Coalescent trees under weak genetic draft: effective population size and the Lewontin's paradox

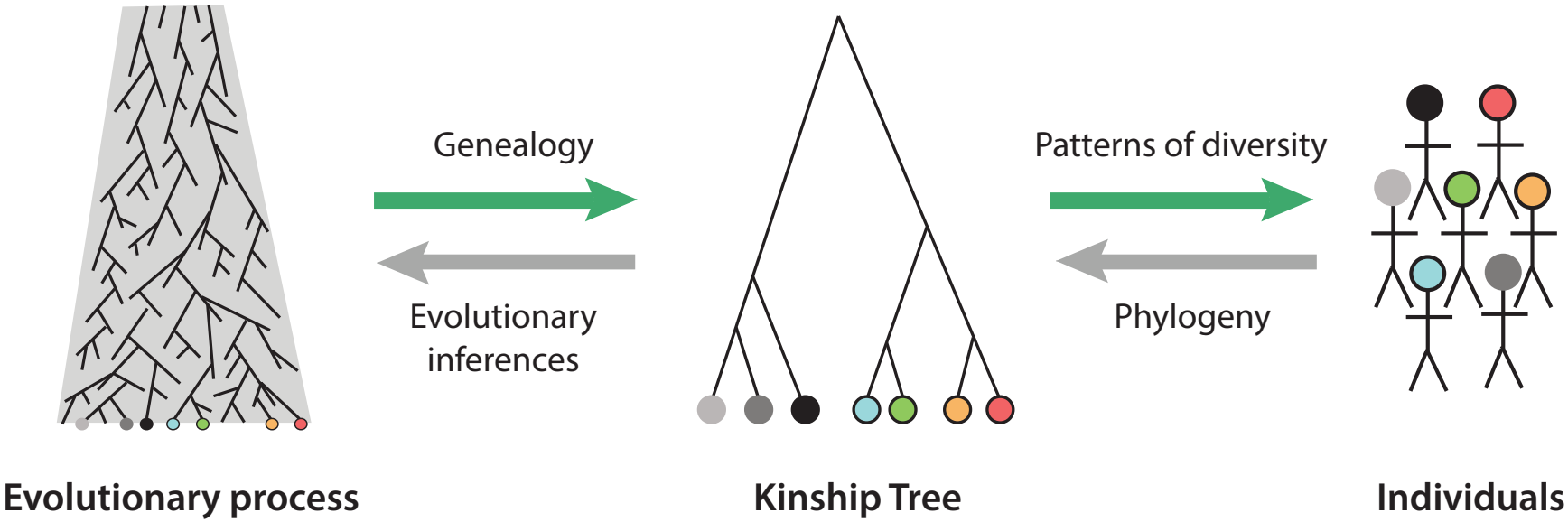
Guillaume Achaz[1] and Emmanuel Schertzer[2]

1: Stochastic Models of the Inference of Life Evolution (CIRB, Collège de France)
2: Dynamical Systems in Biomathematics (University of Vienna)

with the complicity of Jean-Baptiste Grodwohl[3]

3: Science, Philosophie, Histoire (SPHERE, Université Paris-Cité)

Beyond the trees



From processes to patterns and *vice-versa*

Molecular polymorphism

Early key dates

Zuckerkandl and Pauling, 1962

Genetic differences scales with divergence time between species

Lewontin and Hubby, 1966

Many loci are polymorphic within species

Kimura, 1968; King and Jukes 1969

Patterns of polymorphisms are compatible with neutrality

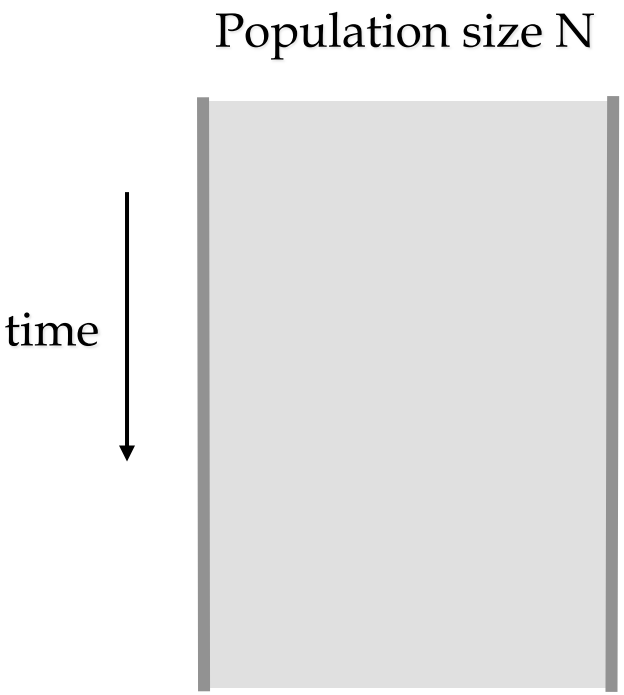
Lewontin, 1974

Range of population sizes does not reflect range of diversity

A neutralist *vs* selectionnist long-standing debate

(still ongoing, e.g. Kern et Hahn, 2018 and responses)

Standard Neutral Models (SNM)



Assumptions

Constant Population Size
Strict Panmixia
No selection

Consequences

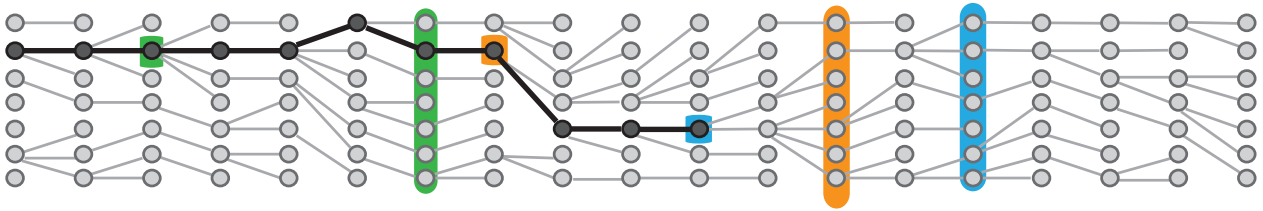
X: the number of descendants
is Poisson distributed with

E[X] = 1, for all N individuals

Genetic drift in two classics

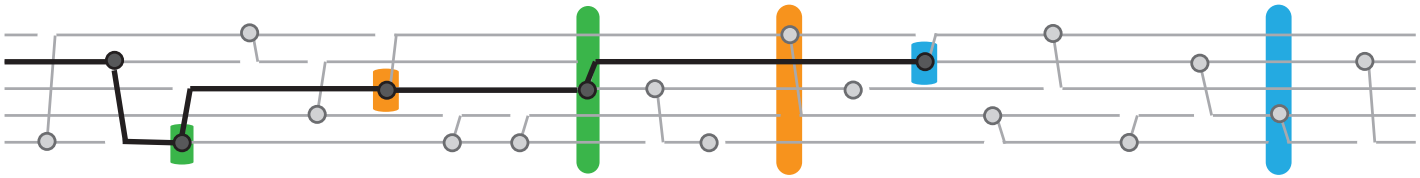
Wright-Fisher (20's) – drift time scale : N generations

1 generation = all individuals die and are replaced by random sampling



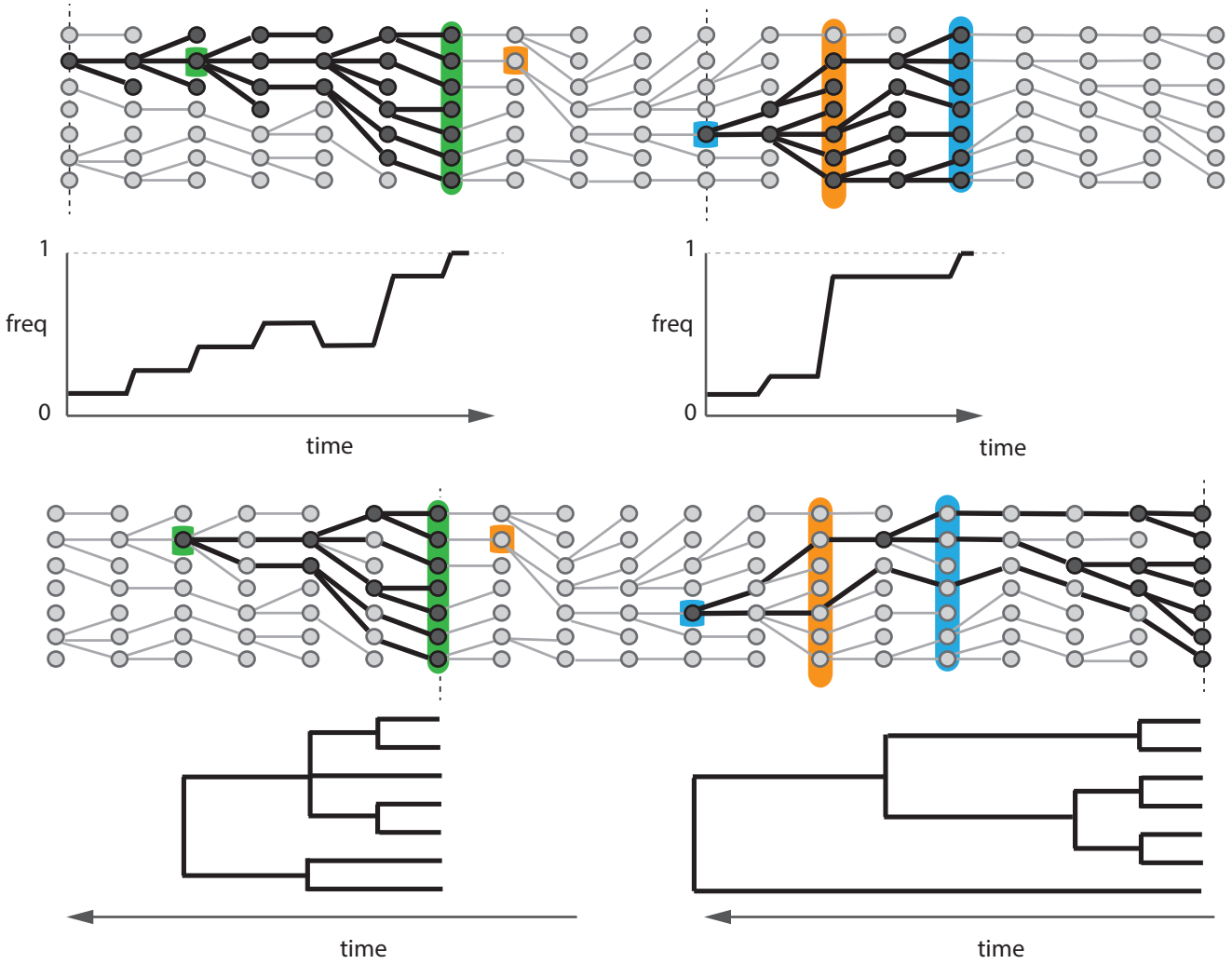
Moran (1958) - drift time scale : N/2 generations

1 time step = one random new-born replaces one random dead
1 generation = N time steps



