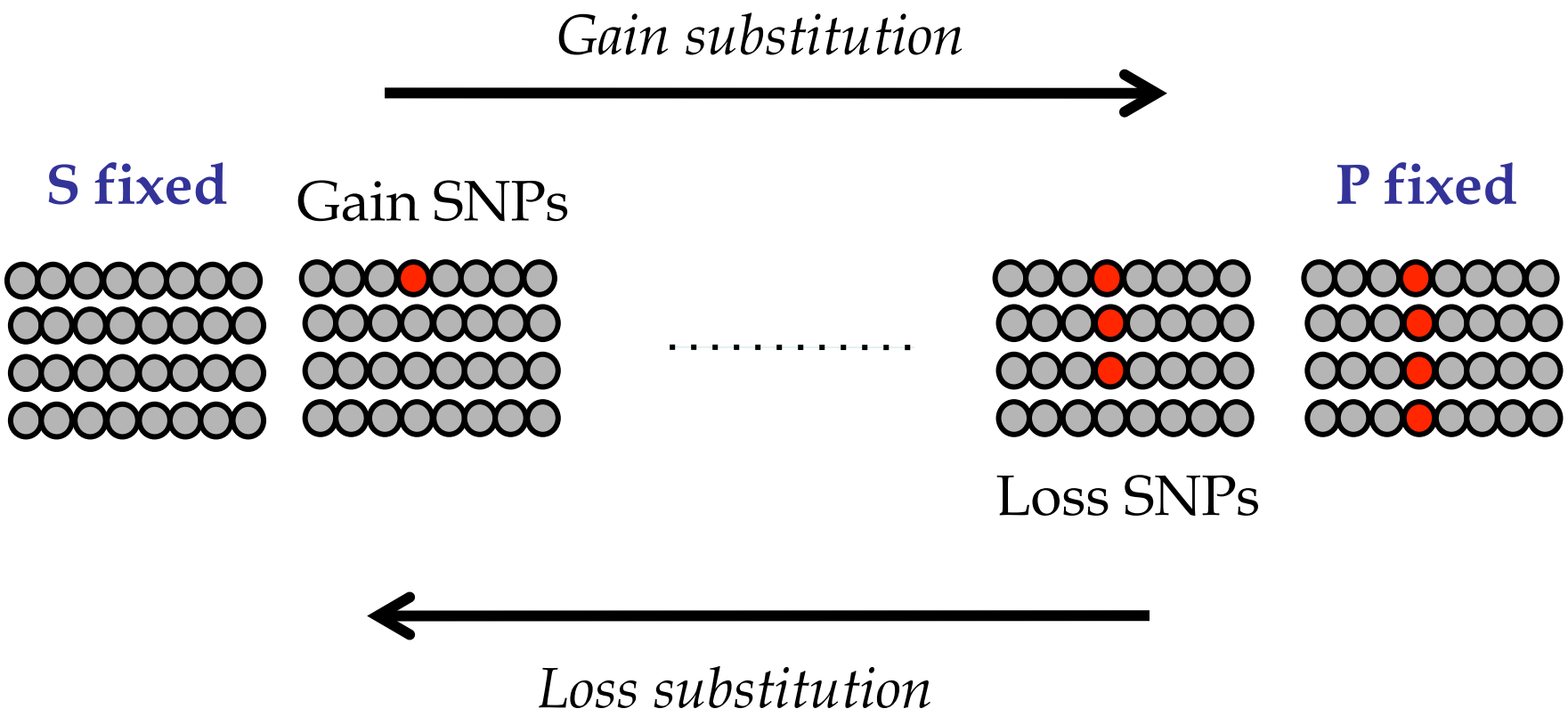
The selection-mutation balance model is right out !

Gains *vs* Losses

SSRs	Gains	Losses	Ratio
mono-, di-, tri-, tetra-, penta-SSRs	128	110	1.16
Tri-, hexa-SSRs	30	104	0.29

Stable *vs* unstable SSRs

SNPs in dynamic SSRs



Gain Substitutions = Loss Substitutions

Estimating the selective coefficient

Gain/Loss SNPs depends on the fixation probability

$$\text{Ratio} = P_{fix}(S)/P_{fix}(P)$$

Estimate $2N_e s$ for mono-SSRs of 8 units

h	1 genome	2,200 genomes
1	0,2	0,05
0,1	2	0.5
0	?	?

Concluding thought

« *Y'en a pas un sur cent et pourtant ils existent !* »