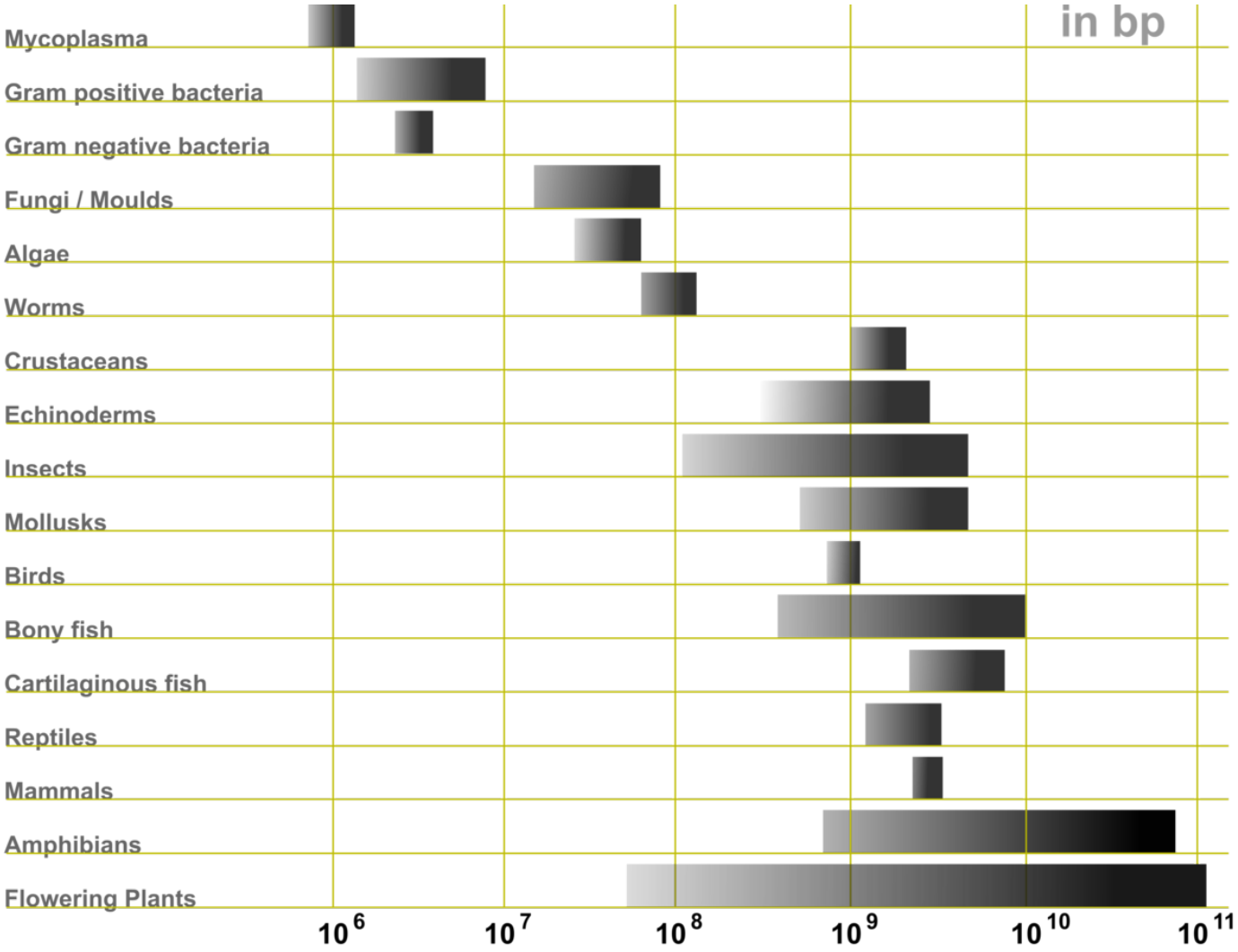


Genome size evolution: challenging intuition with modelling

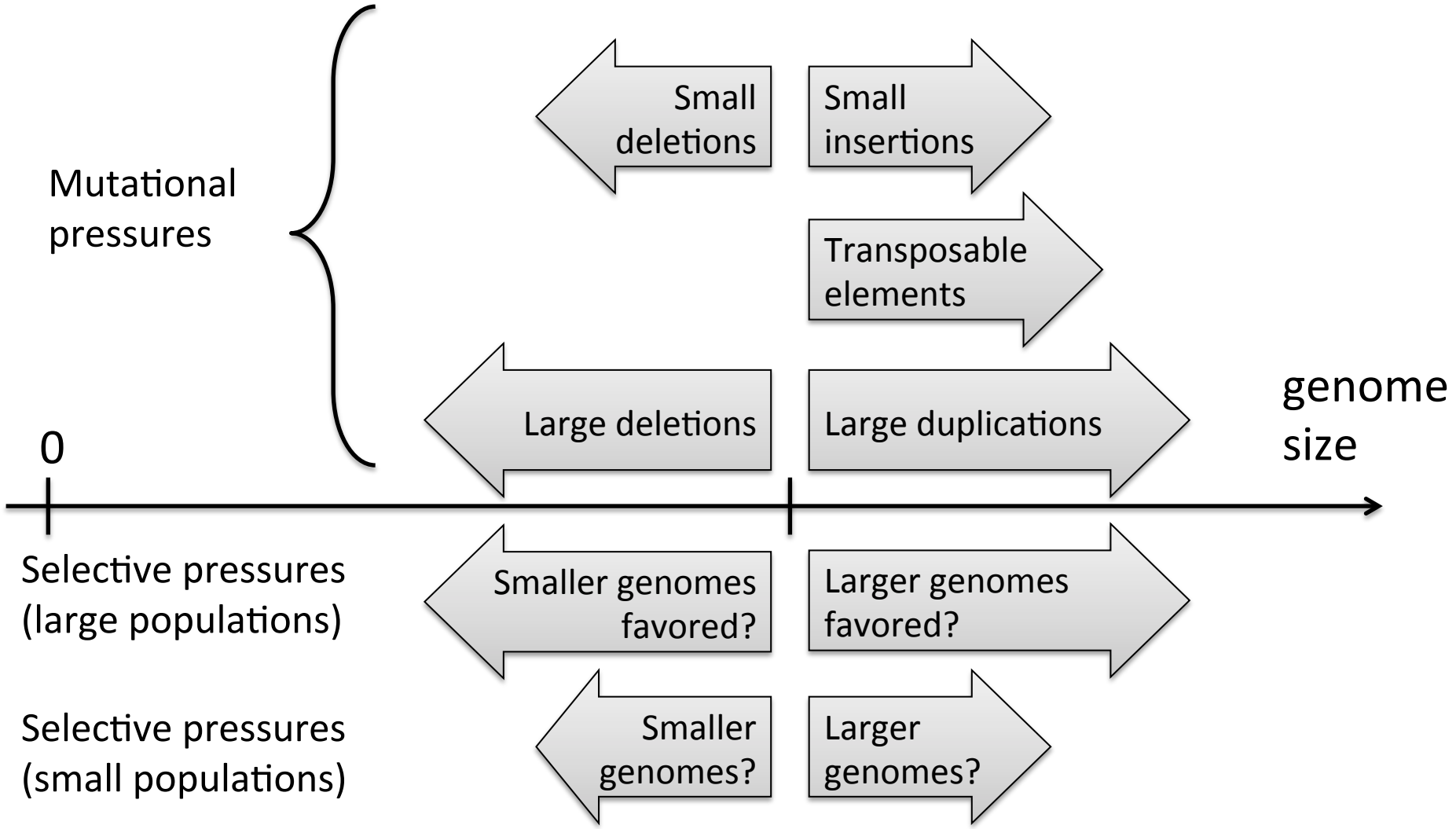
Stephan Fischer, Samuel Bernard,
Guillaume Beslon, Carole Knibbe

Equipe Inria Beagle, Lyon

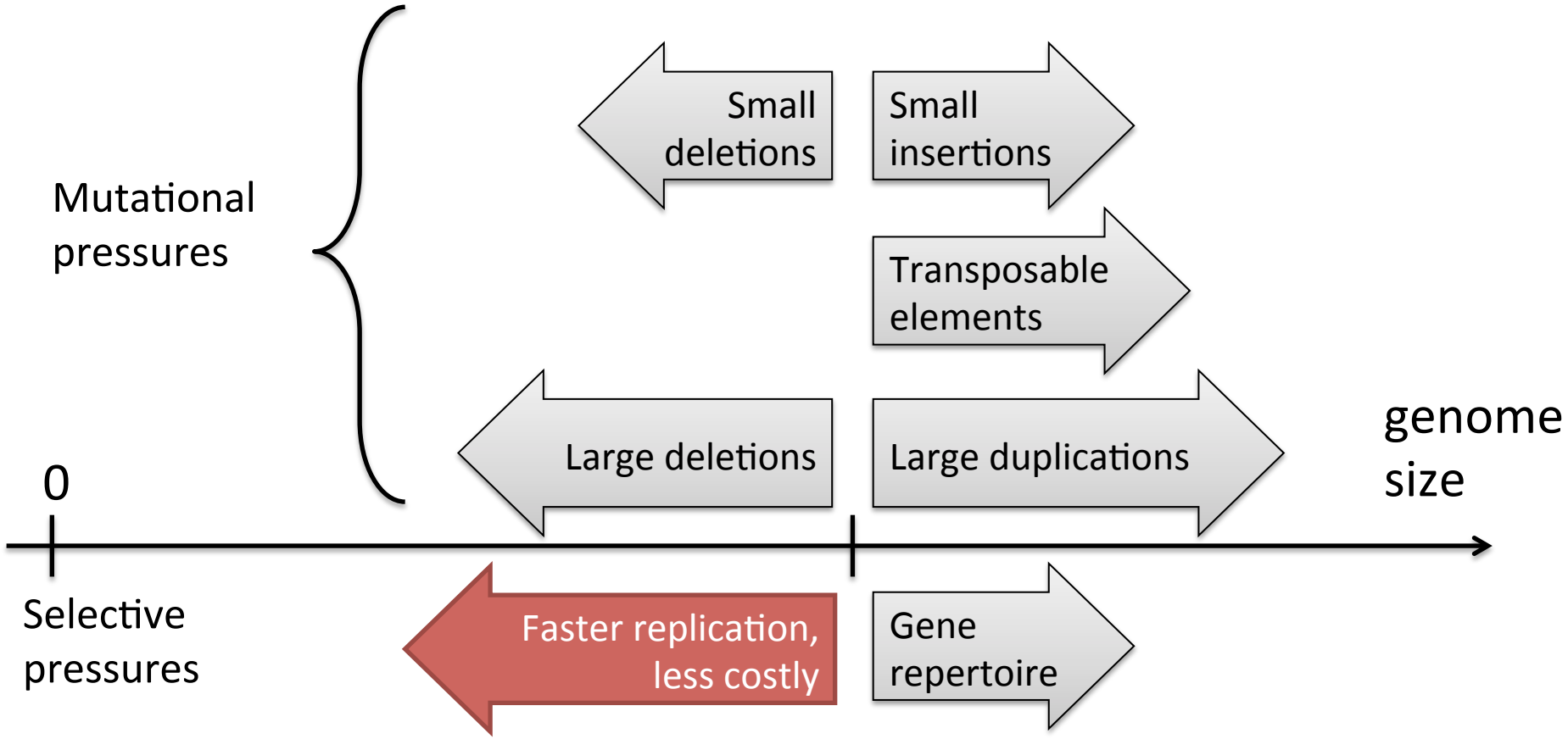
What determines genome size?