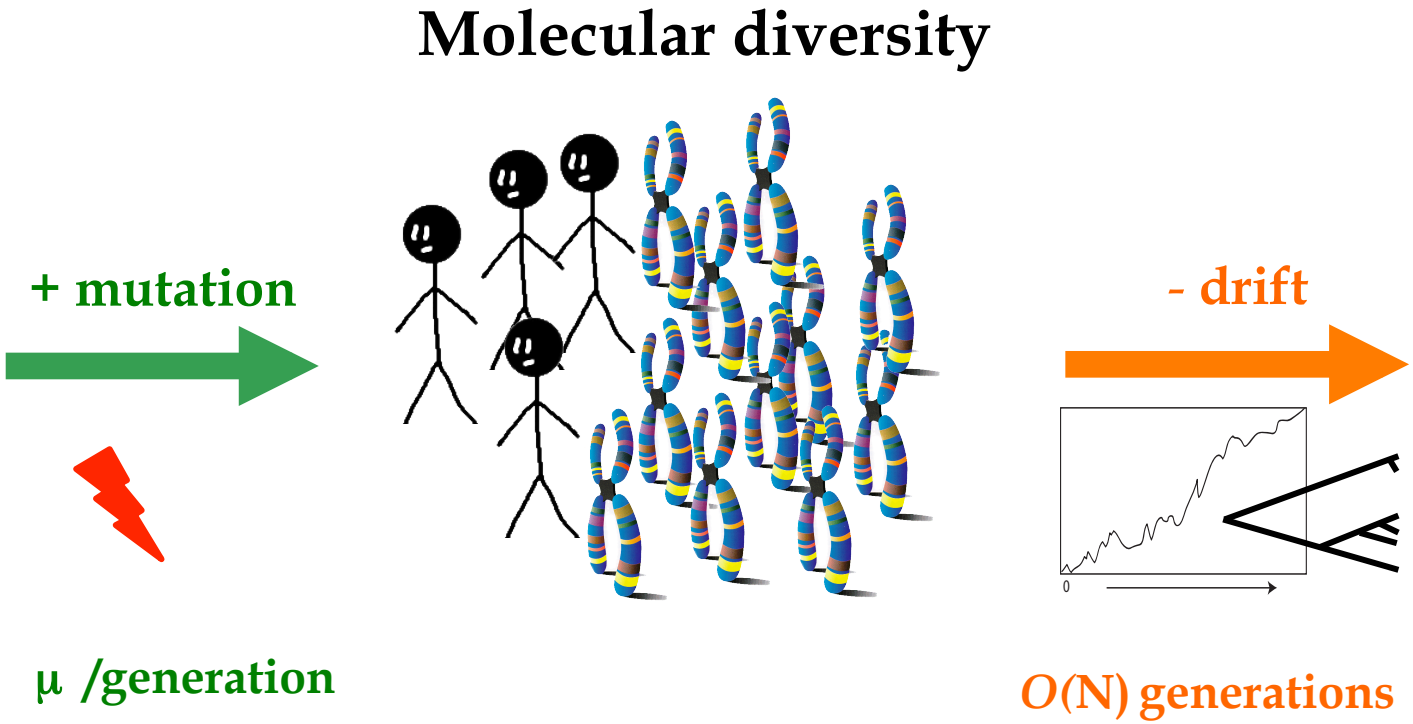
Drift time scale is *O*(N) generations for all *SN models*

A glimpse at the duality



(Achaz, Lambert and Schertzer, Adv. Appl. Prob. 2018)

The mutation-drift paradigm (H0)



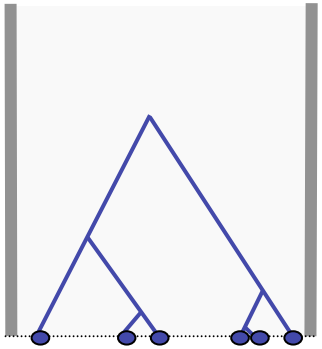
At equilibrium, diversity is $O(N\mu)$ for all S.N.M.

Wait a minute…

do we really have the “right” model?

From model to real populations

Population size N



Model Population

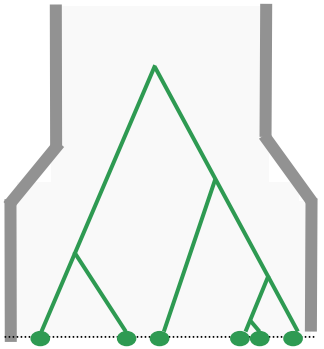
Constant Population Size
Strict Panmixia
No selection

$E[t_2] = N$
$E[\pi] = 2\,N\,\mu$

(in a Wright-Fisher model)

π: pairwise differences
μ: mutation rate

"Relevant" population size N_e



Realistic Population

Demography
Structure
Selection

$\overline{t_2} = N_e$
$\hat{N}_e = \pi_{obs} / 2\,\mu$

N= *expected* time scale *vs* N_e = *observed* time scale

Effective population size

(ongoing work with JB Grodwhol)

Is it a *magic number*?

How to define it in the model world?

How to measure it?

How to infer it from real data?

What does it mean in models? and in data?

How history (1931-present) guides us in the labyrinth?

Is it a walking stick or a source of confusion?

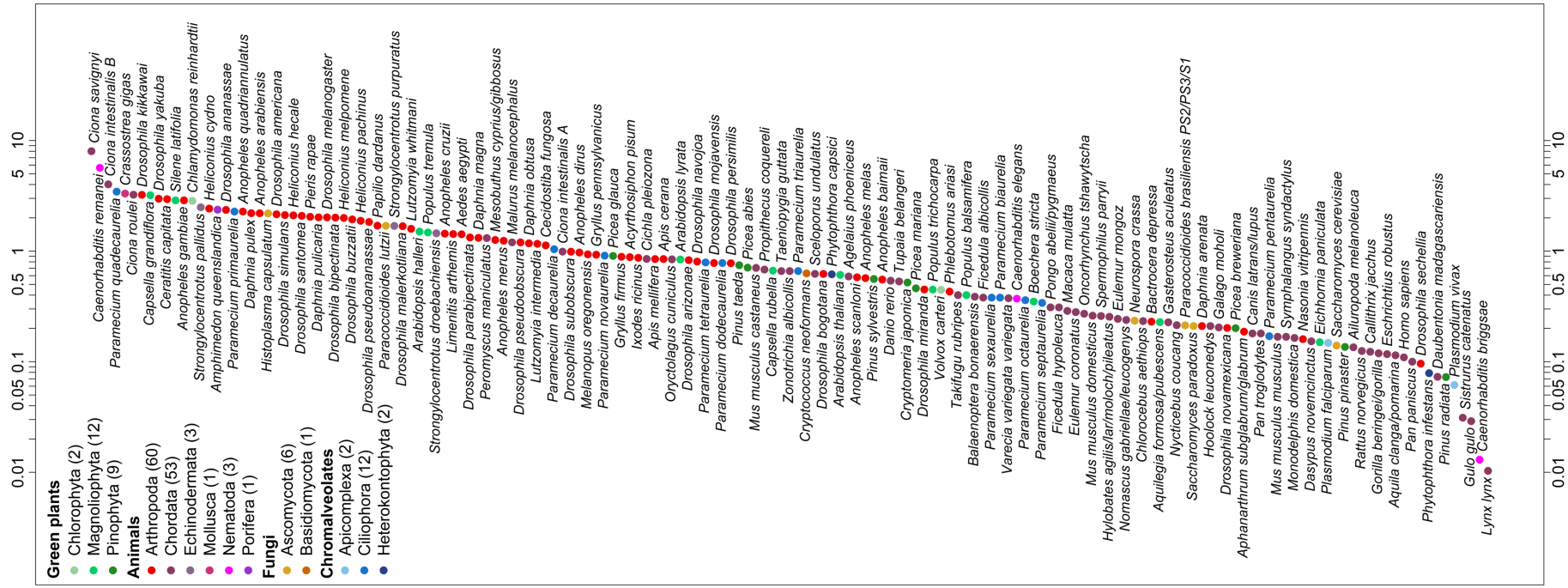
Please, let's discuss about it over coffee

Ancestry within species

Species	N	T_2 (aka N_e)
H. sapiens	10^{10}	10^4
G. gorilla	10^5	10^3
D. melanogaster	?	10^6
C. elegans	?	10^5
A. thaliana	?	10^5
P. kergelensis	?	10
F. psychrophilum	10^9/ml of cult.	10^6
E. coli	10^9/ml of cult.	10^8
HIV (within patient)	10^{10}	10^3

Why N_e (T_2/diversity) does not scale with N ?
(Lewontin 1974 variation paradox ; see also Leffler et al., 2012)

Diversity on a large dataset

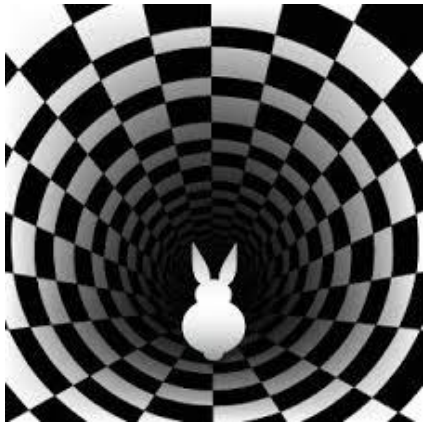


(from Leffler et al., 2012)

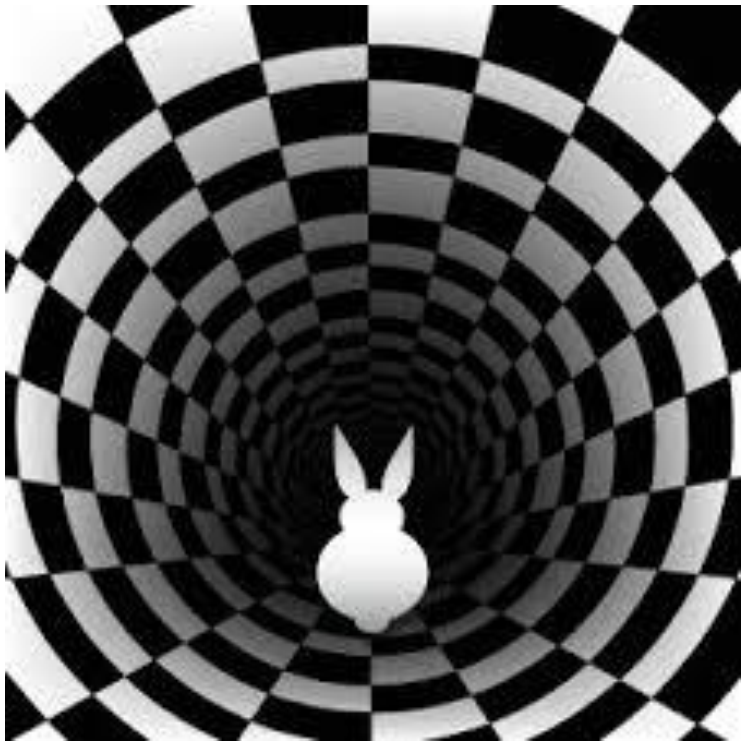
Why N_e does not scale with N? [Lewontin paradox, 1974]

Hmmm… are we confused?

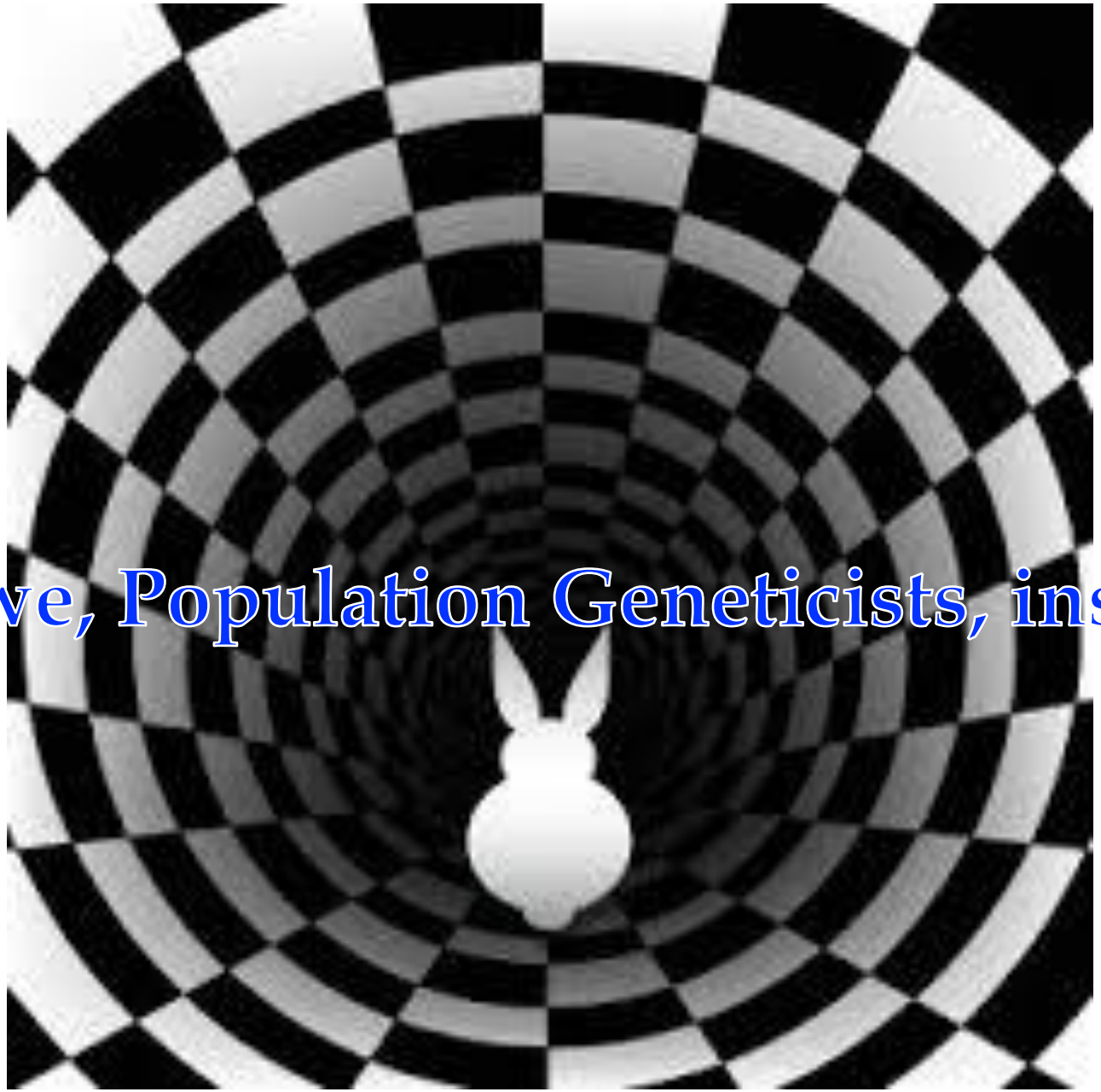
What factors limit *diversity* / T_2 / N_e?



