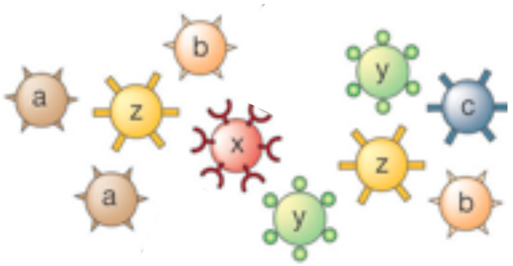
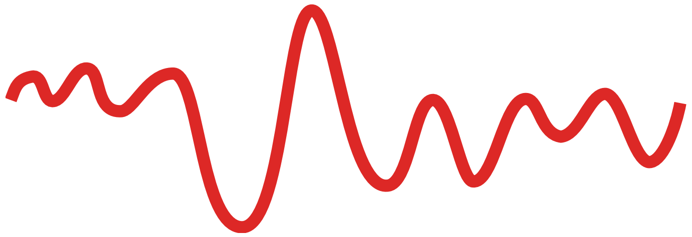
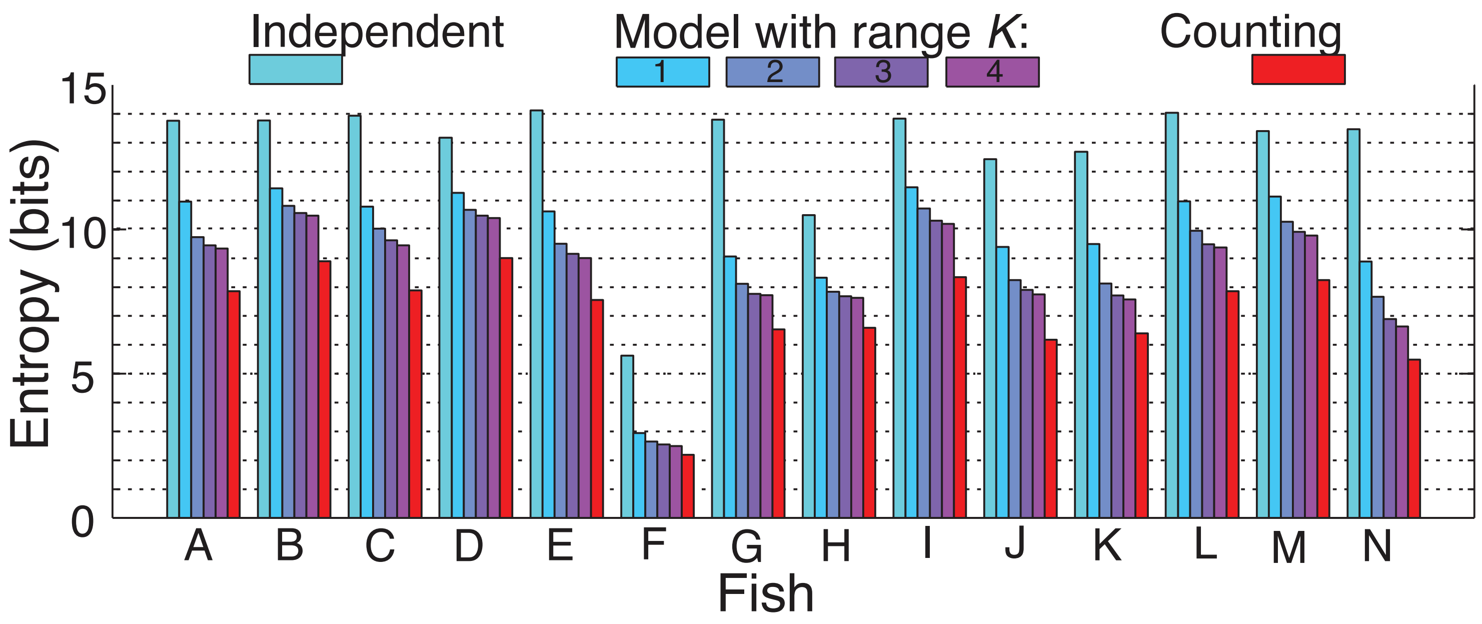
Statistically characterizing antibody diversity

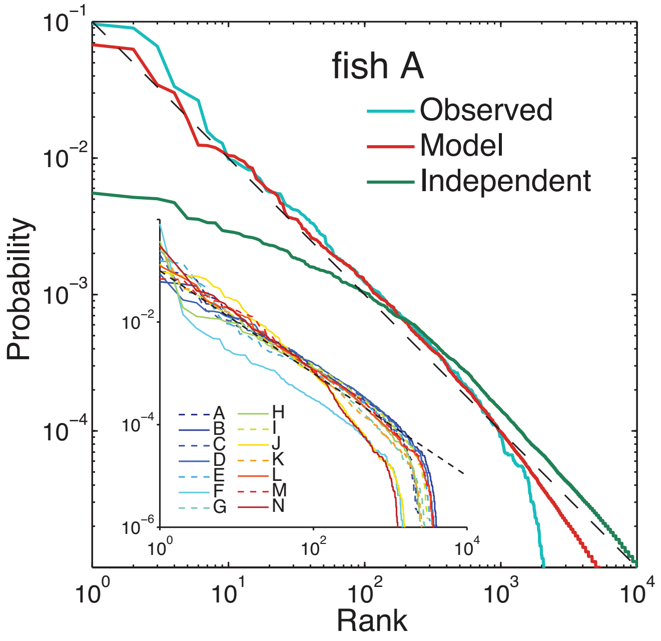
work with Thierry Mora, William Bialek, Curt Callan

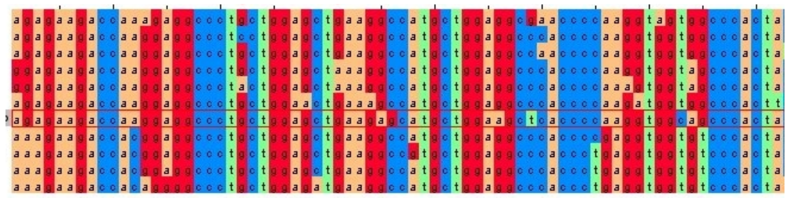


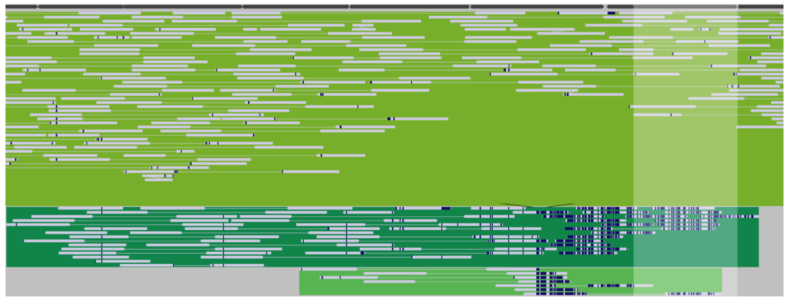
$$\xrightarrow{p^m(\sigma_1,..\sigma_L)}$$

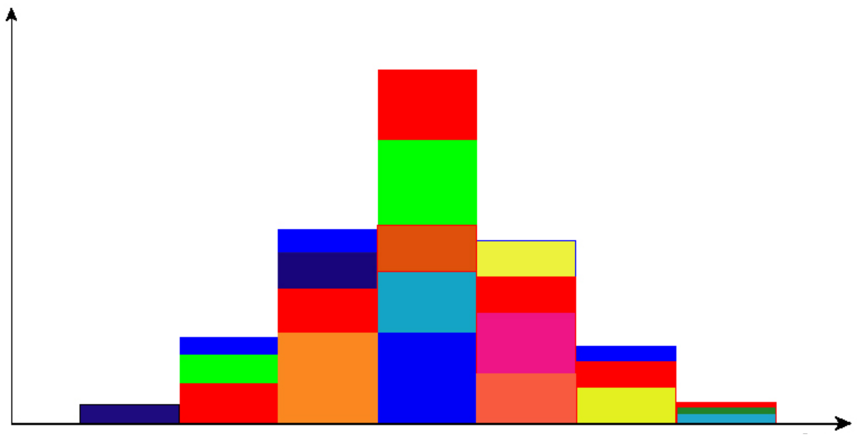










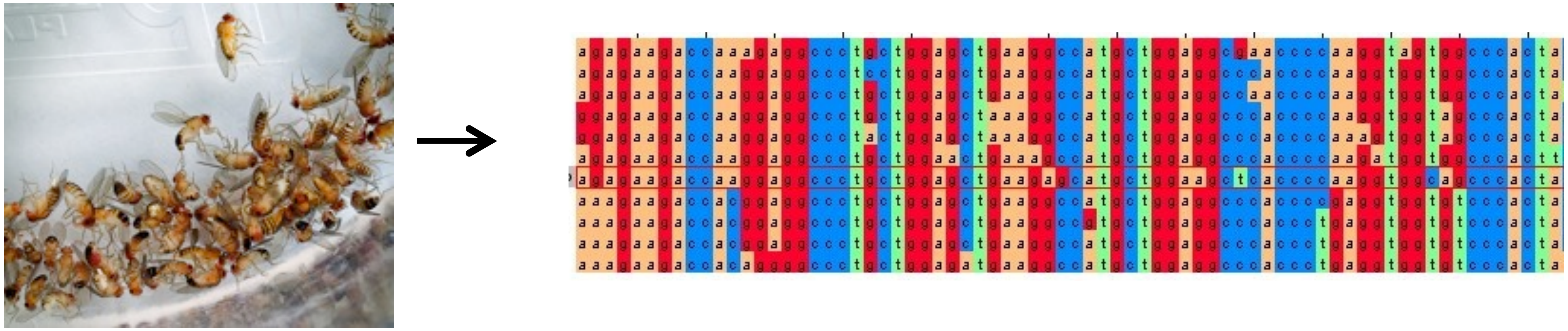


The effects of negative selection on the evolution of linked sites

Aleksandra M Walczak
Laboratoire de Physique Théorique - ENS, CNRS

Michael Desai, Harvard
Joshua Plotkin, University of Pennsylvania
Lauren Nicolaisen, Harvard

Inferring evolutionary processes



Understanding evolutionary processes:

- test consistency of data with null models
- currently: easy to use neutral or weak selection models
- disagreement: selection, demography, geography ...

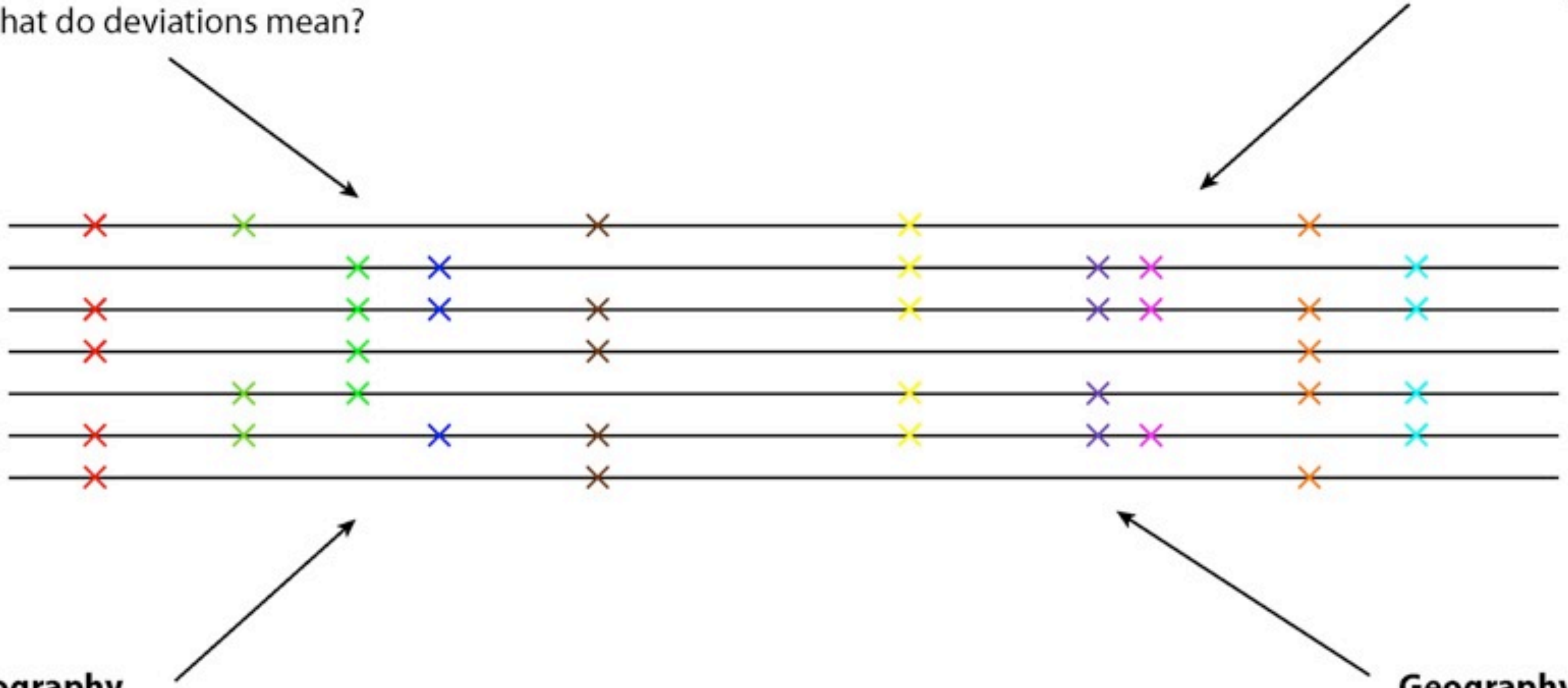
Goal: develop null models with selection

- test consistency of data with null models with selection
- rule out models also when neutrality does not apply
- infer selective parameters from data

Evolutionary scenarios

Genetic Drift
Well understood
But what do deviations mean?

Natural Selection
Reduces diversity

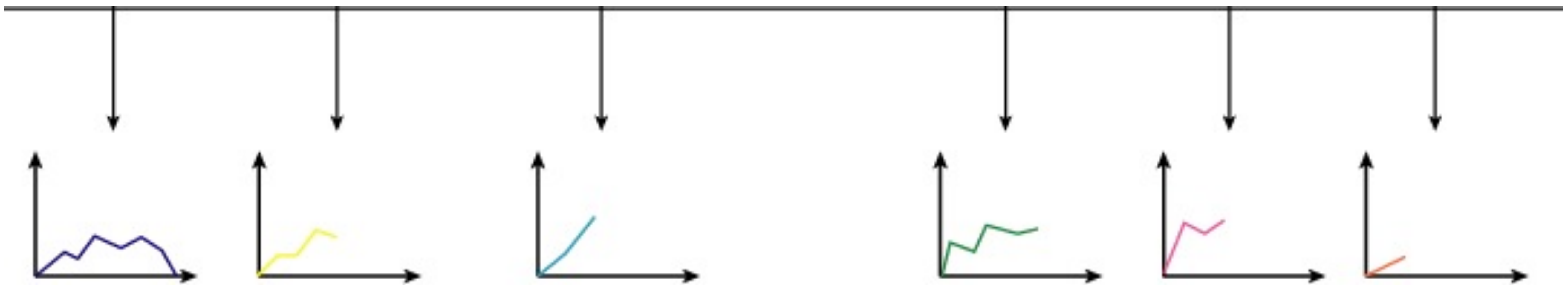


Demography
Bottlenecks, expansions reduce diversity

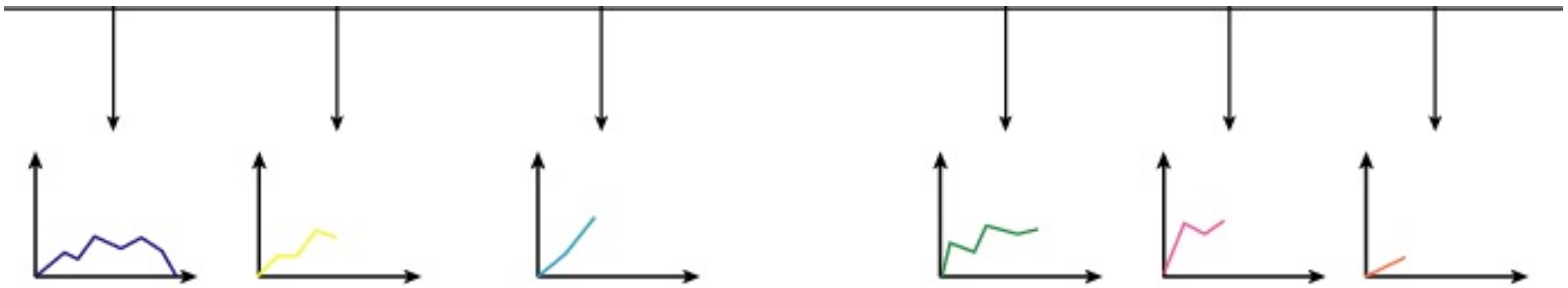
Geography
Environmental structure increases diversity

What should we look at? What do we expect?

Model the fate of each site in the genome

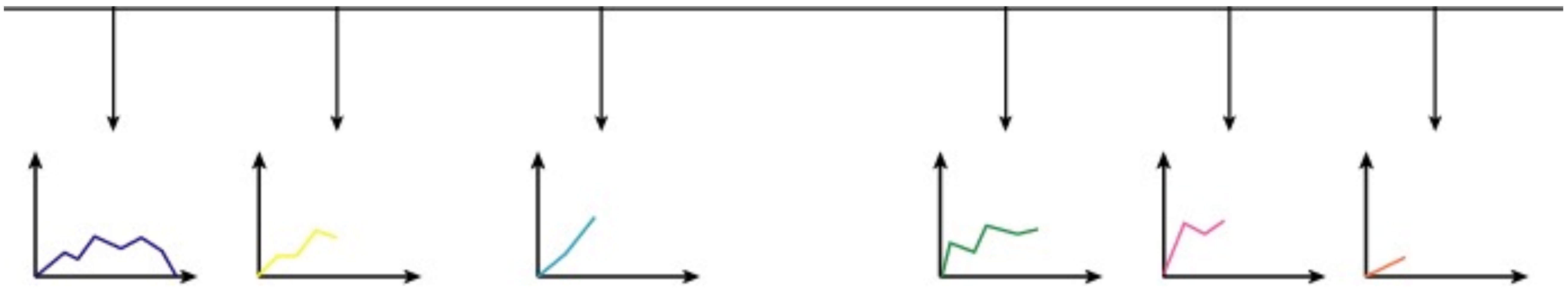


Model the fate of each site in the genome



Calculate the fate of each mutant forward in time.

Model the fate of each site in the genome

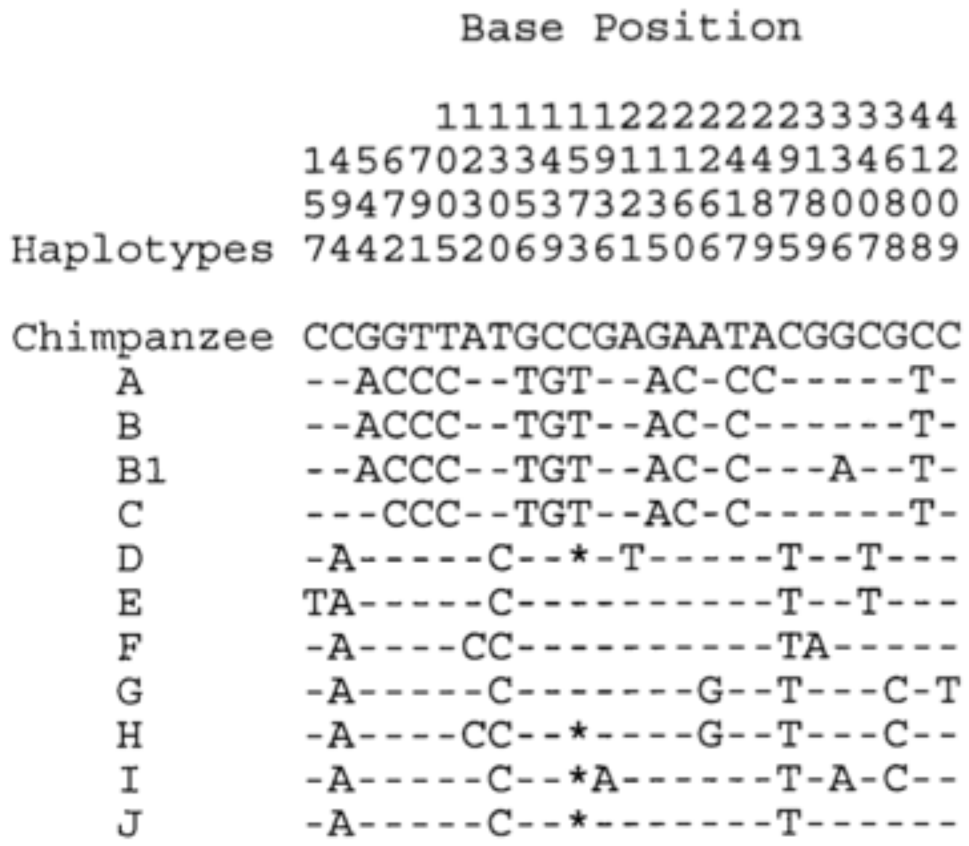


Calculate the fate of each mutant forward in time.

Fate of each mutation is not in steady state

But there is a steady state distribution of the distribution of mutant frequencies

Nearby mutations are not independent



```
                      Base Position

                  1111111222222233344
            1456702334591112449134612
            5947903053732366187800800
Haplotypes  7442152069361506795967889

Chimpanzee  CCGGTTATGCCGAGAATACGGCGCC
    A       --ACCC--TGT--AC-CC-----T-
    B       --ACCC--TGT--AC-C------T-
    B1      --ACCC--TGT--AC-C---A--T-
    C       ---CCC--TGT--AC-C------T-
    D       -A-----C--*-T-----T--T---
    E       TA-----C----------T--T---
    F       -A----CC----------TA-----
    G       -A-----C-------G--T---C-T
    H       -A----CC--*----G--T---C--
    I       -A-----C--*A------T-A-C--
    J       -A-----C--*-------T------
```

[Harris and Hey 1999]

Strong correlations between mutations.

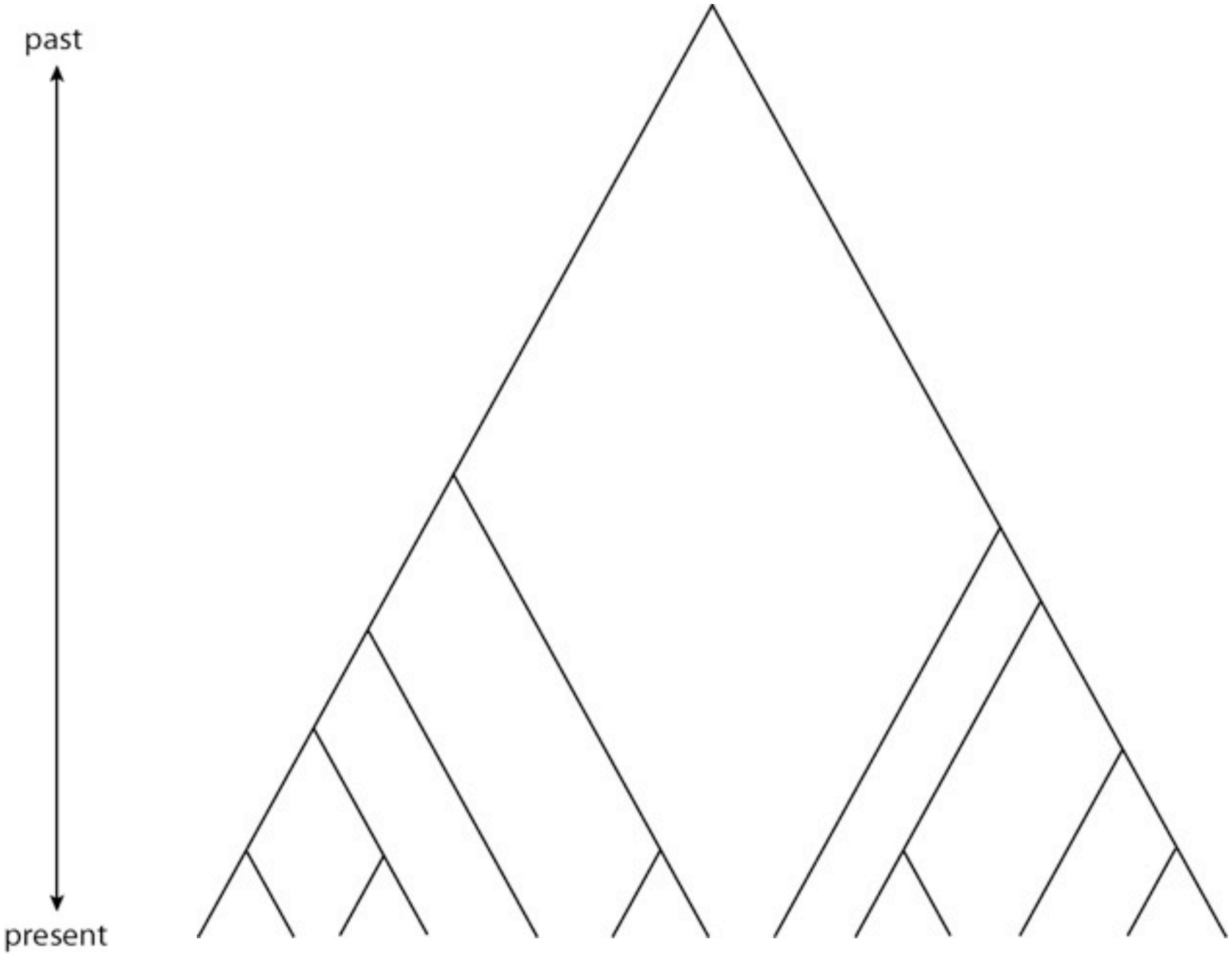
Mutations are physically *linked.*

Recombination breaks linkage.

No recombination - fully linked sites

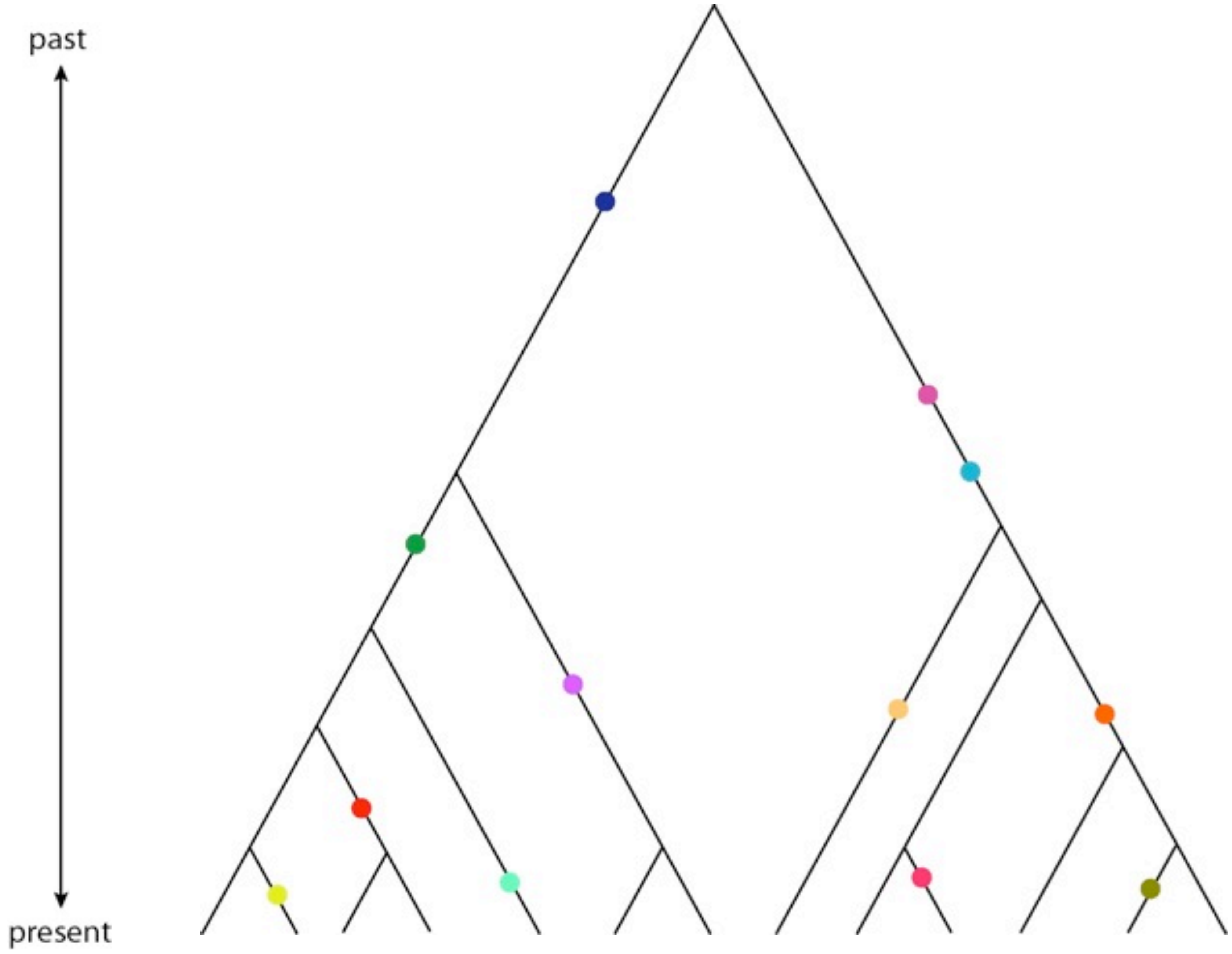
No selection: *Coalescent Theory*

No selection: *Coalescent Theory*



The whole sequence shares a common genealogy.

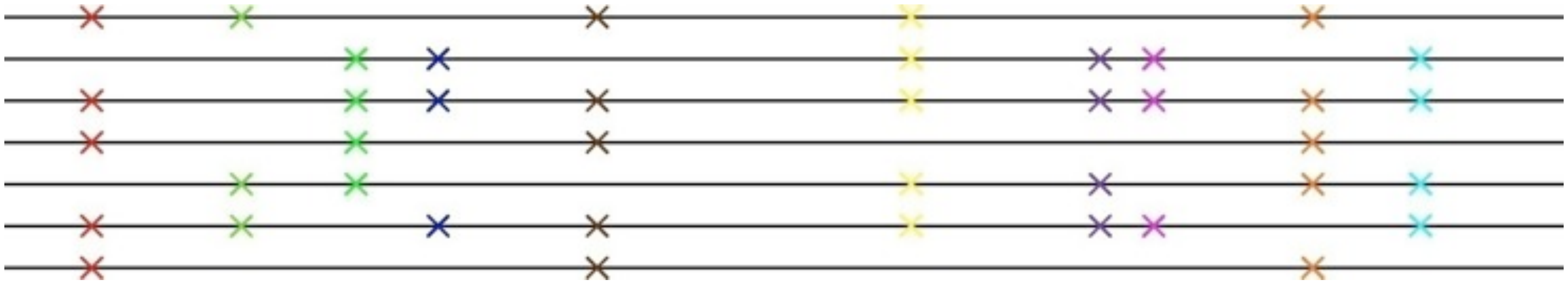
No selection: *Coalescent Theory*



The whole sequence shares a common genealogy.

Cannot easily handle selection, despite 20 years of effort.

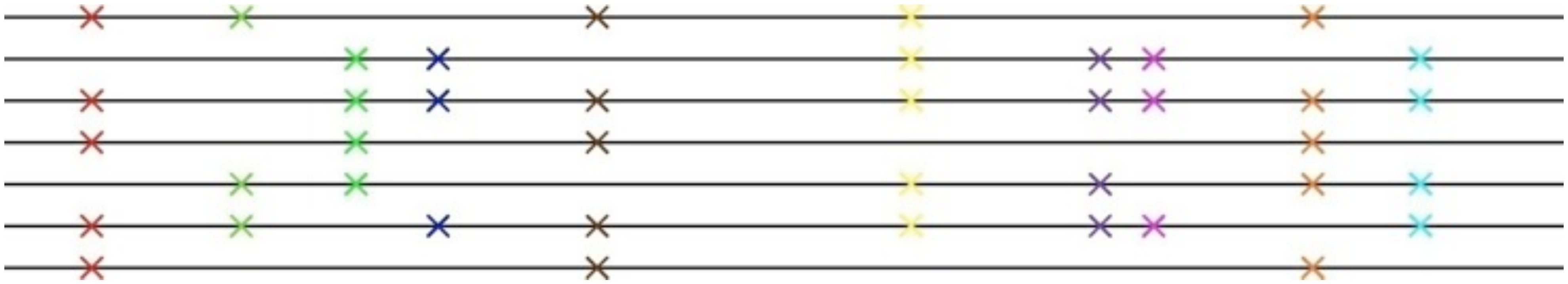
Comparison to the neutral null model



Is this data consistent with neutral well-mixed random-mating population?