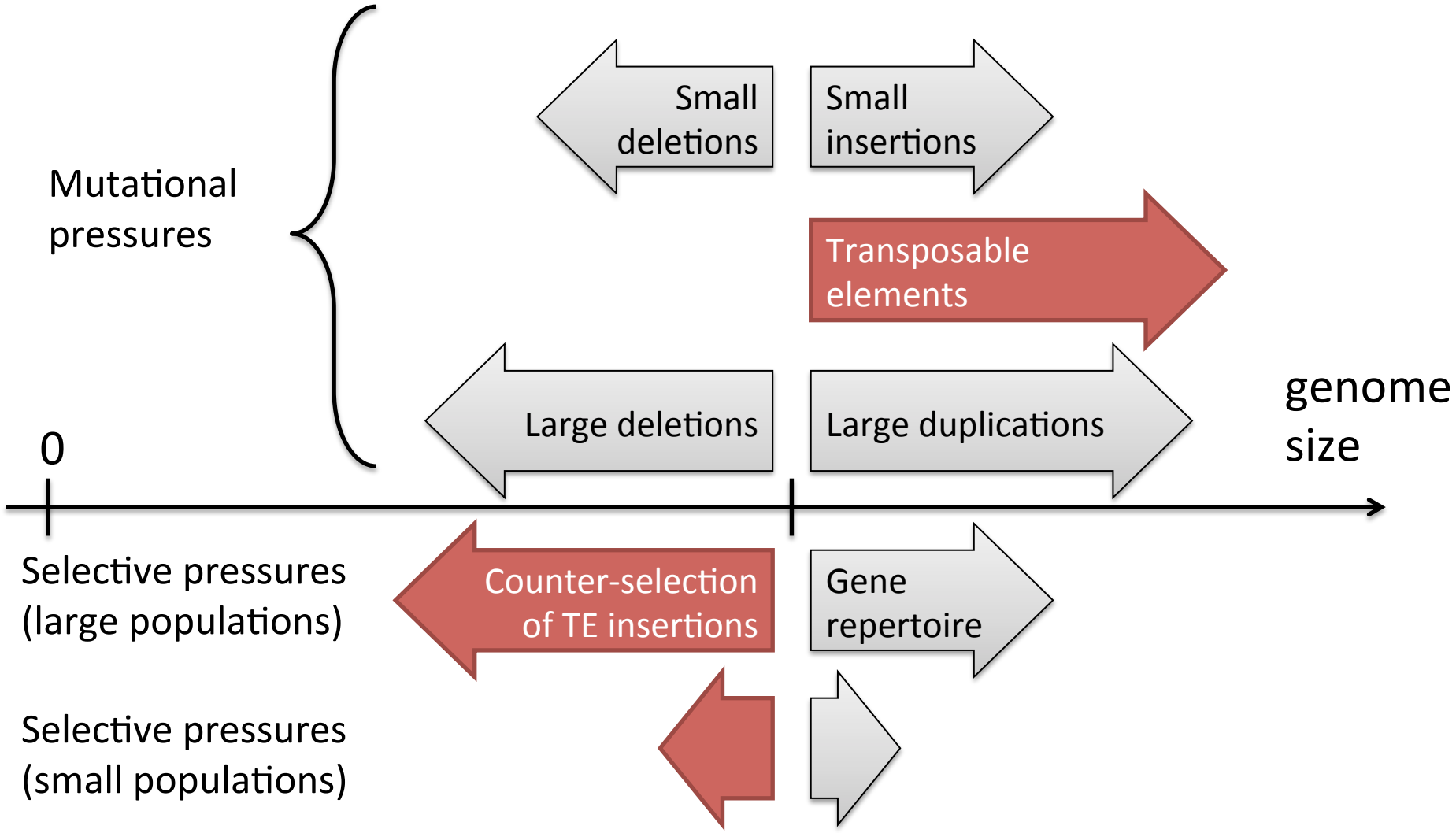
The usual « thought experiment »