What means N_e anyway?



What is genetic drift?



Are we, Population Geneticists, insane?

Some ideas

(recent review: Charlesworth and Jensen, 2023)

Structure

No, as it inflates *global* diversity

Demography

Perhaps, assuming strong deviations / founder effects
we need "N_e *individuals* N_e *generations ago*"

Speciation

Larger populations are more prone to speciation

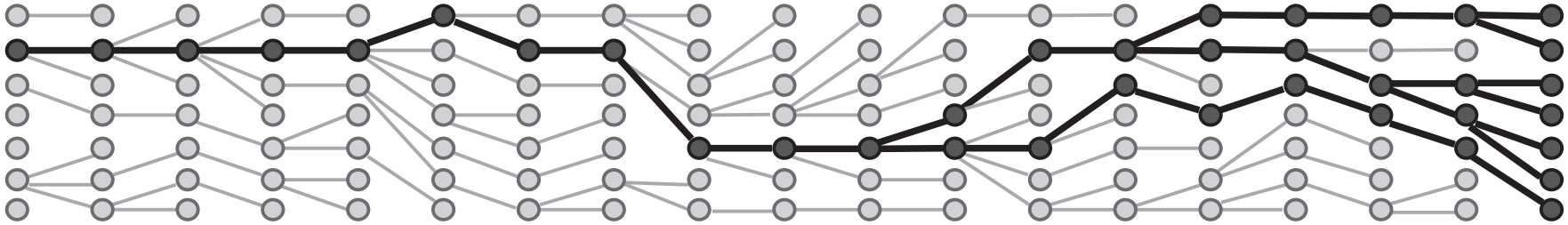
Selection through linkage

Background selection [Charlesworth et al., 1993]
Genetic draft [Gillespie, 2000]

…

Mutation, divergence and diversity

standard Wright-Fisher model

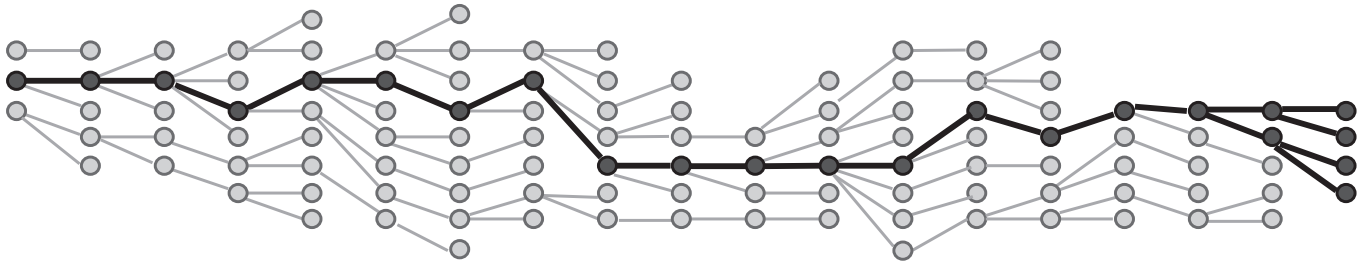


As only a single lineage persists, neutral mutations accumulate (linearly) with generations (evolutionary time)

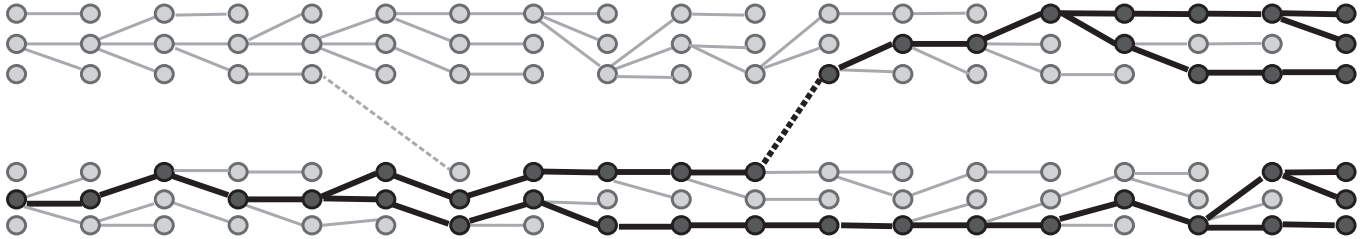
a) Between species (molecular clock, species divergence)
b) Within species (genetic diversity)

Genealogies across processes

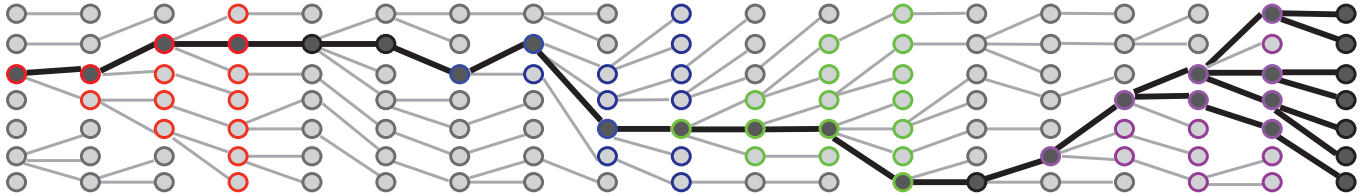
with demography



with structuration



with recurrent selection



Mutation, divergence and diversity

Regardless of the evolutionary process,

neutral mutations accumulates in lineages
with generations at their rate of appearance
(usually constant rate)

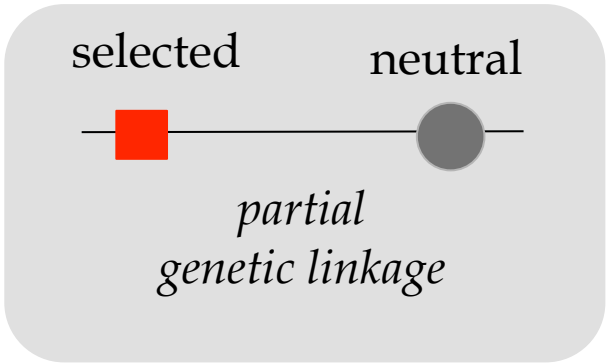
for both
divergence & diversity

If the vast majority of mutations are neutral, we can simply characterize genealogies and lineages

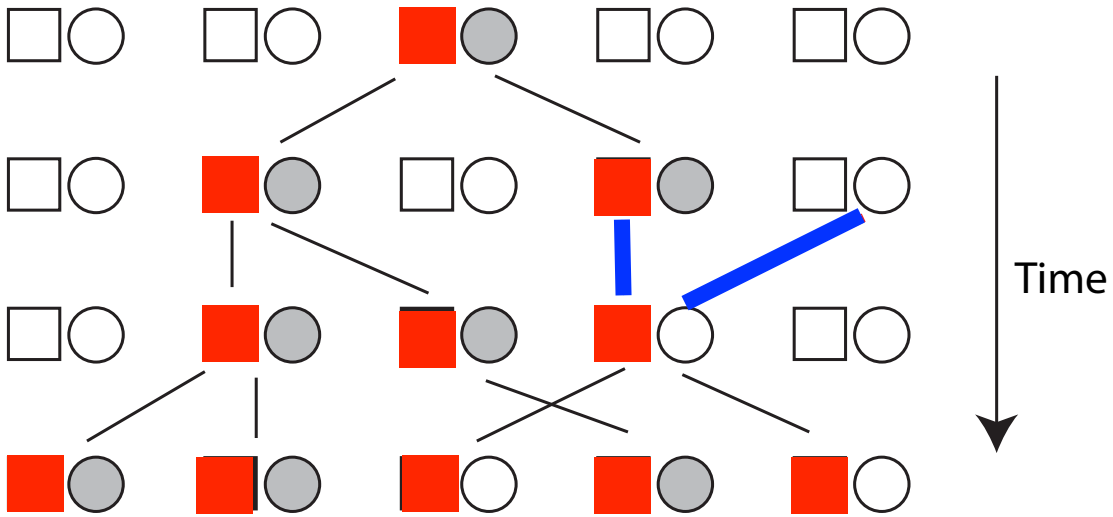
The weak genetic draft

The hitchhiking effect, forward time

(Maynard Smith and Haig, 1974; Wiehe and Stephan 1993)



forward Wright-Fisher



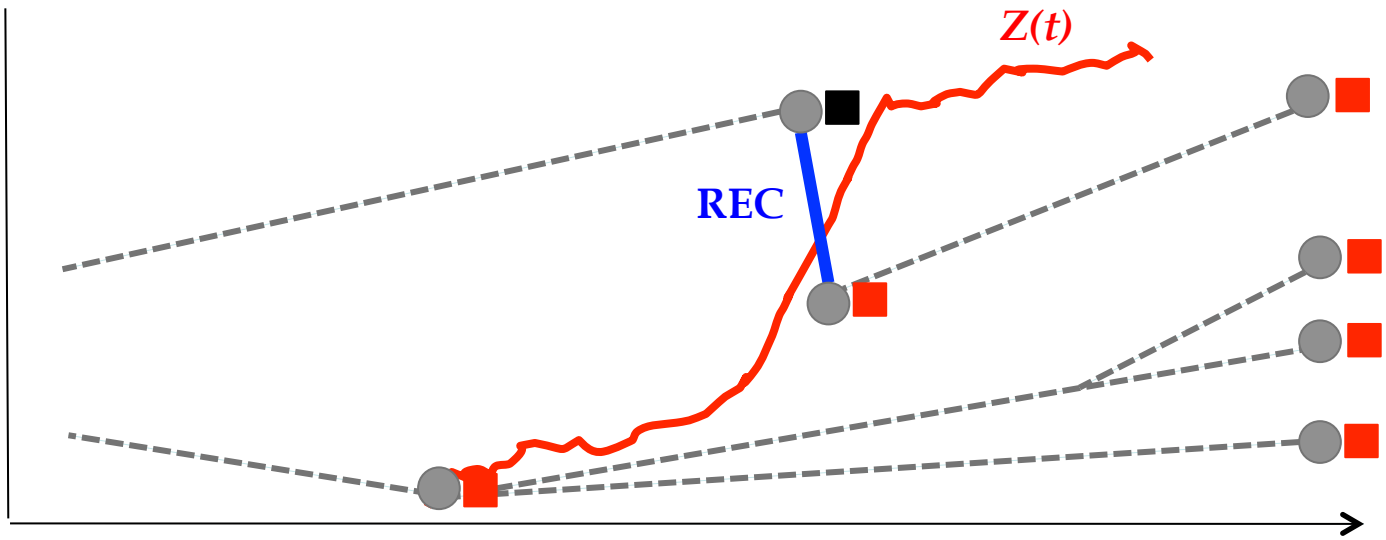
s : selection coefficient
c : recombination rate