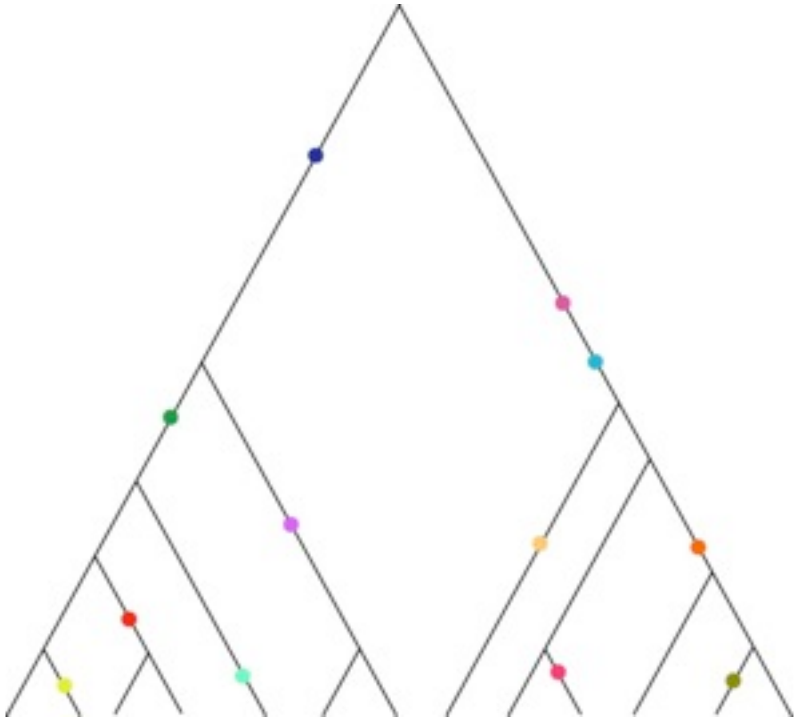
What can we infer about the evolutionary history of this population?

Comparison to the neutral null model



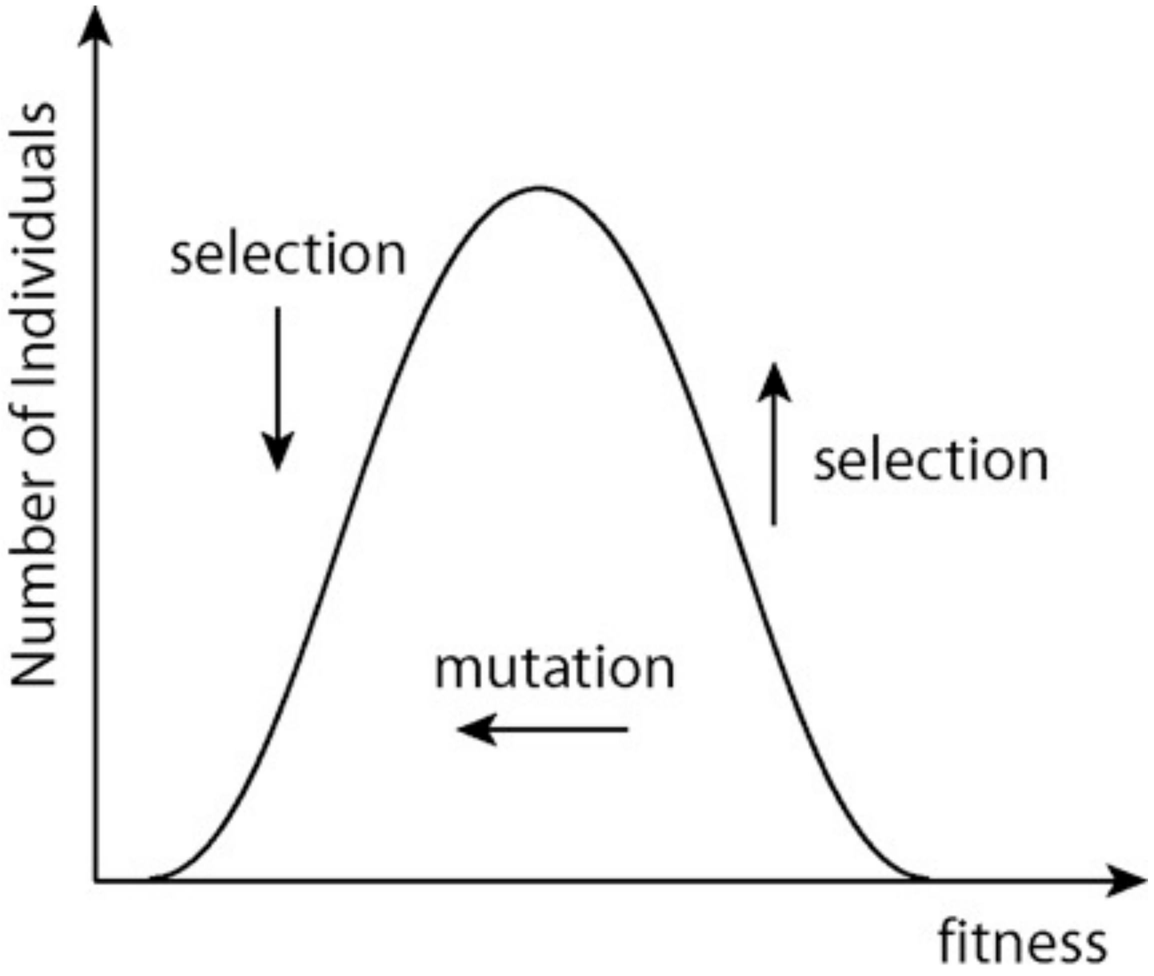
Is this data consistent with neutral well-mixed random-mating population?

What can we infer about the evolutionary history of this population?

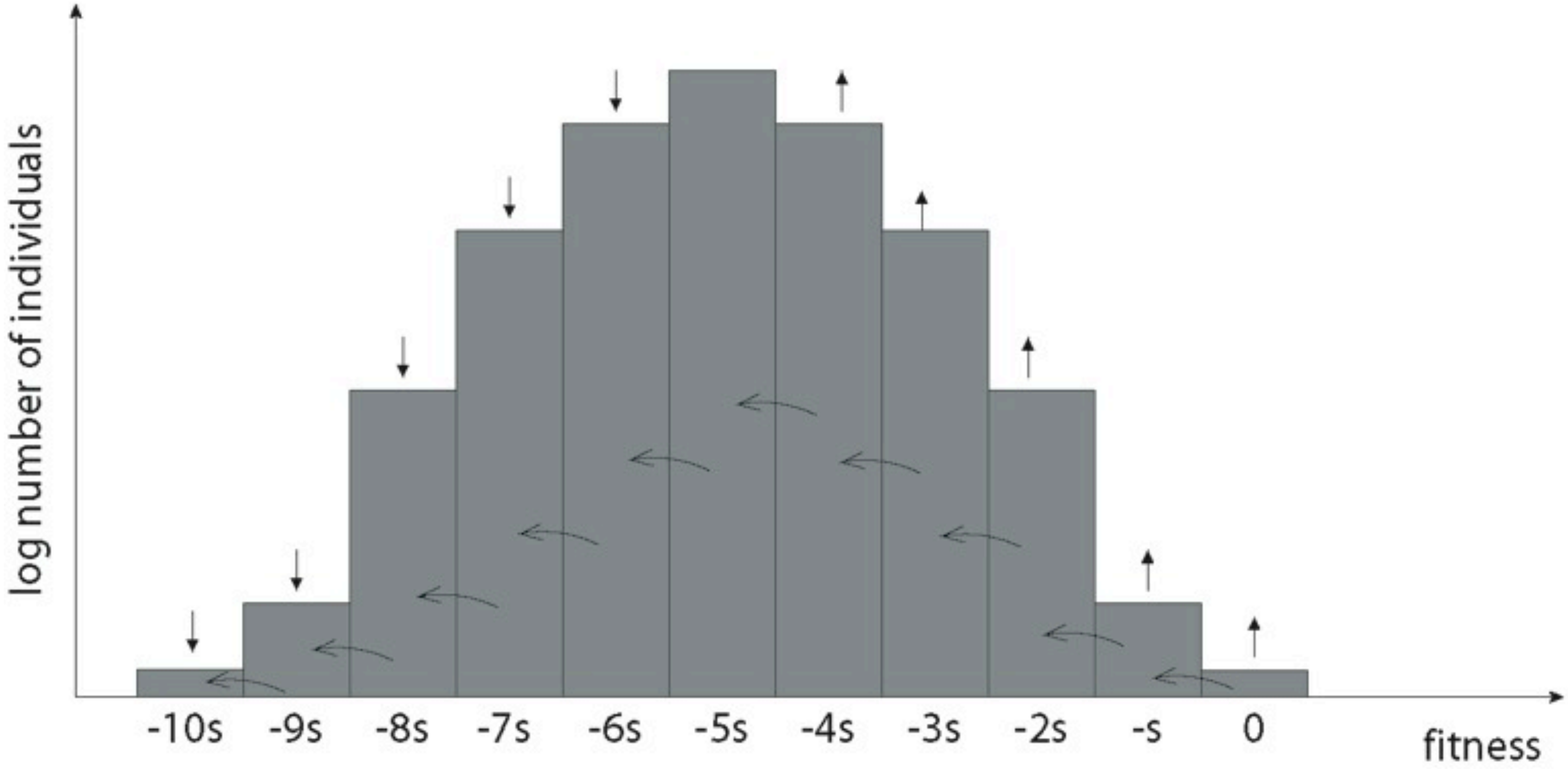


Are the coalescent trees that lead to some aspect of the observed diversity likely?

Evolution of the fitness distribution

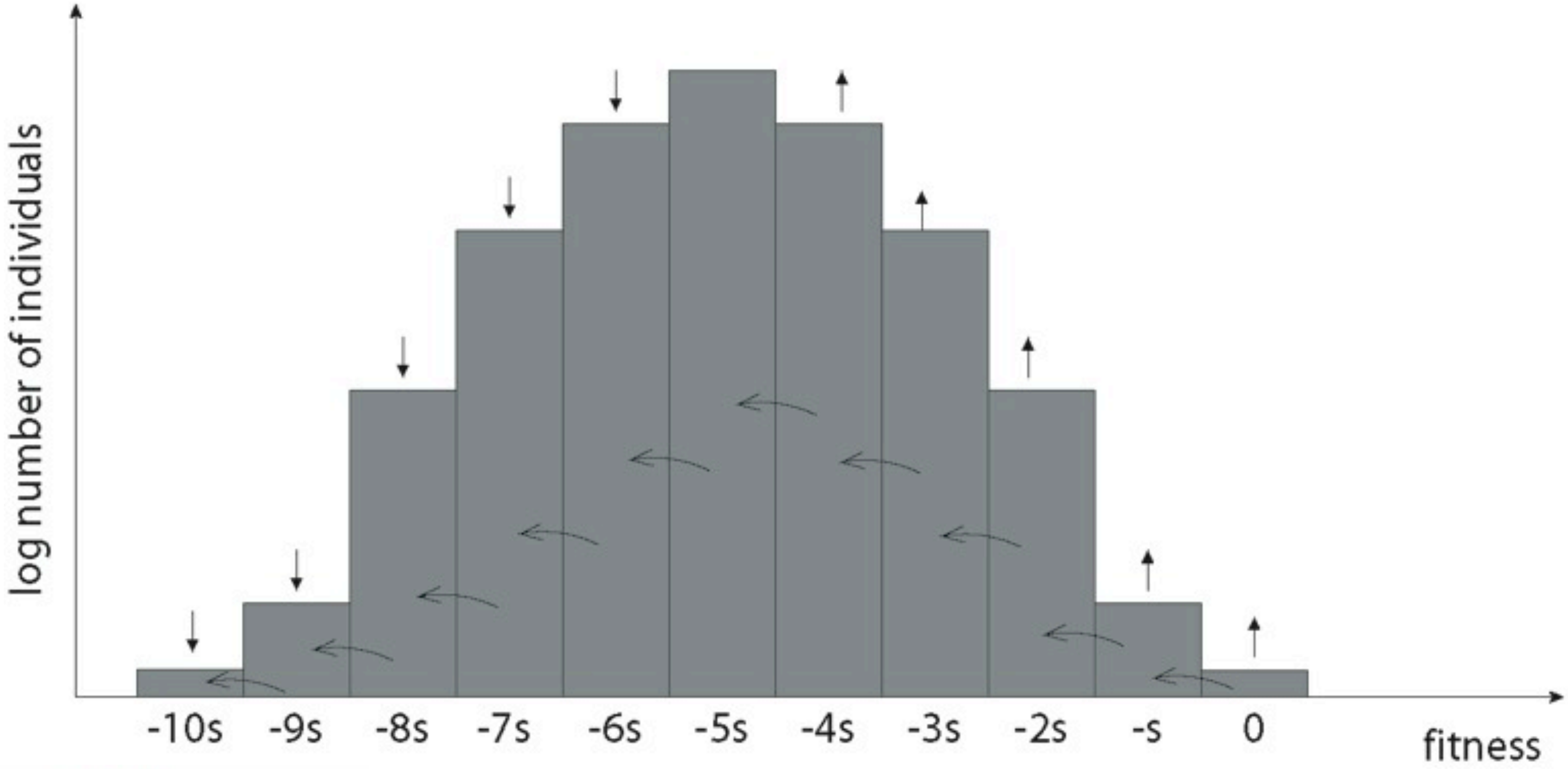


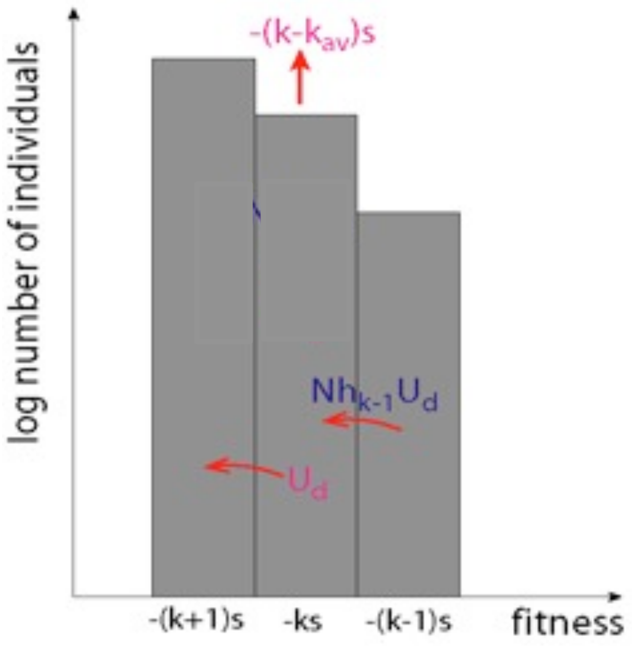
Evolution of the fitness distribution



Balance between mutations and selection in each class:
Deterministic steady state fitness distribution.

Evolution of the fitness distribution





Balance between mutations and selection in each class: Deterministic steady state fitness distribution.

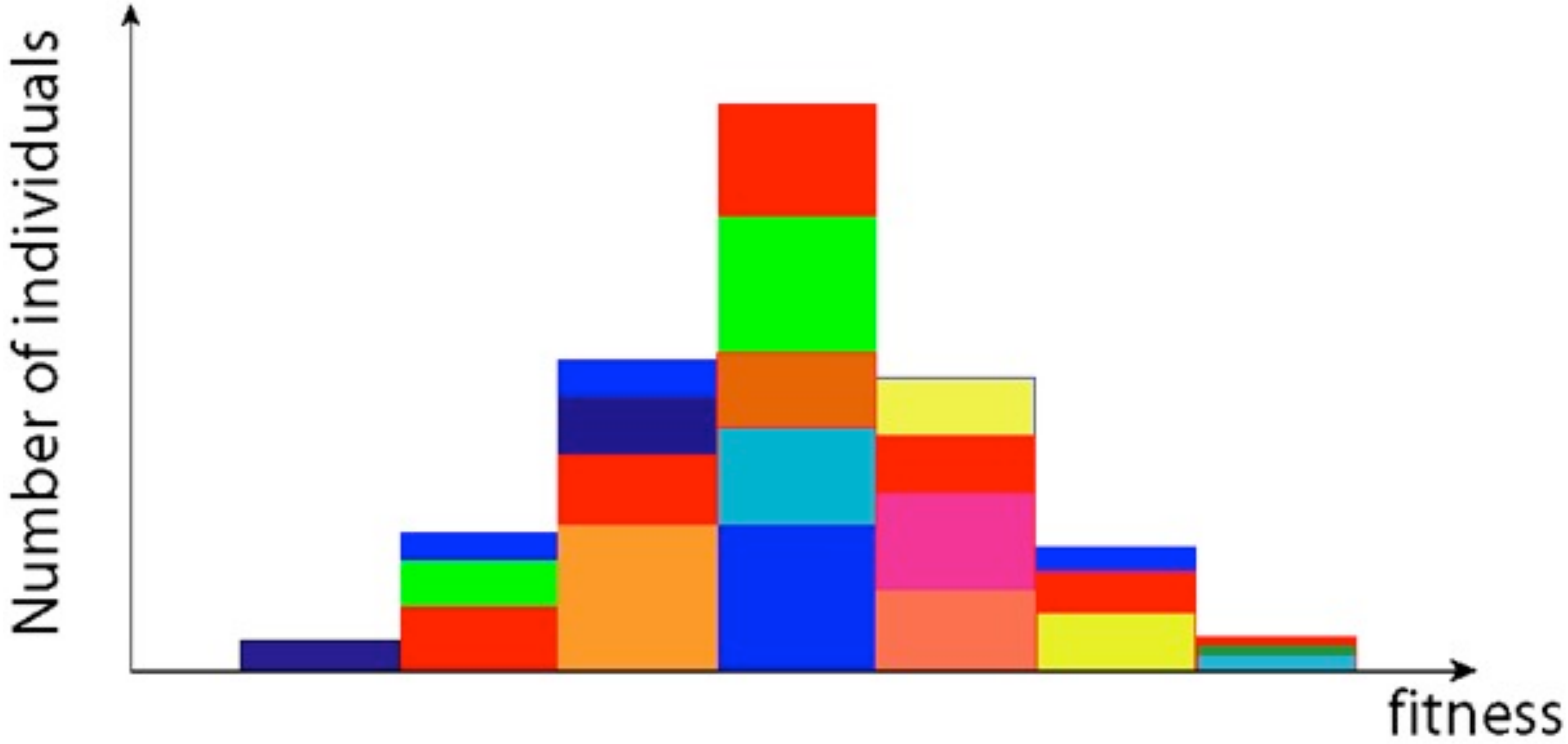
$$\frac{dh_k(t)}{dt} = U_d h_{k-1} - U_d h_k - s(k - k_{av})h_k$$

In steady state:

$$\hat{h}_k = e^{-U_d/s}\frac{U_d^k}{k!s}$$

Many fluctuating lineages maintain the balance

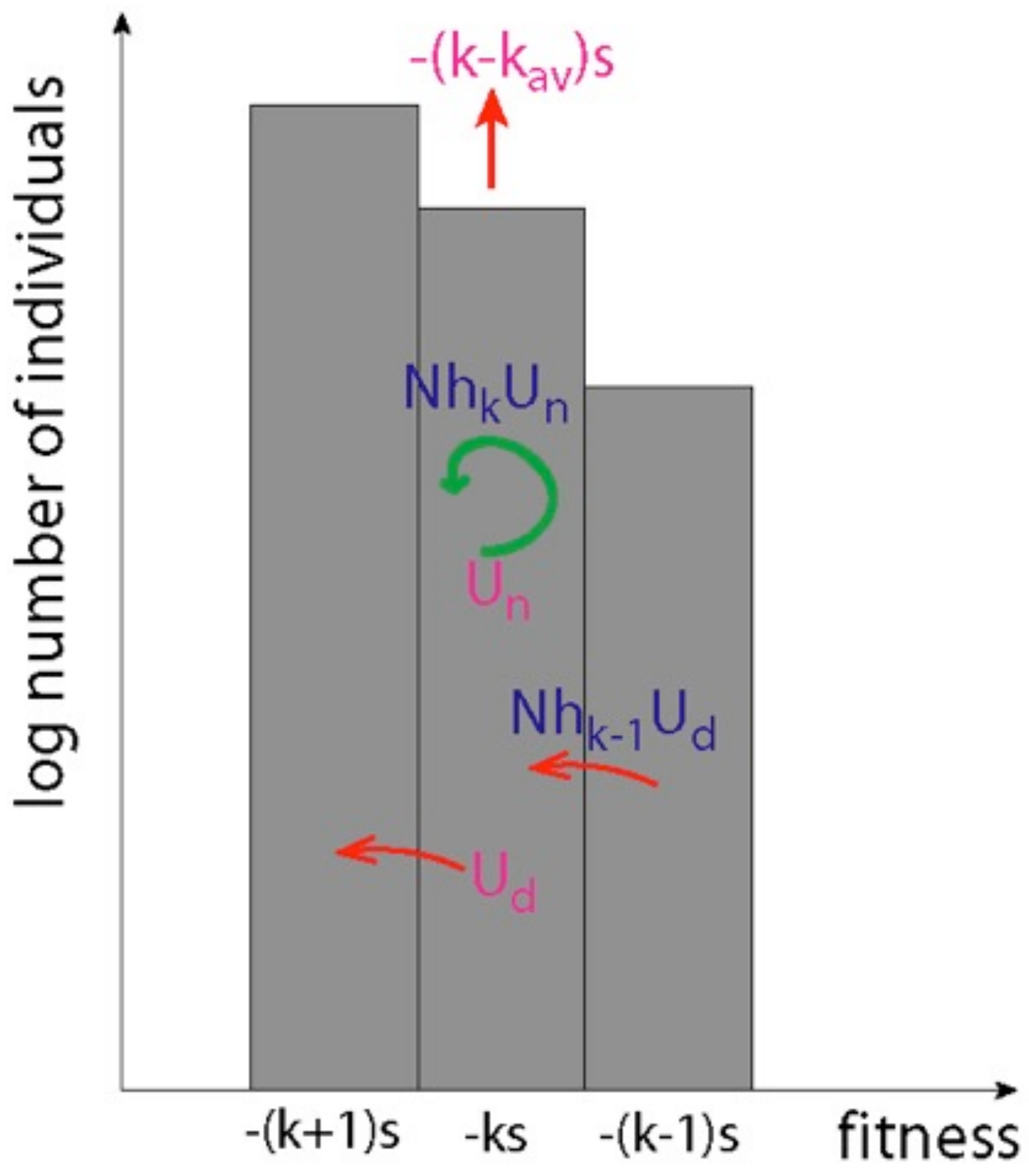
- each fitness class is not genetically homogenous
- each class composed of many lineages
- different alleles with the same total fitness



Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.

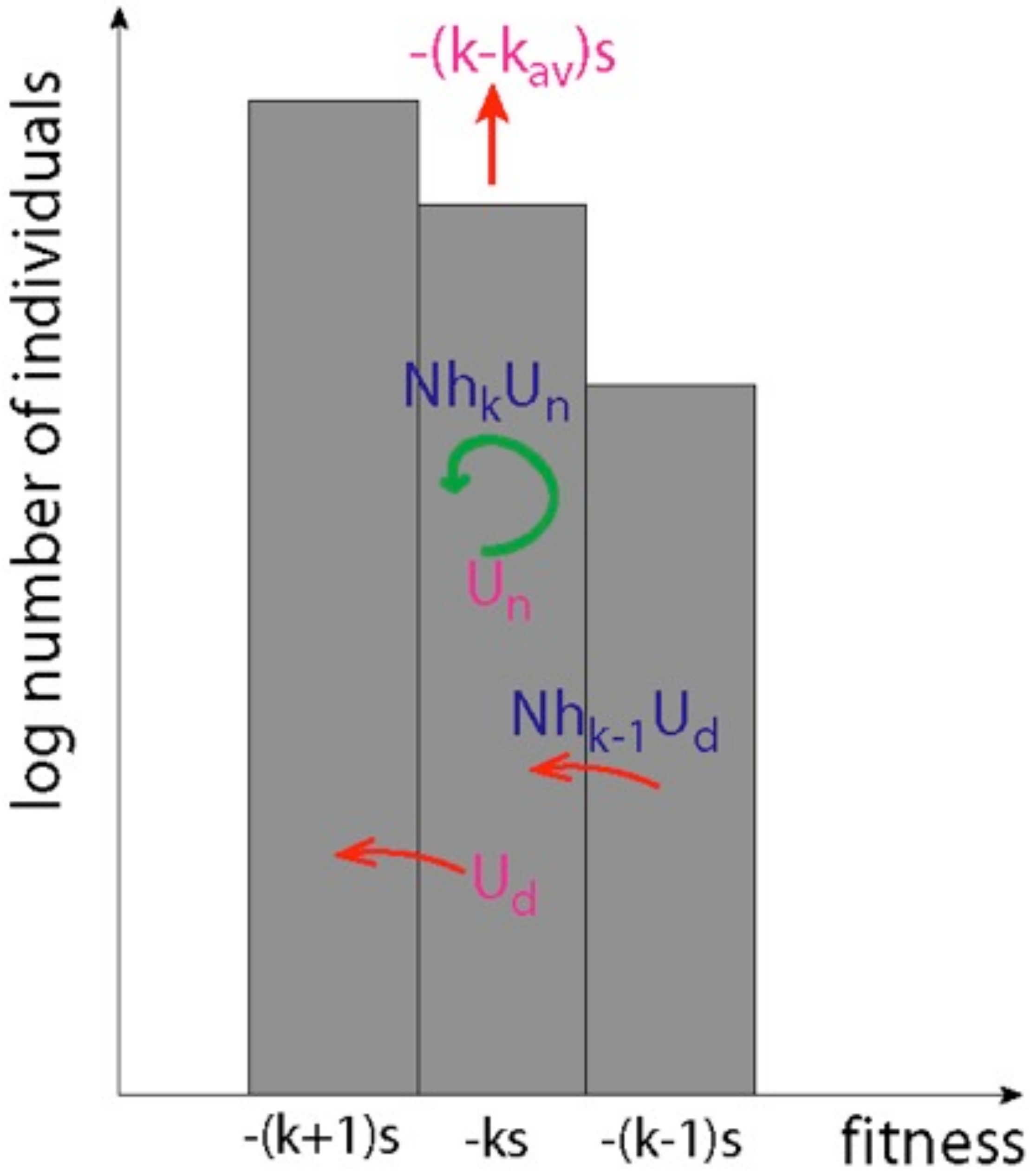


- diffusion limit of Wright-Fisher model
- mutation decoupled from selection
- perfect linkage

- infinite alleles model, but keeps track of how many deleterious mutations each individual has

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.



- diffusion limit of Wright-Fisher model
- mutation decoupled from selection
- perfect linkage

New alleles created at (mutation) rate:

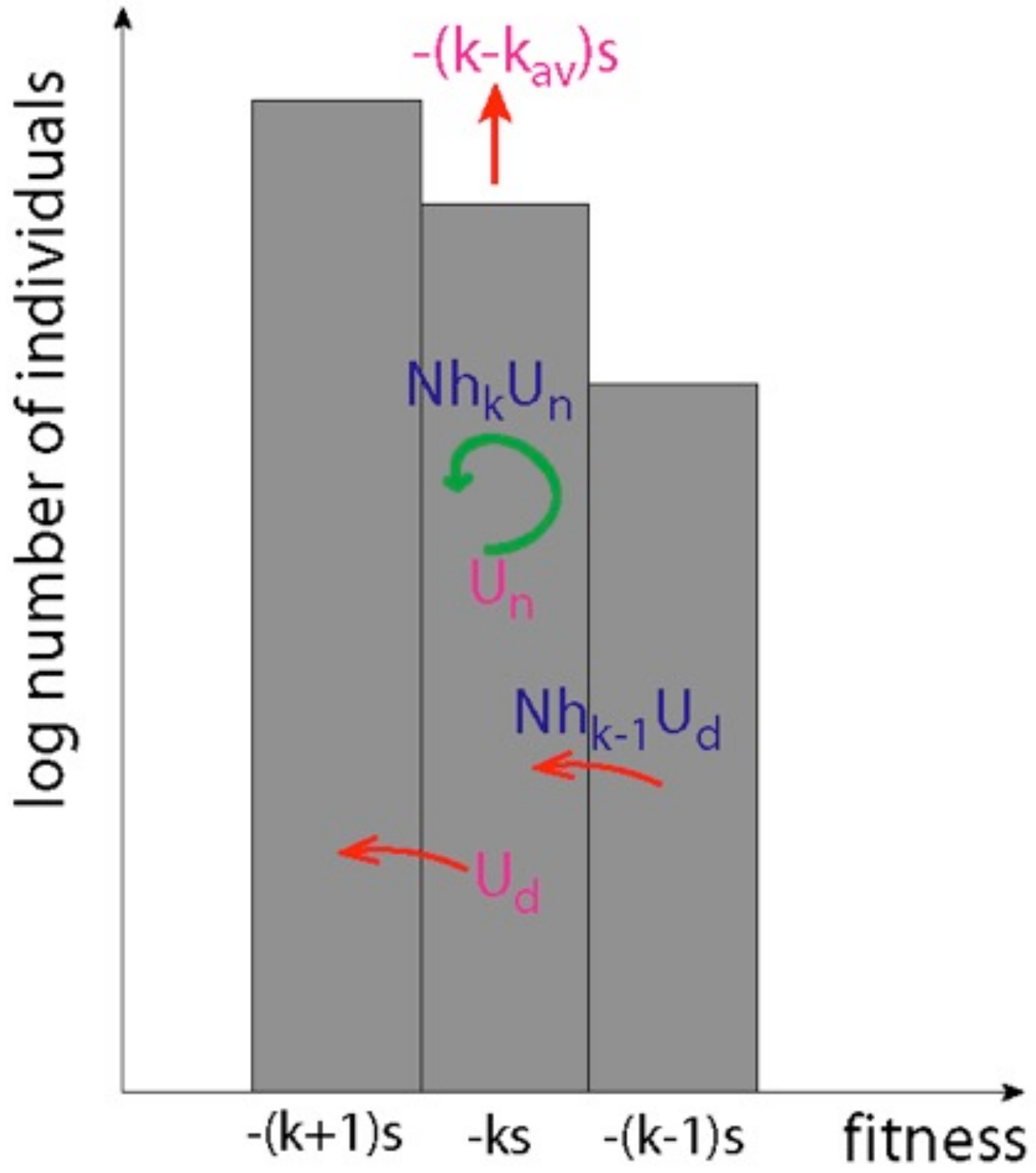
$$\frac{\theta_k}{2} = Nh_{k-1}U_d + Nh_kU_n$$

per genome
per generation

- infinite alleles model, but keeps track of how many deleterious mutations each individual has

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.