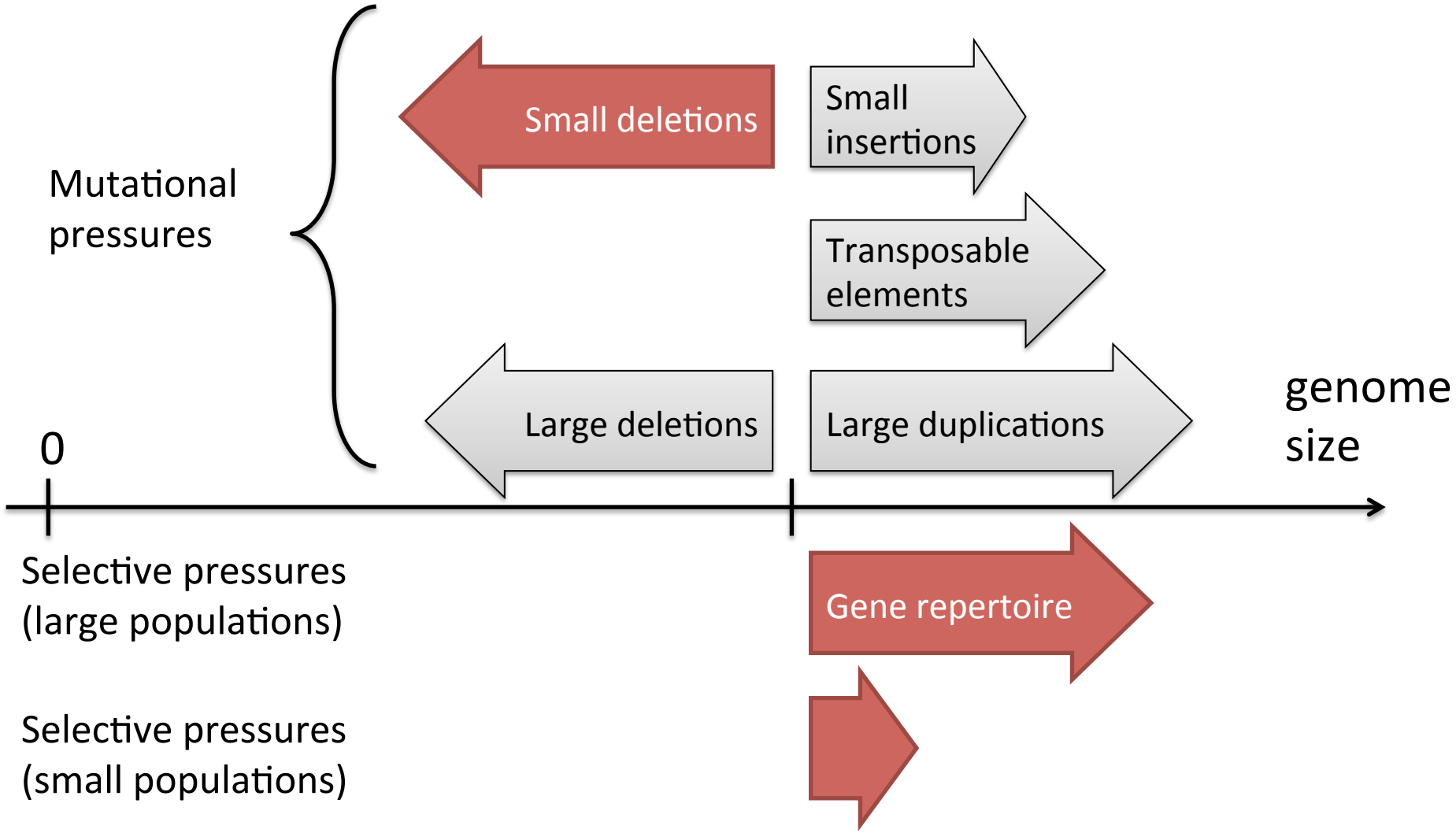
Theory 1: Shorter is better



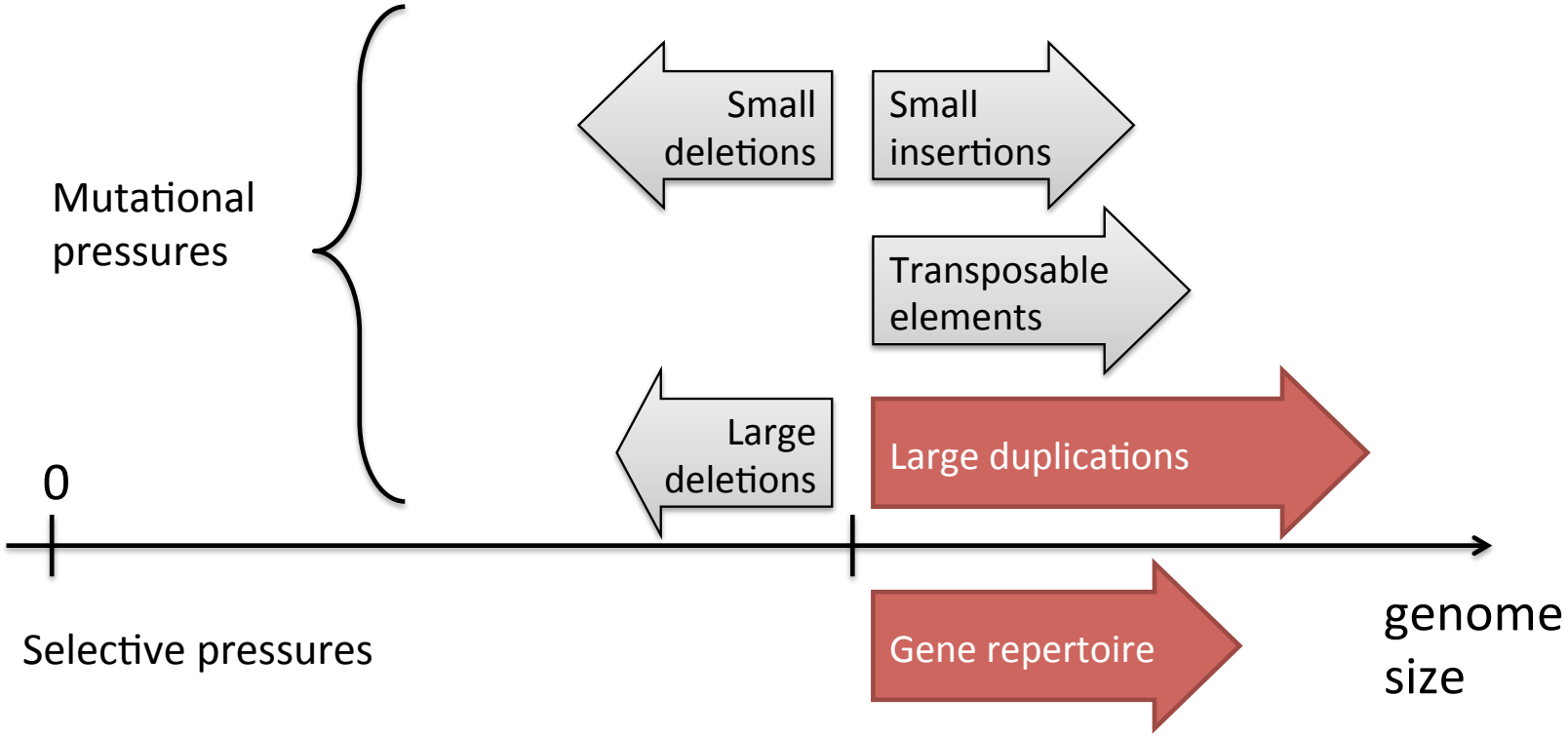
Theory 2: Small populations cannot get rid of transposable elements (Lynch & Conery 2003)



Theory 3: Biases in the small indels mechanisms drive genome size evolution (Petrov 2002, Kuo et al. 2009)



The question we ask:
is the intuitive reasoning correct?



How will genome size evolve if :

- duplications are twice as frequent as deletions,
- transposable elements proliferate,
- and selection systematically favors the highest gene numbers?

Let’s build a minimal model for genome size evolution

A minimal model for genome size evolution

Space of all possible genome sizes: $\mathbb{N}^*$

Infinite population, density vector at time t: ν_t

Transition matrix due to
small and large mutations: $\mathbf{M}_G = ((\mathbf{M}_G)_{ij})_{i,j\in\mathbb{N}\setminus\{0\}}$

Evolution in discrete time,
without selection (Markov chain): $\nu_{t+1} = \nu_t \mathbf{M}_G$

Possible mutations for a genome of size s_0

- Small insertions: + 1 to + l_{ins} bases
- Small deletions: - 1 to - l_{sdel} bases (if possible)

The transition probabilities can be defined arbitrarily, but should not depend on the starting size s_0 .

- Duplications: + 1 to + s_0 bases
- Large deletions: - 1 to - s_0 bases

Each possible final state is reached with probability $1/s_0$. (But we will generalize later).

⇨ Elementary matrices M_{ins}, M_{sdel}, M_{dup}, M_{ldel},

where e.g. $(M_{ins})_{ij}$ is the probability that a genome of initial size i ends up with size j after exactly one small insertion

Why a uniform distribution for the size of the duplications and deletions?

- Assumption on the underlying mechanism: uniformly distributed breakpoints
- Observations in bacteria: single deletions up to more than 200 kb ≈ 180 genes (Porwollik et al, 2004; Nilsson et al, 2005)
- Observations in humans (Lupski, 2007):
  - in 50% of the cases, the Charcot-Marie-Tooth disease is caused by a 1.4 Mb duplication
  - In 90% of the cases, the Smith-Magenis syndrom is caused by a partial deletion of chromosome 17, spanning from 950 kb to 9 Mb… (9Mb is twice the size of the complete *E. coli* genome)
- And we cannot observe the lethal events, which may be even larger…

The mutation rates: From the elementary matrices to the full transition matrix M_G

- 4 mutation rates : μ_{ins}, μ_{sdel}, μ_{ldel}, μ_{dup}
- Expressed per bp per generation
- Total mutation rate: $\mu = \mu_{ins} + \mu_{sdel} + \mu_{ldel} + \mu_{dup}$
- Assumption: mutations follow independent Poisson processes, no preferred order

⇨ Intermediate matrix $M_1 = \frac{\mu_{ins}}{\mu} M_{ins} + \frac{\mu_{sdel}}{\mu} M_{sdel} + \frac{\mu_{dup}}{\mu} M_{dup} + \frac{\mu_{ldel}}{\mu} M_{ldel}$

where $(M_1)_{ij}$ is the probability that a genome of initial size *i* ends up with size *j* after exactly one mutation

The mutation rates: From the elementary matrices to the full transition matrix M_G

- 4 mutation rates : μ_{ins}, μ_{sdel}, μ_{ldel}, μ_{dup}
- Expressed per bp per generation
- Total mutation rate: $\mu = \mu_{ins} + \mu_{sdel} + \mu_{ldel} + \mu_{dup}$
- Assumption: mutations follow independent Poisson processes, no preferred order

⇨ Intermediary matrix $M_1 = \frac{\mu_{ins}}{\mu} M_{ins} + \frac{\mu_{sdel}}{\mu} M_{sdel} + \frac{\mu_{dup}}{\mu} M_{dup} + \frac{\mu_{ldel}}{\mu} M_{ldel}$

⇨ Finally $\left(M_G\right)_{ij} = \sum_{n=0}^{+\infty} \frac{e^{-\mu i}(\mu i)^n}{n!} \left(M_{\mathbb{1}}^n\right)_{ij}$

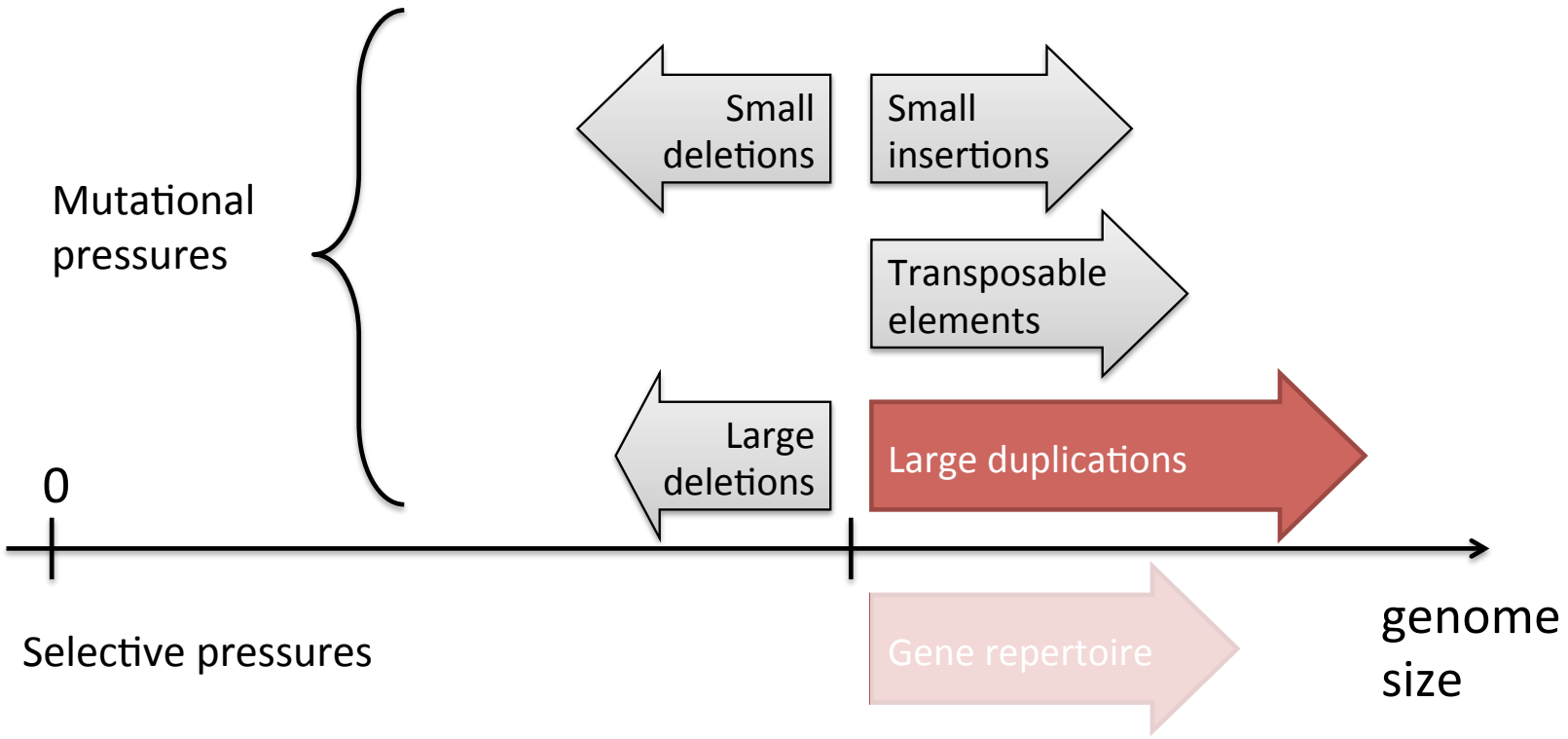
which is the probability that a genome of initial size *i* ends up with size *j* at the end of the reproduction

Related models

- Quasispecies models
  - initial model was very general [Eigen, 1971]
  - but most results were obtained for the special case of fixed genome length and point mutations only [Eigen, 1971; Nowak & Schuster, 1989; Barbosa et al., 2012].