The hitchhiking effect is tuned by the ratio c/s

The hitchhiking effect, backward time

(Hudson and Kaplan, 1988; Kaplan et al., 1989; Fay and Wu, 2000)

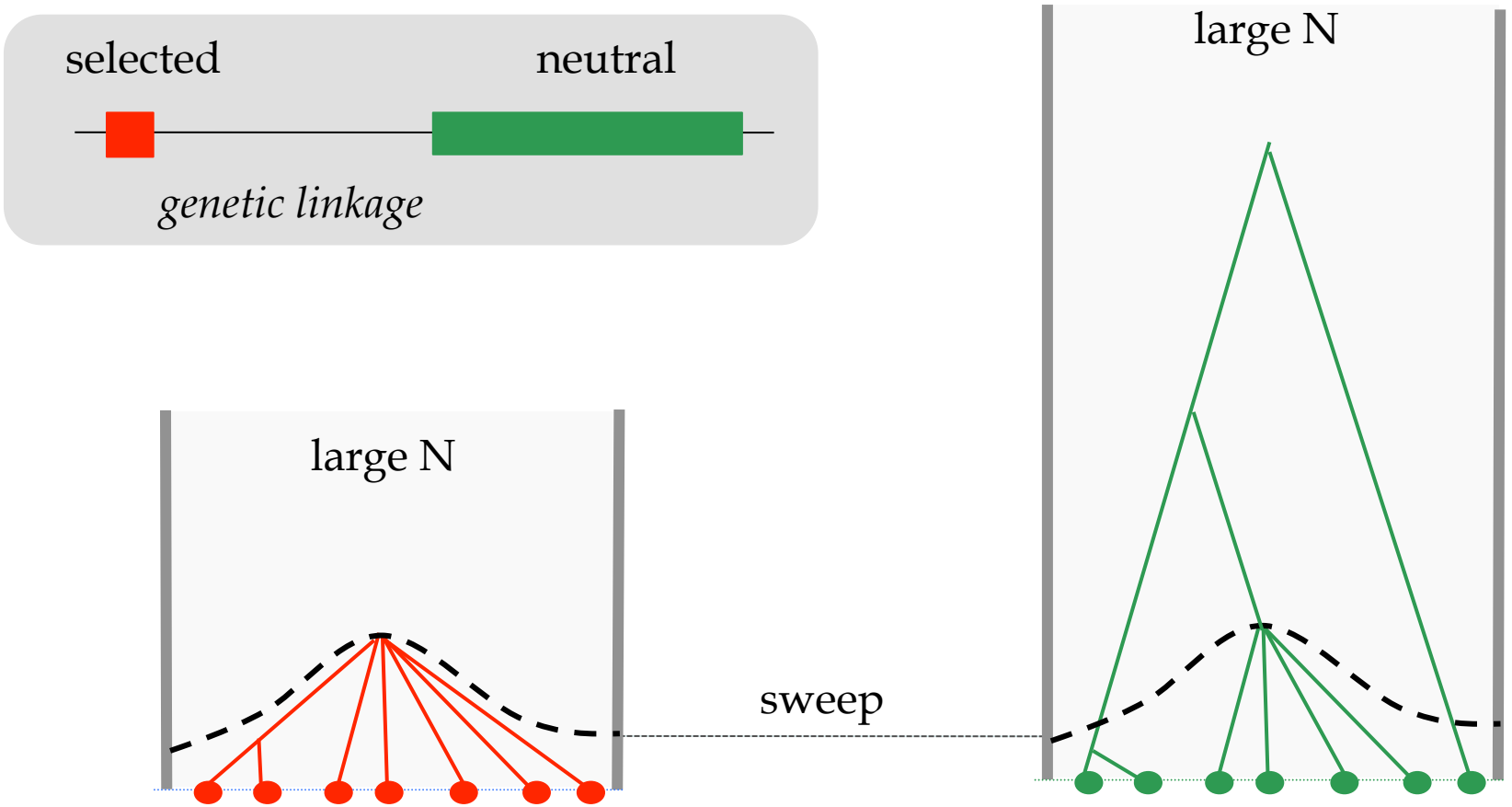
	in the mutant	**in the resident**
coal rates:	1 / N Z(t)	1 / N (1-Z(t))
eff rec rate:	c (1-Z(t))	c Z(t)



A structured coalescent
where recombination is 'migration' between 'compartments'

The hitchhiking effect, backward time

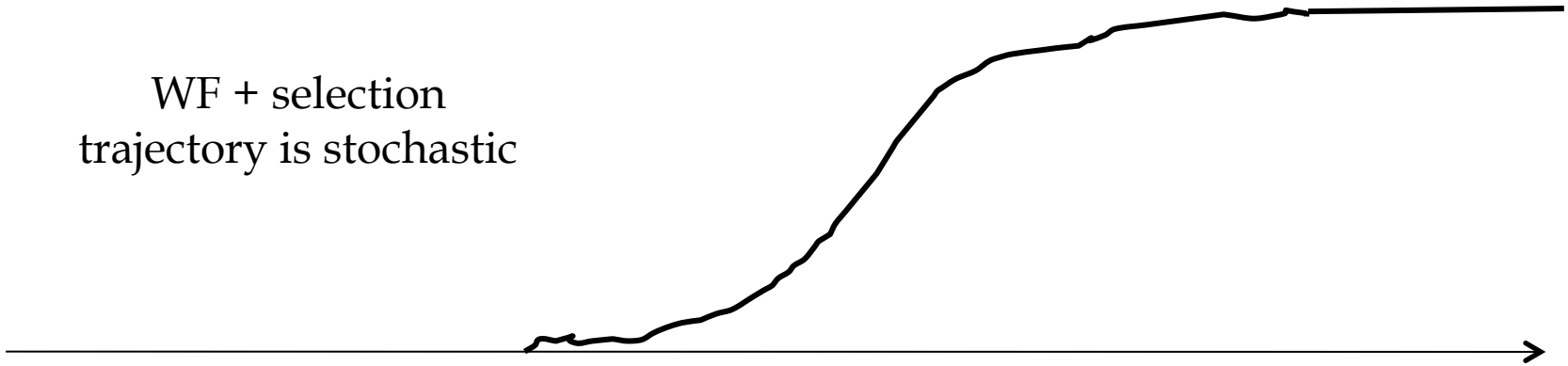
(Hudson and Kaplan, 1988; Kaplan et al., 1989; Fay and Wu, 2000)



The RIF approximation

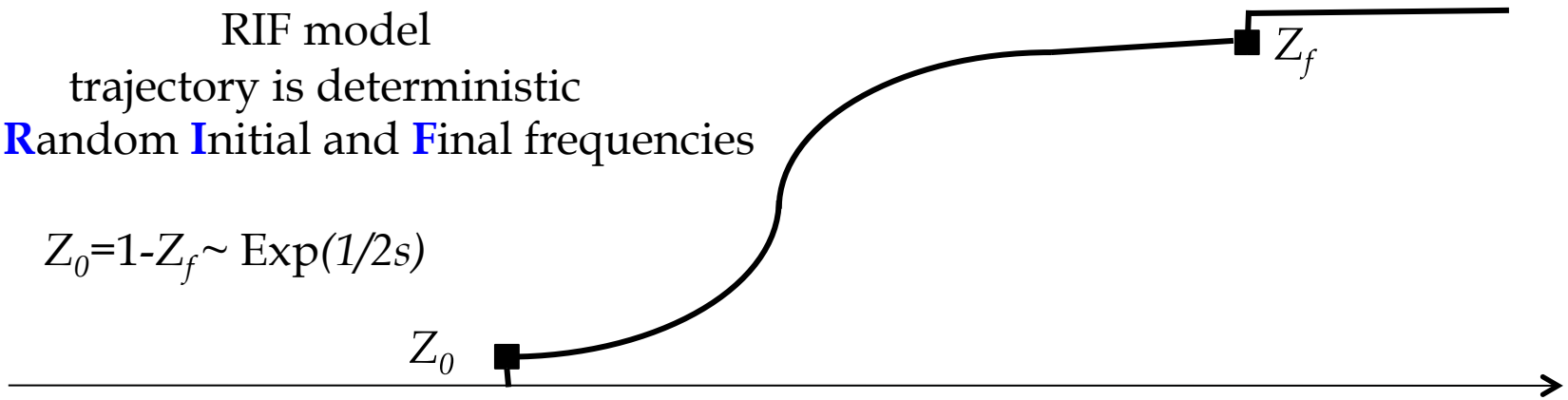
(Martin and Lambert, 2005; This study)

WF + selection
trajectory is stochastic



RIF model
trajectory is deterministic
with **R**andom **I**nitial and **F**inal frequencies

Z_0=1-Z_f~ Exp*(1/2s)*



The RIF approximation (1)

1) Start with a Wright Fisher diffusion with frequency Z_t

$$d\bar{Z}_t = \underbrace{s\bar{Z}_t(1-\bar{Z}_t)dt}_{\text{Logistic growth}} + \underbrace{\sqrt{\frac{1}{N}\bar{Z}_t(1-\bar{Z}_t)}dW_t.}_{\text{Diffusion term}}$$

2) Conditioned on fixation, it becomes

$$dZ_t = \underbrace{sZ_t(1-Z_t)\coth(NsZ_t)dt}_{\text{Gets larger for small } Z_t} + \sqrt{\frac{1}{N}Z_t(1-Z_t)}dW_t$$

3) For low frequency, (1) approximates well to a Feller diffusion

$$d\bar{Z}_t \approx \underbrace{s\bar{Z}_t dt}_{\text{Exp growth}} + \sqrt{\frac{1}{N}\bar{Z}_t}dW_t.$$

The RIF approximation (2)

4) Express time in 1/s units + condition on survival

$$dy_t \approx y_t \coth(y_t)dt + \sqrt{y_t}dw_t.$$

5) A supercritical branching process conditioned on survival

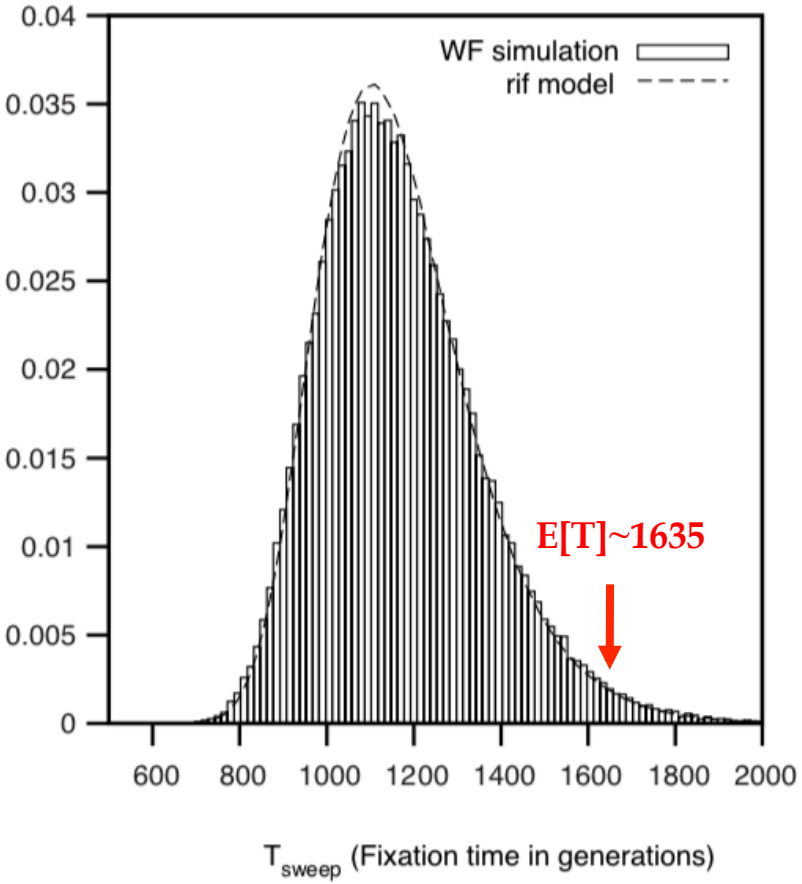
$$\lim_{t\to\infty} y_t e^{-t} = \frac{\mathcal{E}}{2} \text{ almost surely.}$$ (Yaglom's law)