- diffusion limit of Wright-Fisher model
- mutation decoupled from selection
- perfect linkage

New alleles created at (mutation) rate:

$$\frac{\theta_k}{2} = Nh_{k-1}U_d + Nh_kU_n$$

per genome
per generation

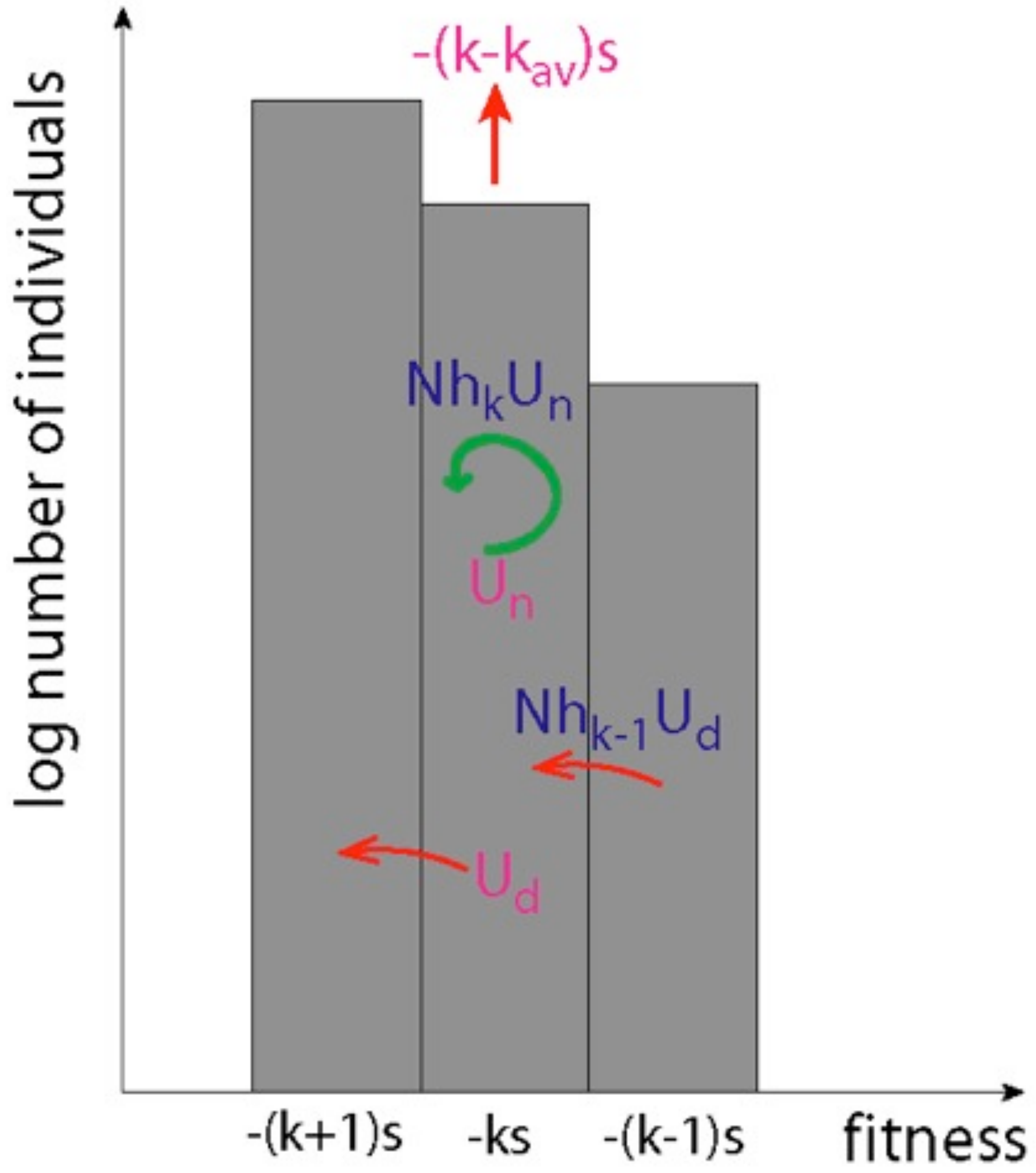
Experience effective selective pressure:

$$s_k = -U_d - U_n - (k - k_{av})s$$

- infinite alleles model, but keeps track of how many deleterious mutations each individual has

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.



- diffusion limit of Wright-Fisher model
- mutation decoupled from selection
- perfect linkage

New alleles created at (mutation) rate:

$$\frac{\theta_k}{2} = Nh_{k-1}U_d + Nh_kU_n$$

per genome per generation

Experience effective selective pressure:

$$s_k = -U_d - U_n - (k-k_{av})s$$

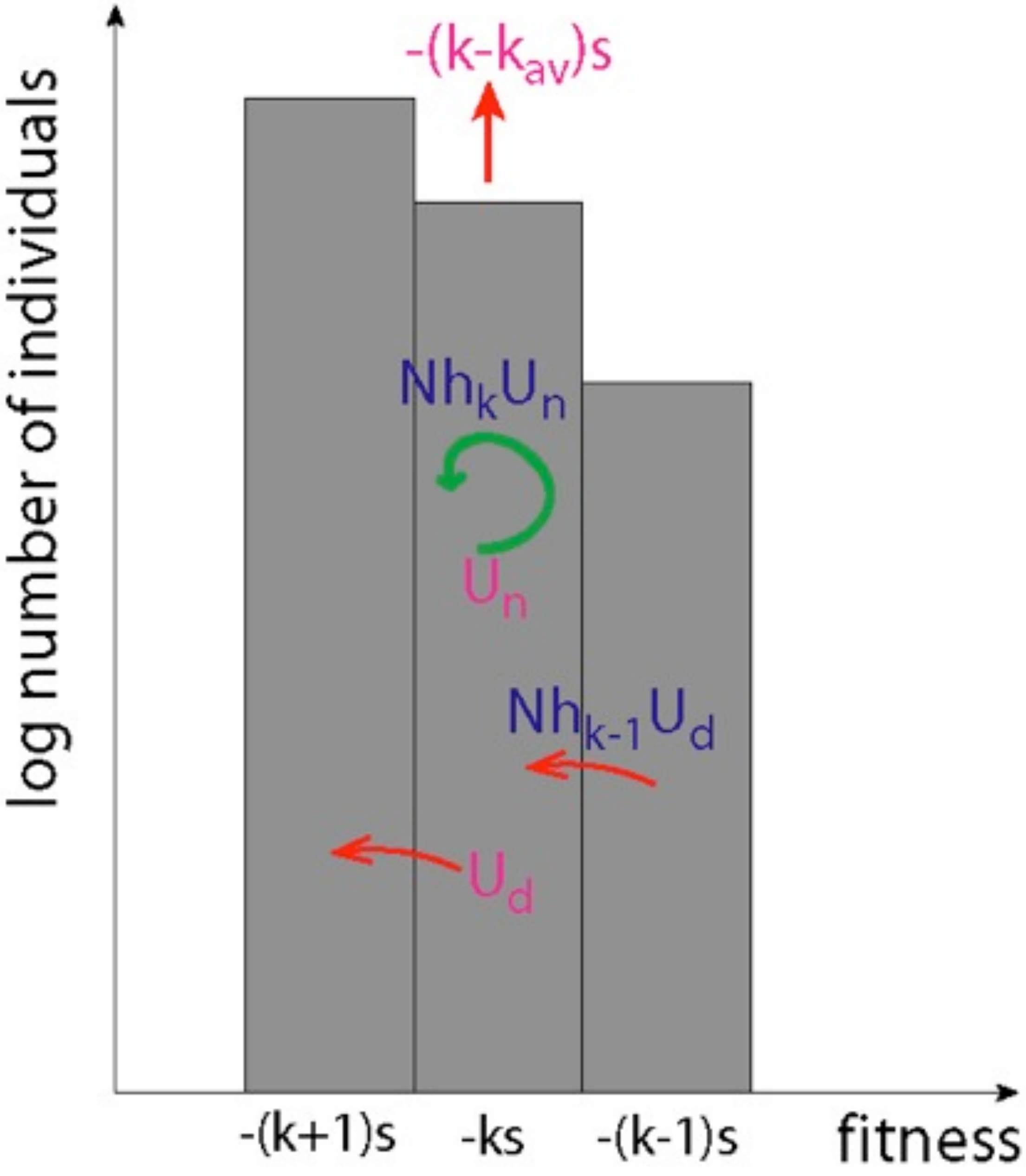
- infinite alleles model, but keeps track of how many deleterious mutations each individual has

$$\hat{h}_k = e^{-U_d/s}\frac{U_d^k}{k!s}$$

θ_k and s_k determined by state of other fluctuating alleles: self-consistency.

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.



New alleles created at (mutation) rate:

$$\frac{\theta_k}{2} = Nh_{k-1}U_d + Nh_kU_n$$

per genome
per generation

Experience effective selective pressure:

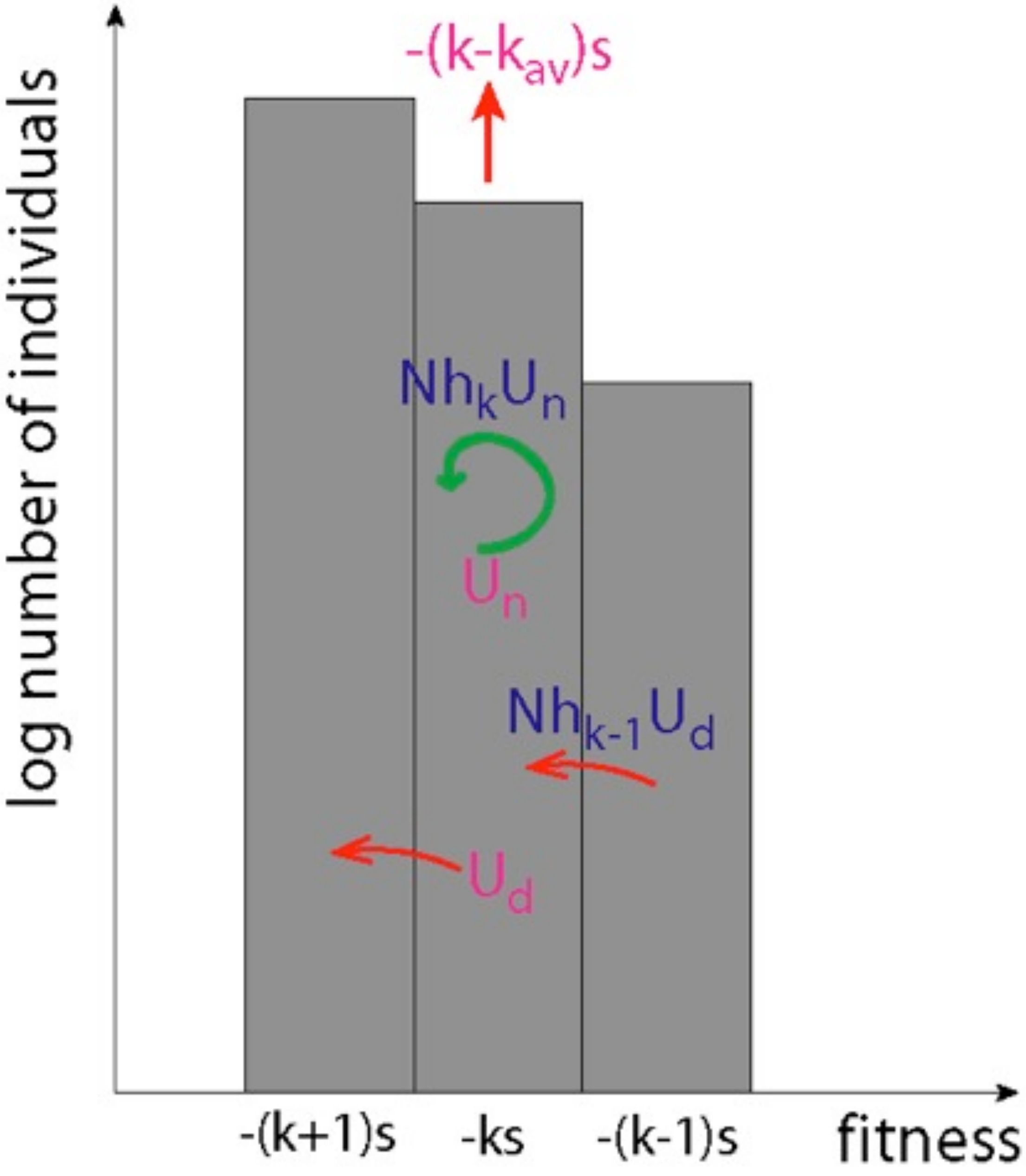
$$s_k = -U_d - U_n - (k - k_{av})s$$

θ_k and s_k determined by state of other fluctuating alleles: self-consistency:

$$\hat{h}_k = e^{-U_d/s}\frac{U_d^k}{k!s}$$

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.



New alleles created at (mutation) rate:

$$\frac{\theta_k}{2} = Nh_{k-1}U_d + Nh_kU_n$$

per genome per generation

Experience effective selective pressure:

$$s_k = -U_d - U_n - (k - k_{av})s$$

θ_k and s_k determined by state of other fluctuating alleles: self-consistency:

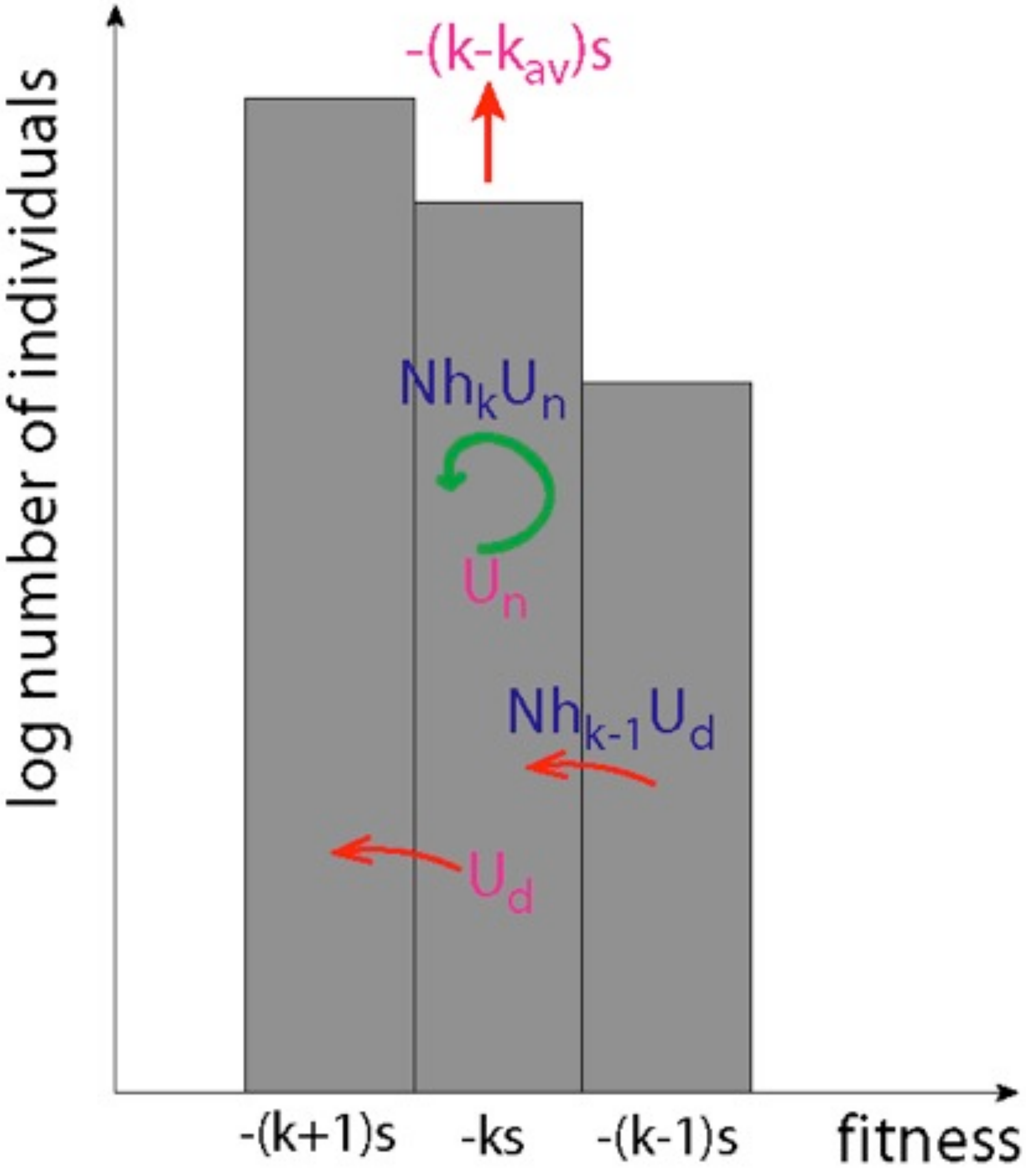
$$\hat{h}_k = e^{-U_d/s}\frac{U_d^k}{k!s}$$

no neutral mutations:

- $s_k < 0$, each class except for k=0 is always receiving new individuals due to mutations
- older individuals must die out to conserve steady state fitness distribution
- k=0 class drifts neutrally - fitness advantage balanced by loss of individuals to less fit classes

Many fluctuating lineages maintain the balance

Each class is maintained by flux in of new mutant alleles as old alleles drift and go extinct.



New alleles created at (mutation) rate:

$$\frac{\theta_k}{2} = Nh_{k-1}U_d + Nh_kU_n$$

per genome per generation

Experience effective selective pressure:

$$s_k = -U_d - U_n - (k - k_{av})s$$

θ_k and s_k determined by state of other fluctuating alleles: self-consistency:

$$\hat{h}_k = e^{-U_d/s}\frac{U_d^k}{k!s}$$

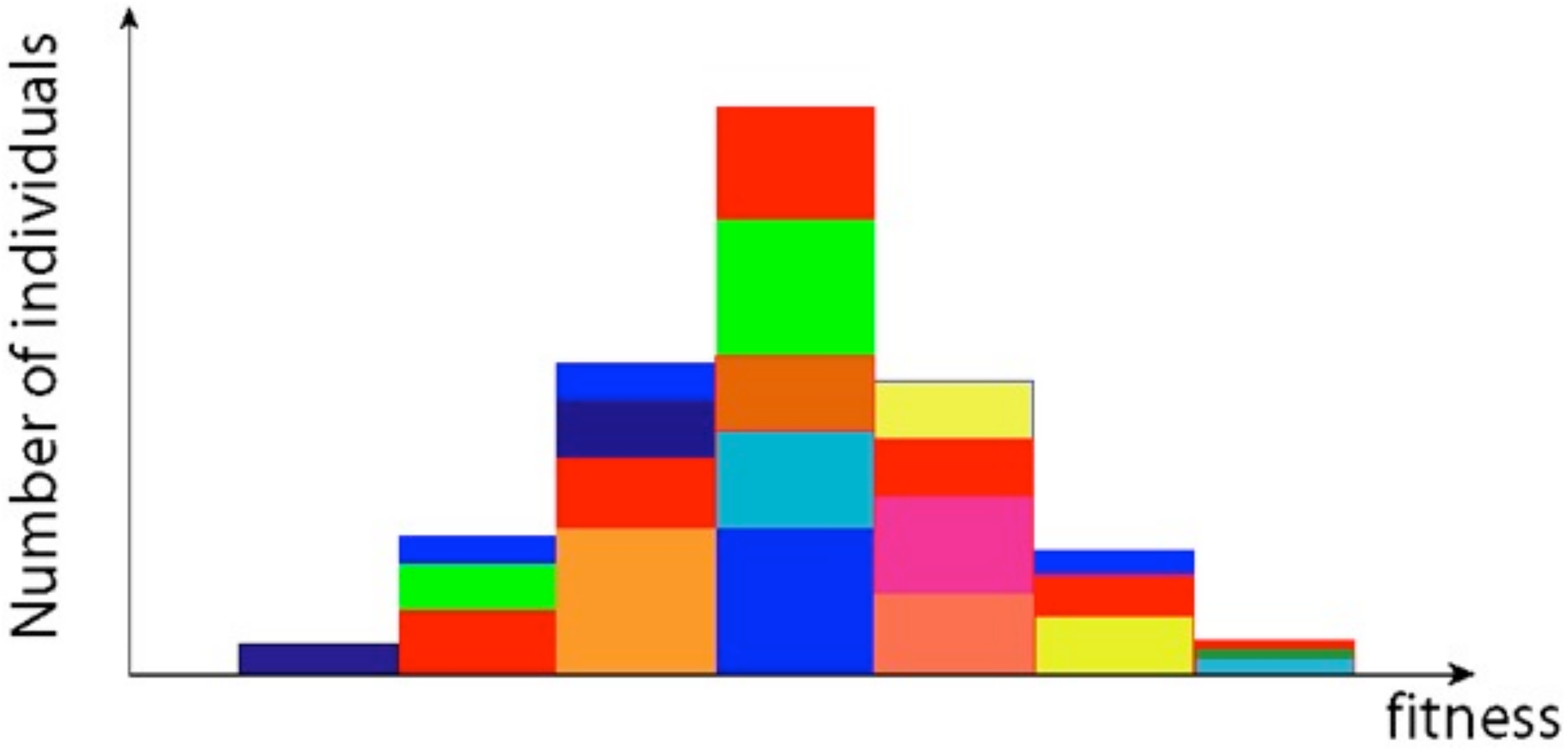
no neutral mutations:

- $s_k < 0$, each class except for k=0 is always receiving new individuals due to mutations
- older individuals must die out to conserve steady state fitness distribution
- k=0 class drifts neutrally - fitness advantage balanced by loss of individuals to less fit classes

with neutral mutations:

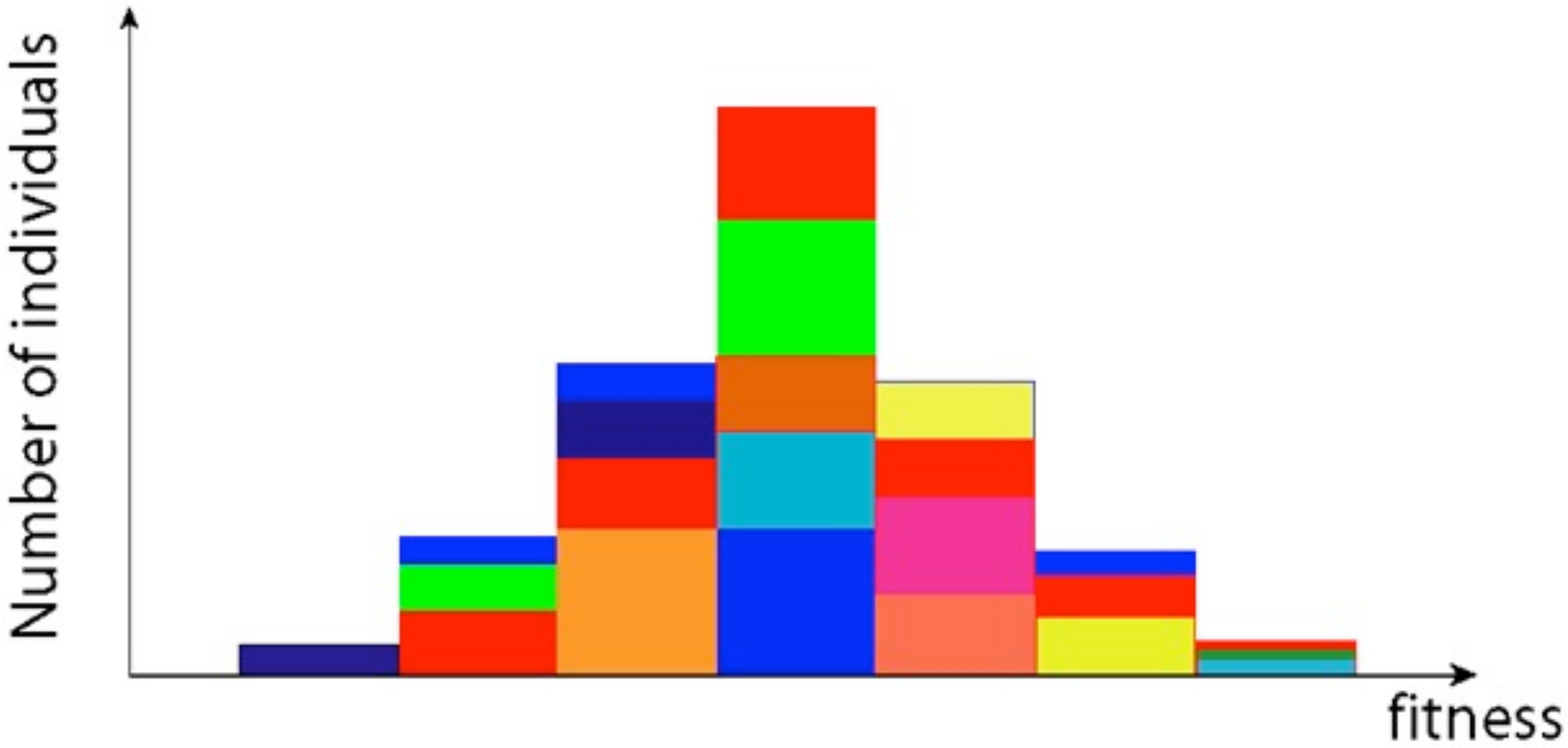
- $s_k < 0$, effective selection even more negative
- even $s_0 < 0$, all classes effectively selected against!

Allelic diversity within each class



Balance between creation and destruction of alleles

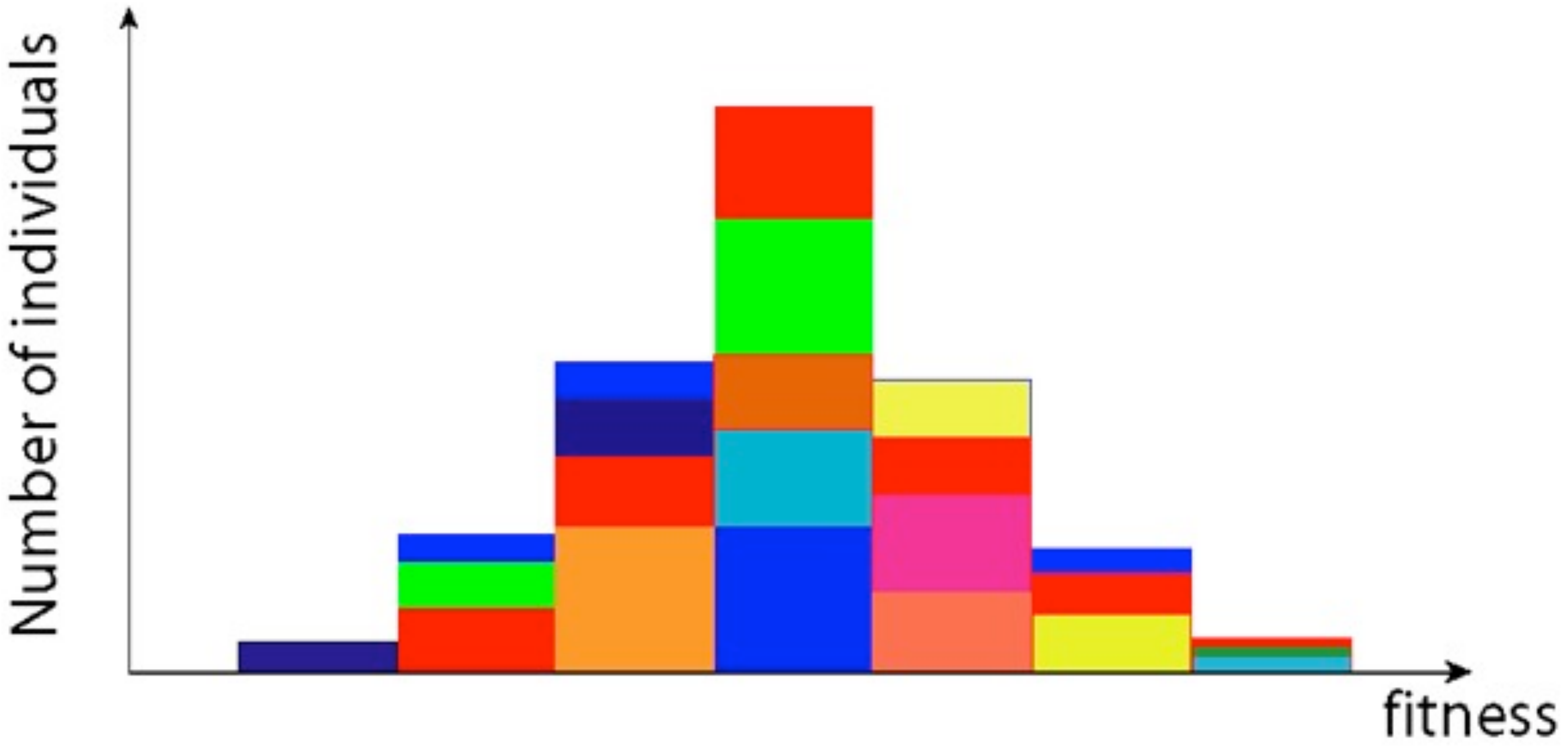
Allelic diversity within each class



Balance between creation and destruction of alleles

Fluctuations of particular mutations are ***not*** independent. Fluctuations of alleles ***are***.

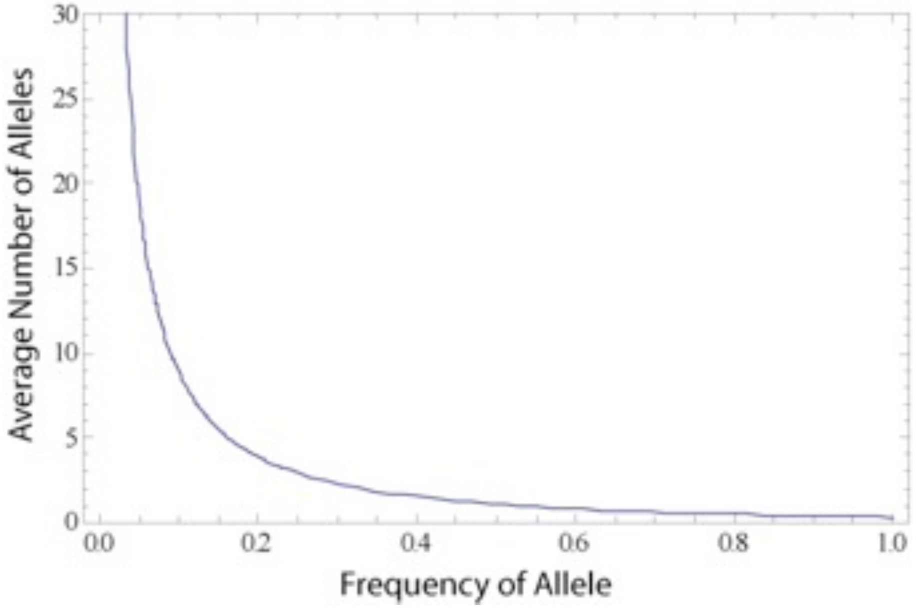
Allelic diversity within each class



Balance between creation and destruction of alleles

→ Distribution of probability of seeing an allele frequency x:

$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)} dx$$



Poisson Random Field (PRF) gives
distribution of lineages in given fitness class

+

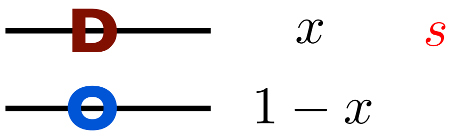
self-consistency condition - fluctuations of alleles affect the mean fitness and the rate of mutations to less-fit alleles

$$h_k = \int_0^1 x f_k(x) dx$$

Fluctuations of particular mutations are ***not*** independent. Fluctuations of alleles ***are***.