- Population genetics models for microsatellite and transposable elements
  - number of elements not bounded [Falush & Iwasa, 1999]
  - additive and multiplicative effects [Stephan, 1987; Falush & Iwasa, 1999]
  - several mutations can occur during the reproduction [Ohta & Kimura, 1981; Stephan, 1987]
  - but no model combines those three features

What does this model answer
to our original question ?



How will genome size evolve if :

- duplications are twice as frequent as deletions ($\mu_{dup} = 2\mu_{ldel}$),
- transposable elements proliferate ($\mu_{ins} > \mu_{sdel}$)
- and selection systematically favors the highest gene numbers?

Result 1: Condition for non-infinite growth without selection

Theorem 2 (Stationary distribution for genome size without selection). *If* $(2\log 2 - 1)\mu_{dup} < \mu_{ldel}$, *then the Markov chain* $(\mathbb{N}^*, \mathbf{M}_G)$ *has a unique asymptotic stationary probability vector* ν_∞. *For any initial distribution* ν_0, *the distribution of genome sizes converges to* ν_∞. *Mathematically,*

$$\lim_{t\to\infty} \|\nu_0 \mathbf{M}_G^t - \nu_\infty\| = 0$$

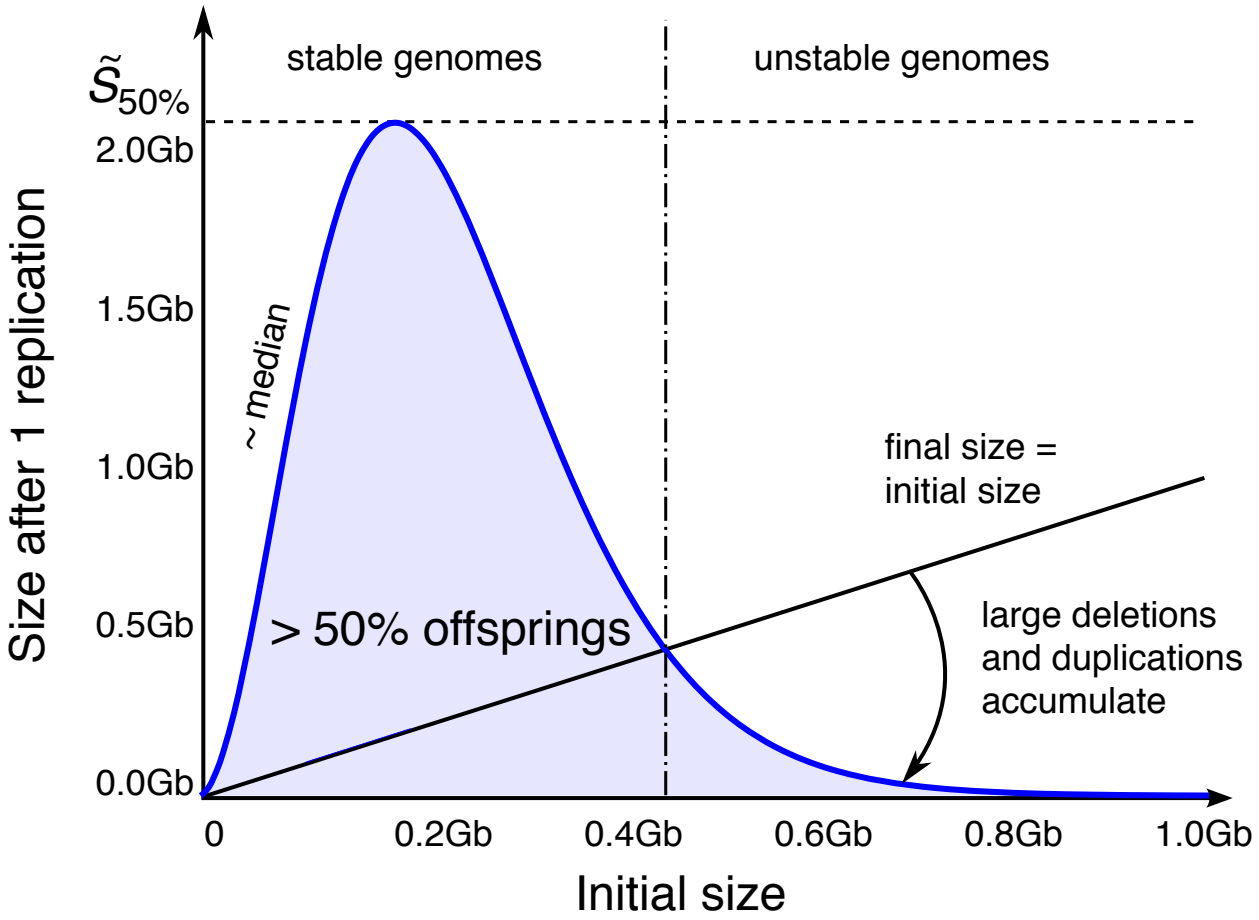
Biologically, the convergence of the distribution implies that, even after a long time of evolution, genome size does not tend to infinity: an arbitrary large part of genomes is located beneath a finite size.

No infinite growth if $\mu_{dup} < 2.6\ \mu_{ldel}$
This condition is independent from the rates of small insertions (eg transposable elements) and small deletions

The proof uses Doeblin's condition.

[Fisher et al. , submitted]

Result 2: Even a caricatural selection cannot push the genomes towards an infinite size



Upper bound for the median of the distribution at t+1

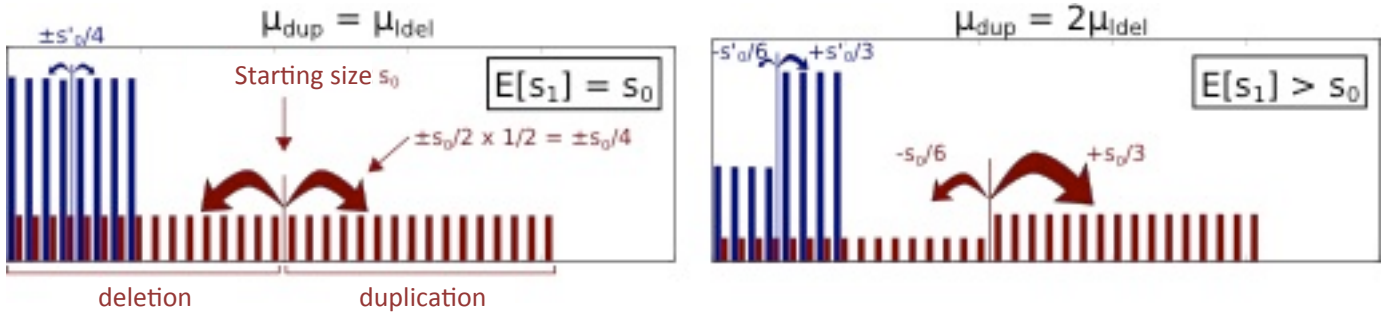
At each generation, >50% of the population is below some threshold

Large genomes can be selected but are too unstable…

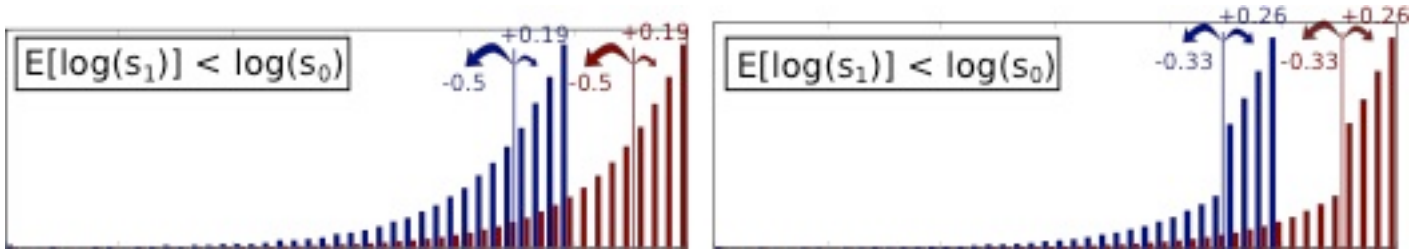
Which properties
underlie these results?

Property 1: Asymmetry of duplications and deletions in logarithmic scale

Linear scale: apparent symmetry but ***no*** scale invariance. The behavior is hard to predict intuitively.



Log scale: scale-invariance, no symmetry anymore. A mutational bias towards shrinkage is revealed, even for $\mu_{dup} = 2\mu_{ldel}$.



Property 1: Asymmetry of duplications and deletions in logarithmic scale

Property 1. Let $\Delta(s) = \mathbb{E}\left[\log(S_{n+1})|S_n = s\right] - \mathbb{E}\left[\log(S_n)|S_n = s\right]$, the average size of one-mutation jumps in logarithmic scale, starting from s.

- if the (n+1)th mutation is a large deletion, $\Delta(s) \underset{s\to+\infty}{\longrightarrow} -1$.
- if the (n+1)th mutation is a duplication, $\Delta(s) \underset{s\to+\infty}{\longrightarrow} 2\log 2 - 1$.
- if the (n+1)th mutation is an indel, $\Delta(s) \underset{s\to+\infty}{\longrightarrow} 0$.