6) Exponential growth with random starting frequency $\varepsilon/2Ns$

$$Z_t \approx \frac{\mathcal{E}}{2Ns} e^{st}, \text{ when } Z_t \ll 1.$$

7) Trajectory is entirely reversible, so it ends as it begins

Evaluation of the RIF approximation



N=10^5, s=0.01, 10^6 replicates

CPU time:

RIF model : 0.2 sec

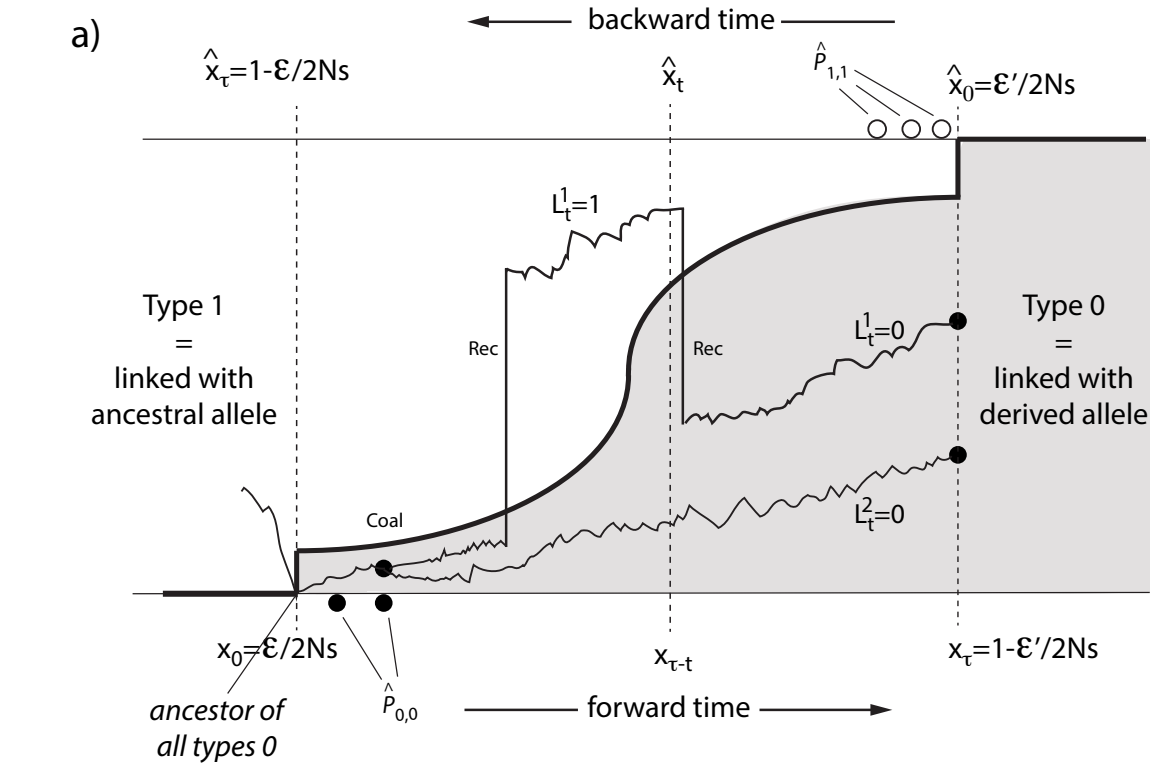
WF model: 38 min (10^4 slower)

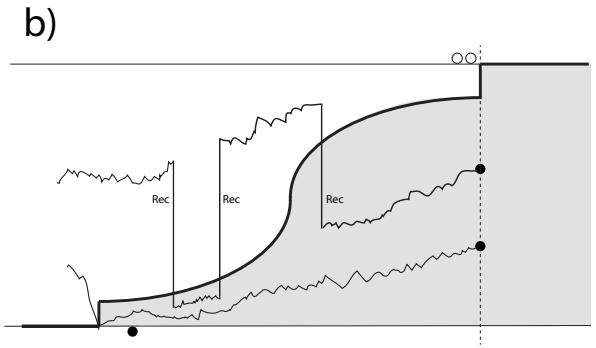
$$E[T_{sweep}] = 2(\ln(2Ns) + \gamma) / s$$

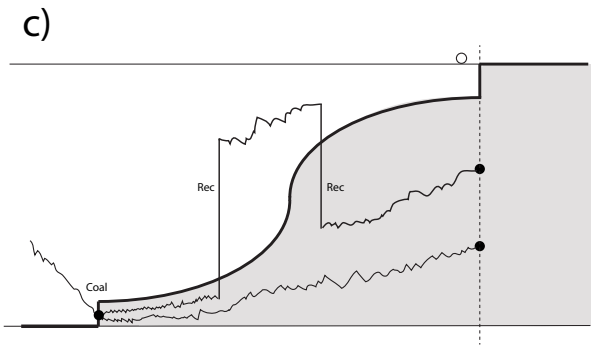
($\gamma \sim 0.5772$)

A efficient excellent approximation (see also Martin & Lambert, 2005)

Coalescent under the RIF model







Two new coalescent approximations

AS1 : Coalescent under the RIF model

$$\mathbb{P}(T_c < \tau) \approx (2Ns)^{-2A} \left(\frac{\pi A}{\sin(\pi A)}\right)^2 \left[2^{2A} \int_0^\infty \gamma(1-2A, \frac{2}{x}) e^{-x} dx \; + \; \int_0^\infty x^{2A} e^{-x-\frac{2}{x}} dx\right]$$

Pop. size | Two types 0 | Coal before Z_0 | Coal at Z_0

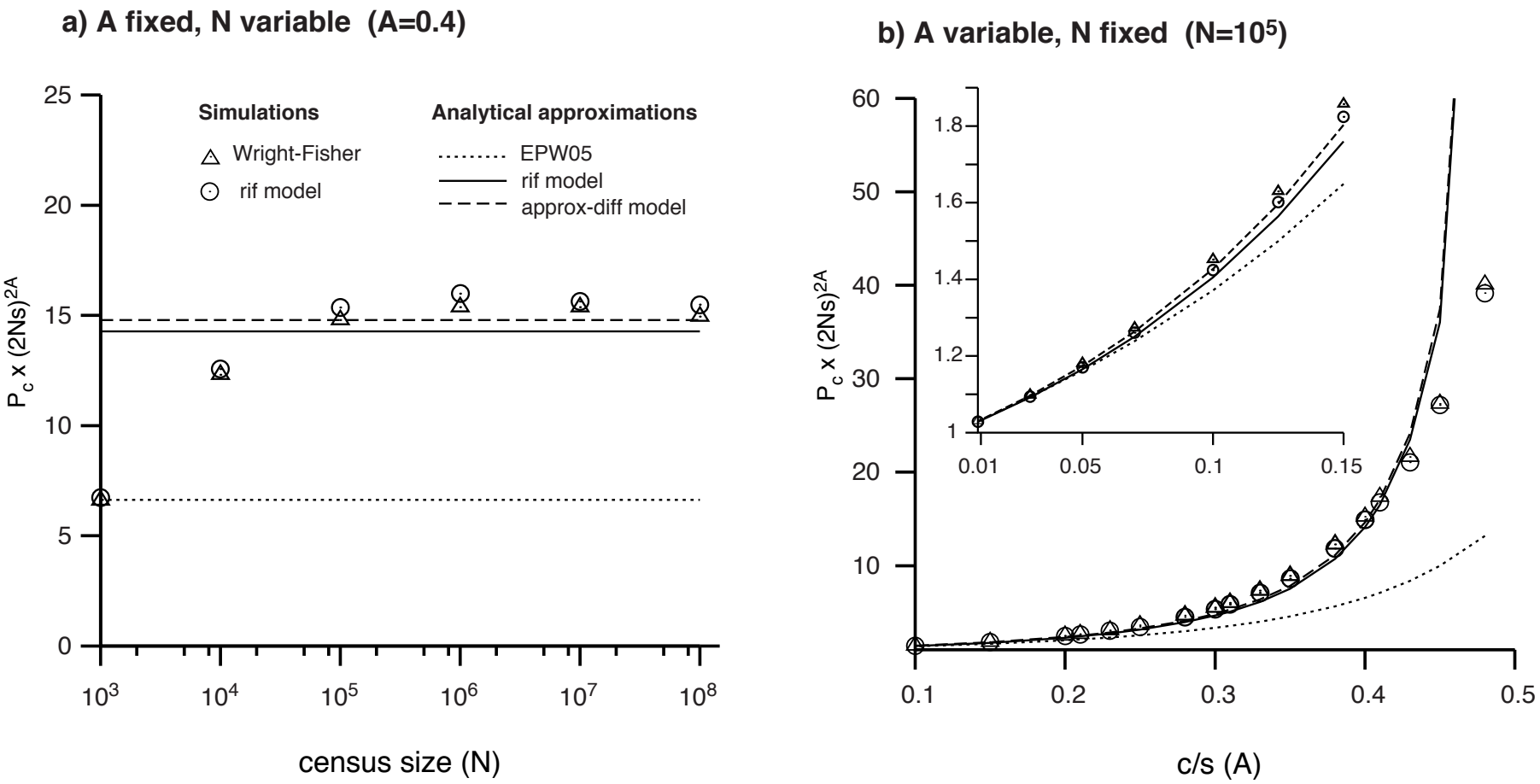
AS2 : Coalescent using another more elaborate diffusion approximation

$$\mathbb{P}(T_c < \tau) \approx (2Ns)^{-2A} \left(\frac{\pi A}{\sin(\pi A)}\right)^2 \frac{\Gamma(2(A+1))}{1-2A}.$$

Both approximations scale with $(Ns)^{-2A}$

Two new coalescent approximations

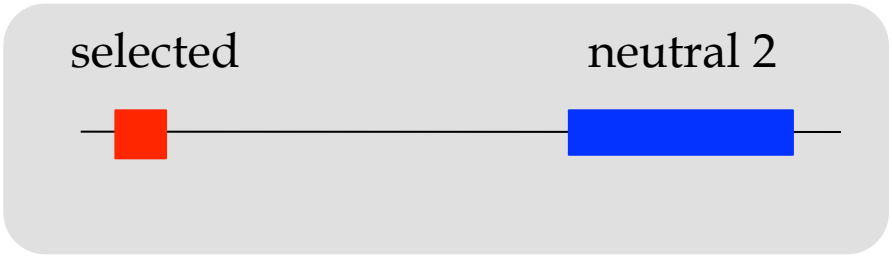
EPW = Etheridge Pfaffelhuber and Wakolbinger, 2006