Poisson Random Field traditionally

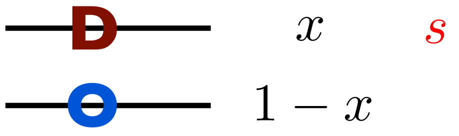
PRF - qualitatively determines the intensity of selection on a particular gene



The model: $p(x; x_0, t)$ probability distribution of derived allele frequency x at time t, given x_0 at time t_0

Poisson Random Field traditionally

PRF - qualitatively determines the intensity of selection on a particular gene



The model: $p(x; x_0, t)$ probability distribution of derived allele frequency x at time t, given x_0 at time t_0

$q(x_0; x, t)$ backward equation:

+

absorbing boundary conditions at $x = 1$ or $x = 0$

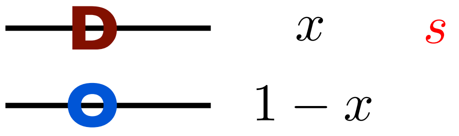
$$\partial_t q(x_0; x, t) = v(x_0)\frac{\partial q(x_0; x, t)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 q(x_0; x, t)}{\partial x_0^2}$$

$$v(x_0) = 2Nsx_0(1-x_0)$$

$$D(x_0) = x_0(1-x_0)$$

Poisson Random Field traditionally

PRF - qualitatively determines the intensity of selection on a particular gene



The model: $p(x;x_0,t)$ probability distribution of derived allele frequency x at time t, given x_0 at time t_0

$q(x_0;x,t)$ backward equation:

+

absorbing boundary conditions at $x=1$ or $x=0$

$$\partial_t q(x_0;x,t) = v(x_0)\frac{\partial q(x_0;x,t)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 q(x_0;x,t)}{\partial x_0^2}$$

$$v(x_0) = 2Nsx_0(1-x_0)$$

$$D(x_0) = x_0(1-x_0)$$

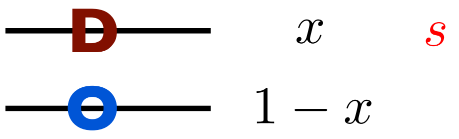
$T(x_0) = \int_0^1 \tilde{f}(x;x_0)dx$ - mean time until absorption (MFPT)

→ mean time derived allele frequency spends in the interval $(x, x+dx)$:

$$\tilde{f}(x) = \frac{1-e^{2Ns(1-x)}}{1-e^{2Ns}}\frac{2}{x(1-x)}$$

Poisson Random Field traditionally

PRF - qualitatively determines the intensity of selection on a particular gene



The model: $p(x;x_0,t)$ probability distribution of derived allele frequency x at time t, given x_0 at time t_0

$q(x_0;x,t)$ backward equation:

+

absorbing boundary conditions at $x=1$ or $x=0$

$$\partial_t q(x_0;x,t) = v(x_0)\frac{\partial q(x_0;x,t)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 q(x_0;x,t)}{\partial x_0^2}$$

$$v(x_0) = 2Nsx_0(1-x_0)$$

$$D(x_0) = x_0(1-x_0)$$

$T(x_0) = \int_0^1 \tilde{f}(x;x_0)dx$ - mean time until absorption (MFPT)

→ mean time derived allele frequency spends in the interval $(x, x+dx)$:

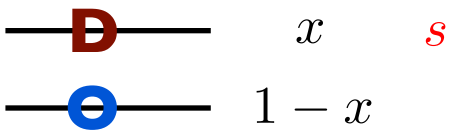
$$\tilde{f}(x) = \frac{1-e^{2Ns(1-x)}}{1-e^{2Ns}}\frac{2}{x(1-x)}$$

Generalize to multiple alleles, assume:

- mutations arise at Poisson times
- each mutation forms a new allele
- independent alleles - each mutant follows an independent Wright-Fisher process

Poisson Random Field traditionally

PRF - qualitatively determines the intensity of selection on a particular gene



The model: $p(x;x_0,t)$ probability distribution of derived allele frequency x at time t, given x_0 at time t_0

$q(x_0;x,t)$ backward equation: + absorbing boundary conditions at $x=1$ or $x=0$

$$\partial_t q(x_0;x,t) = v(x_0)\frac{\partial q(x_0;x,t)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 q(x_0;x,t)}{\partial x_0^2}$$

$$v(x_0) = 2Nsx_0(1-x_0)$$

$$D(x_0) = x_0(1-x_0)$$

$T(x_0) = \int_0^1 \tilde{f}(x;x_0)dx$ - mean time until absorption (MFPT)

→ mean time derived allele frequency spends in the interval $(x, x+dx)$:

$$\tilde{f}(x) = \frac{1-e^{2Ns(1-x)}}{1-e^{2Ns}}\frac{2}{x(1-x)}$$

Generalize to multiple alleles, assume:

- mutations arise at Poisson times
- each mutation forms a new allele
- independent alleles - each mutant follows an independent Wright-Fisher process

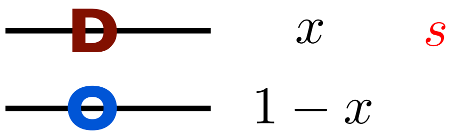
$\int_{x_1}^{x_2}\theta\tilde{f}(x)dx = \int_{x_1}^{x_2} f(x)dx$ - expected number of sites with derived allele/lineage frequency in a given range:

$$f(x)dx = \theta\frac{1-e^{2Ns(1-x)}}{(1-e^{2Ns})x(1-x)}dx$$

θ - per site mutation rate

Poisson Random Field traditionally

PRF - qualitatively determines the intensity of selection on a particular gene



The model: $p(x;x_0,t)$ probability distribution of derived allele frequency x at time t, given x_0 at time t_0

$q(x_0;x,t)$ backward equation:
+
absorbing boundary conditions
at $x=1$ or $x=0$

$$\partial_t q(x_0;x,t) = v(x_0)\frac{\partial q(x_0;x,t)}{\partial x_0} + \frac{D(x_0)}{2}\frac{\partial^2 q(x_0;x,t)}{\partial x_0^2}$$

$$v(x_0) = 2Nsx_0(1-x_0)$$

$$D(x_0) = x_0(1-x_0)$$

$T(x_0) = \int_0^1 \tilde{f}(x;x_0)dx$ - mean time until absorption (MFPT)

→ mean time derived allele frequency spends in the interval $(x, x+dx)$:

$$\tilde{f}(x) = \frac{1-e^{2Ns(1-x)}}{1-e^{2Ns}}\frac{2}{x(1-x)}$$

Generalize to multiple alleles, assume:

- mutations arise at Poisson times
- each mutation forms a new allele
- independent alleles - each mutant follows an independent Wright-Fisher process

$\int_{x_1}^{x_2}\theta\tilde{f}(x)dx = \int_{x_1}^{x_2} f(x)dx$ - expected number of sites with derived allele/lineage frequency in a given range:

$$f(x)dx = \theta\frac{1-e^{2Ns(1-x)}}{(1-e^{2Ns})x(1-x)}dx$$

θ - per site mutation rate

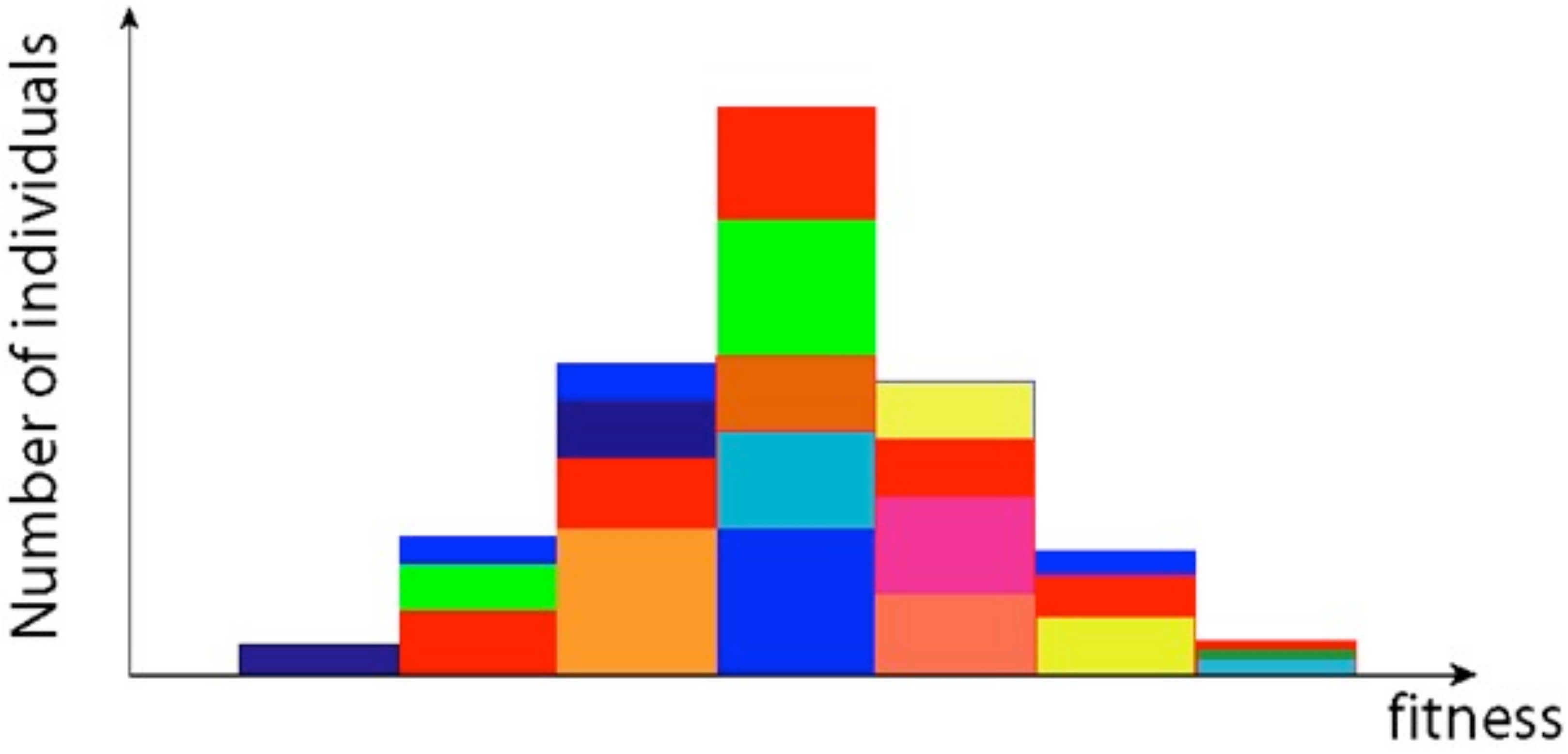
The number of sites that have i copies of the derived allele are Poisson distributed with mean:

probability that the site has i copies in the sample

$$\int\binom{n}{i}x^i(1-x)^{n-i}f(x)dx$$

[hence Poisson Random Field]

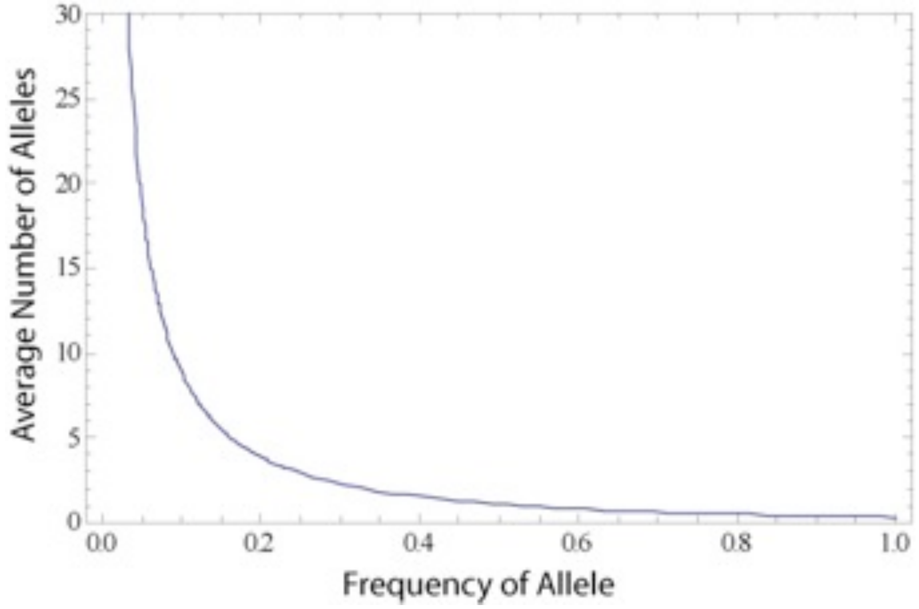
Allelic diversity within each class



Balance between creation and destruction of alleles

→ Distribution of probability of seeing an allele frequency x:

$$f_k(x)dx = \theta_k \frac{1-e^{-2Ns_k(1-x)}}{(1-e^{-2Ns_k})x(1-x)}dx$$



Poisson Random Field (PRF) gives distribution of lineages in given fitness class

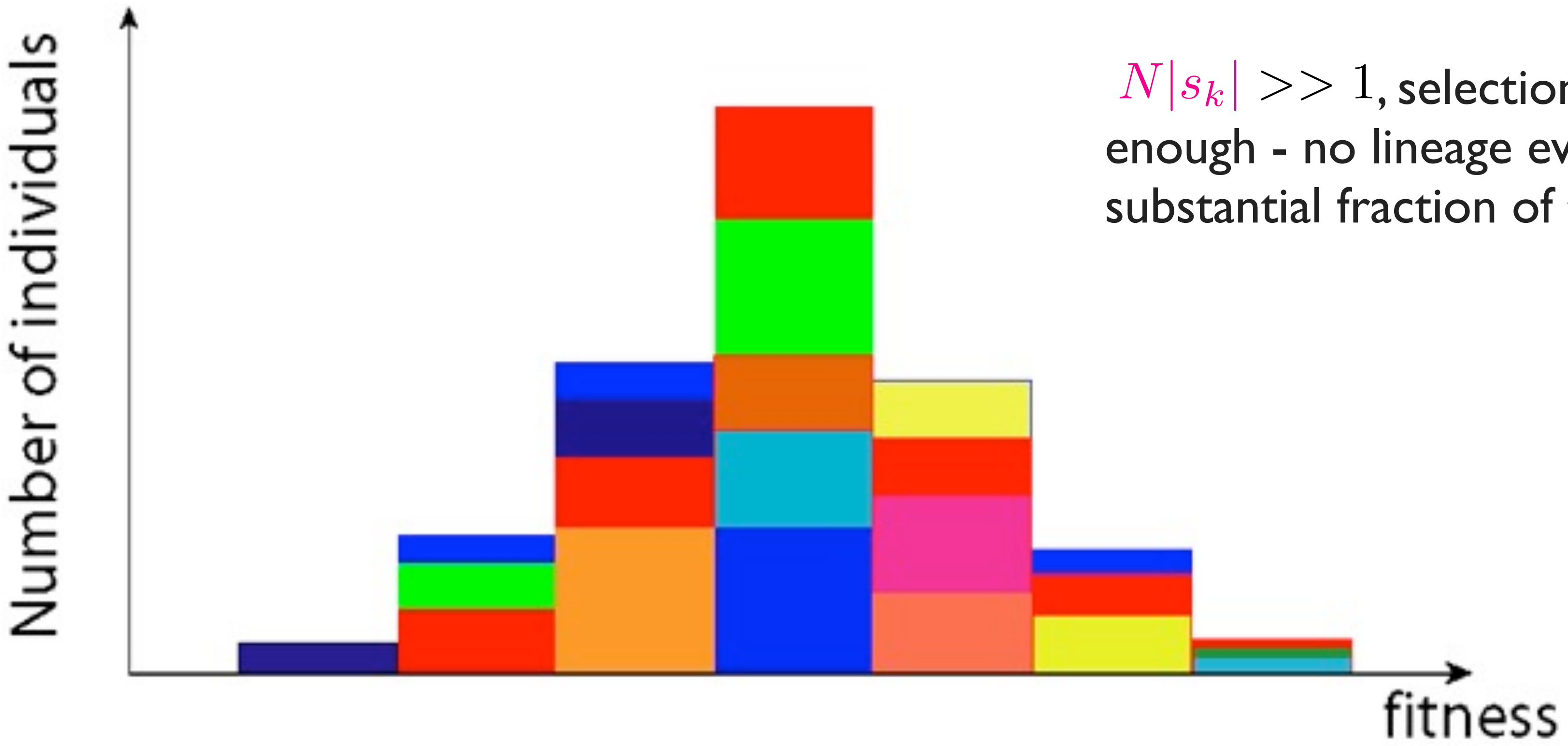
+

self-consistency condition - fluctuations of alleles affect the mean fitness and the rate of mutations to less-fit alleles

$$h_k = \int_0^1 x f_k(x)dx$$

Fluctuations of particular mutations are ***not*** independent. Fluctuations of alleles ***are***.

Allelic diversity within each class



$N|s_k| >> 1$, selection is strong enough - no lineage ever becomes a substantial fraction of the population

Balance between creation and destruction of alleles

→ Distribution of probability of seeing an allele frequency x:

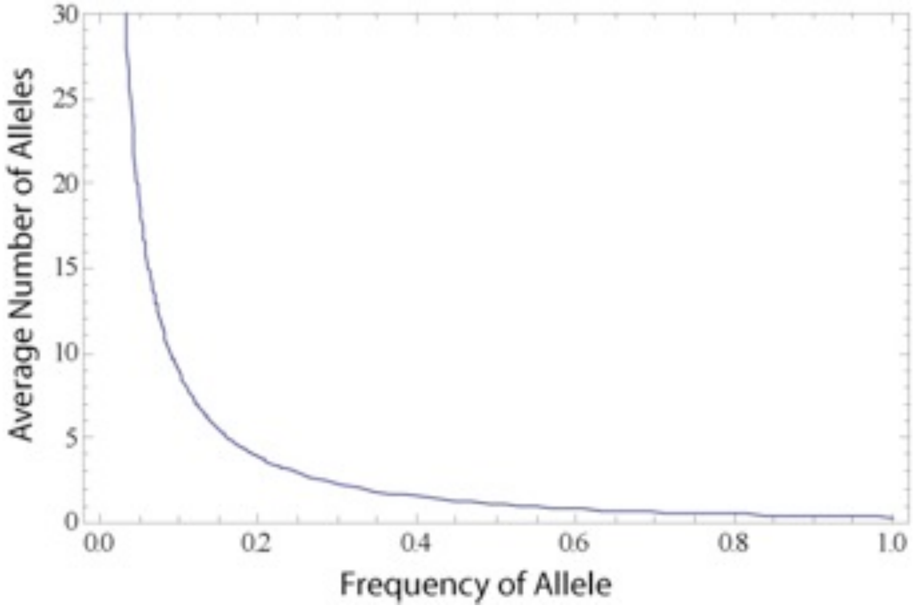
$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)}dx$$

Poisson Random Field (PRF) gives distribution of lineages in given fitness class

+

self-consistency condition - fluctuations of alleles affect the mean fitness and the rate of mutations to less-fit alleles

$$h_k = \int_0^1 x f_k(x)dx$$



Fluctuations of particular mutations are ***not*** independent. Fluctuations of alleles ***are***.

Self-consistency condition

Poisson Random Field (PRF) gives distribution of lineages in given fitness class

+

steady state distribution of fitness classes

$$h_k = \int_0^1 x f_k(x) dx$$

$$\Updownarrow$$

$$\int_0^1 \frac{1 - e^{-2\gamma_k x}}{x} dx = \frac{1 - e^{-2\gamma_k}}{2|\gamma_k|}$$

$$\downarrow$$

$N|s_k| >> 1$

$s_k = -U_d - U_n - (k - k_{av})s$

Self-consistency condition

Poisson Random Field (PRF) gives distribution of lineages in given fitness class

+

steady state distribution of fitness classes

$$h_k = \int_0^1 x f_k(x) dx$$

$$\Updownarrow$$

$$\int_0^1 \frac{1 - e^{-2\gamma_k x}}{x} dx = \frac{1 - e^{-2\gamma_k}}{2|\gamma_k|}$$

$$\downarrow$$

$$N|s_k| >> 1$$

$$s_k = -U_d - U_n - (k - k_{av})s$$

close to k_{av}:

$$N(U_d + U_n) >> 1$$

Self-consistency condition

Poisson Random Field (PRF) gives distribution of lineages in given fitness class

+

steady state distribution of fitness classes

$$h_k = \int_0^1 x f_k(x) dx$$

$$\Updownarrow$$

$$\int_0^1 \frac{1 - e^{-2\gamma_k x}}{x} dx = \frac{1 - e^{-2\gamma_k}}{2|\gamma_k|}$$

$$\downarrow$$

$N|s_k| >> 1$ $\qquad$ $s_k = -U_d - U_n - (k - k_{av})s$

close to k_{av}: $\quad N(U_d + U_n) >> 1$

$$\downarrow$$

$NU_d >> 1$ or $NU_n >> 1$ ⟵ self-consistency holds

Self-consistency condition

Poisson Random Field (PRF) gives distribution of lineages in given fitness class

+

steady state distribution of fitness classes

$$h_k = \int_0^1 x f_k(x) dx$$

⇕

$$\int_0^1 \frac{1 - e^{-2\gamma_k x}}{x} dx = \frac{1 - e^{-2\gamma_k}}{2|\gamma_k|}$$

↓

$N|s_k| >> 1$ $\qquad$ $s_k = -U_d - U_n - (k - k_{av})s$

close to k_{av}: $N(U_d + U_n) >> 1$

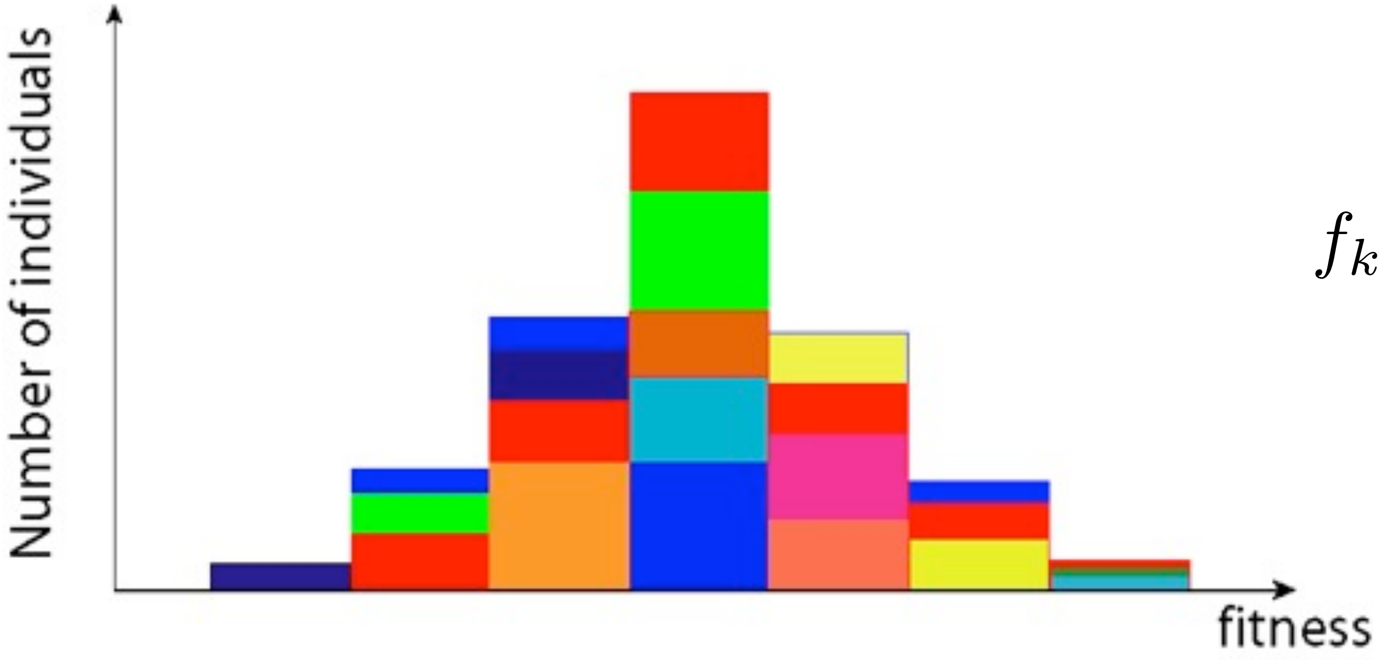
↓

$NU_d >> 1$ or $NU_n >> 1$ ← self-consistency holds

For $N(U_d + U_n) < 1$ PRF breaks down :

- the growth of some mutants is limited by size of population
- lineages are no longer independent

Expected genetic variation



$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)} dx$$

Sample n individuals.
What is the probability of a particular allelic configuration?

(n_1 individuals with allele 1, n_2 individuals with allele 2,....)

Homozygosity: $$Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$$

Sample n=2 individuals. What is the probability that they have the same genotype?

"Bizygosity": $$\begin{aligned} Q_{2,1} &= \sum_k \int 3x^2(1-x) f_k(x)dx \\ &= 3\sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}(1 - \frac{1}{Ns_k}) \end{aligned}$$

Sample n=3 individuals. What is the probability that two have the same alleles and one is different?

Comparison to known results

Sample n individuals.
What is the probability of a particular allelic configuration?

(n_1 individuals with allele 1, n_2 individuals with allele 2,....)

→ generalization of Ewens Sampling Formula (ESF)

$$P(n_1, ..., n_2) = \frac{n!}{\theta(\theta+1)...(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{n_j}}{j^{n_j} n_j!}$$

- neutral model
- steady state with respect to mutation and drift
- infinite alleles
- sample size n<<N - population size

Comparison to known results

Sample n individuals.
What is the probability of a particular allelic configuration?

(n_1 individuals with allele 1, n_2 individuals with allele 2,....)

➡ generalization of Ewens Sampling Formula (ESF)

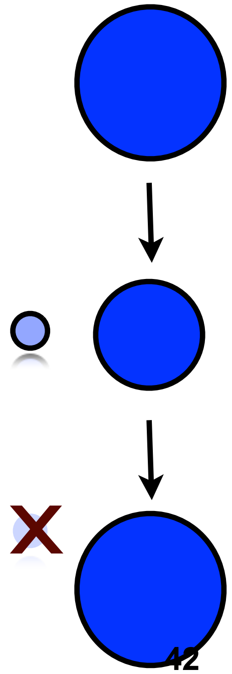
$$P(n_1, ..., n_2) = \frac{n!}{\theta(\theta+1)...(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{n_j}}{j^{n_j} n_j!}$$

- neutral model
- steady state with respect to mutation and drift
- infinite alleles
- sample size n<<N - population size

Effective Population Size Approximation (EPS):

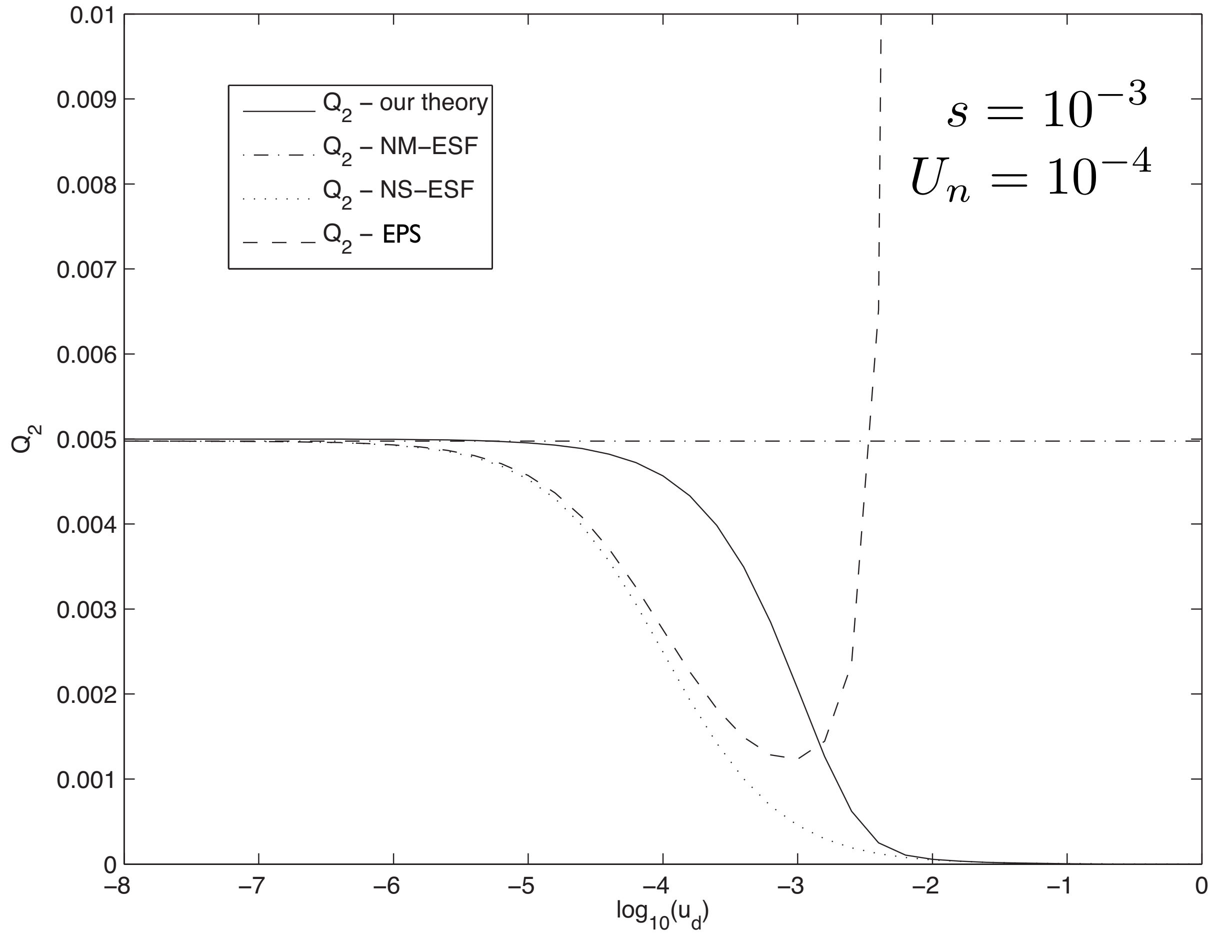
- deleterious mutations are purged quickly from the population
- all individuals are recently descended from neutral individuals
- only the zero-class matters
- results in neutral population with an effective reduced population size
- makes predictions about diversity at individual sites
- only makes predictions for neutral sites

$$N_e = Nh_0 = Ne^{-U_d/s}$$



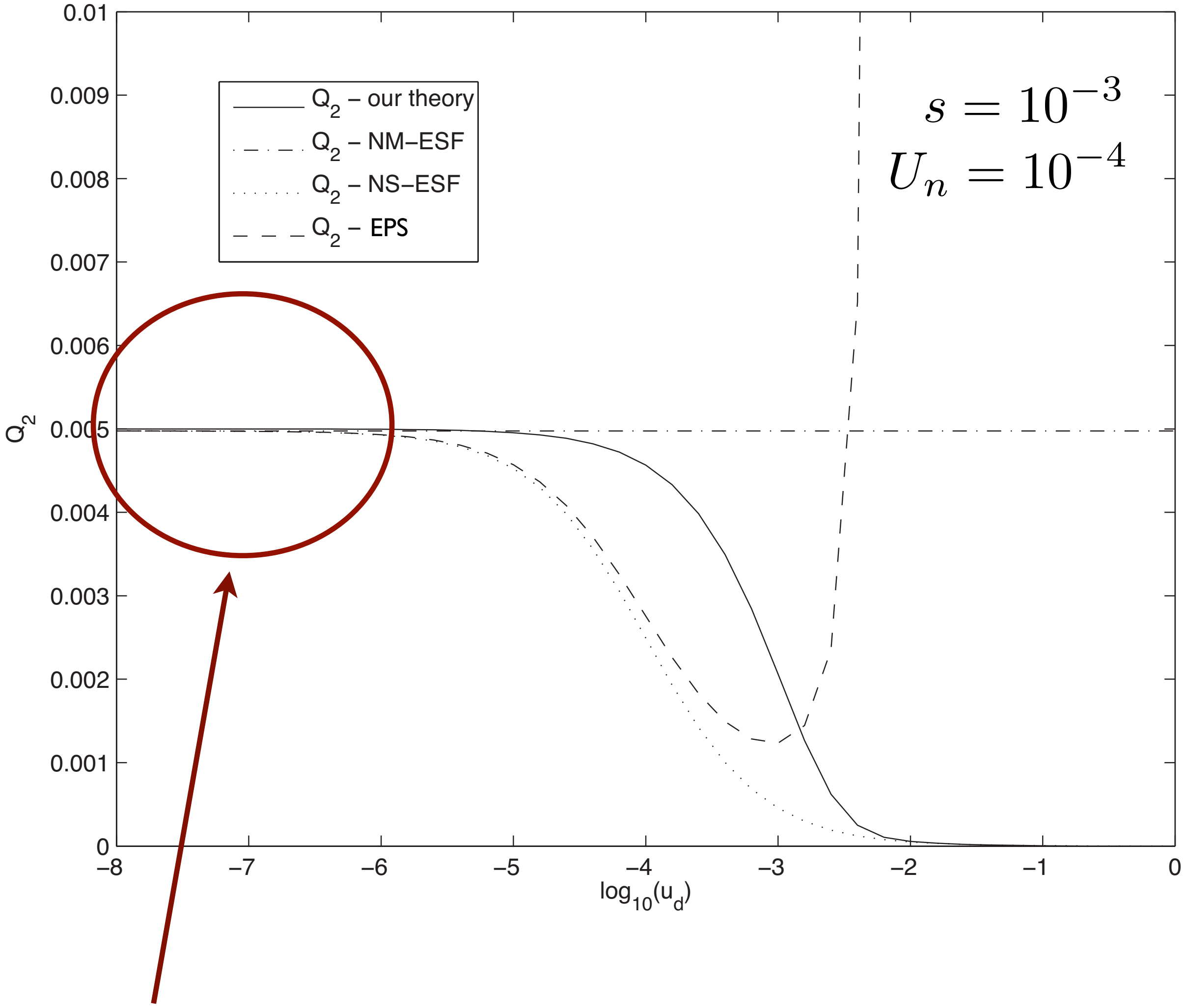
Expected genetic variation

Homozygosity: $$Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$$



Expected genetic variation

Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$



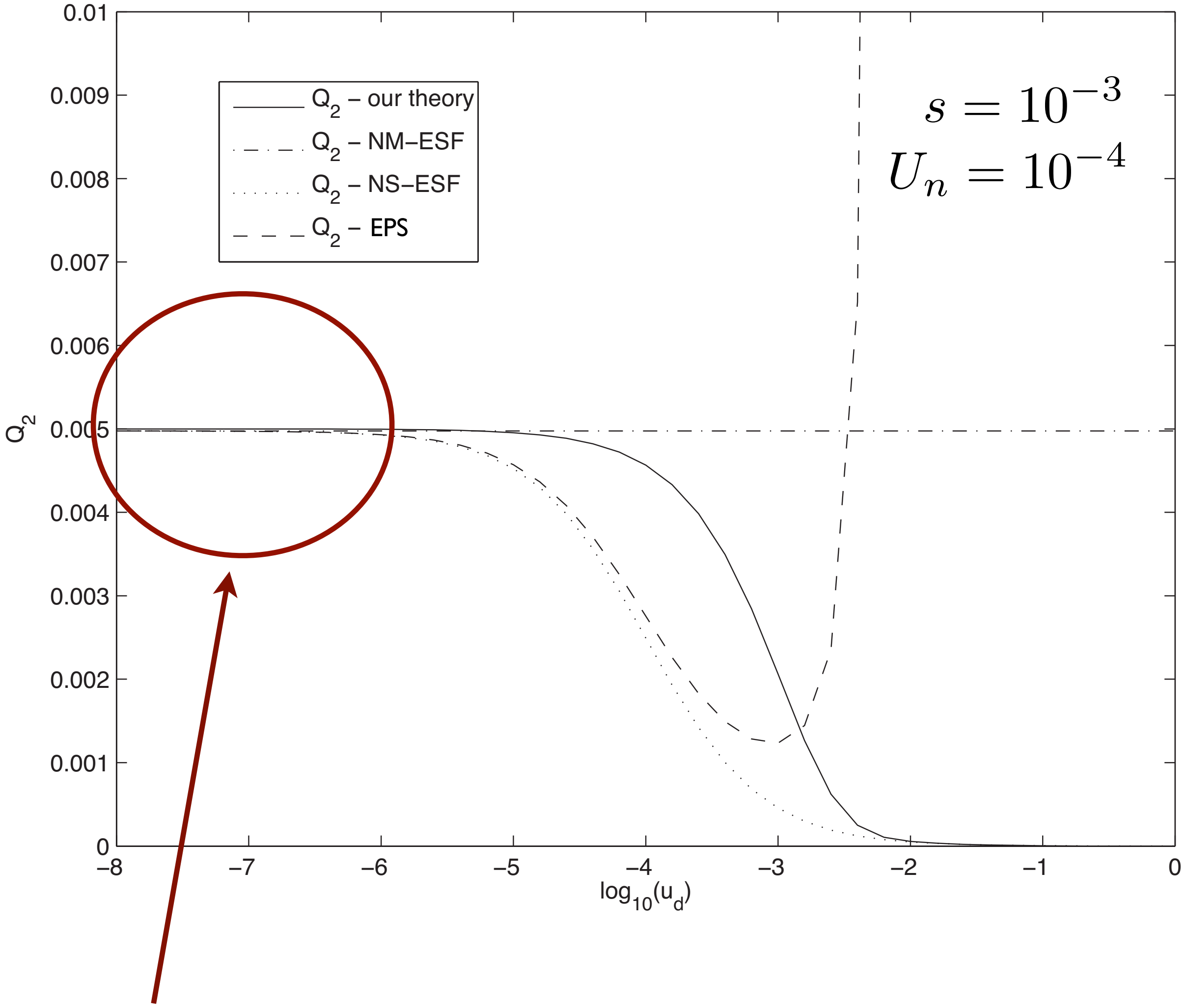
neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

$$Q_2 = \frac{1}{\theta}$$
$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