This property is important in the proof of the first result (condition for the existence of a stationary distribution).

Generalization to non-uniform distributions for the size of duplications and deletions

Corollary 1. *(Generalization of Theorem 2) Suppose we have distributions of duplications, large deletions and indels, such that there exists a positive and increasing scaling function f that verifies the following conditions.*

For $\Delta(s) = \mathbb{E}\left[f(S_{n+1}) - f(S_n) | S_n = s\right]$*:*

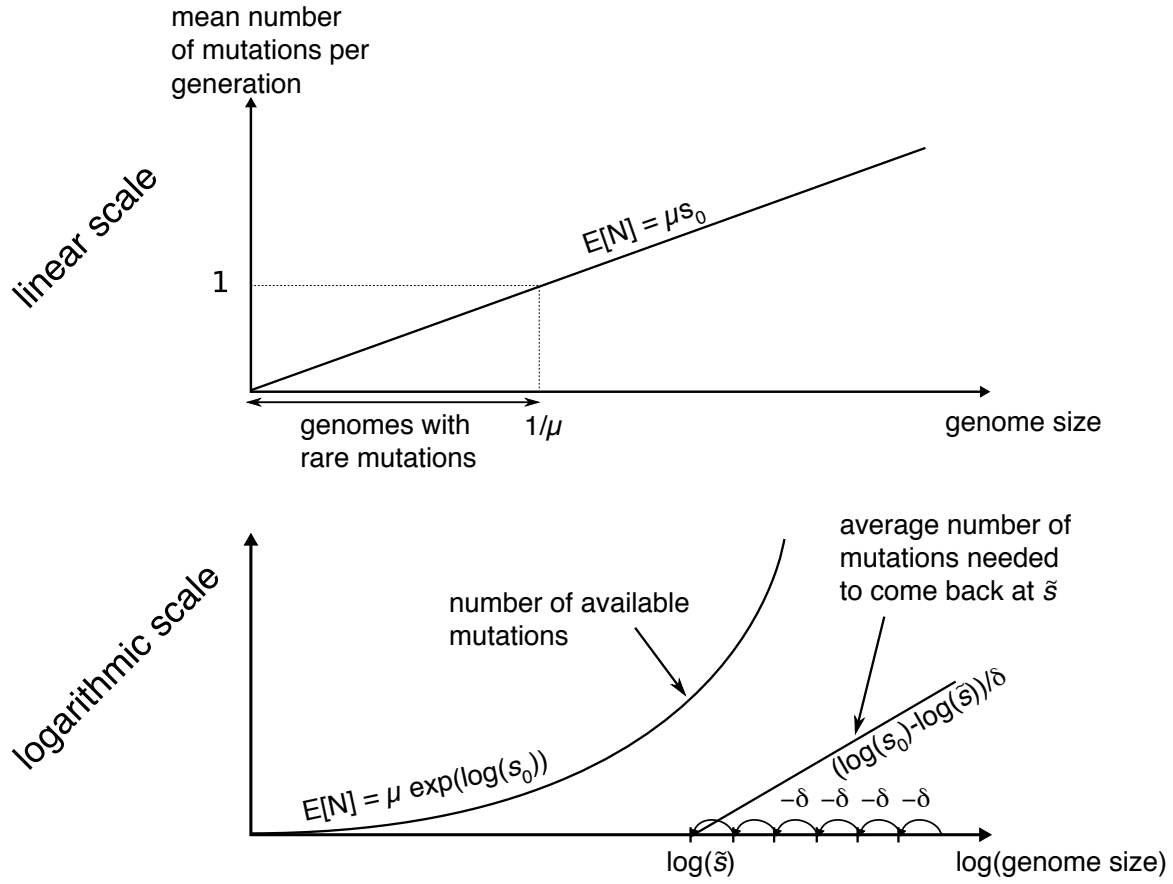
- *if the (n+1)th mutation is a deletion,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{ldel}$.
- *if the (n+1)th mutation is a duplication,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{dup}$.
- *if the (n+1)th mutation is an small insertion,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{ins}$.
- *if the (n+1)th mutation is an small deletion,* $\Delta(s) \underset{s\to+\infty}{\longrightarrow} \delta_{sdel}$.

where $\delta_{ldel} \leq 0$, $\delta_{dup} \geq 0$, $\delta_{ins} \geq 0$ *and* $\delta_{sdel} \leq 0$ *are constants among which at least one is nonzero.*

Then the Markov chain $(\mathbb{N}^*, \mathbf{M}_G)$ *has a unique stationary probability vector* ν_∞ *if*

$$\mu_{ldel}\delta_{ldel} + \mu_{dup}\delta_{dup} + \mu_{ins}\delta_{ins} + \mu_{sdel}\delta_{sdel} < 0 \tag{4}$$

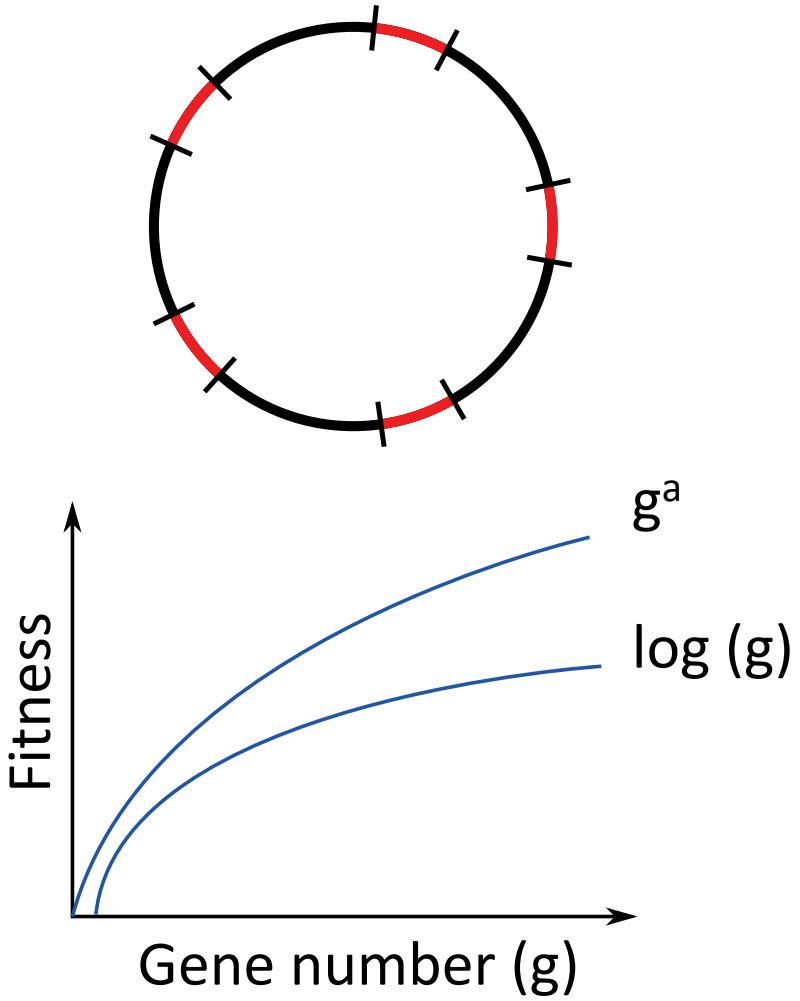
Property 2: Larger genomes undergo more mutations



This property is important in the proof of the second result (selection cannot overcome the spontaneous mutational dynamics).