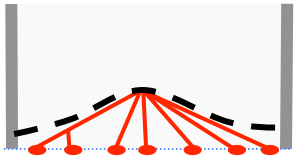


The new approximations are good for large N and A<0.45

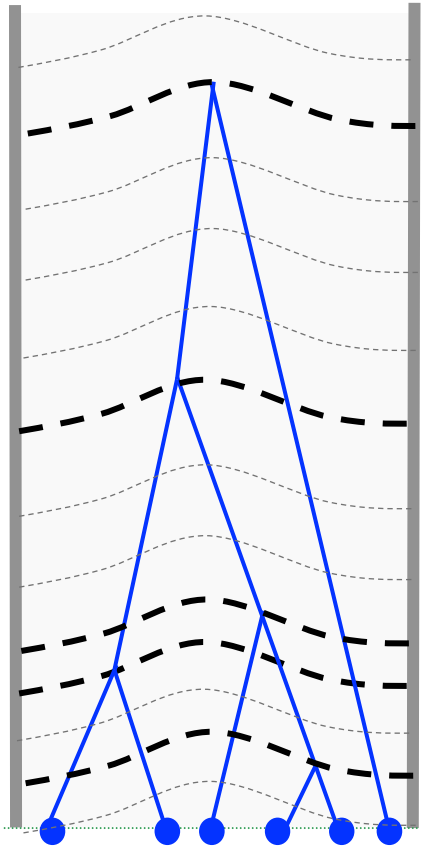
The genetic draft

(Gillespie 2000a, 2000b; Neher & Shraiman, 2011; Coop & Ralph, 2012 ; …)





radiation

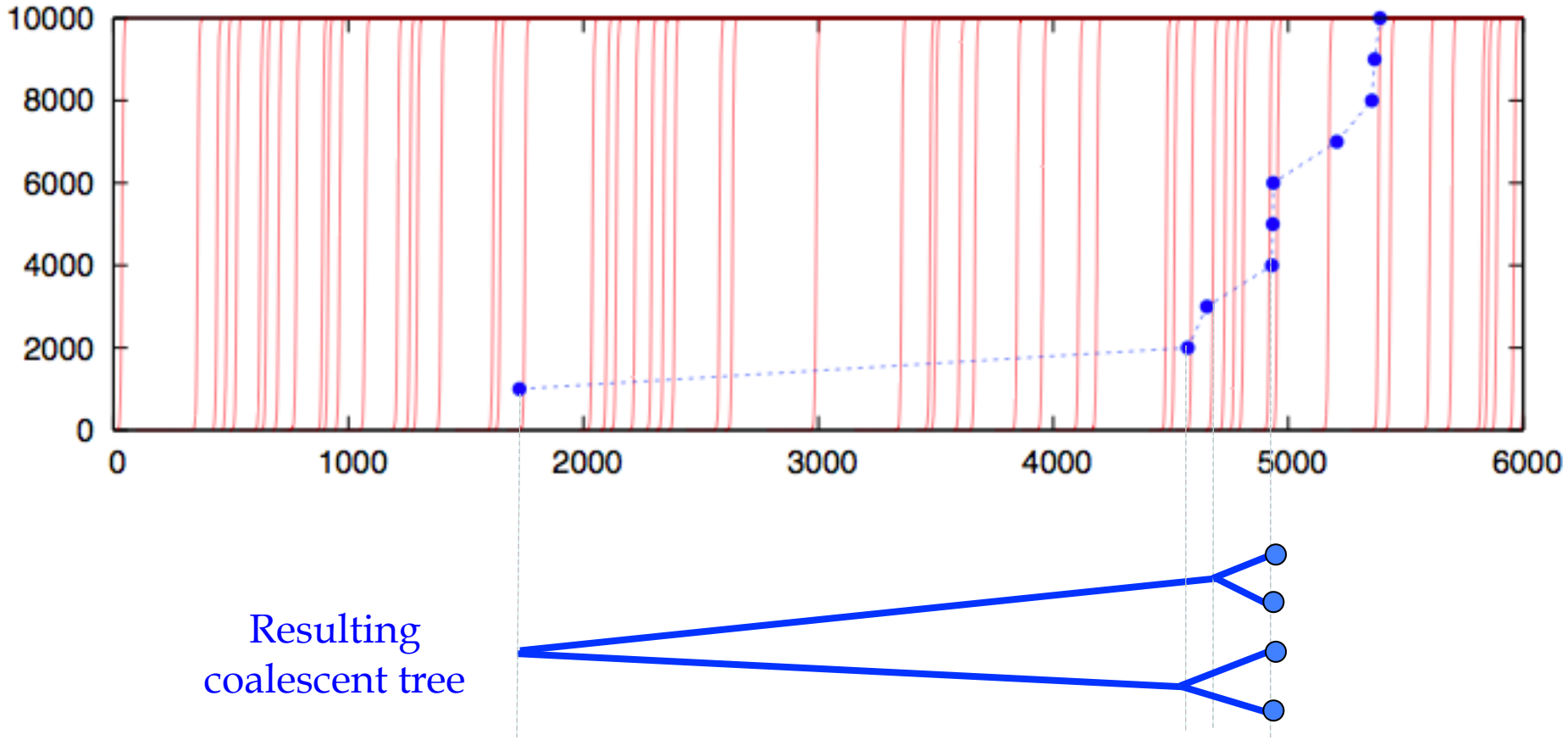


Sweep's coalescent

Selection tunes neutral diversity through genetic linkage

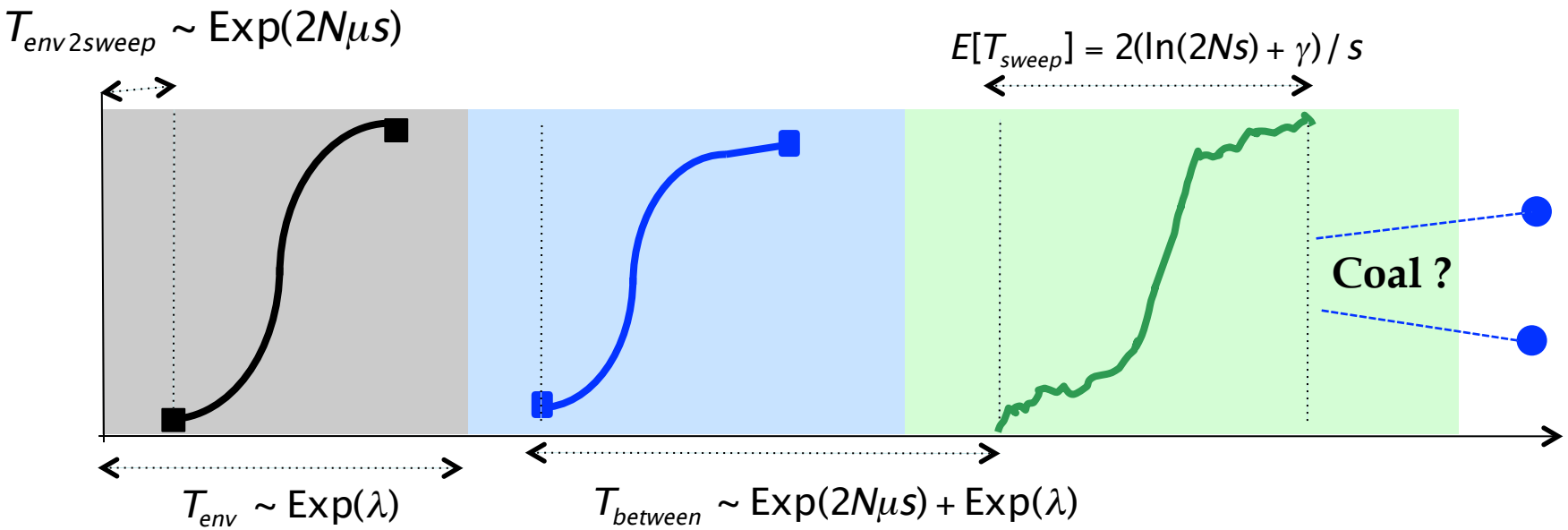
Visual of one realization

(Individual based simulation: $N=10^4$; $s=0.025$, $N\mu=0.025$, $c=0.005$)



$E[T_2] \sim 3{,}500$
(10,000 with drift)

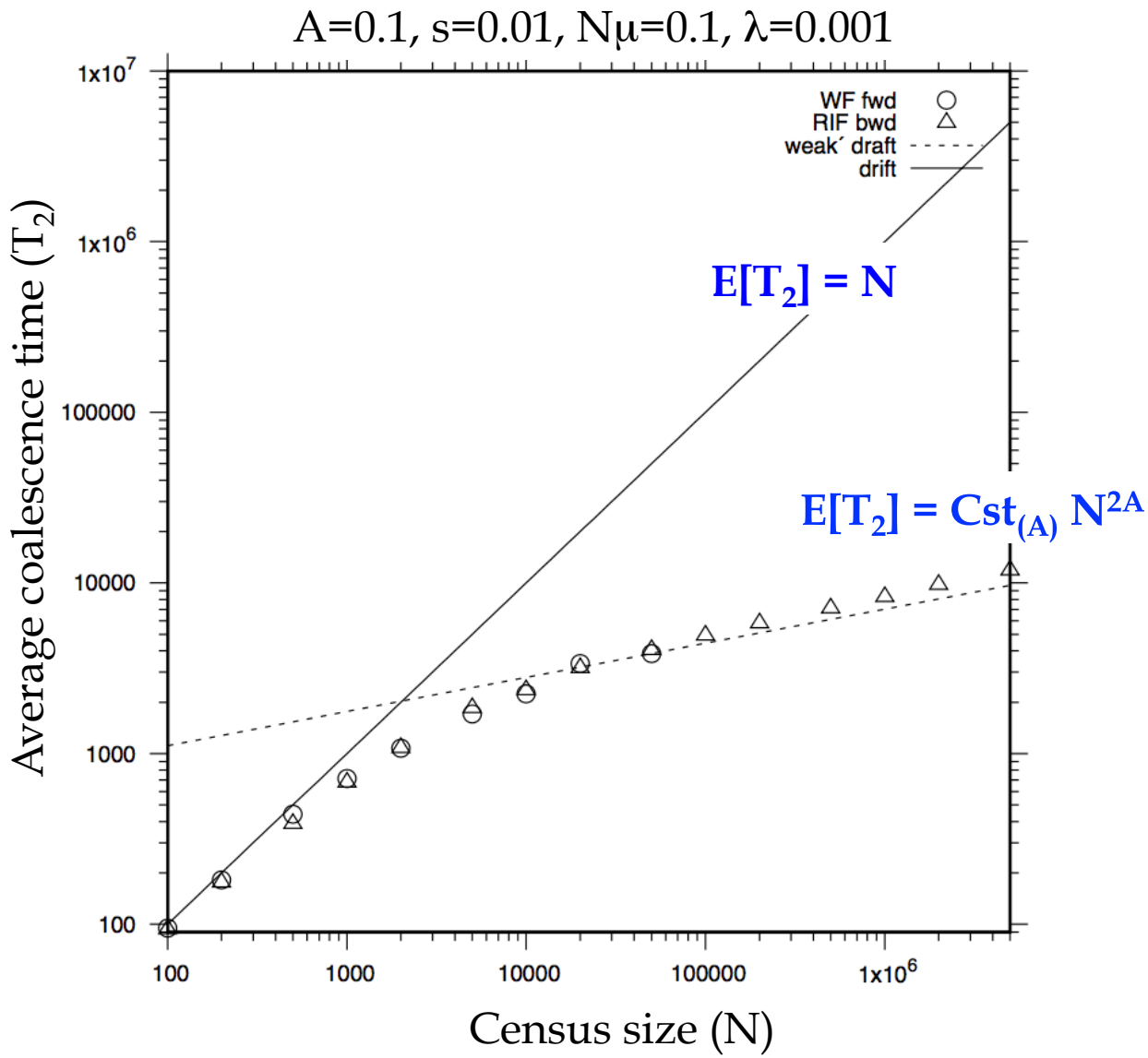
Ever changing environment



$$P_{coal} \underset{N \to \infty}{\to} K_A / (2Ns)^{2A} \qquad T_2 \sim \mathrm{Geometric}(P_{coal})$$