Expected genetic variation

Homozygosity: $$Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$$



EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

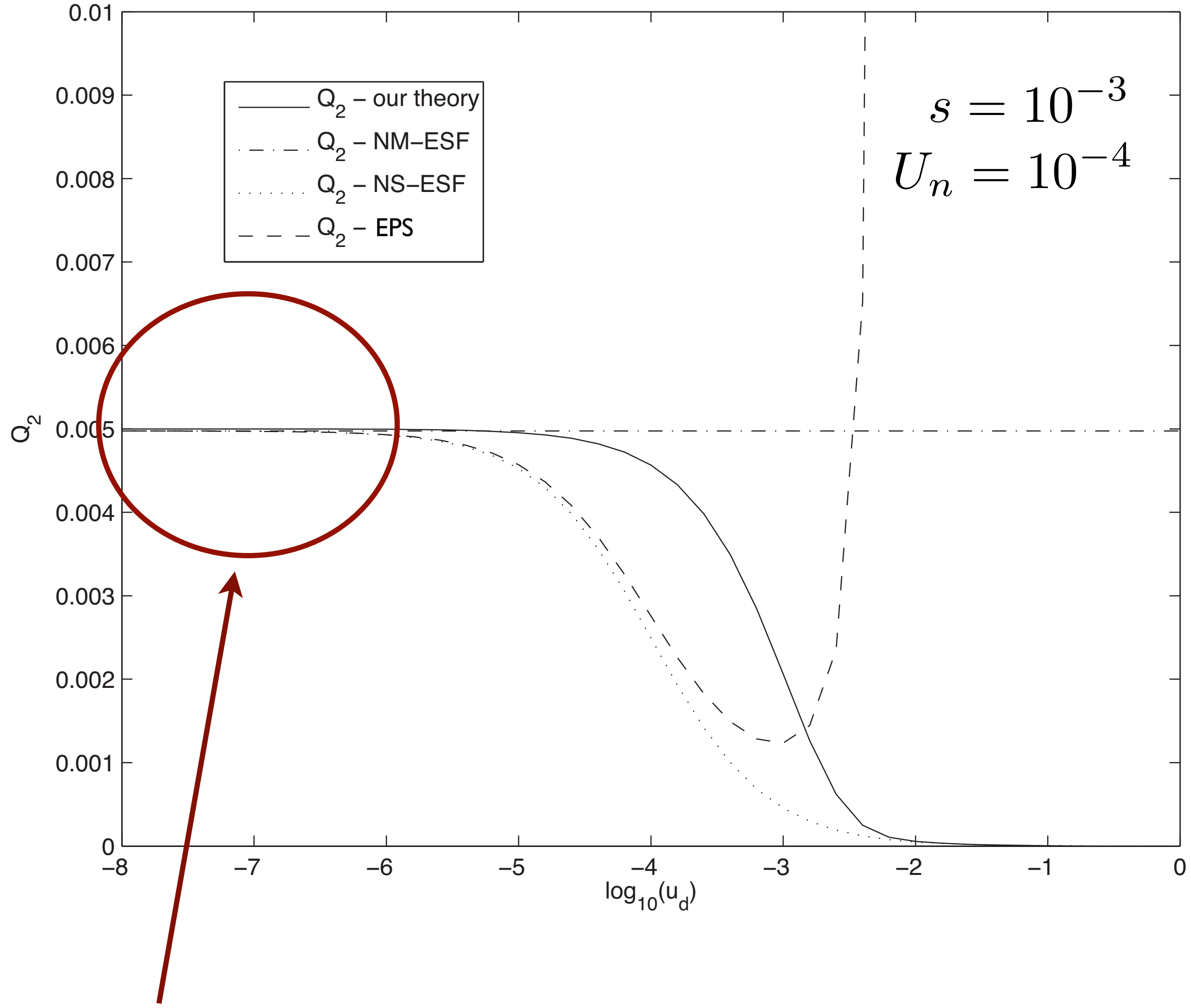
$$Q_2 = \frac{1}{\theta}$$

$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

Expected genetic variation

Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x) dx = \sum_{k=0}^{\infty} \frac{h_k}{2N s_k}$



EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

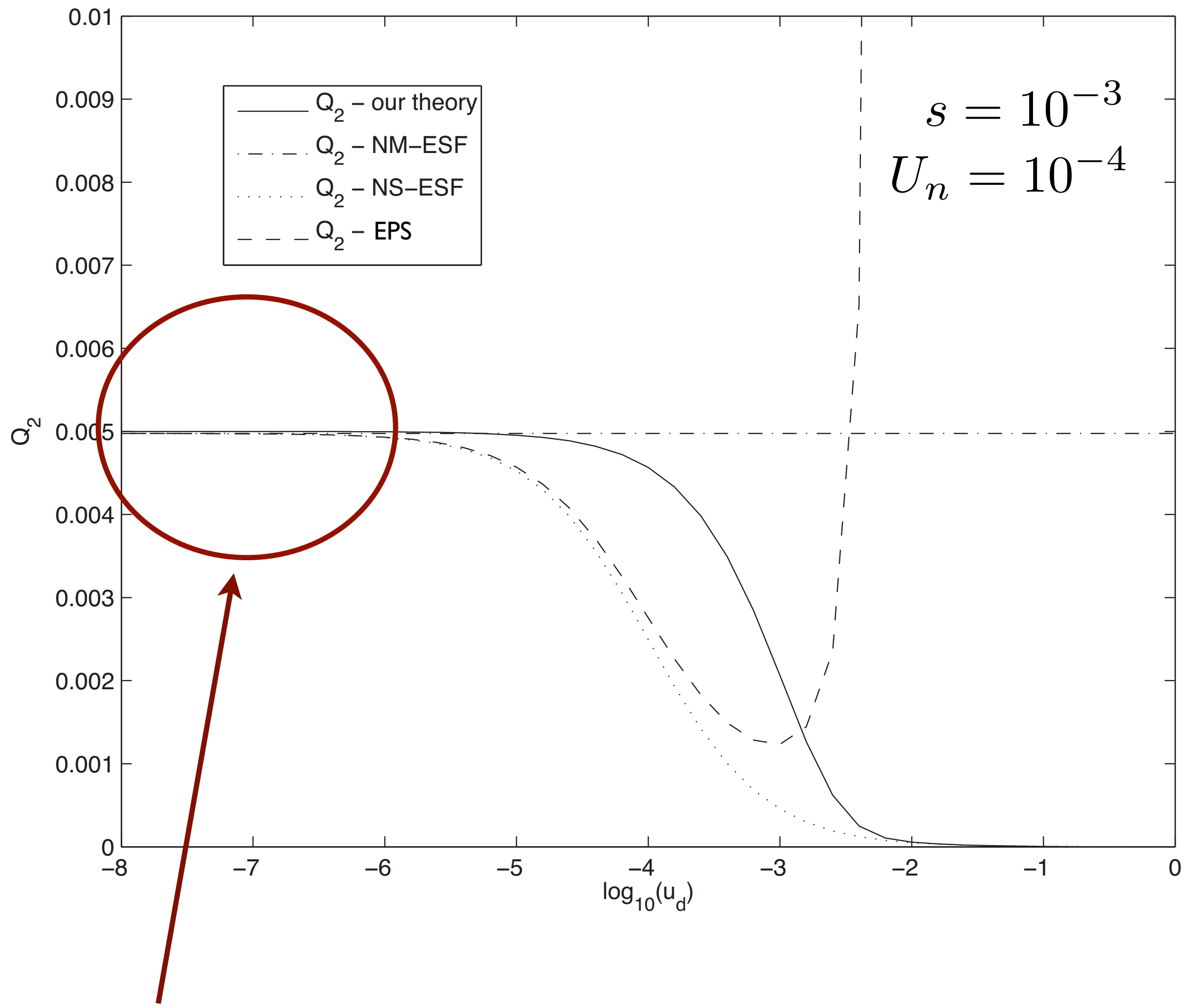
$$Q_2 = \frac{1}{\theta}$$

$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

Expected genetic variation

Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x) dx = \sum_{k=0}^{\infty} \frac{h_k}{2N s_k}$



neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

$$Q_2 = \frac{1}{\theta}$$

$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

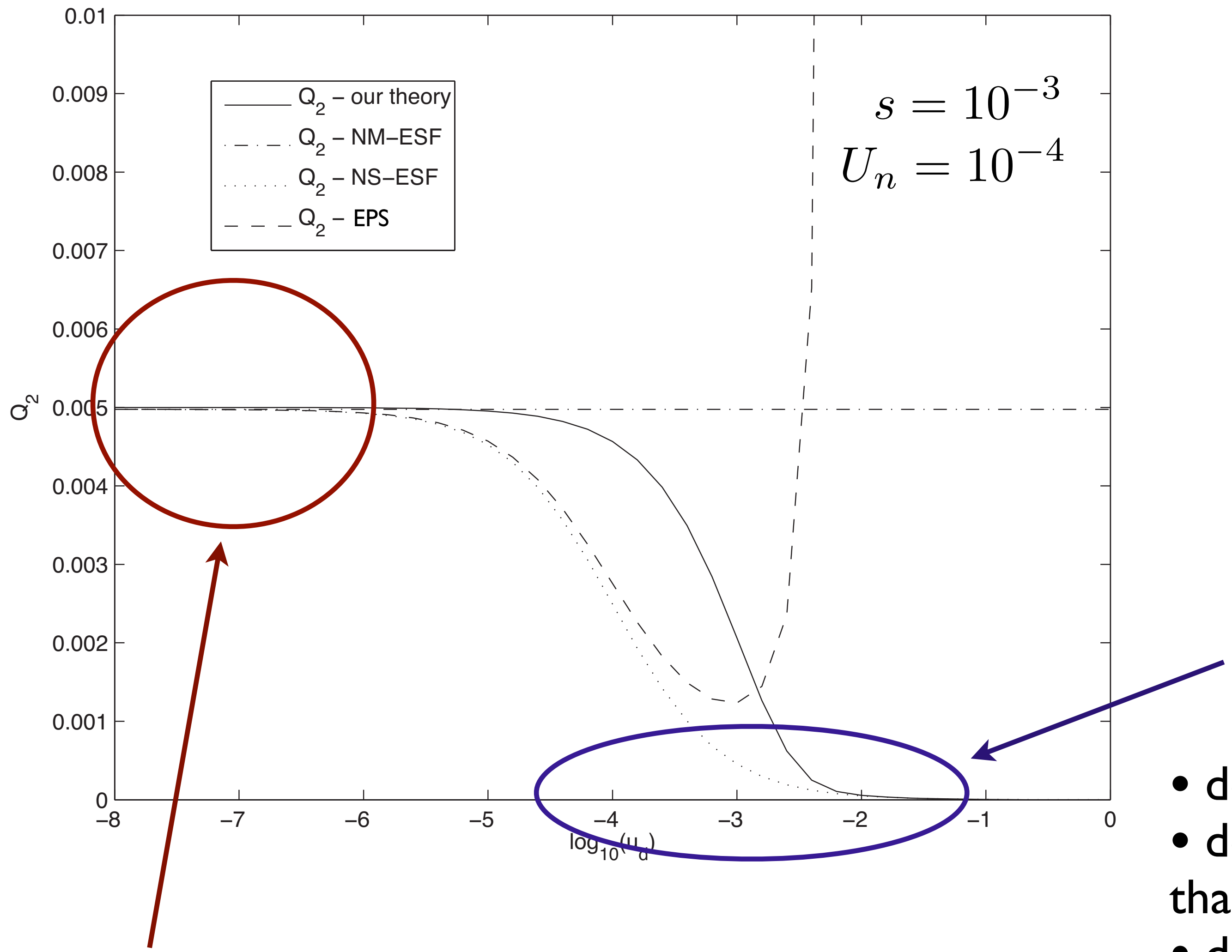
NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

- deleterious mutations decrease the homozygozity, $U_d \approx s$
- deleterious mutations decrease homozygosity less than neutral ones (they must eventually die)
- deleterious mutations are not rare for $U_d > s$, NM-ESF breaks down
- for $U_d > s$ still significant difference between NS-ESF and our results
- important parameter: are mutations purged slowly enough to matter $U_d \approx s$
- contrary to intuition from EPS, more deleterious mutations cannot decrease diversity

Expected genetic variation

Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$



EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

$$Q_2 = \frac{1}{\theta}$$

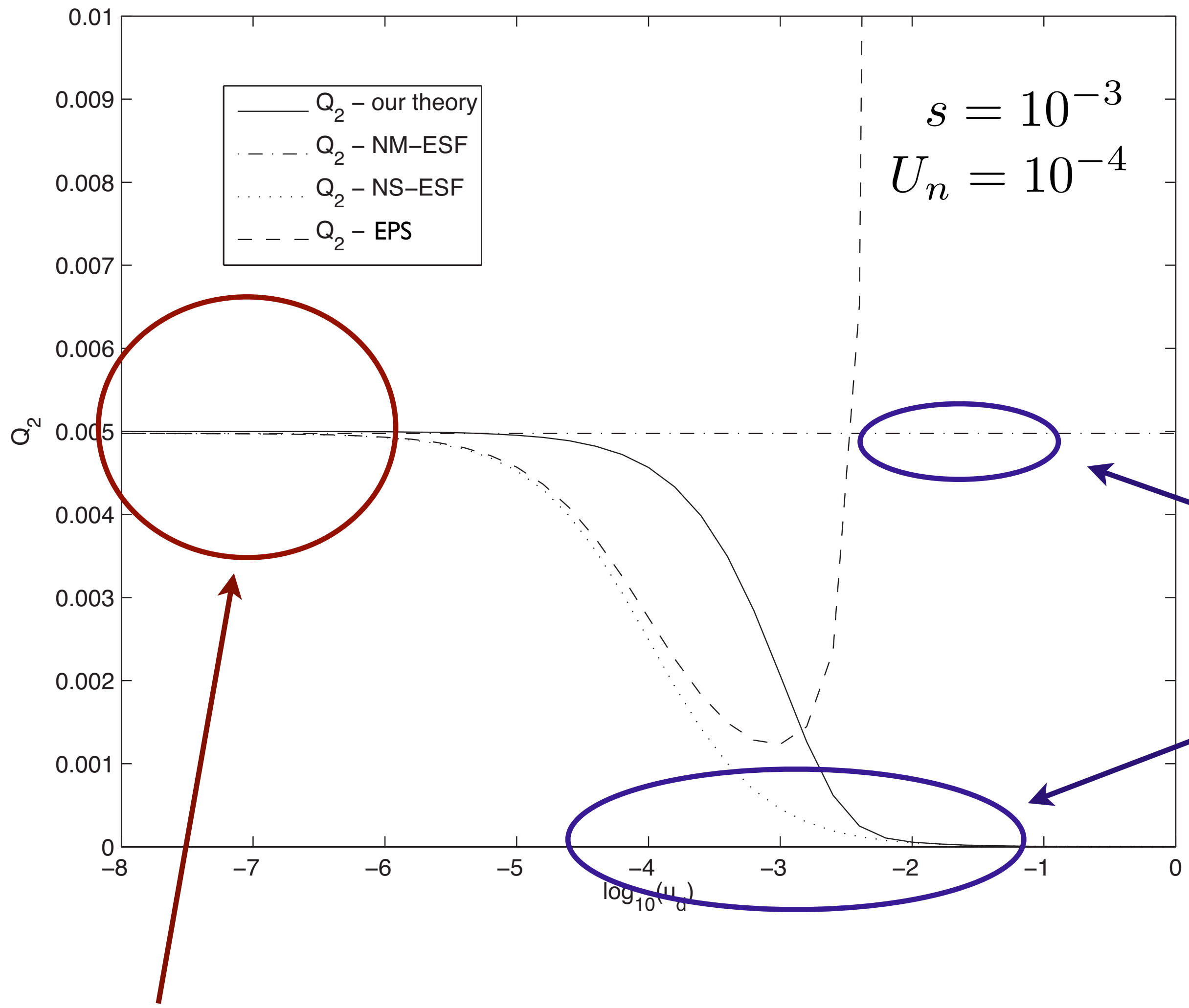
$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

- deleterious mutations decrease the homozygozity, $U_d \approx s$
- deleterious mutations decrease homozygosity less than neutral ones (they must eventually die)
- deleterious mutations are not rare for $U_d > s$, NM-ESF breaks down
- for $U_d > s$ still significant difference between NS-ESF and our results
- important parameter: are mutations purged slowly enough to matter $U_d \approx s$
- contrary to intuition from EPS, more deleterious mutations cannot decrease diversity

Expected genetic variation

Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$



EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

- deleterious mutations decrease the homozygozity, $U_d \approx s$
- deleterious mutations decrease homozygosity less than neutral ones (they must eventually die)
- deleterious mutations are not rare for $U_d > s$, NM-ESF breaks down
- for $U_d > s$ still significant difference between NS-ESF and our results
- important parameter: are mutations purged slowly enough to matter $U_d \approx s$
- contrary to intuition from EPS, more deleterious mutations cannot decrease diversity

neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

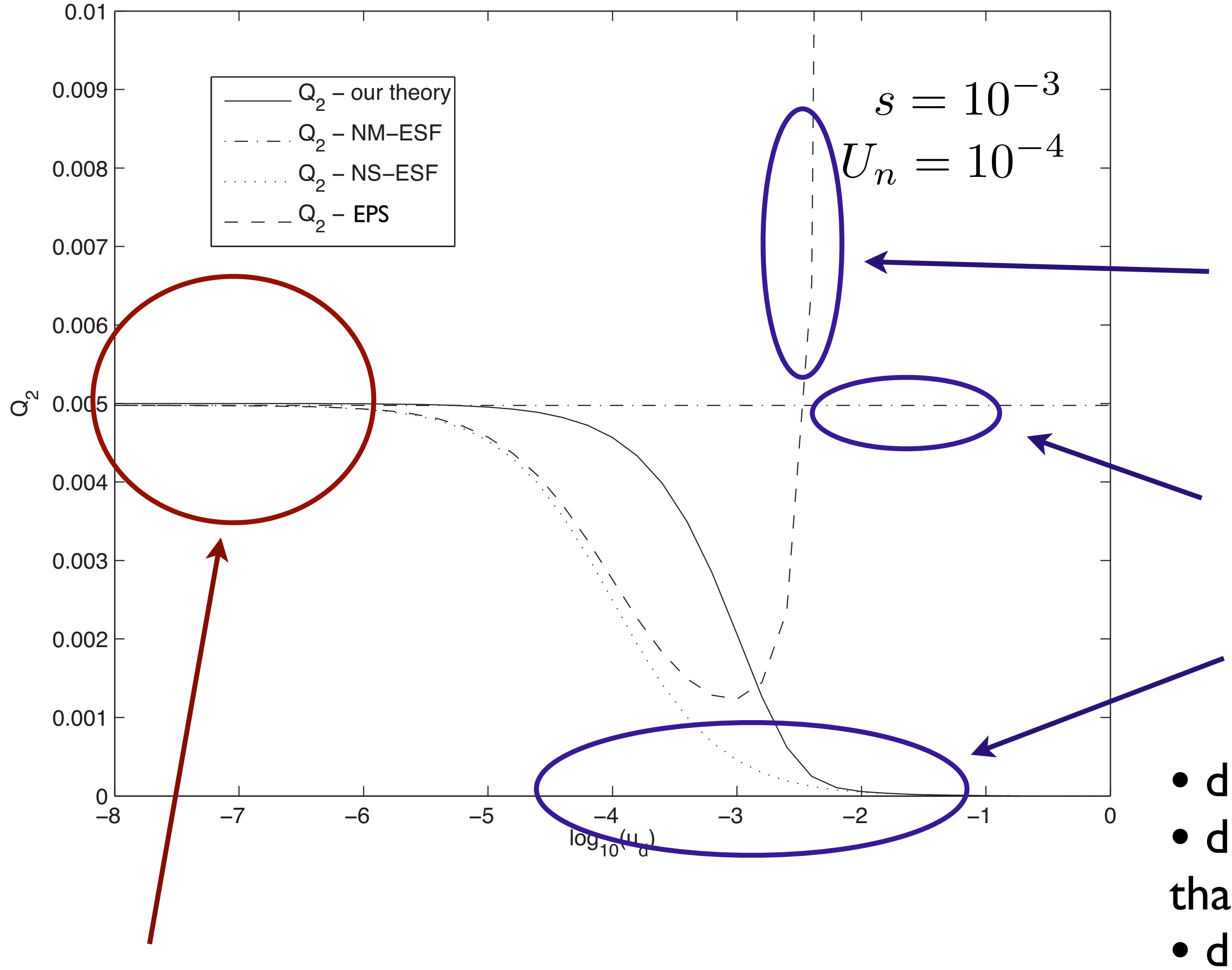
$$Q_2 = \frac{1}{\theta}$$

$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

Expected genetic variation

Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x) dx = \sum_{k=0}^{\infty} \frac{h_k}{2N s_k}$



EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

neutral case, $U_d = 0$: $\theta = 2NU_n >> 1$

$$Q_2 = \frac{1}{\theta}$$

$$Q_2^{ESF} = \frac{1}{1+\theta}$$

- all neutral models agree: ESF, BGS

- deleterious mutations decrease the homozygozity, $U_d \approx s$
- deleterious mutations decrease homozygosity less than neutral ones (they must eventually die)
- deleterious mutations are not rare for $U_d > s$, NM-ESF breaks down
- for $U_d > s$ still significant difference between NS-ESF and our results
- important parameter: are mutations purged slowly enough to matter $U_d \approx s$
- contrary to intuition from EPS, more deleterious mutations cannot decrease diversity

Expected genetic variation

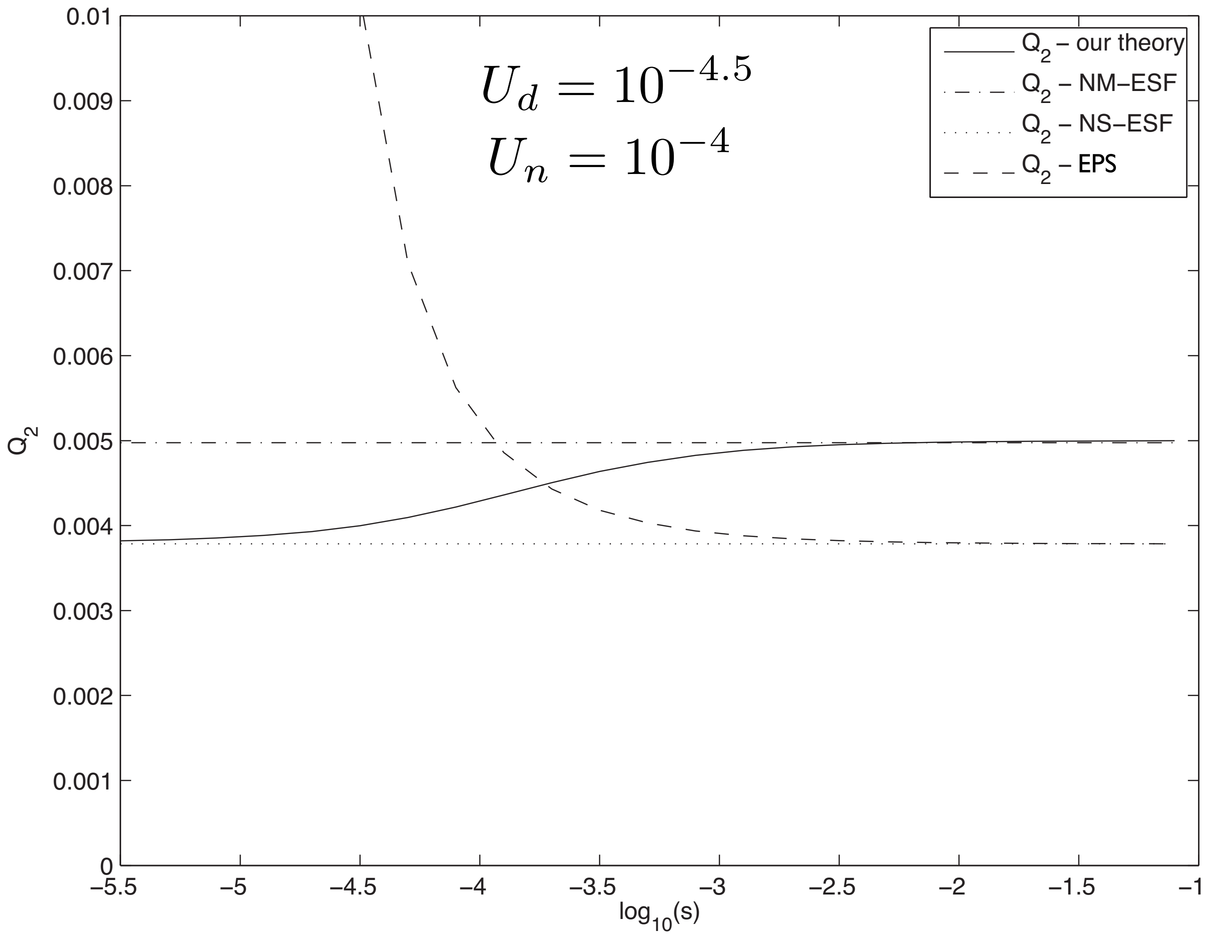
Homozygosity: $$Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

EPS - change in reduced effective population size of “neutral” population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$



Expected genetic variation

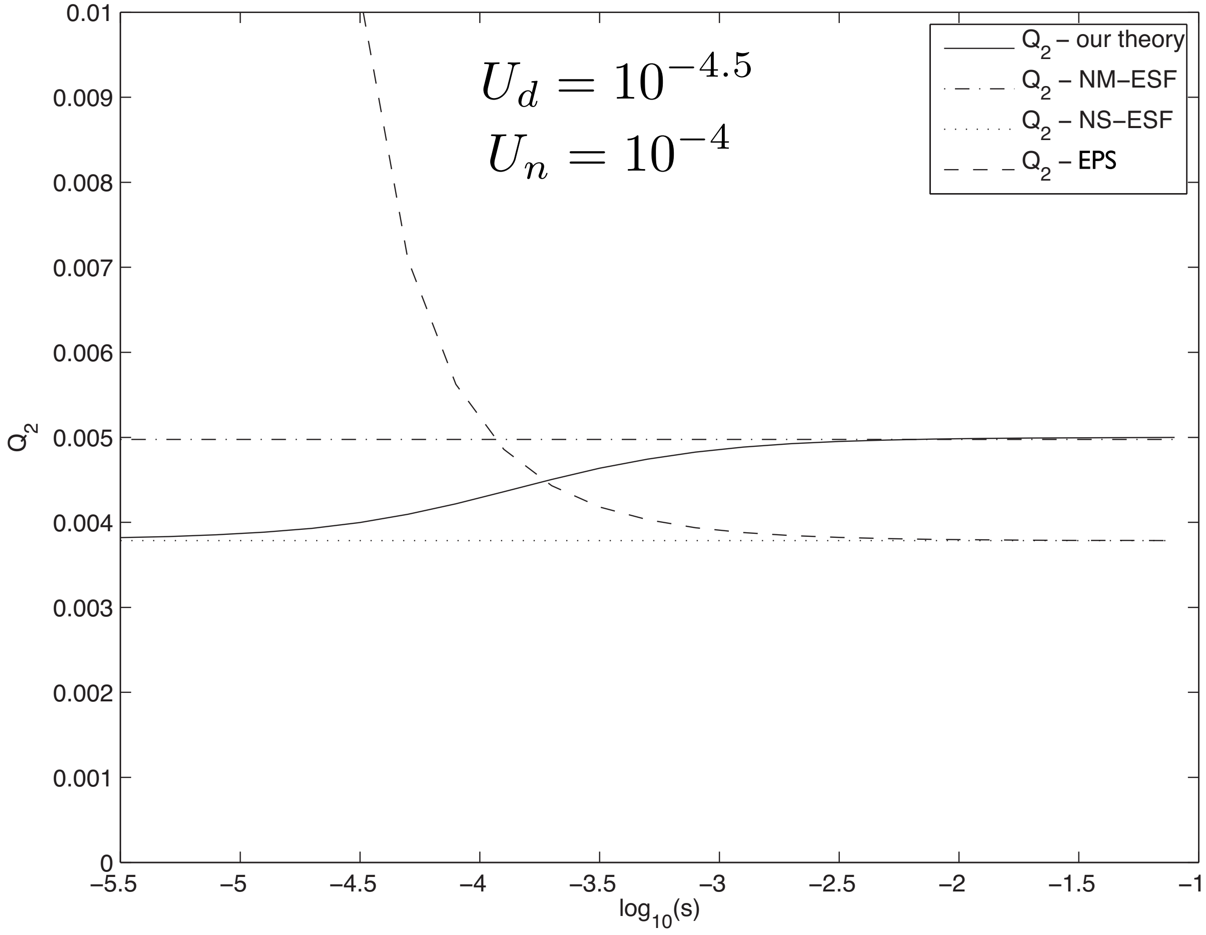
Homozygosity: $$Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$



- for strong selection mutations eliminated quickly - neutral mutations dominate - NM-ESF holds
- for very weak selection, deleterious mutations are like neutral - NS-ESF holds
- regions where no neutral theory holds
- EPS underestimates size of most fit for weak selection