Let’s simulate the model with (a rather brutal) selection

Introducing fitness : coding versus non-coding DNA

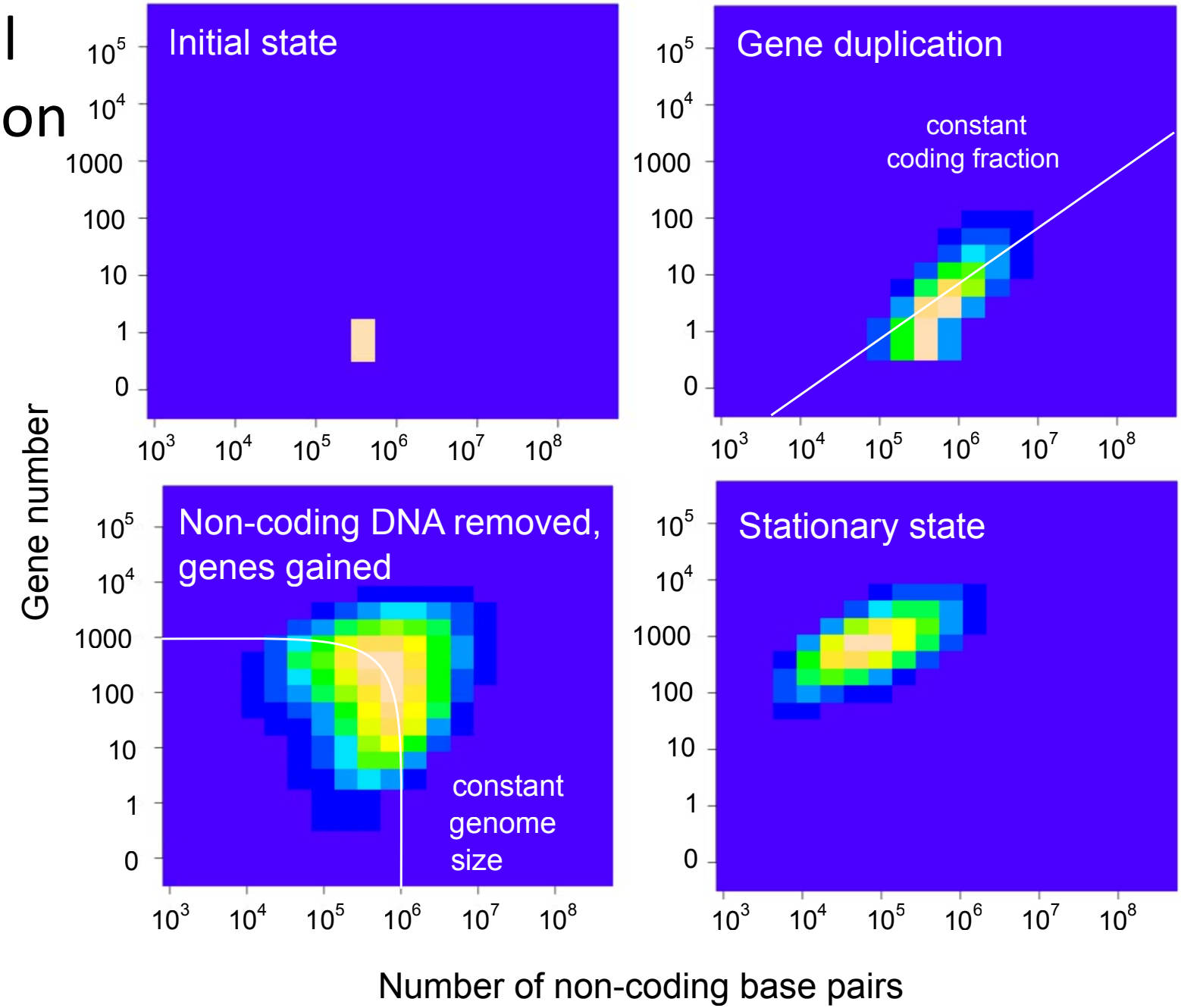


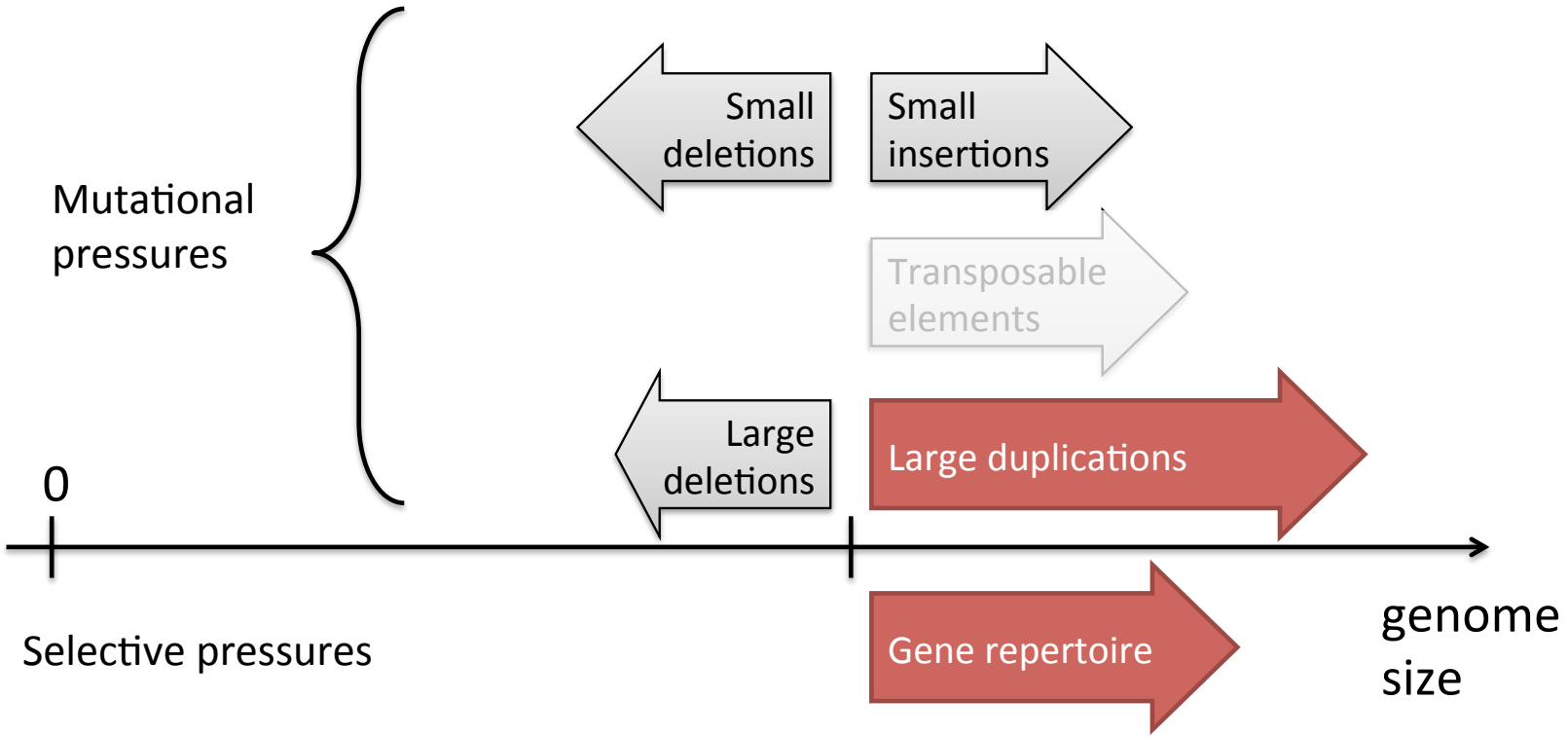
Hypotheses:

- circular genome described by its gene number g and the number of non-coding bases,
- Genes all have the same (fixed) length
- Non-coding bases are equally distributed between genes
- The fitness is a monotonically increasing, not bounded, function of g