Coalescence corresponds to changes of environment

Diversity as a power law

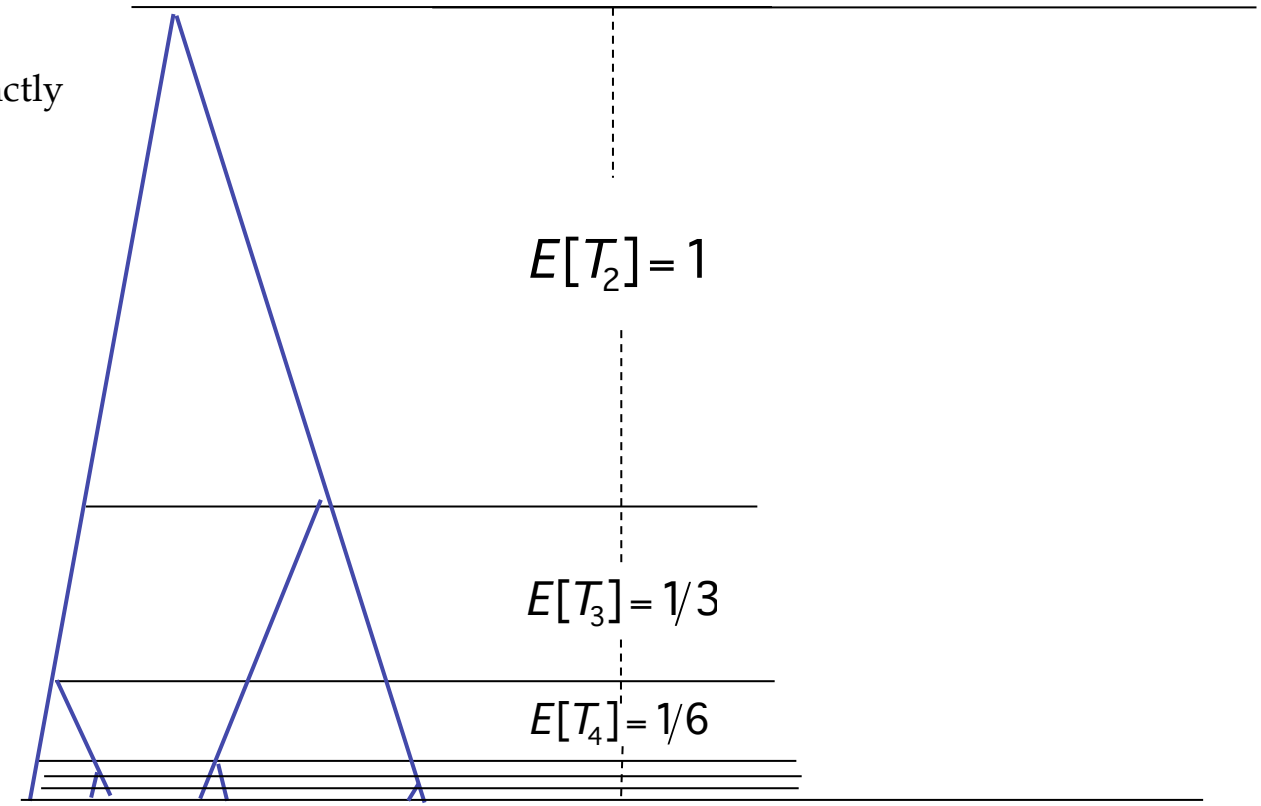


Kingman coalescent tree

T_k = time while there are exactly k lineages.

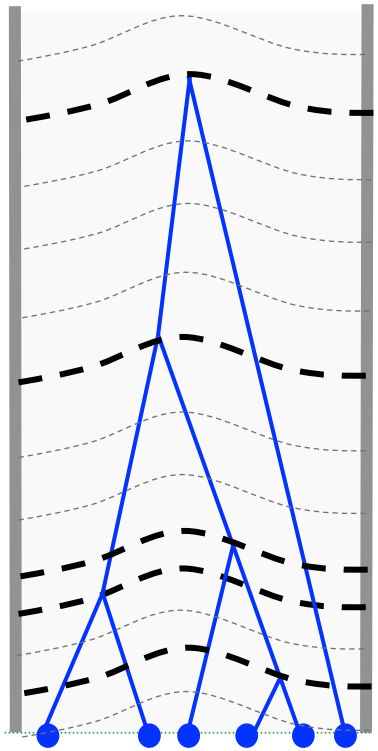
$$f(T_k) = \lambda_k e^{-\lambda_k T}$$

$$\lambda_k = \frac{k(k-1)}{2}$$

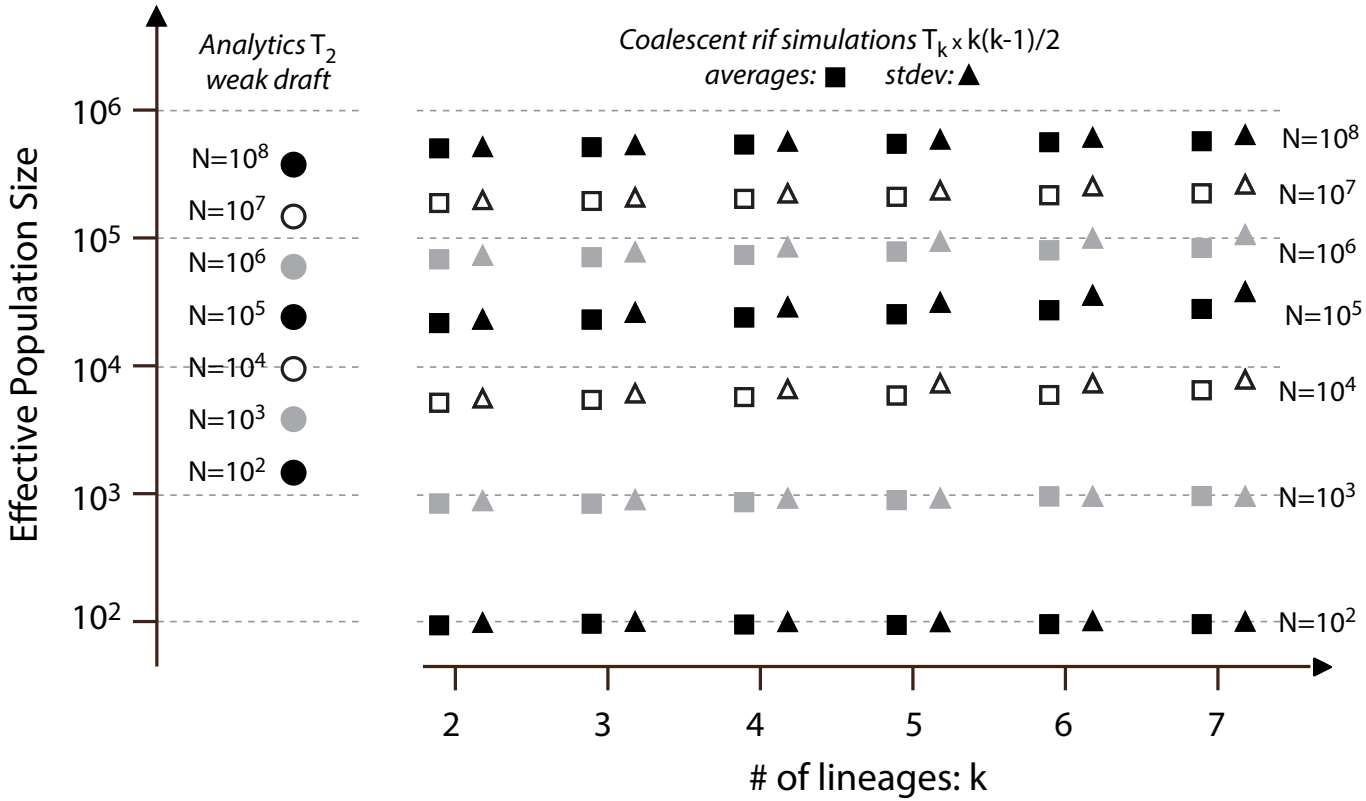


For each step T_k x k(k-1)/2 has constant mean and stdev.

Convergence to a Kingman coalescent

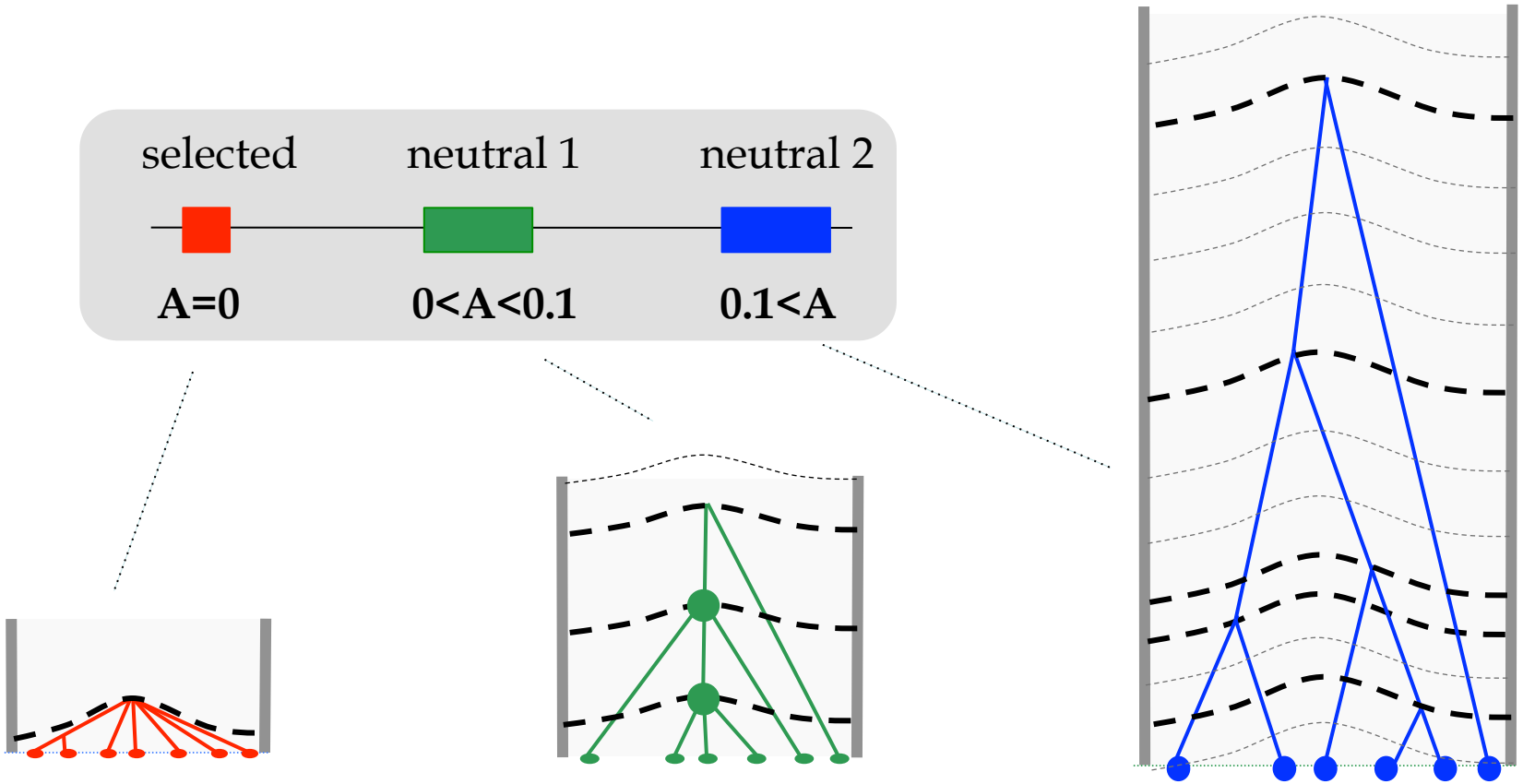


Weak draft coalescent
(A>0.1 ; P_{coal}<<1)