Expected genetic variation

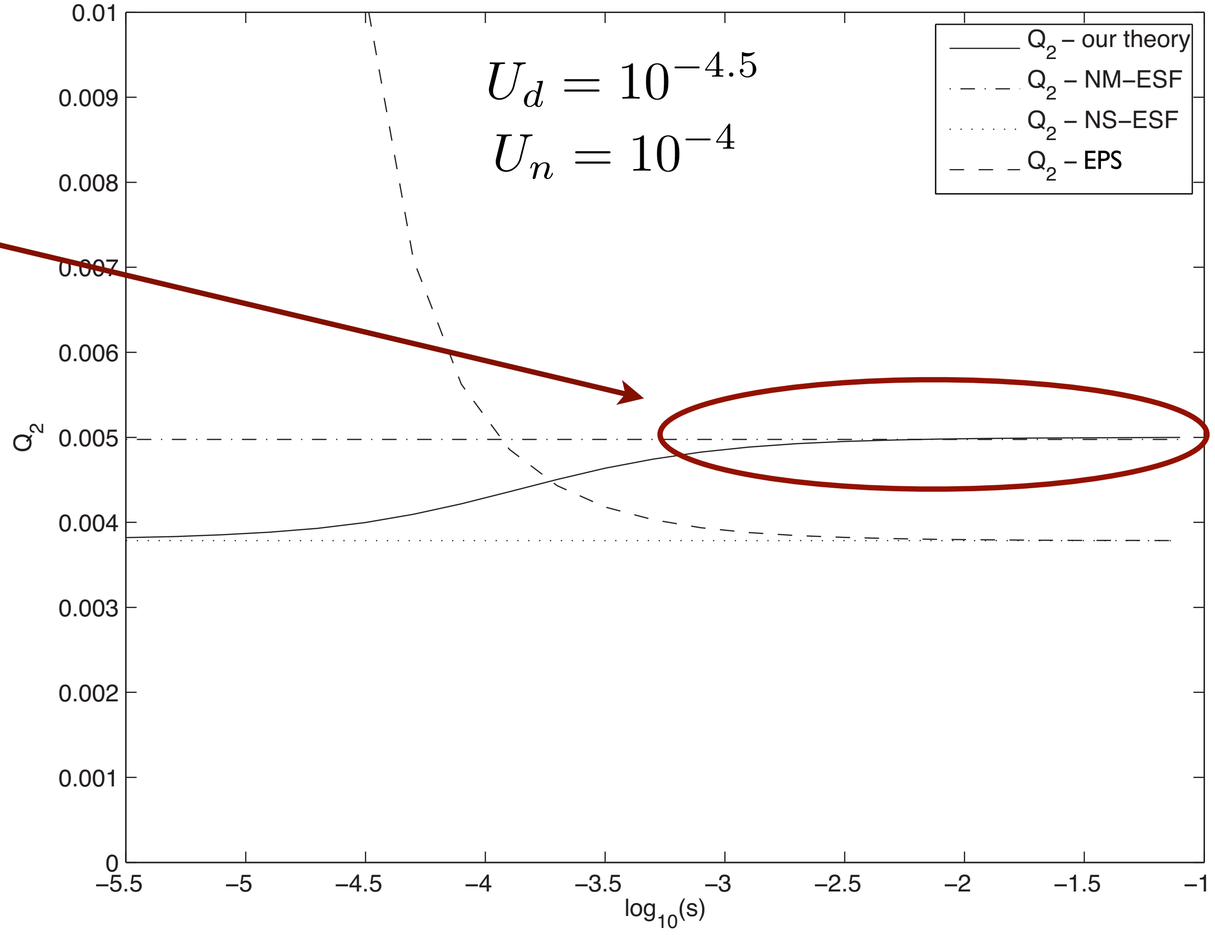
Homozygosity: $Q_2 = \sum_k \int x^2 f_k(x)dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$

EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$



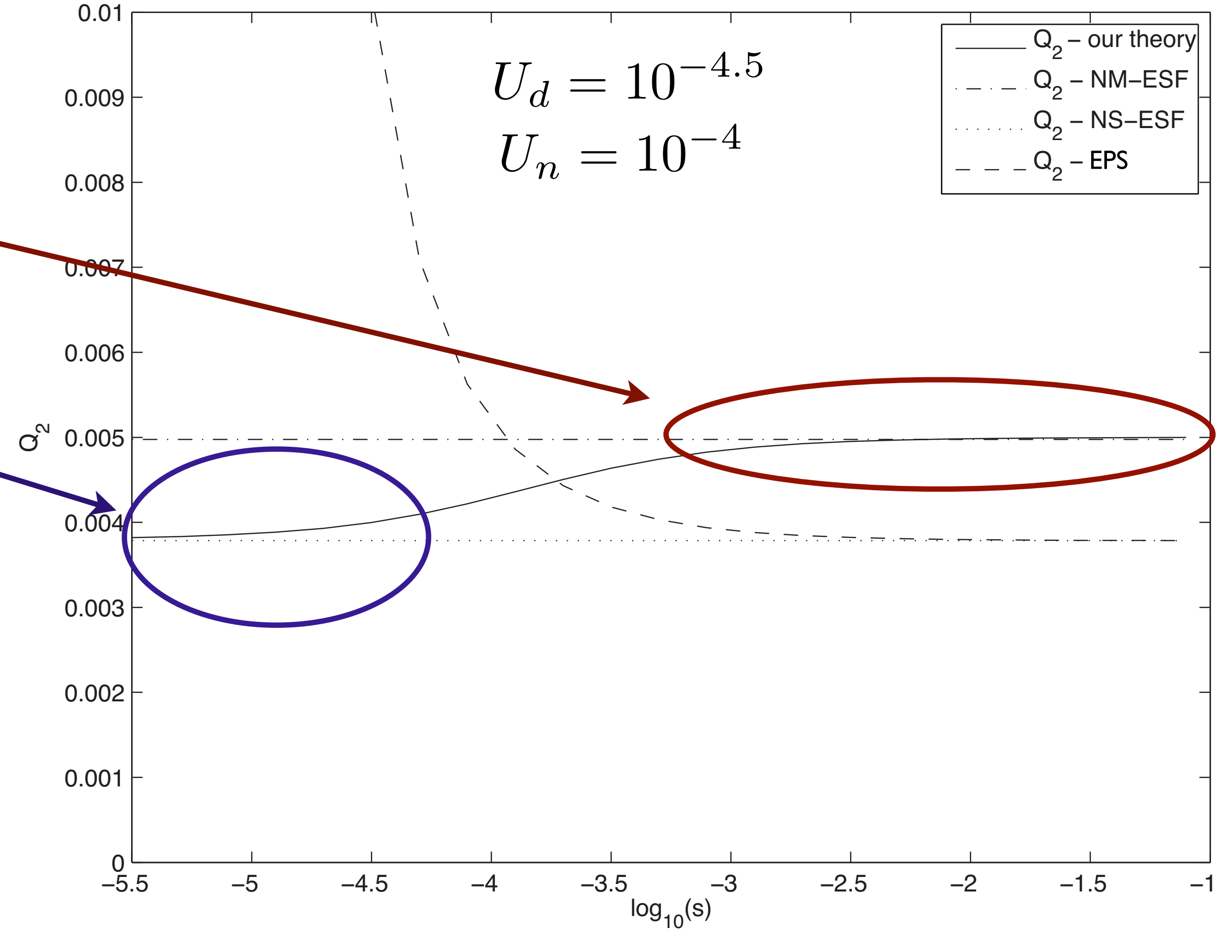
- for strong selection mutations eliminated quickly - neutral mutations dominate - NM-ESF holds
- for very weak selection, deleterious mutations are like neutral - NS-ESF holds
- regions where no neutral theory holds
- EPS underestimates size of most fit for weak selection

Expected genetic variation

Homozygosity: $$Q_2 = \sum_k \int x^2 f_k(x) dx = \sum_{k=0}^{\infty} \frac{h_k}{2Ns_k}$$

NM-ESF - neglect deleterious mutations: $\theta = 2NU_n$

NS-ESF - neglect selection against deleterious mutations: $\theta = 2N(U_n + U_d)$



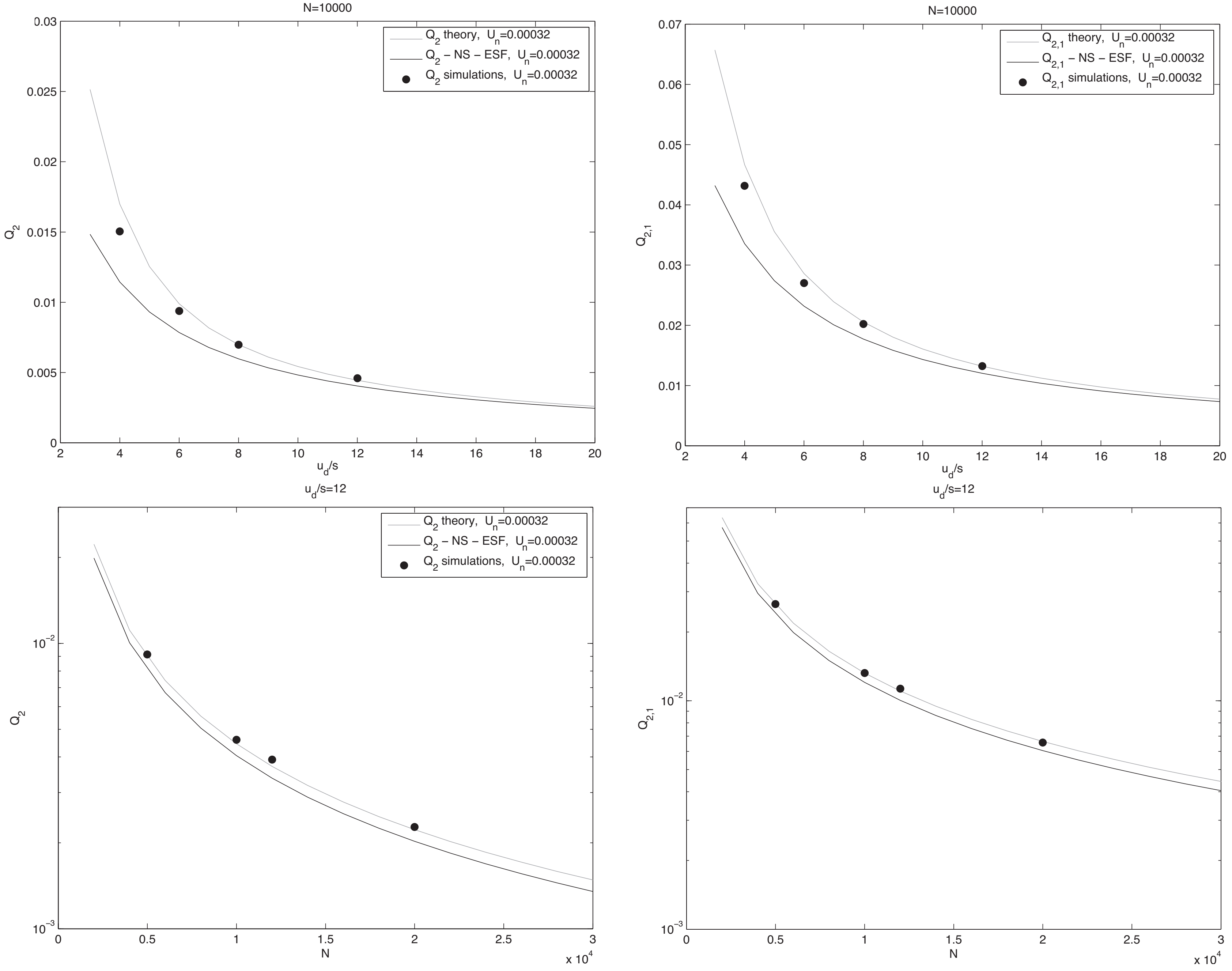
EPS - change in reduced effective population size of "neutral" population:

$$\theta = 2N(U_n + U_d)e^{-U_d/|s|}$$

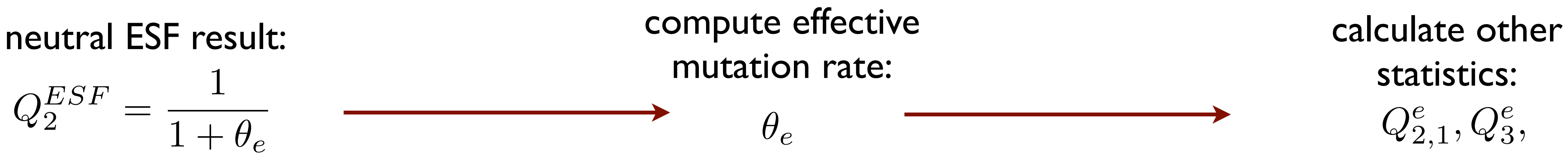
- for strong selection mutations eliminated quickly - neutral mutations dominate - NM-ESF holds
- for very weak selection, deleterious mutations are like neutral - NS-ESF holds
- regions where no neutral theory holds
- EPS underestimates size of most fit for weak selection

Comparison with simulations

- MC of Wright-Fisher population
- constant size N
- N individuals sampled with replacement in each generation
- sampling according to relative fitness in the population
- Poisson number of deleterious and neutral mutations introduced in each generation
- mutations introduced randomly and independently among individuals
- keep track of frequencies of all genotypes
- genotype - set of mutation sites



Statistics to describe deviation from neutrality



neutral ESF result:

$$Q_2^{ESF} = \frac{1}{1+\theta_e}$$

compute effective mutation rate:

θ_e

calculate other statistics:

$Q_{2,1}^e, Q_3^e,$

Compute expected $Q_{2,1}$ or Q_3 Given Q_2

Expected deviation from neutral ratio $Q_{2,1}^e / Q_{2,1}$ Q_3^e / Q_3

Statistics to describe deviation from neutrality

neutral ESF result:

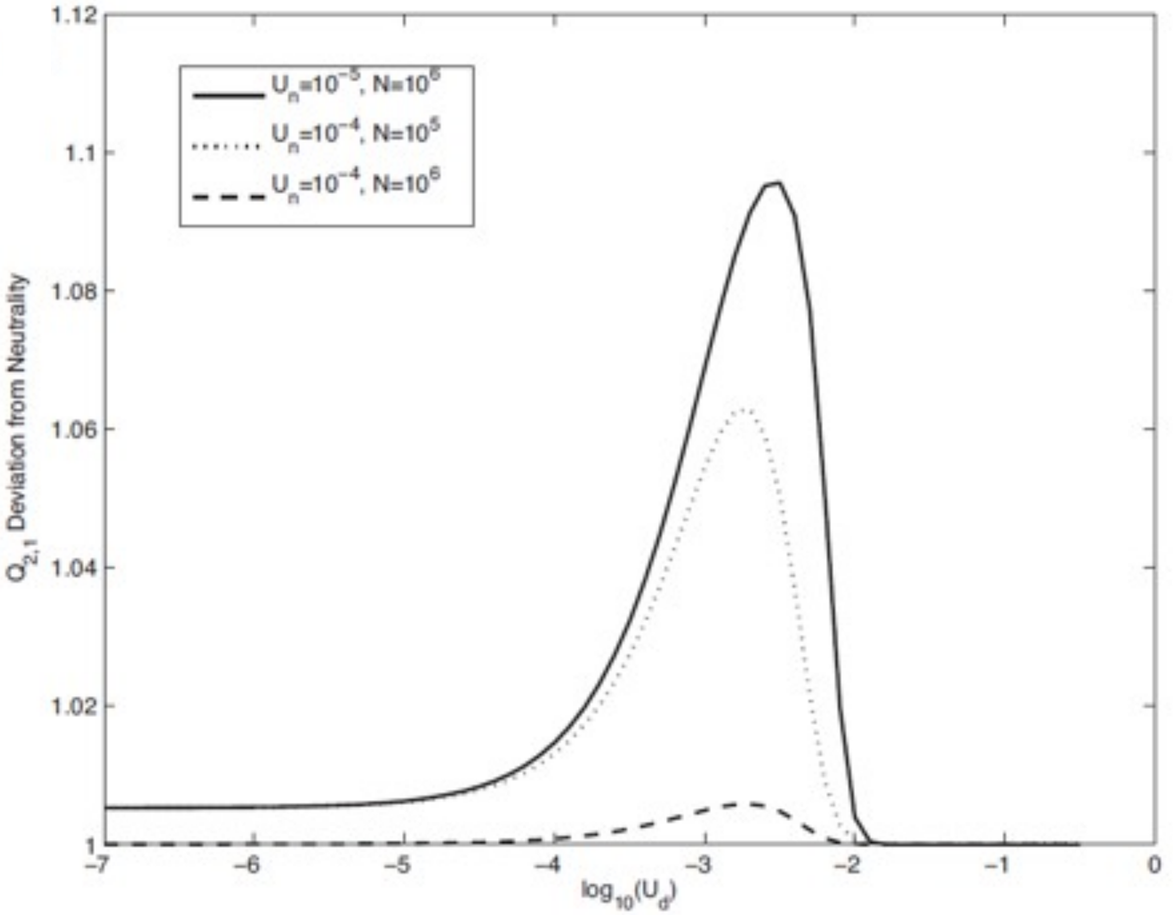
$$Q_2^{ESF} = \frac{1}{1+\theta_e}$$

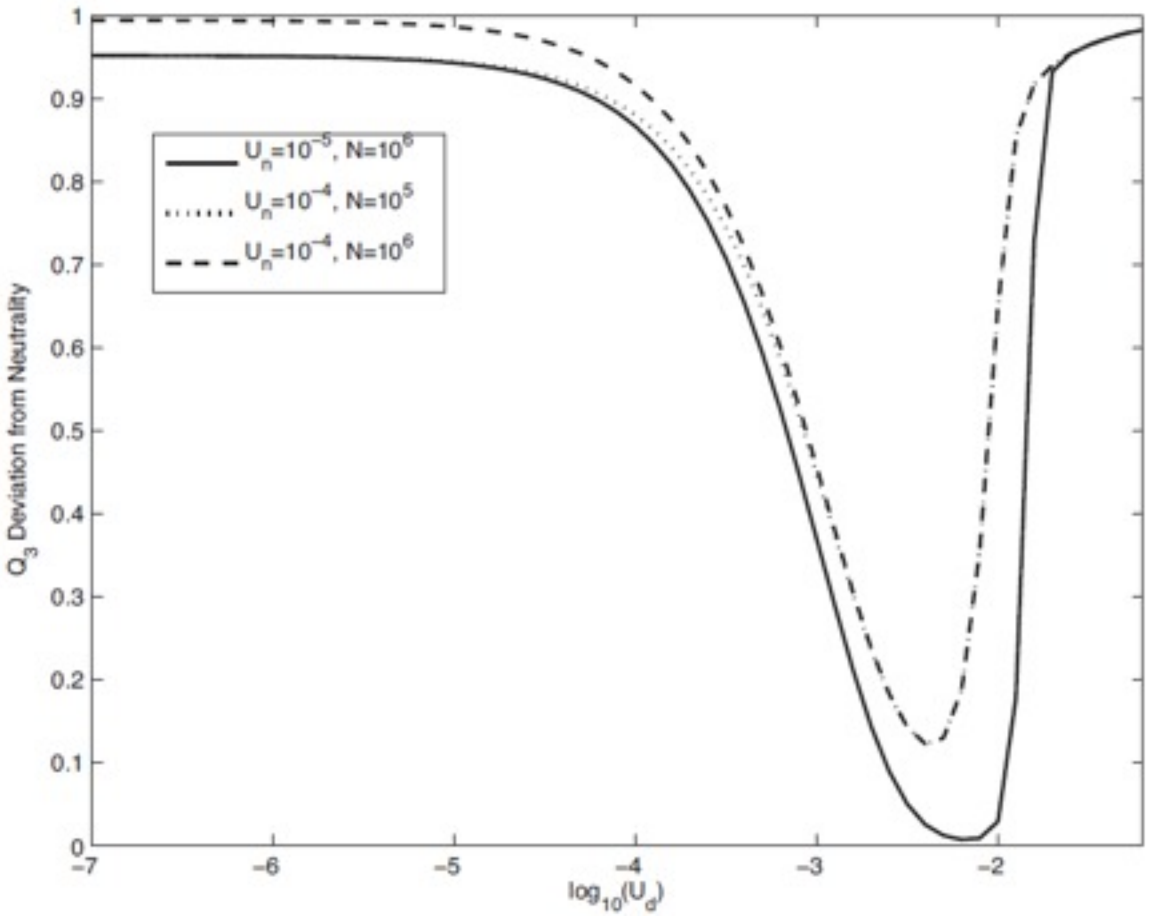
compute effective mutation rate:

$$\theta_e$$

calculate other statistics:

$$Q_{2,1}^e, Q_3^e,$$





Compute expected $Q_{2,1}$ or Q_3 Given Q_2

Expected deviation from neutral ratio $Q_{2,1}^e/Q_{2,1}$ Q_3^e/Q_3

Statistics to describe deviation from neutrality

neutral ESF result:

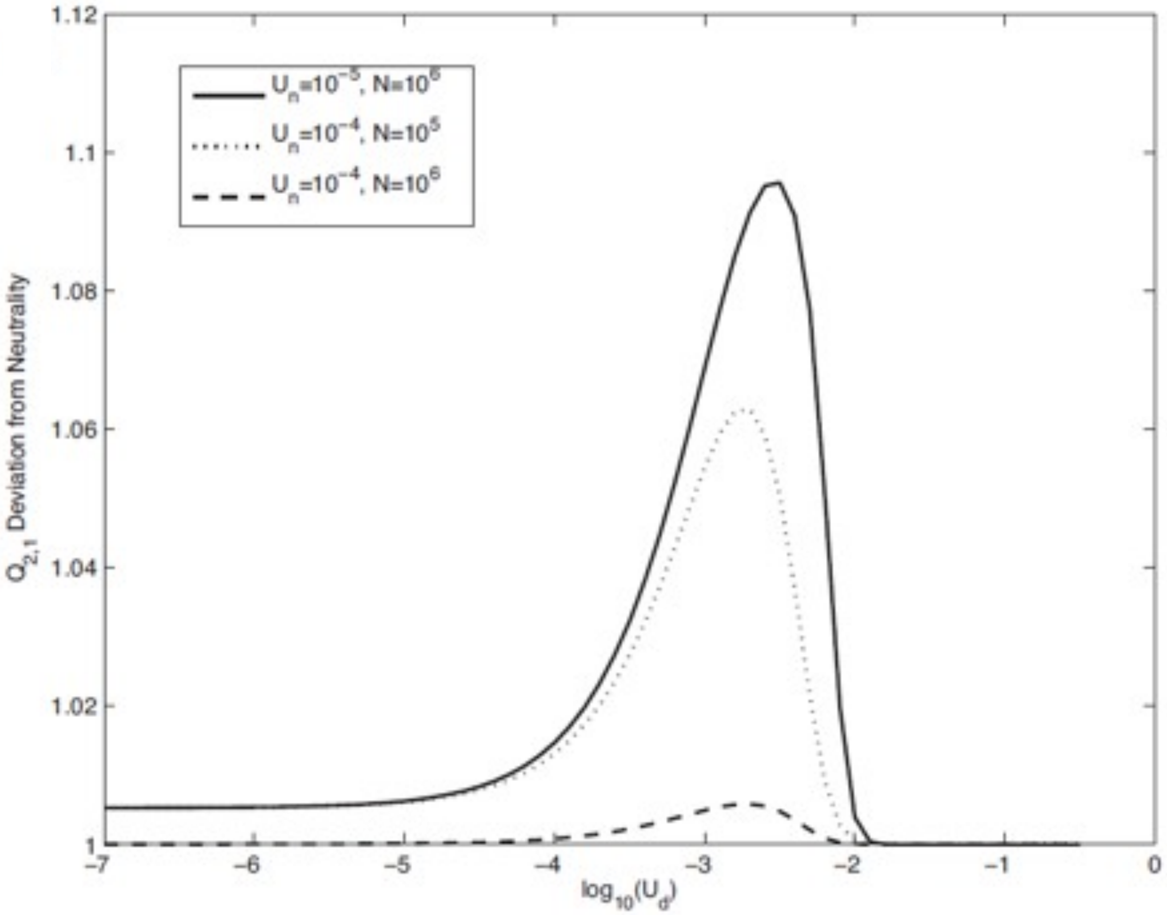
$$Q_2^{ESF} = \frac{1}{1+\theta_e}$$

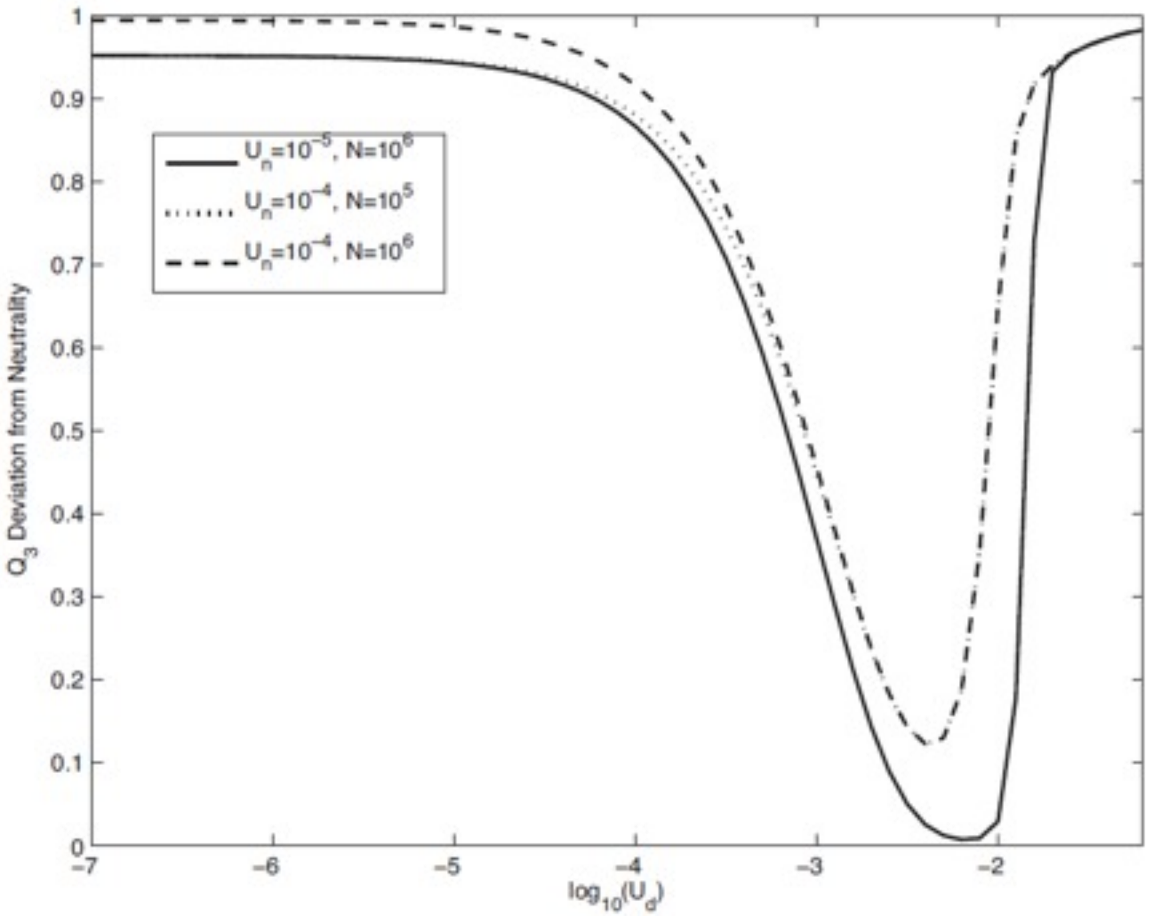
compute effective mutation rate:

$$\theta_e$$

calculate other statistics:

$$Q_{2,1}^e, Q_3^e,$$



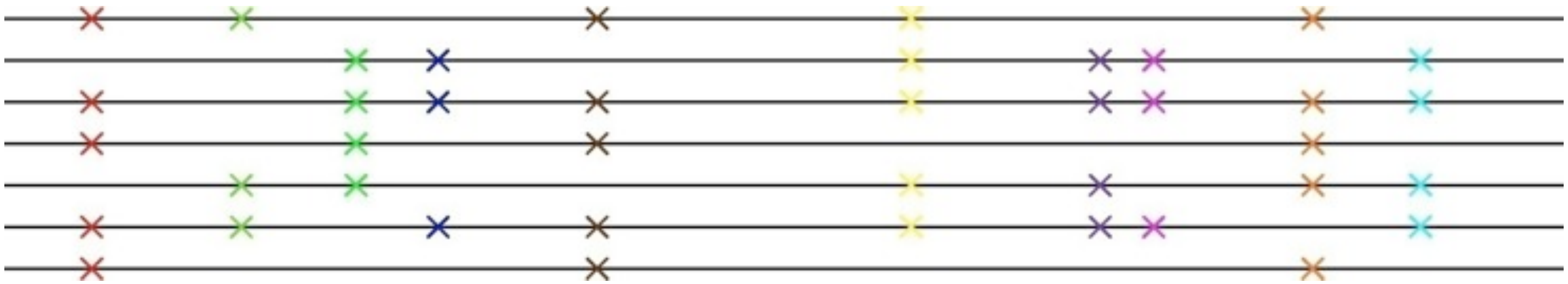


Compute expected $Q_{2,1}$ or Q_3 Given Q_2

Expected deviation from neutral ratio $Q_{2,1}^e/Q_{2,1}$ Q_3^e/Q_3

There is no effective population size that reproduces the statistics consistently

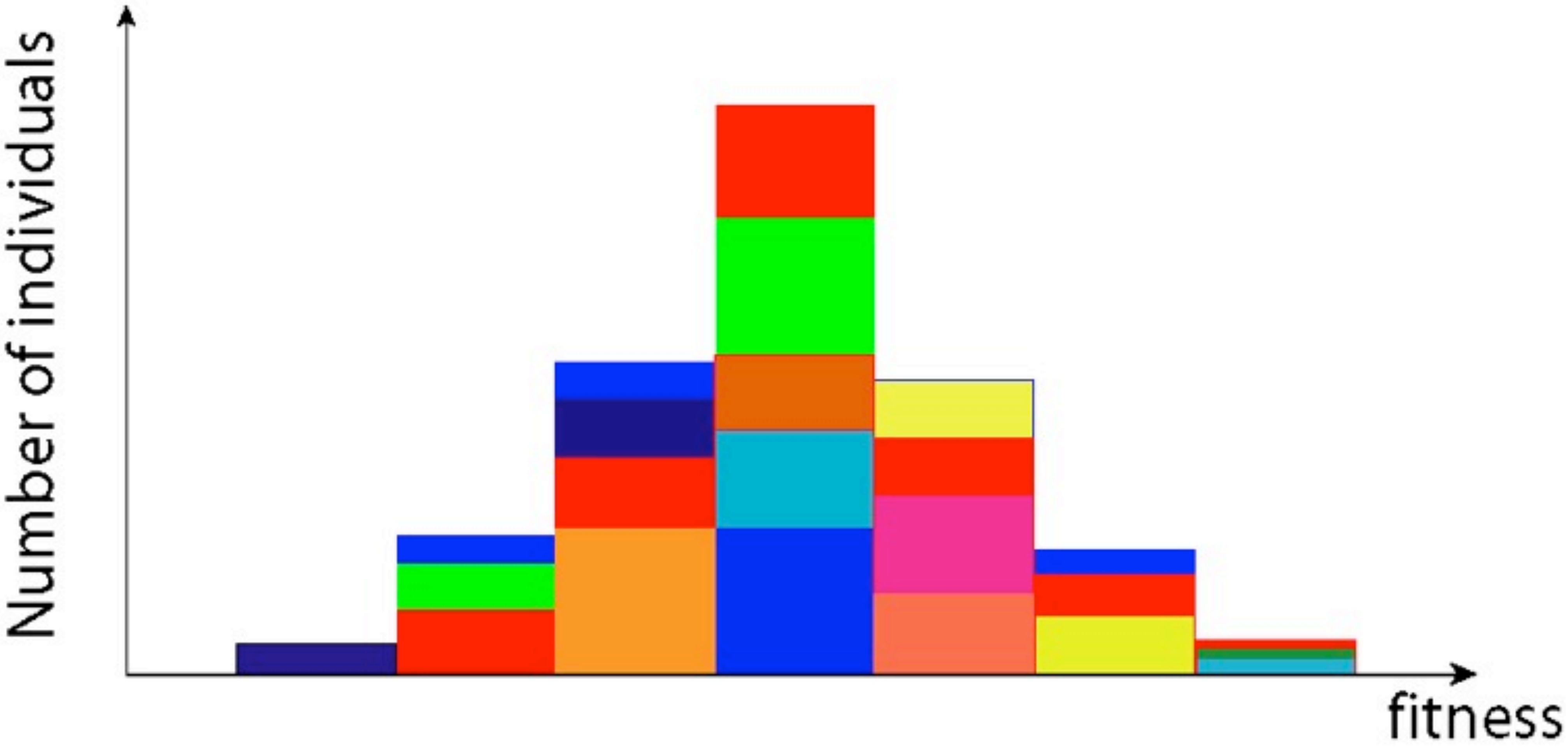
Tracing the genealogies



We now know the probability of different allelic configurations

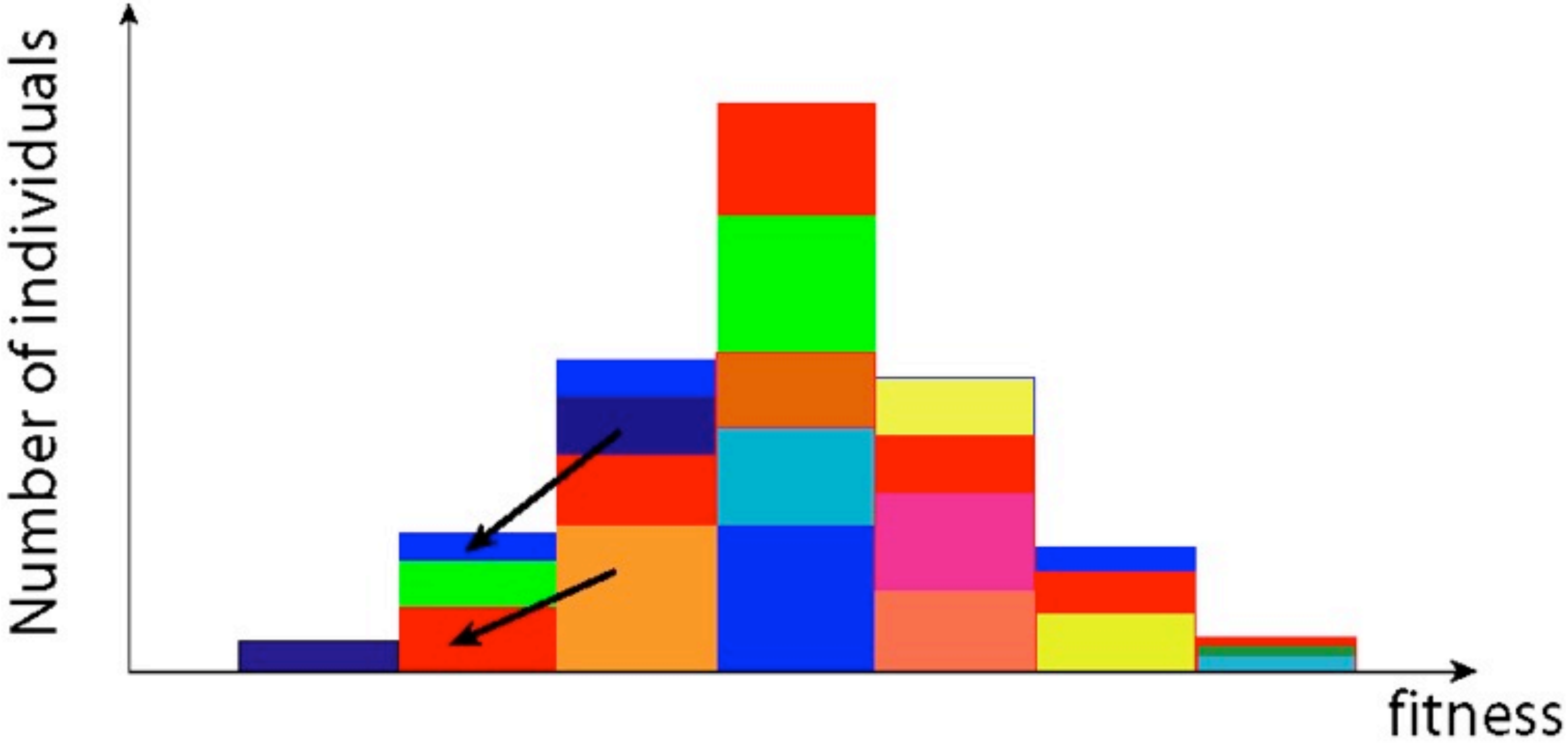
What is the relationship among alleles?