$$n_{t+1} = n_t \frac{F}{\|n_t F\|} M$$

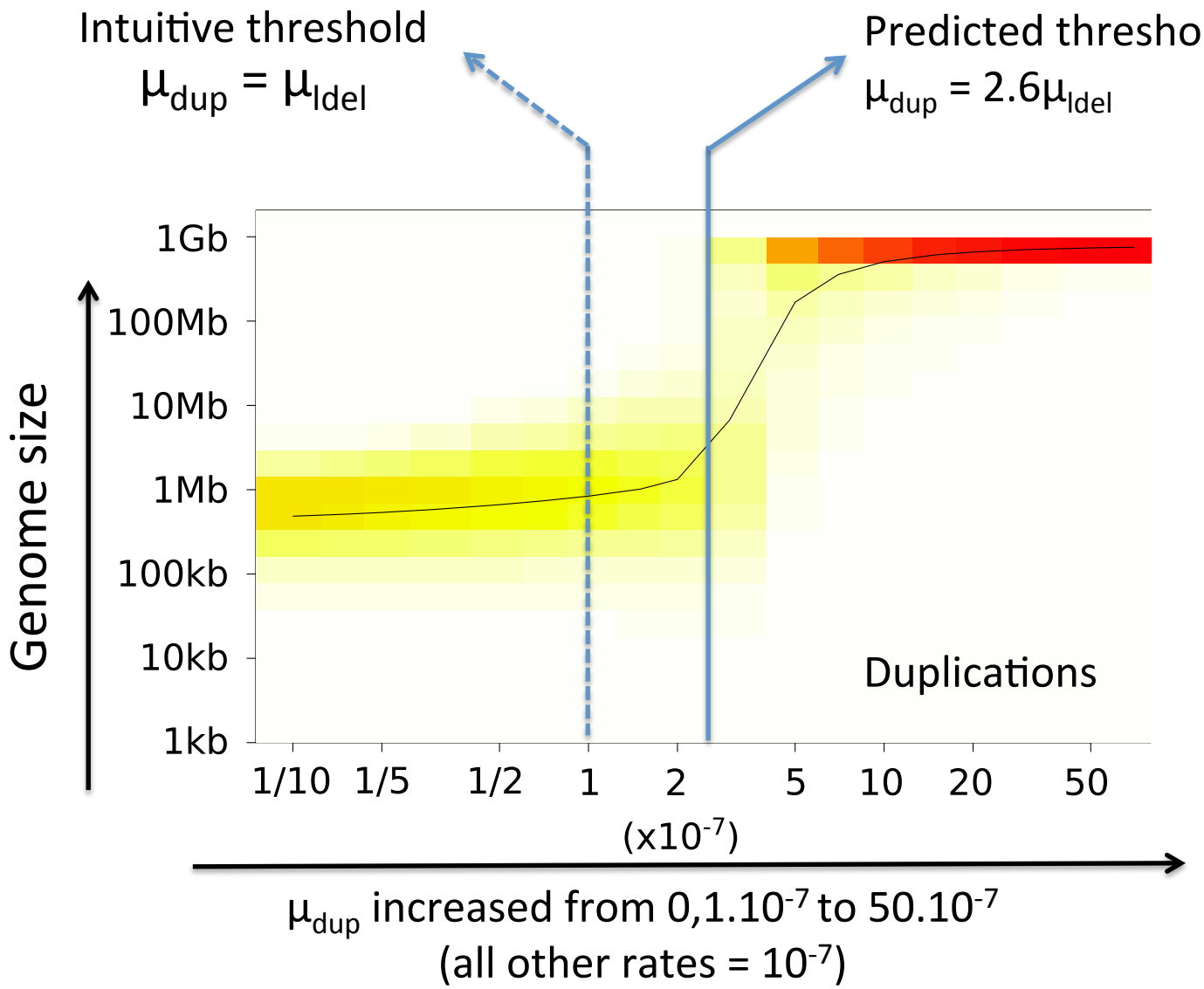
A typical simulation



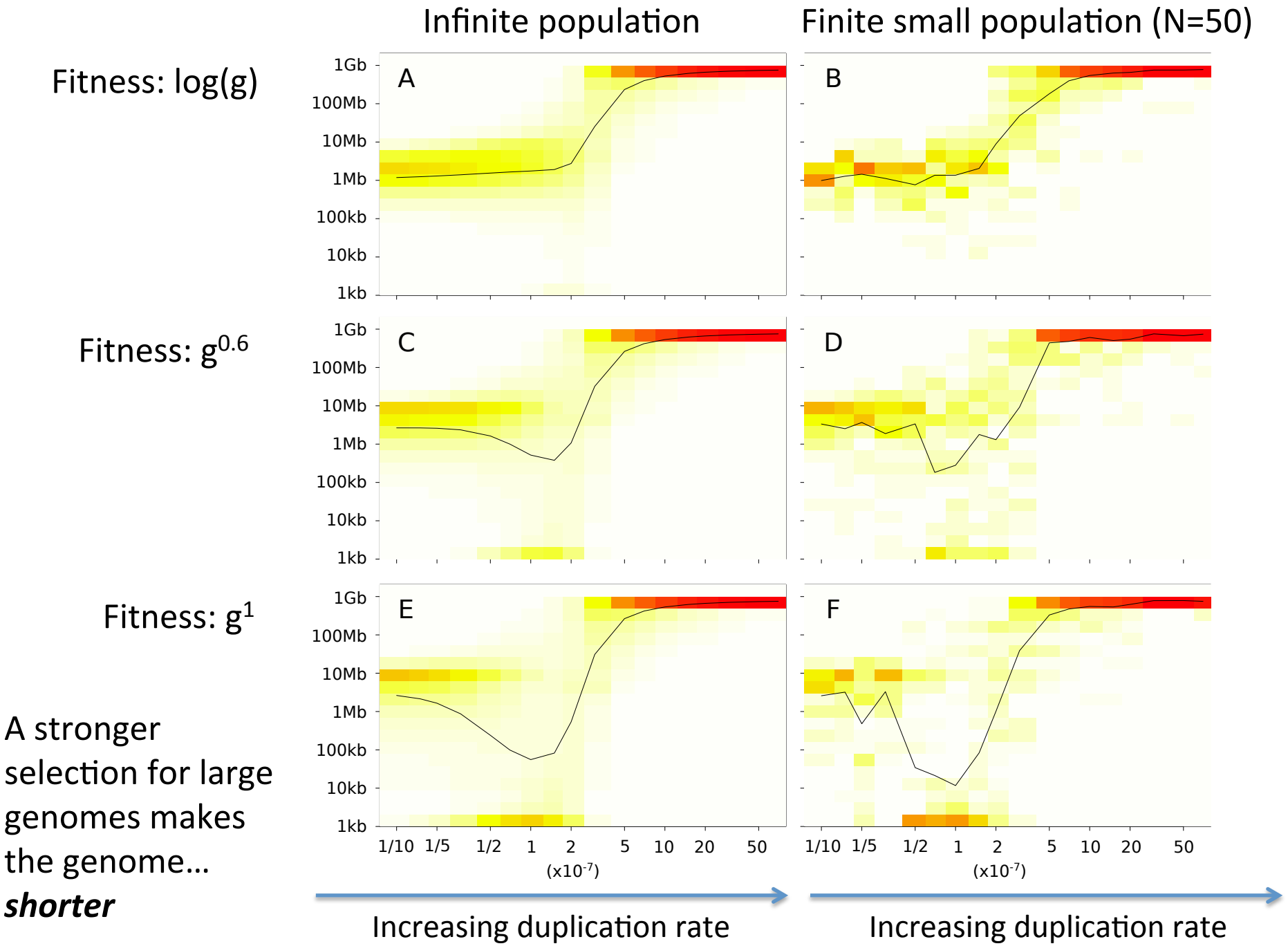


How will genome size evolve if :

- duplications are twice as frequent as deletions,
- transposable elements proliferate,
- and selection systematically favors the highest gene numbers?



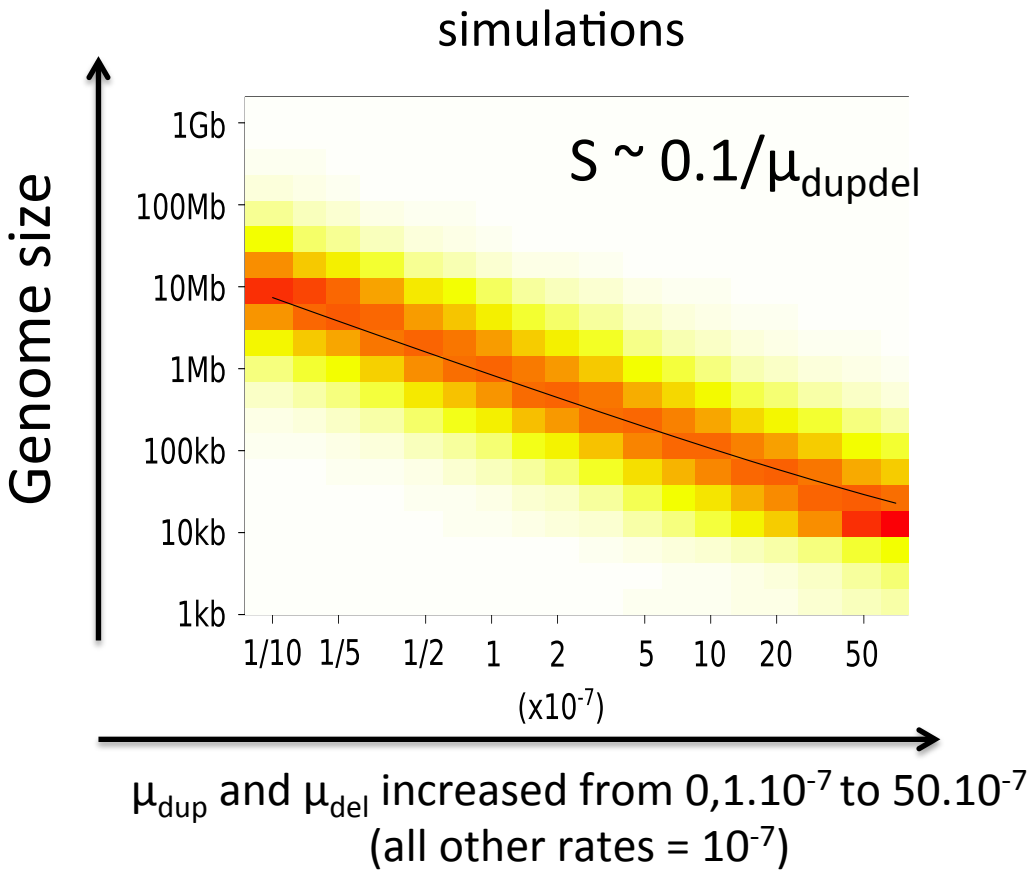
In these runs, fitness was proportional to log(gene number)

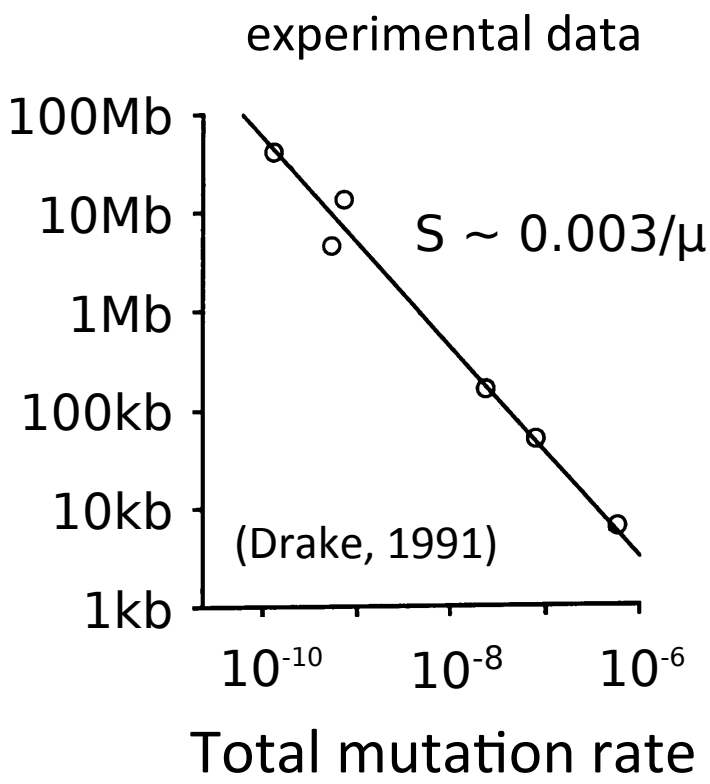

Infinite population
Finite small population (N=50)
Fitness: log(g)
Fitness: g^0.6
Fitness: g^1
A
B
C
D
E
F
1Gb
100Mb
10Mb
1Mb
100kb
10kb
1kb
1/10 1/5 1/2 1 2 5 10 20 50
(x10^-7)
Increasing duplication rate
Increasing duplication rate
A stronger selection for large genomes makes the genome…
shorter