Weak draft mimics closely neutrality

‘A’ modulates the intensity of draft



radiation

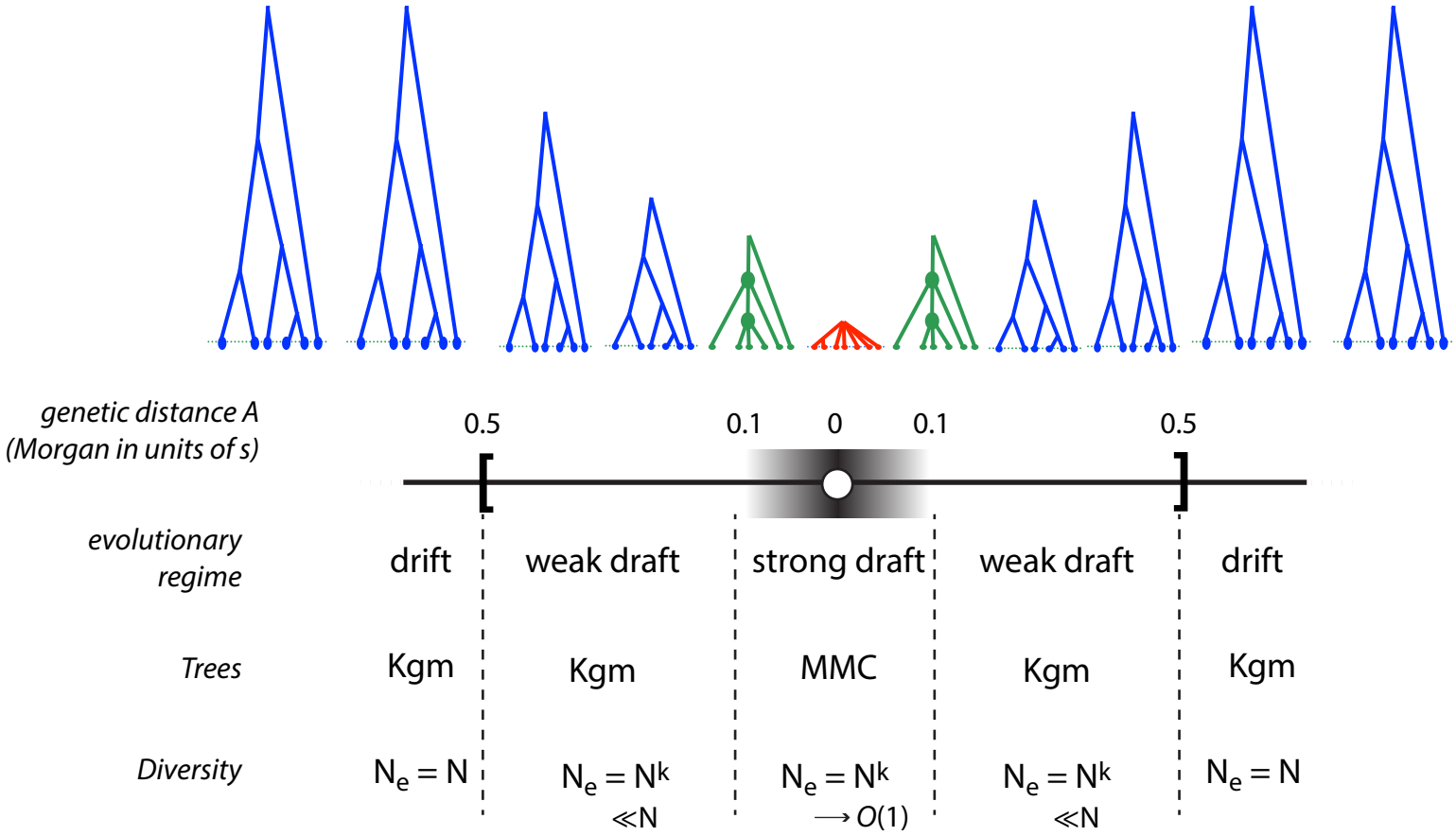
Multiple Merger Coalescent

In cod-fish (Arnasson et al., 2023)
In manyspecies (Freund et al., 2023)

Kingman coalescent

A<0.5 : weak draft
A>0.5: drift

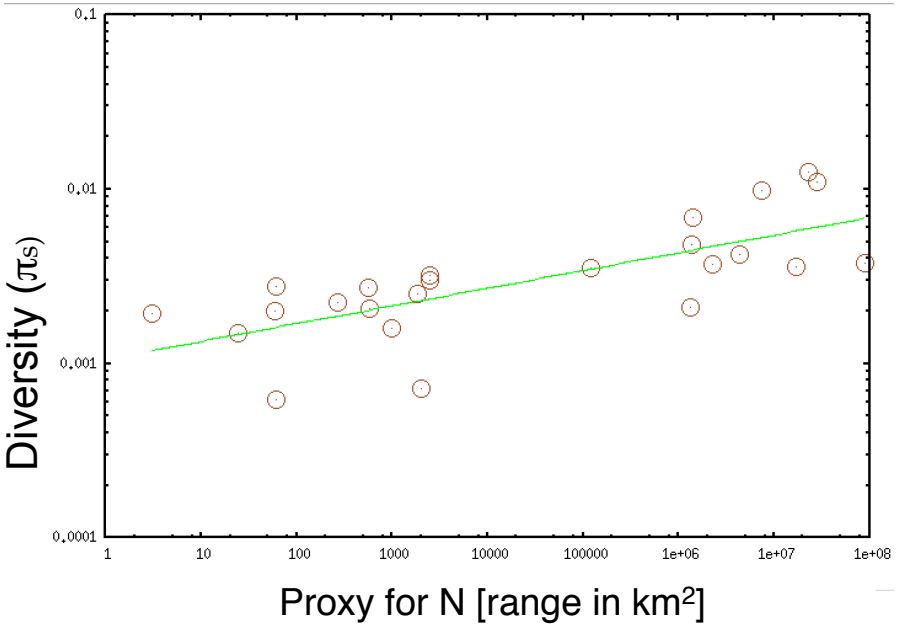
Genomic pattern



Average pattern should be 'almost' neutral, with a reduced N_e and a zest of MMC:
Genome-wide SFS support MMC across the Tree of Life (*Arnasson et al., 2023 ; Freund et al., 2023*)

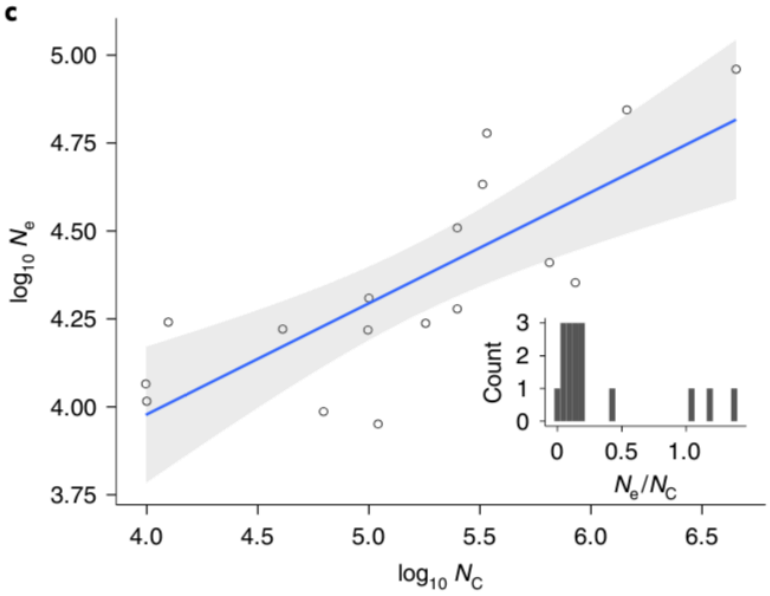
What data say?

Island birds



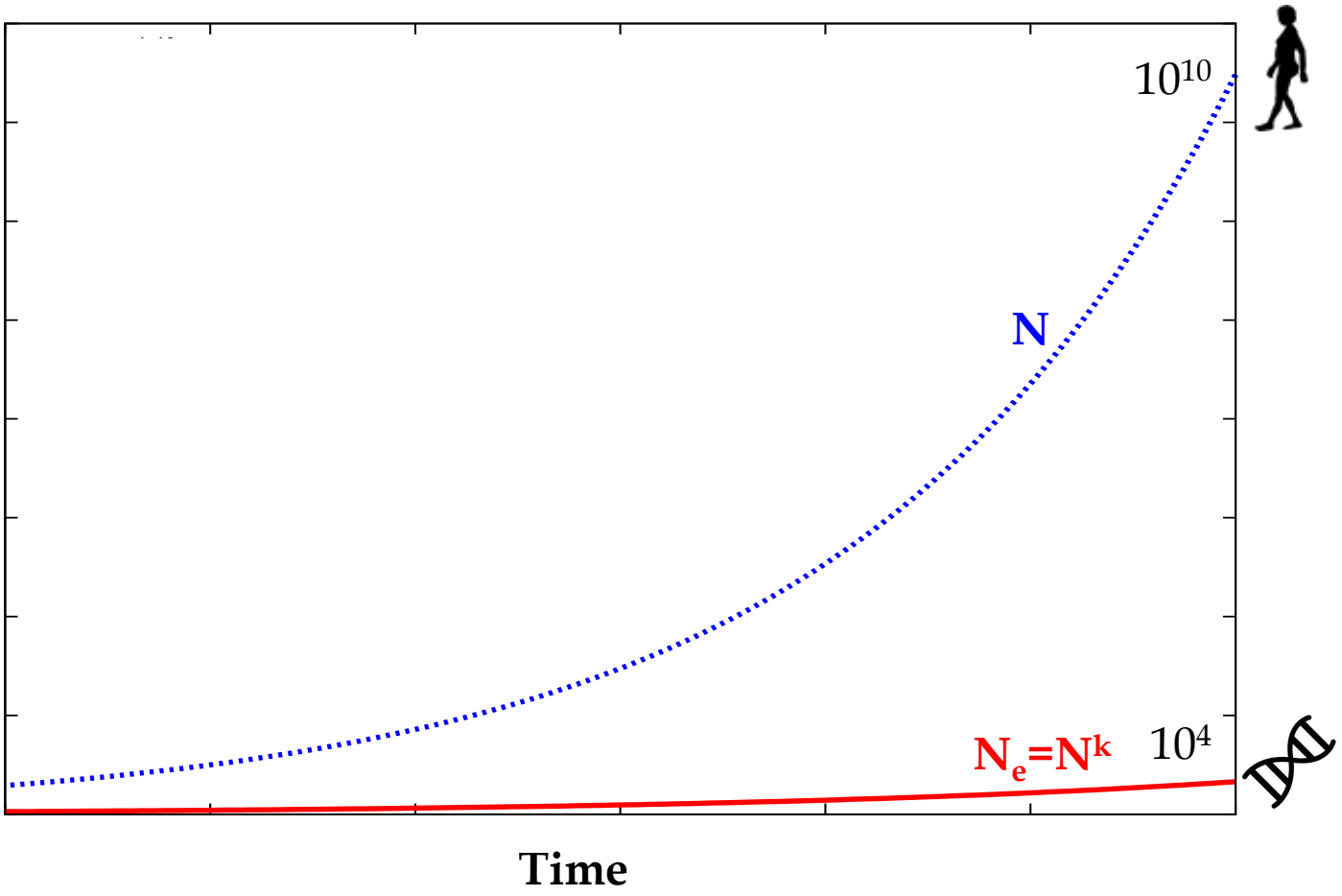
(unpublished, courtesy of B Nabholz (ISEM))

Pinipedes



(from Peart et al., Nat Ecol Evol, 2020)

Population genetics and demography?



"Large" variation in N <=> "moderate" variation in N_e

About this work

On the methodological side

> The N_e riddle can be recast as a time scale problem

> RIF model can be used to analyze and simulate very efficiently selection in finite population

On the biological side

> H0 systematically underpredicts the amount of polymorphism

> Selection

- alter diversity through linkage
- effect depends on distance to selected site
- at medium distance : MMC with $N_e = N^k$
- at far distance: Kingman with $N_e = N^k$

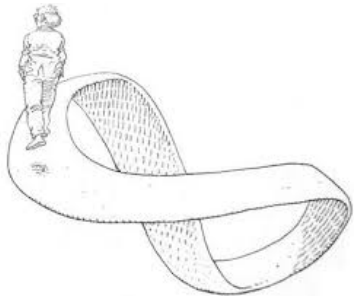
Extensions, open questions

Can the selection be less of a caricature in real cases?

What happens when:

- Ns is on the order of 1… or even less (mild sweep)
- Nμ is very large (multiple origins of the benefical allele)
- Multiple sweeps occur in the genome (Hill-Robertson effect)
- The beneficial alleles pre-exist (standing variation)
- The trait under selection is polygenic (several contributing loci)
- … any other suggestion is most welcome!

More generally



Observations

Sequences do change
Homologous loci show diversity

The (unknown) cause of Molecular Evolution

Neutral theory

=> diversity scales with N
=> N_e is inferred using diversity (tautology!)

Linked selection

=> Do we see light at the end of the tunnel?