OK, the intuition fails in this scenario…

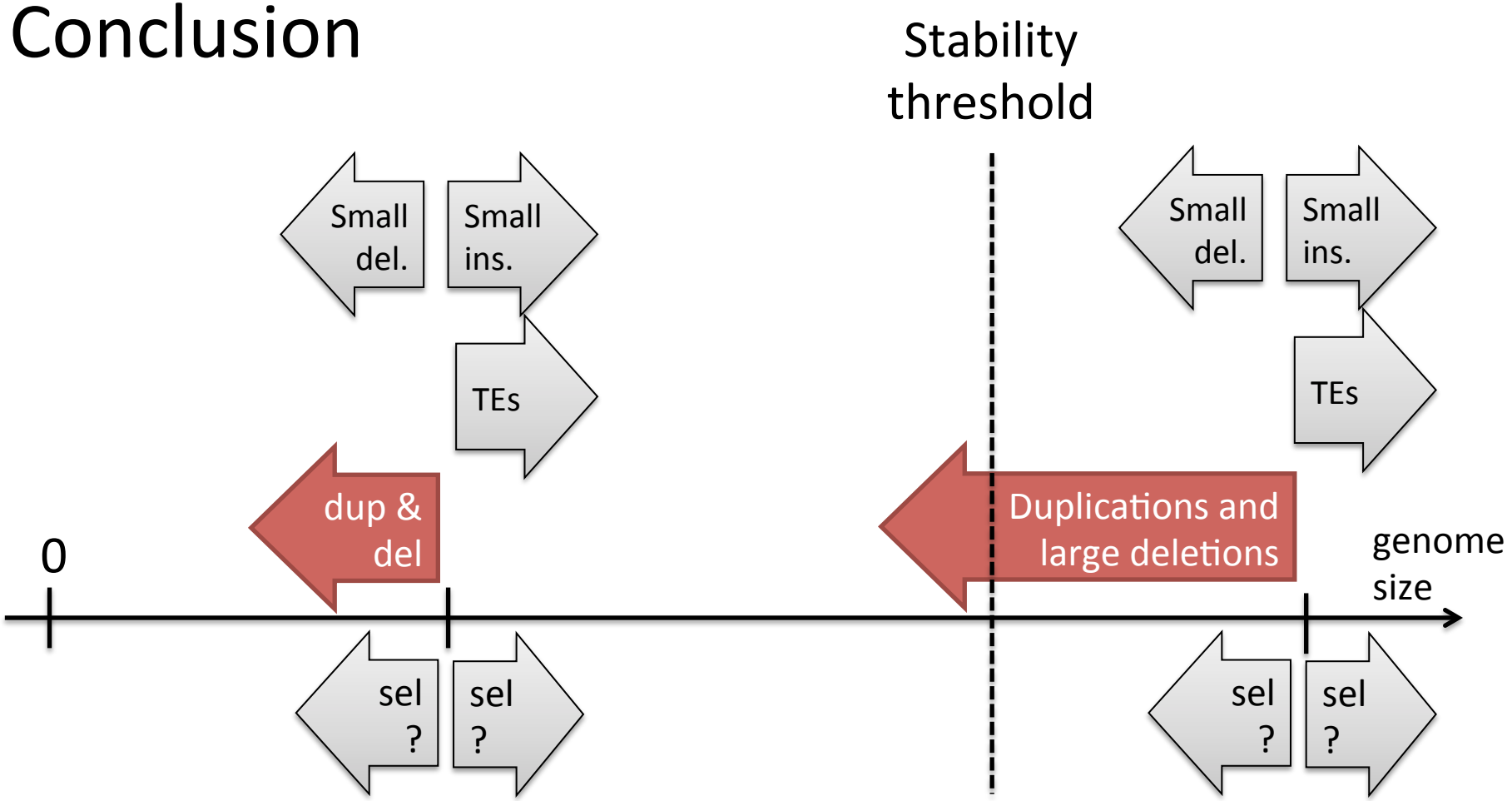
but in reality, the rates of duplication and deletion do not evolve independently

Evolved genome size is inversely proportional to the rate of multiplicative events





Conclusion



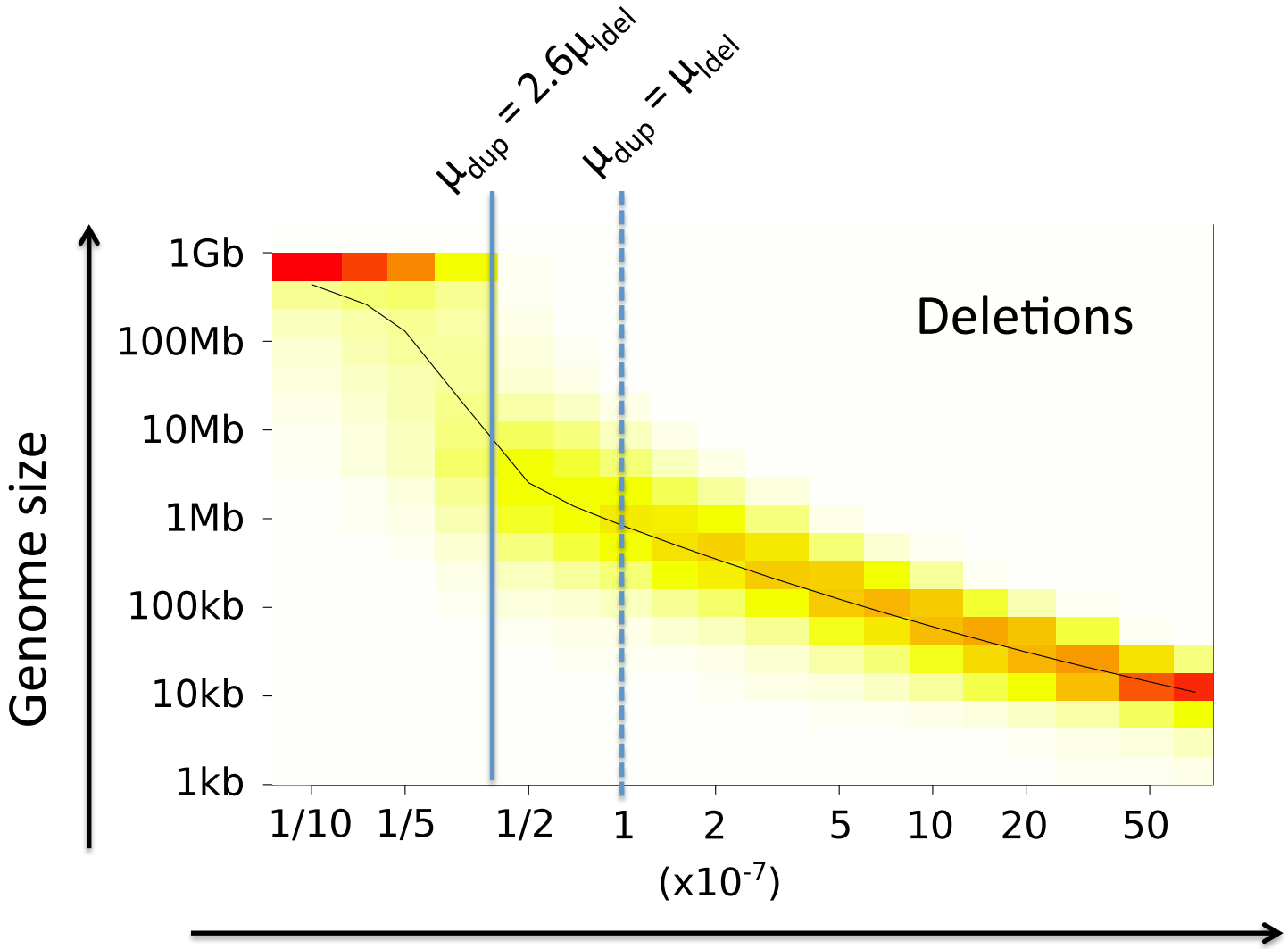
- The instability caused by duplications and deletions prevents genomes from growing above a certain threshold
- Below this threshold, this mutational bias is weaker and other pressures can play a role

Perspectives

- Incorporate transposable elements in the simulations too, not just in the formal analysis
- Update after each duplication or deletion the number of other events to be done, not just at each replication
- Try more realistic fitness landscapes, where not every duplication is beneficial and where deletions can be lethal

Take-home message

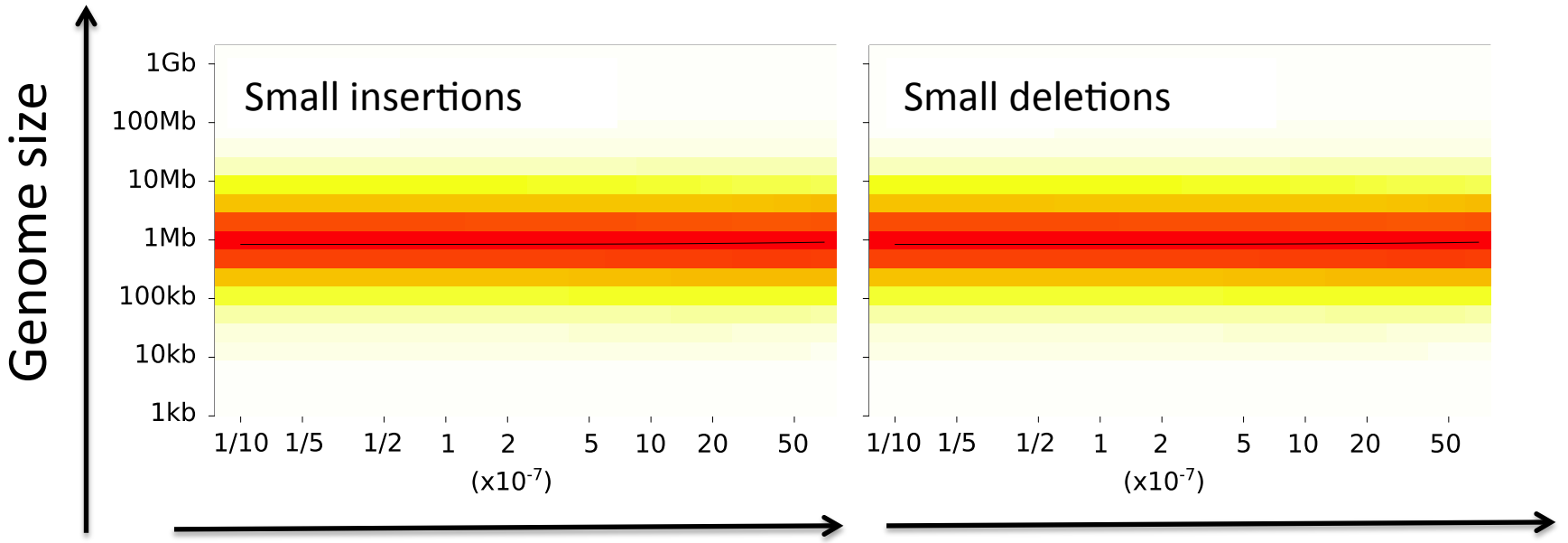
Evolution can be more subtle than we think, do not trust « thought experiments »



μ_{Idel} increased from 0,1.10^{-7} to 50.10^{-7}
(all other rates = 10^{-7})

In these runs, fitness was proportional to log(gene number)

(Non) effect of small insertions and small deletions on the stationary distribution of genome size



μ_{ins} increased from 0,1.10^{-7} to 50.10^{-7}
(all other rates = 10^{-7})

μ_{sdel} increased from 0,1.10^{-7} to 50.10^{-7}
(all other rates = 10^{-7})

In these runs, fitness was proportional to log(gene number)