An effective coalescent approach



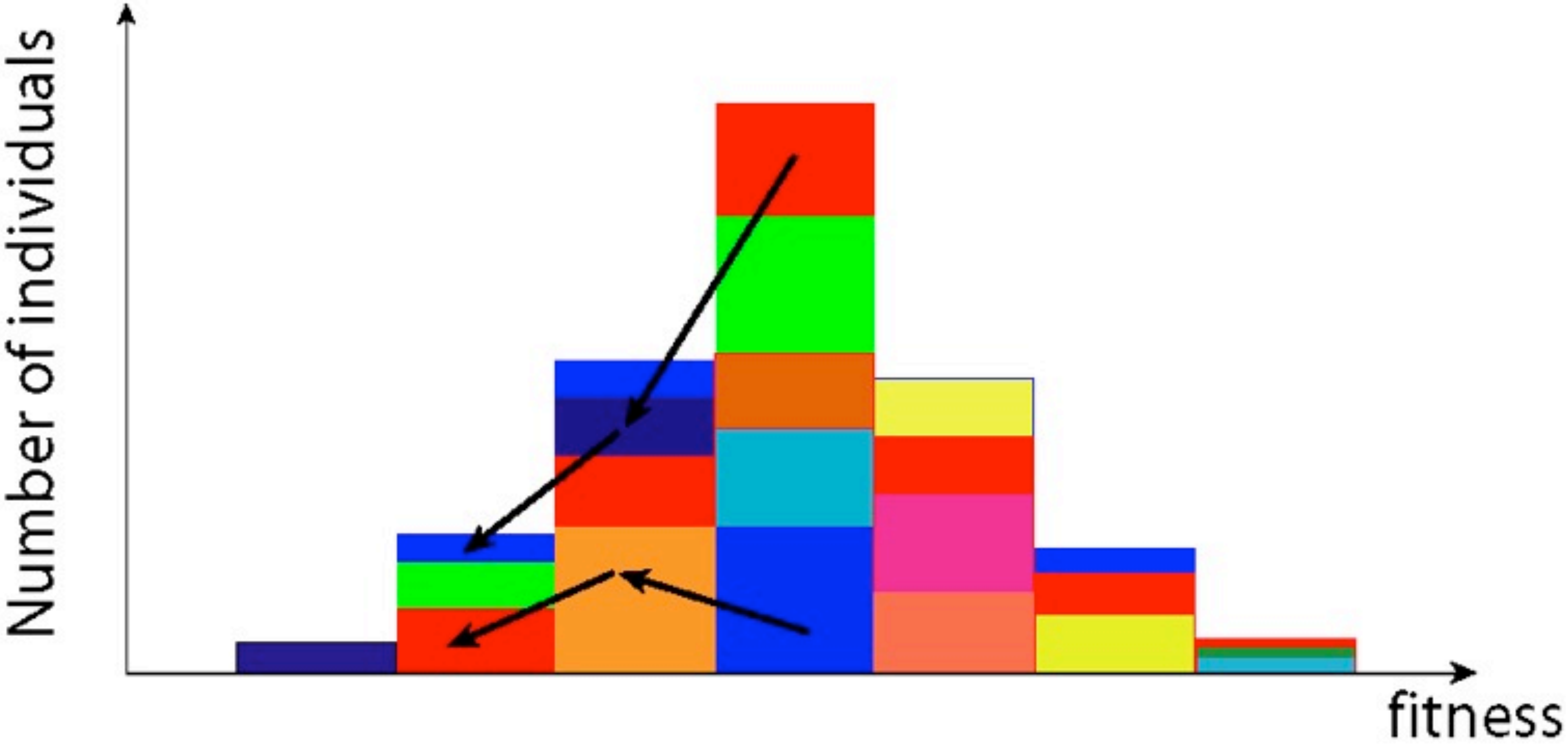
Trace the ancestry of each individual through the fitness distribution

An effective coalescent approach



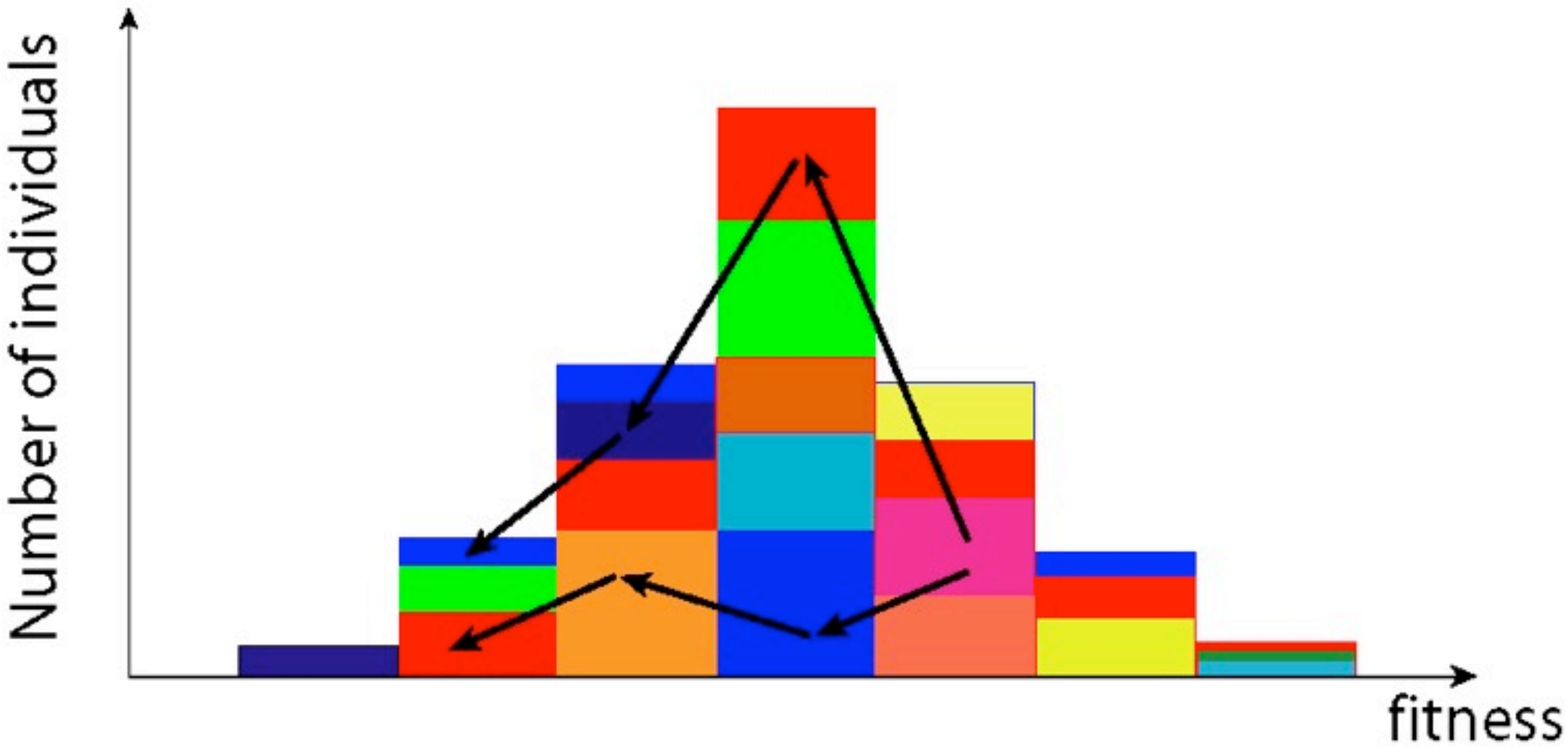
Trace the ancestry of each individual through the fitness distribution

An effective coalescent approach



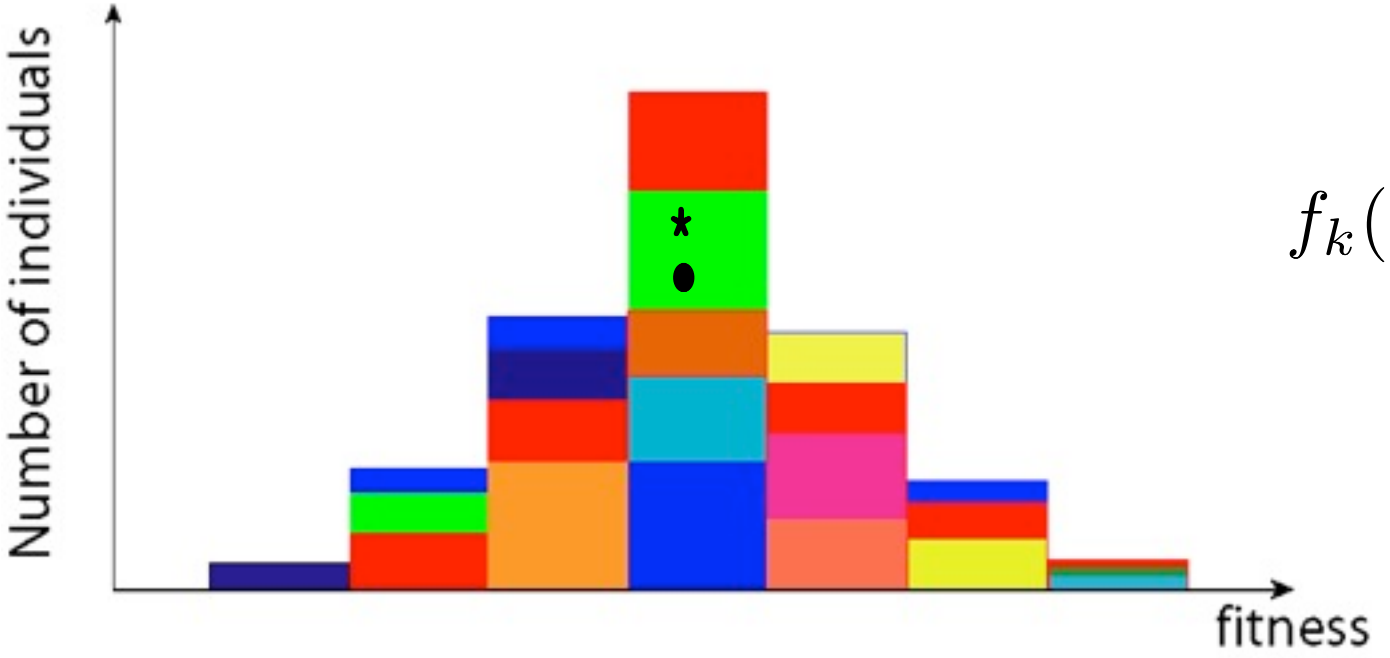
Trace the ancestry of each individual through the fitness distribution

An effective coalescent approach



Trace the ancestry of each individual through the fitness distribution

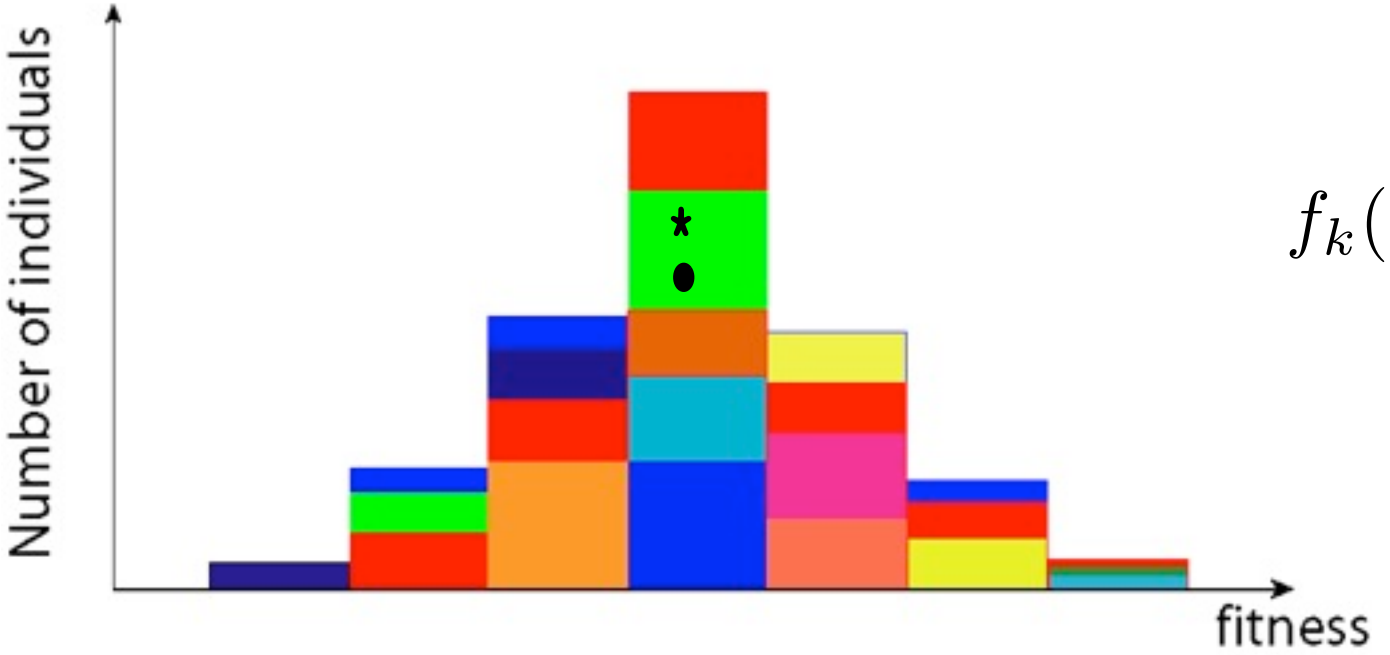
Effective coalescent probabilities



$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)}dx$$

Sample 2 individuals from class k

Effective coalescent probabilities

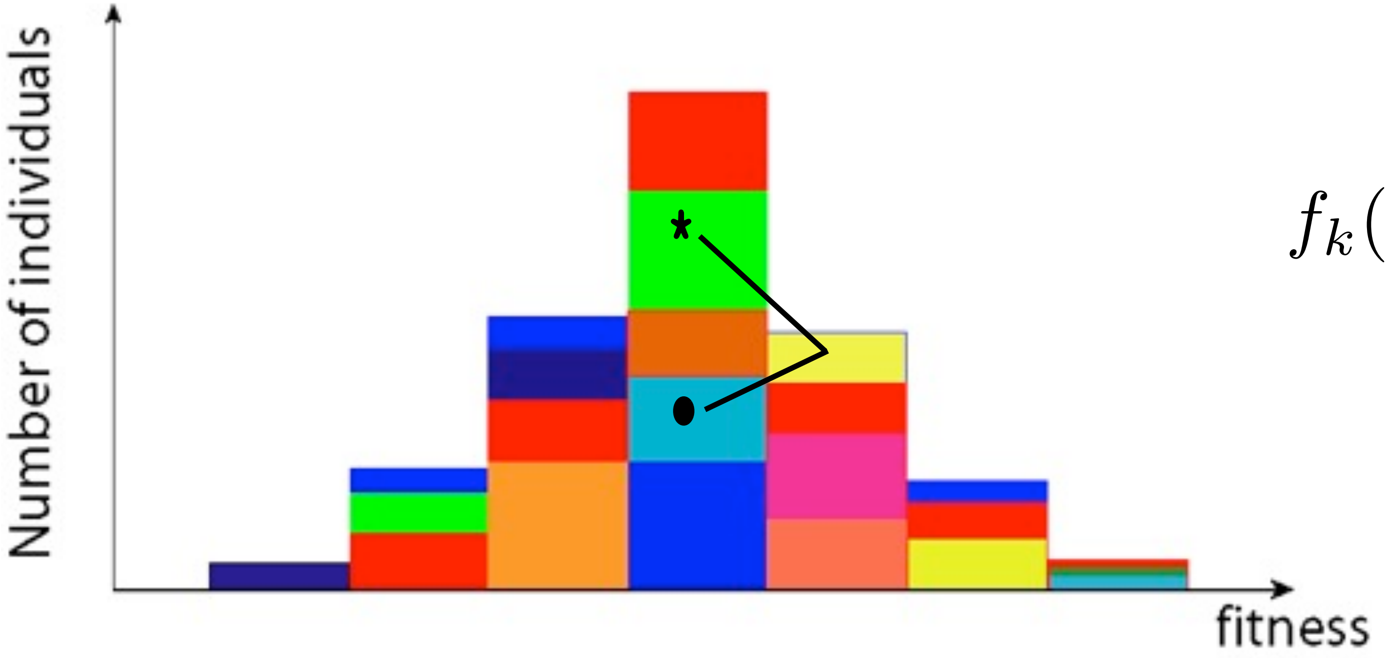


$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)}dx$$

Sample 2 individuals from class k

Coalescent probability in class k: $P_c^{k,k\rightarrow k} = \int \frac{x^2}{h_k^2} f_k(x)dx$

Effective coalescent probabilities



$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)}dx$$

Sample 2 individuals from class k

Coalescent probability in class k: $P_c^{k,k\to k} = \int \frac{x^2}{h_k^2} f_k(x)dx$

Coalescent probability in class k-1:

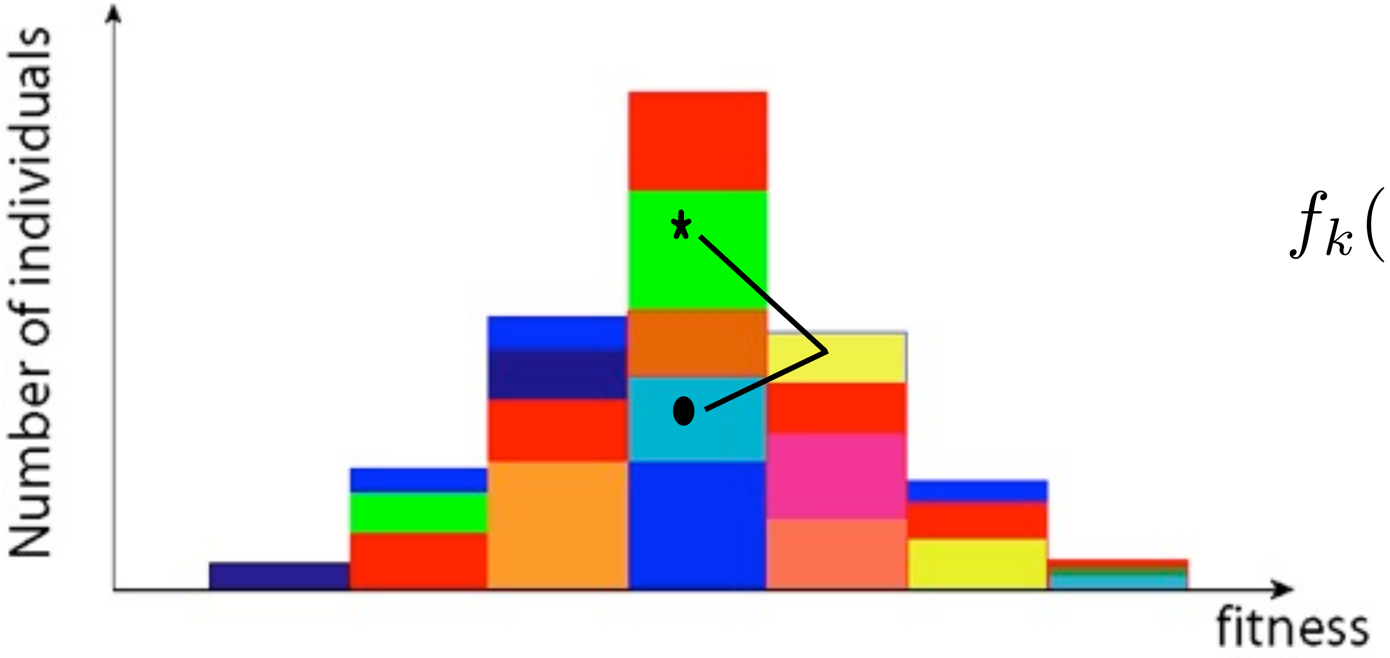
$$P_c^{k,k\to k-1} = \int \frac{x f_{k-1}}{h_{k-1}} \frac{y G_{k-1}(y \to x, |t_2 - t_1|)}{h_{k-1}} Q_{k,k}^{k-1}(t_1, t_2) dx dy dt_1 dt_2$$

probability an individual comes from class k and lineage with frequency x

probability that a lineage in class k-1 changes in frequency from x to y in time lt_2-t_1l

joint distribution of times t_1 and t_2 - times when lineages in class k where founded by mutations

Effective coalescent probabilities



$$f_k(x)dx = \theta_k \frac{1-e^{-2Ns_k(1-x)}}{(1-e^{-2Ns_k})x(1-x)}dx$$

Sample 2 individuals from class k

Coalescent probability in class k: $P_c^{k,k\to k} = \int \frac{x^2}{h_k^2} f_k(x)dx$

Coalescent probability in class k-1:

$$P_c^{k,k\to k-1} = \int \frac{x f_{k-1}}{h_{k-1}} \frac{y G_{k-1}(y \to x, |t_2 - t_1|)}{h_{k-1}} Q_{k,k}^{k-1}(t_1, t_2) dx dy dt_1 dt_2$$

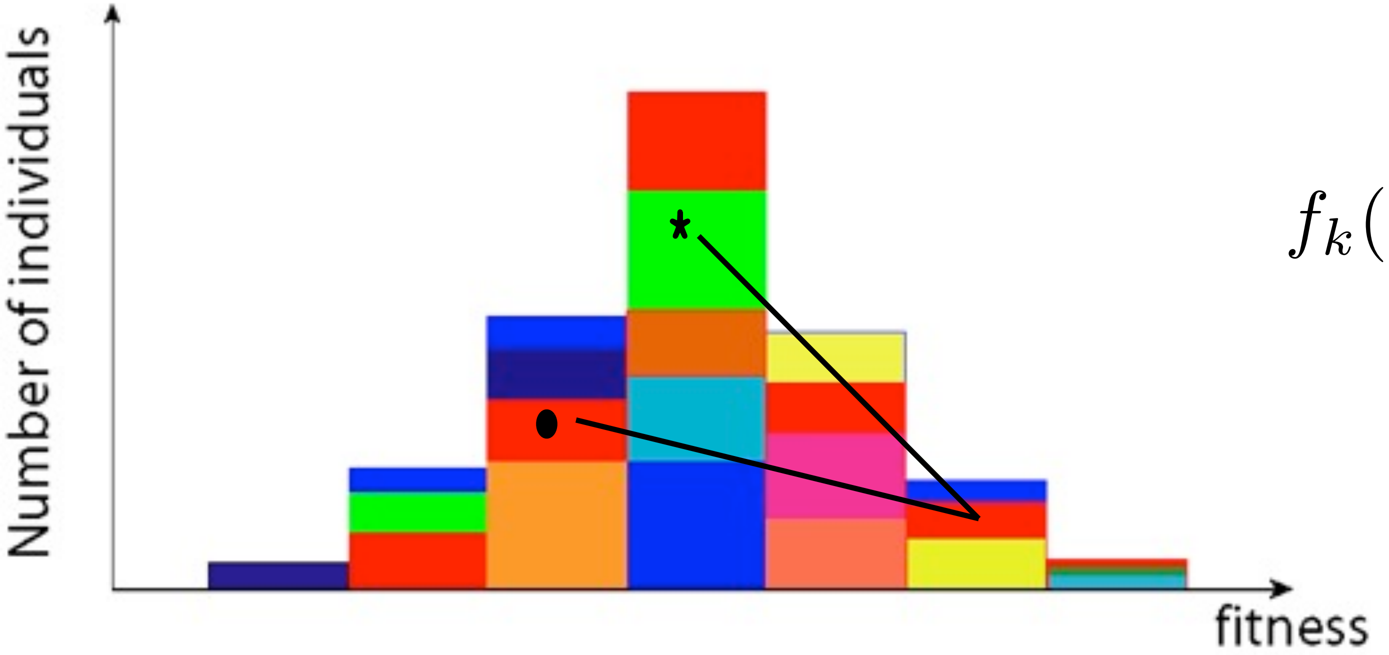
probability an individual comes from class k and lineage with frequency x

probability that a lineage in class k-1 changes in frequency from x to y in time lt_2-t_1l

joint distribution of times t_1 and t_2 - times when lineages in class k where founded by mutations

$$P_c^{k,k\to k-2} = \dots$$

Effective coalescent probabilities



$$f_k(x)dx = \theta_k \frac{1 - e^{-2Ns_k(1-x)}}{(1 - e^{-2Ns_k})x(1-x)}dx$$

Sample 2 individuals from class k

Coalescent probability in class k: $P_c^{k,k\to k} = \int \frac{x^2}{h_k^2} f_k(x)dx$

General coalescent probability in class k-ℓ:

$$P_c^{k,k+m\to k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \to x, |t_2 - t_1|)}{h_{k-l}} Q^{k-\ell}_{k,k+m}(t_1, t_2) dx dy dt_1 dt_2$$

probability an individual comes from class k and lineage with frequency x

probability that a lineage in class k-ℓ changes in frequency from x to y in time $|t_2-t_1|$

joint distribution of times t_1 and t_2 - times when lineages in class k where founded by mutations

Non-conditional approximation

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{xf_{k-\ell}}{h_{k-l}} \frac{yG_{k-\ell}(y \rightarrow x, |t_2 - t_1|)}{h_{k-l}} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dy dt_1 dt_2$$

Non-conditional approximation

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \rightarrow x, |t_2 - t_1|)}{h_{k-l}} Q^{k-\ell}_{k,k+m}(t_1, t_2) dx dy dt_1 dt_2$$

y integral is just mean y - deterministic result for the change in the frequency of the lineage

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h^2_{k-l}} e^{-s(k-\ell)|t_2-t_1|} Q^{k-\ell}_{k,k+m}(t_1, t_2) dx dt_1 dt_2$$

Non-conditional approximation

$$P_c^{k,k+m\to k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \to x, |t_2 - t_1|)}{h_{k-l}} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dy dt_1 dt_2$$

y integral is just mean y - deterministic
result for the change in the frequency of
the lineage

$$P_c^{k,k+m\to k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dt_1 dt_2$$

Assume:

- non-conditional approximation: the times at which the two individuals moved from one fitness class to another is independent

Non-conditional approximation

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \rightarrow x, |t_2 - t_1|)}{h_{k-l}} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dy dt_1 dt_2$$

y integral is just mean y - deterministic result for the change in the frequency of the lineage

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dt_1 dt_2$$

Assume:

- non-conditional approximation: the times at which the two individuals moved from one fitness class to another is independent

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_{k}^{k-\ell}(t_1) Q_{k+m}^{k-\ell}(t_2) dx dt_1 dt_2$$

[generally not true because moving between fitness classes assumes no coalescence - but small correction]

Non-conditional approximation

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \rightarrow x, |t_2 - t_1|)}{h_{k-l}} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dy dt_1 dt_2$$

y integral is just mean y - deterministic result for the change in the frequency of the lineage

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dt_1 dt_2$$

Assume:

- non-conditional approximation: the times at which the two individuals moved from one fitness class to another is independent

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_k^{k-\ell}(t_1) Q_{k+m}^{k-\ell}(t_2) dx dt_1 dt_2$$

[generally not true because moving between fitness classes assumes no coalescence - but small correction]

+

distribution of mutant timings:

$$Q_k^{k-\ell}(t) = Q_k^{k-1}(t) * Q_k^{k-2}(t) * \ldots Q_{k-\ell+1}^{k-\ell}(t) \quad \text{and} \quad Q_{k-\ell+1}^{k-\ell}(t) = s(k-\ell+1)e^{-s(k-\ell+1)t}$$

Non-conditional approximation

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \rightarrow x, |t_2 - t_1|)}{h_{k-l}} Q^{k-\ell}_{k,k+m}(t_1,t_2)\, dx\, dy\, dt_1\, dt_2$$

y integral is just mean y - deterministic result for the change in the frequency of the lineage

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h^2_{k-l}} e^{-s(k-\ell)|t_2-t_1|} Q^{k-\ell}_{k,k+m}(t_1,t_2)\, dx\, dt_1\, dt_2$$

Assume:

- non-conditional approximation: the times at which the two individuals moved from one fitness class to another is independent

$$P_c^{k,k+m\rightarrow k-\ell} = \int \frac{x^2 f_{k-\ell}}{h^2_{k-l}} e^{-s(k-\ell)|t_2-t_1|} Q^{k-\ell}_{k}(t_1) Q^{k-\ell}_{k+m}(t_2)\, dx\, dt_1\, dt_2$$

[generally not true because moving between fitness classes assumes no coalescence - but small correction]

+

distribution of mutant timings:

$Q_k^{k-\ell}(t) = Q_k^{k-1}(t) * Q_k^{k-2}(t) * \ldots Q^{k-\ell}_{k-\ell+1}(t)$ and $Q^{k-\ell}_{k-\ell+1}(t) = s(k-\ell+1)e^{-s(k-\ell+1)t}$

evaluate many integrals

Non-conditional approximation

$$P_c^{k,k+m\to k-\ell} = \int \frac{x f_{k-\ell}}{h_{k-l}} \frac{y G_{k-\ell}(y \to x, |t_2 - t_1|)}{h_{k-l}} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dy dt_1 dt_2$$

y integral is just mean y - deterministic result for the change in the frequency of the lineage

$$P_c^{k,k+m\to k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_{k,k+m}^{k-\ell}(t_1, t_2) dx dt_1 dt_2$$

Assume:

- non-conditional approximation: the times at which the two individuals moved from one fitness class to another is independent

$$P_c^{k,k+m\to k-\ell} = \int \frac{x^2 f_{k-\ell}}{h_{k-l}^2} e^{-s(k-\ell)|t_2-t_1|} Q_k^{k-\ell}(t_1) Q_{k+m}^{k-\ell}(t_2) dx dt_1 dt_2$$

[generally not true because moving between fitness classes assumes no coalescence - but small correction]

+

distribution of mutant timings:

$$Q_k^{k-\ell}(t) = Q_k^{k-1}(t) * Q_k^{k-2}(t) * \ldots Q_{k-\ell+1}^{k-\ell}(t) \quad \text{and} \quad Q_{k-\ell+1}^{k-\ell}(t) = s(k-\ell+1)e^{-s(k-\ell+1)t}$$

evaluate many integrals

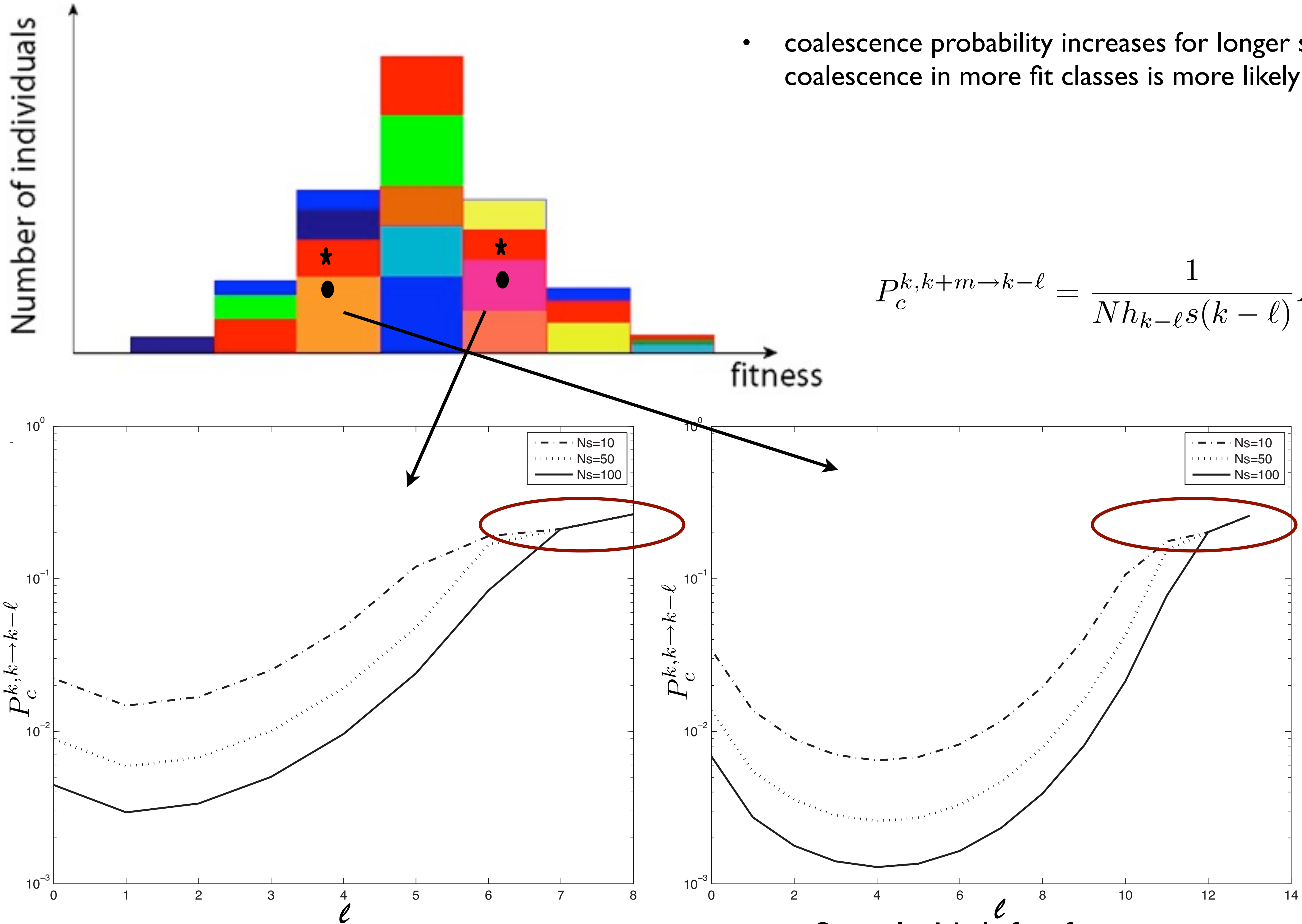
In non-conditional approximation: $P_c^{k,k+m\to k-\ell} = \dfrac{1}{N h_{k-\ell} s(k-\ell)} A_\ell^{k,m}$

Easy formula for coefficient: $A_\ell^{k,m} = \dfrac{\binom{k+m}{k-\ell}\binom{k}{k-\ell}}{\binom{2k+m}{2\ell+m}}$

Effective coalescence probabilities

- coalescence probability increases for longer steptimes - coalescence in more fit classes is more likely

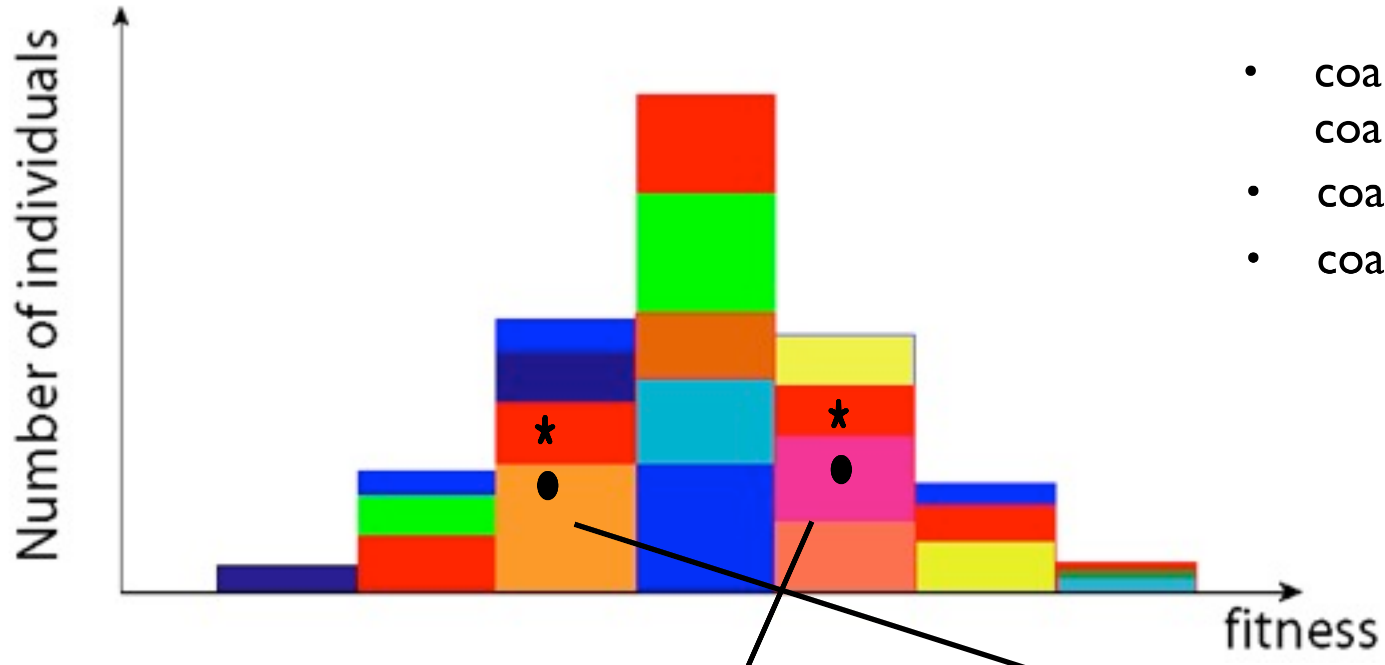
$$P_c^{k,k+m\rightarrow k-\ell} = \frac{1}{Nh_{k-\ell}s(k-\ell)}A_\ell^{k,m}$$



Sampled k just right of mean
(more fit than mean)

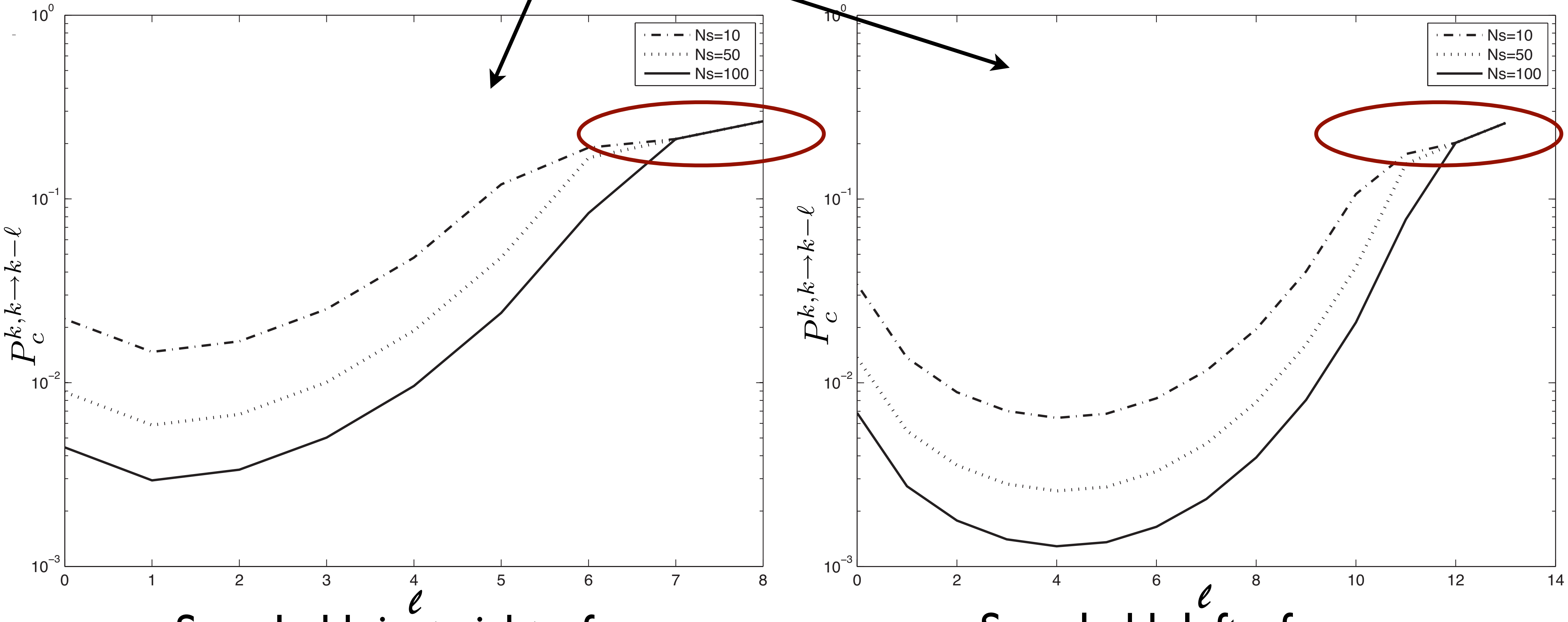
Sampled k left of mean
(less fit than mean)

Effective coalescence probabilities



- coalescence probability increases for longer steptimes - coalescence in more fit classes is more likely
- coalescence probability decreases with selection
- coalescence probability decreases with population size

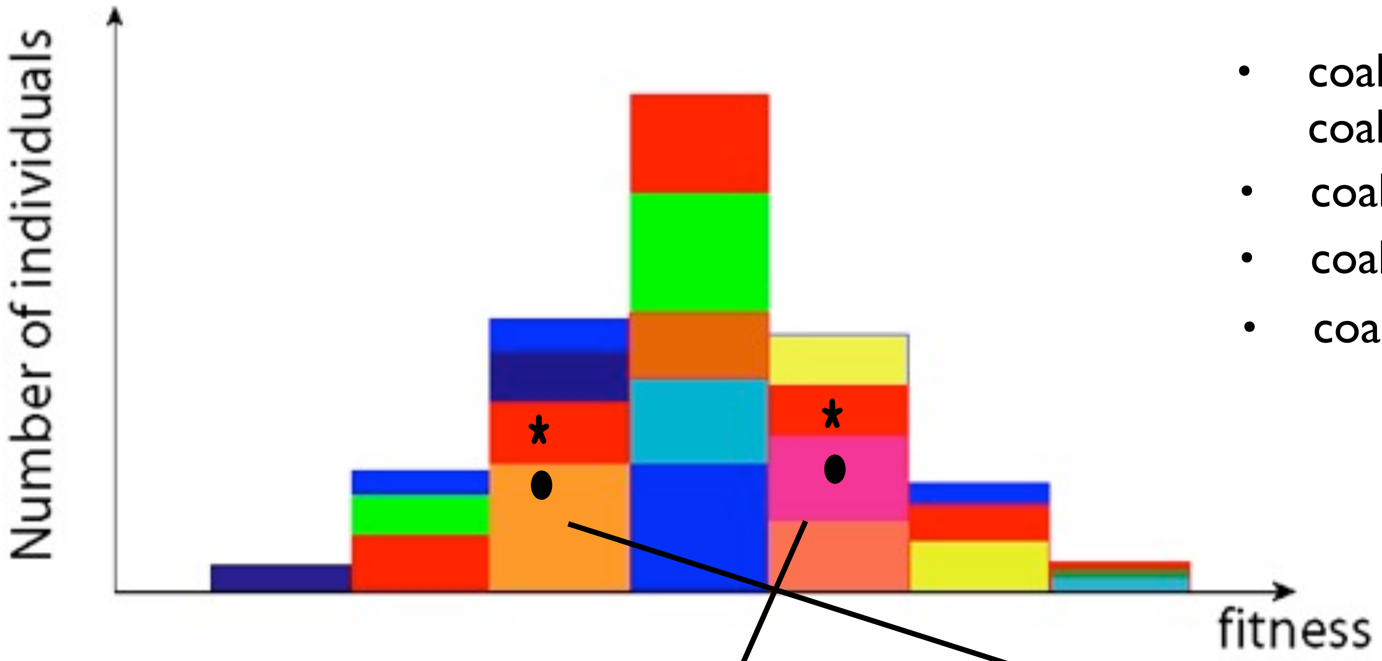
$$P_c^{k,k+m\to k-\ell} = \frac{1}{N h_{k-\ell} s(k-\ell)} A_\ell^{k,m}$$



Sampled k just right of mean
(more fit than mean)

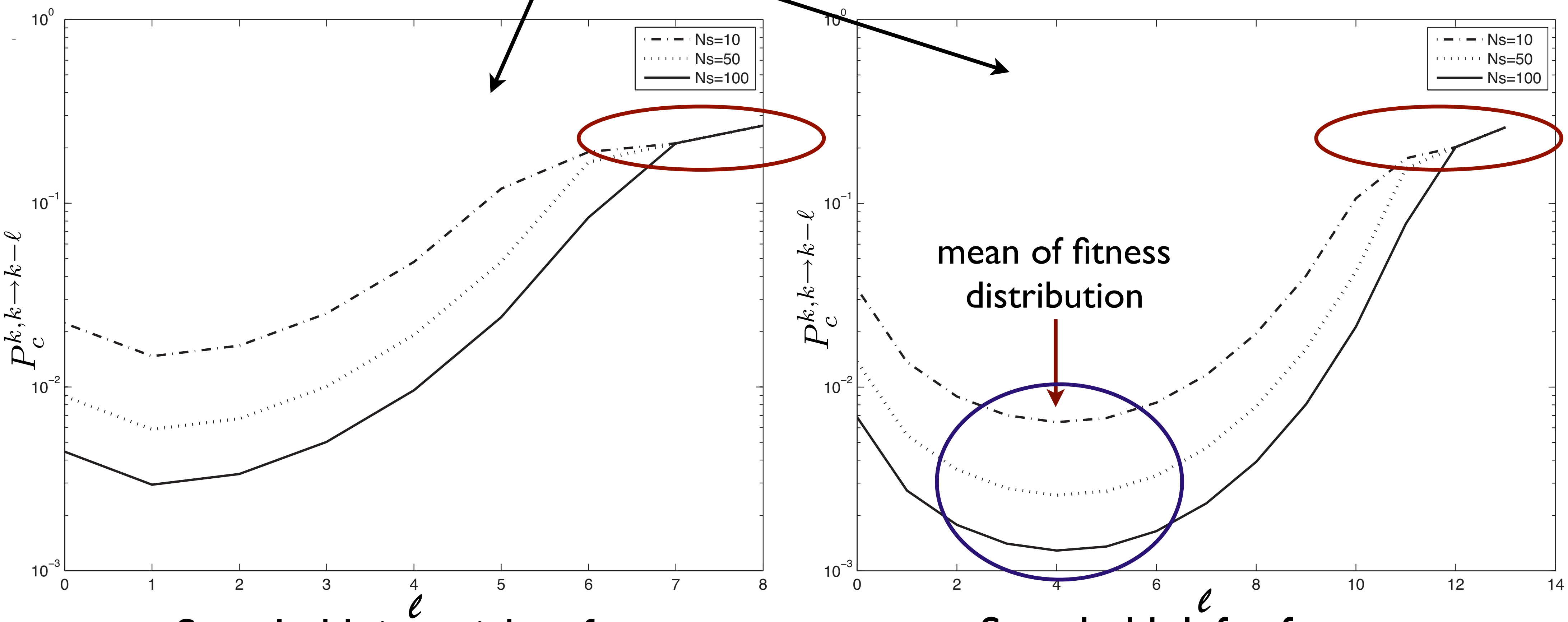
Sampled k left of mean
(less fit than mean)

Effective coalescence probabilities



- coalescence probability increases for longer steptimes - coalescence in more fit classes is more likely
- coalescence probability decreases with selection
- coalescence probability decreases with population size
- coalescence probability is less likely in most probable class

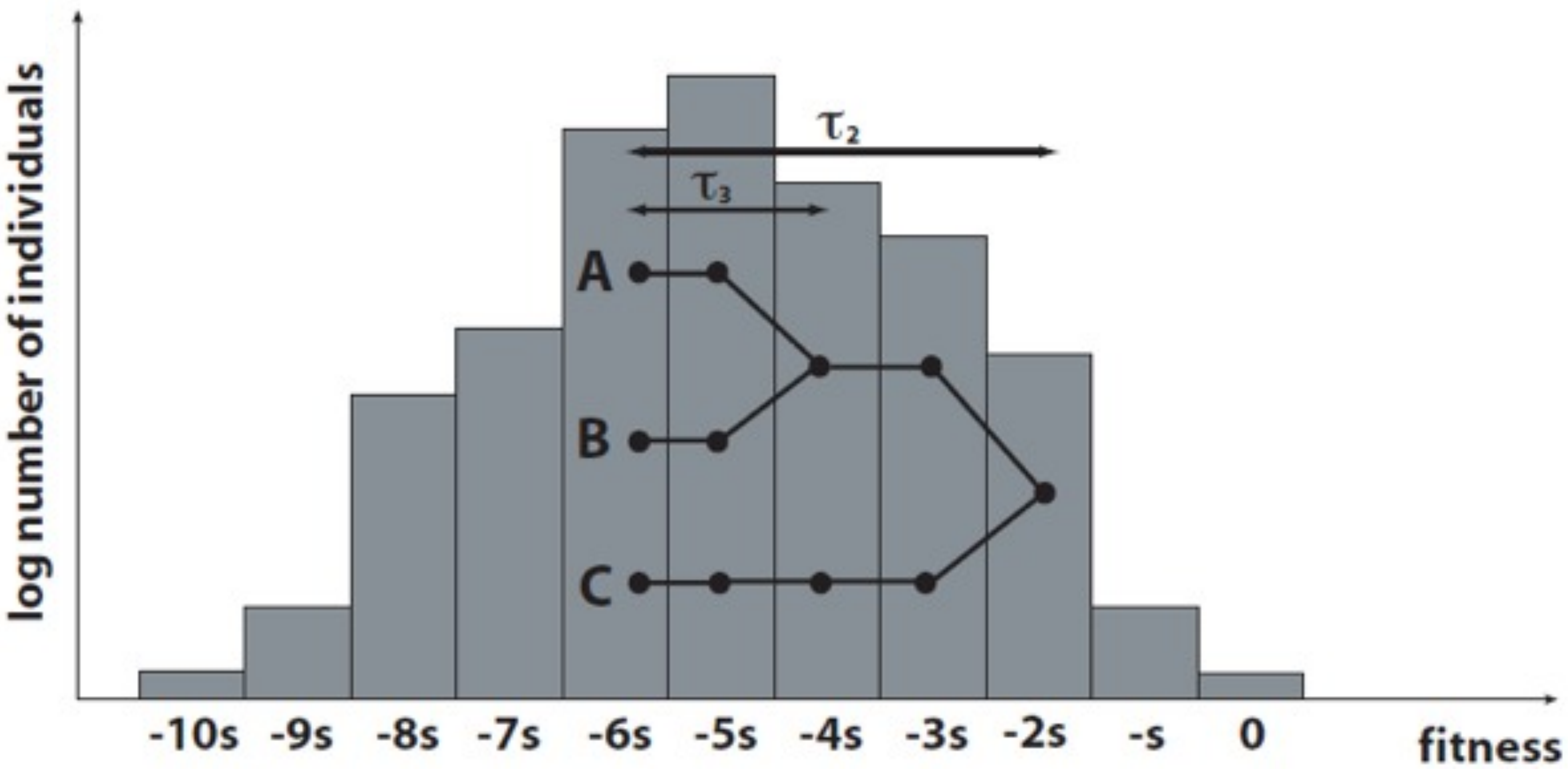
$$P_c^{k,k+m\to k-\ell} = \frac{1}{N h_{k-\ell} s(k-\ell)} A_\ell^{k,m}$$



Sampled k just right of mean (more fit than mean)

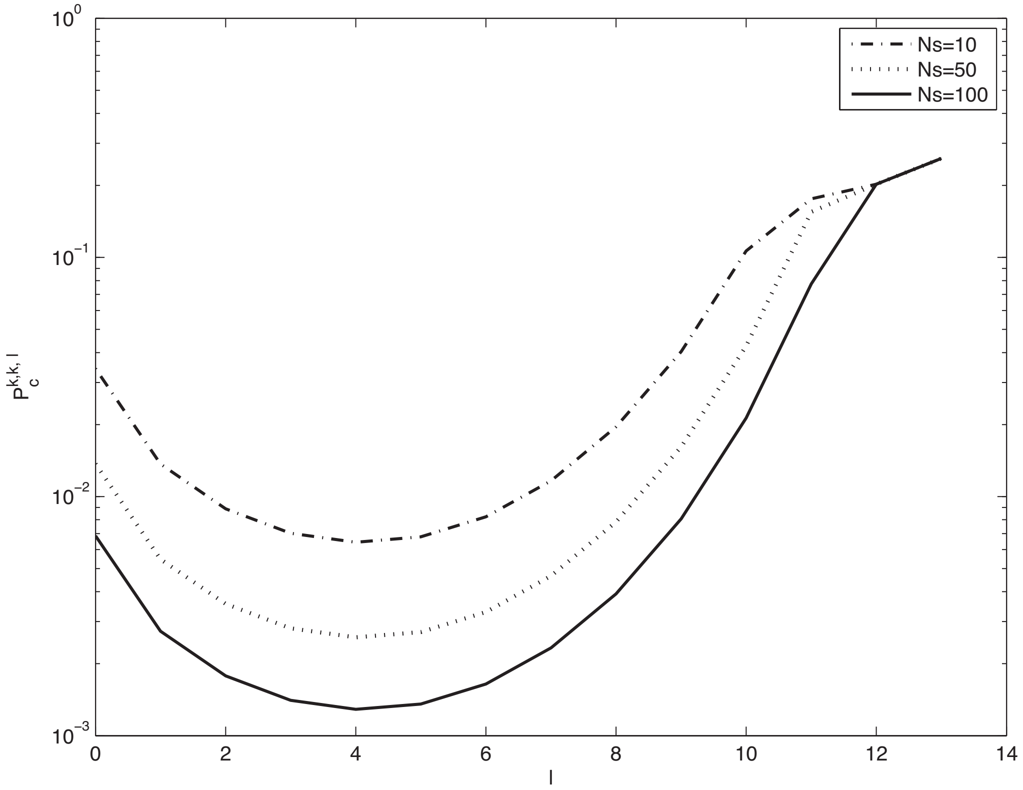
Sampled k left of mean (less fit than mean)

Comparison to variable population size

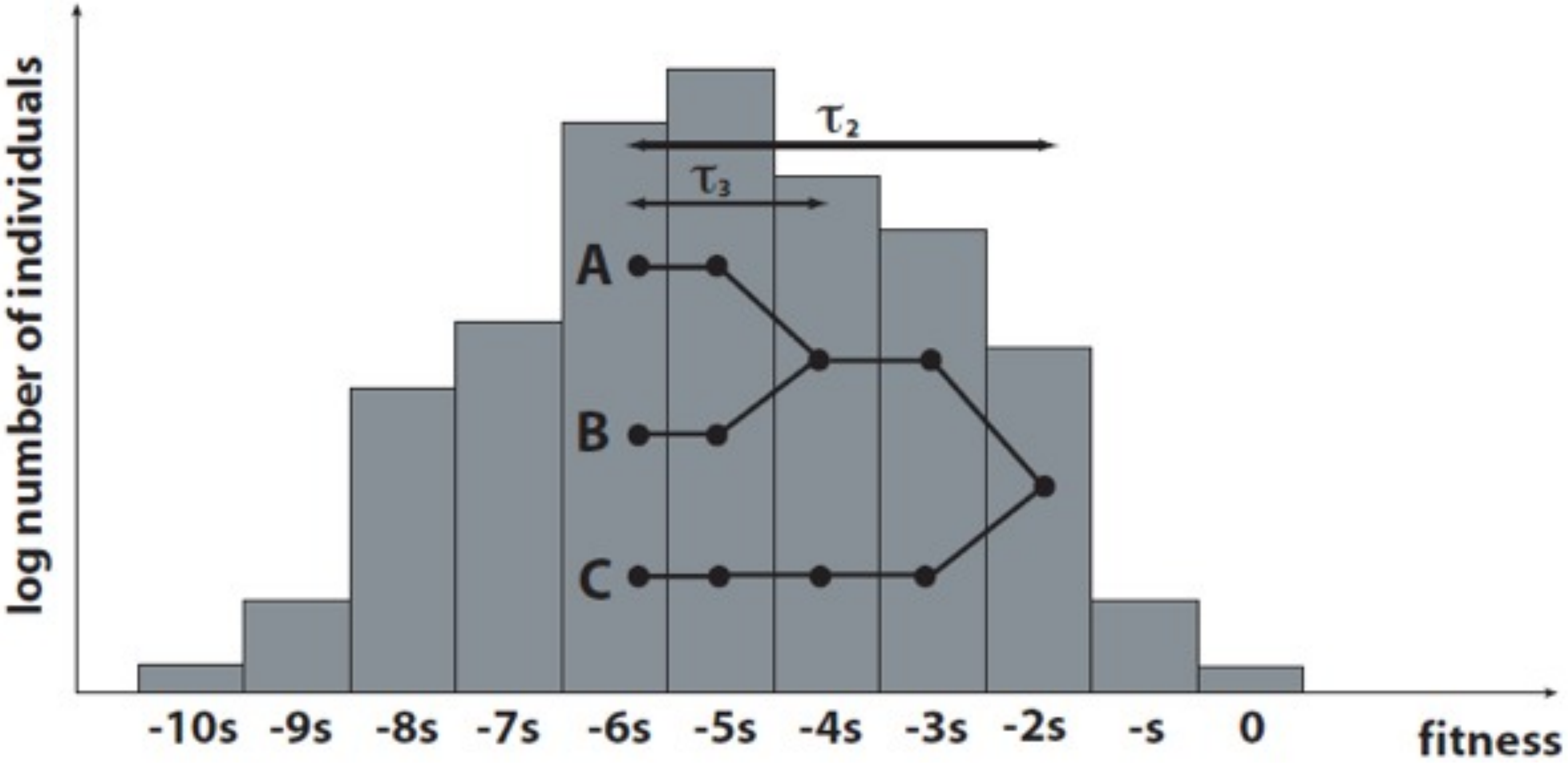


$$P_c^{k,k+m\to k-\ell} = \frac{1}{Nh_{k-\ell}s(k-\ell)}A_\ell^{k,m}$$

$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell}s_{k-\ell}}A_\ell^{k,m}$$

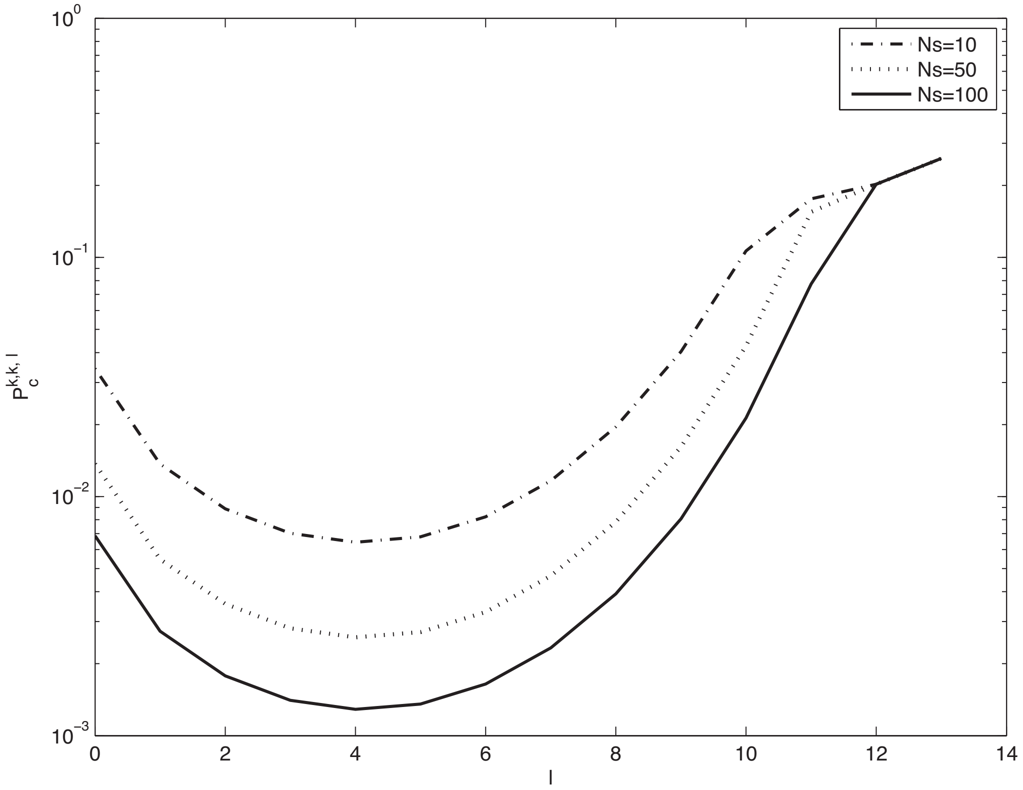


Comparison to variable population size

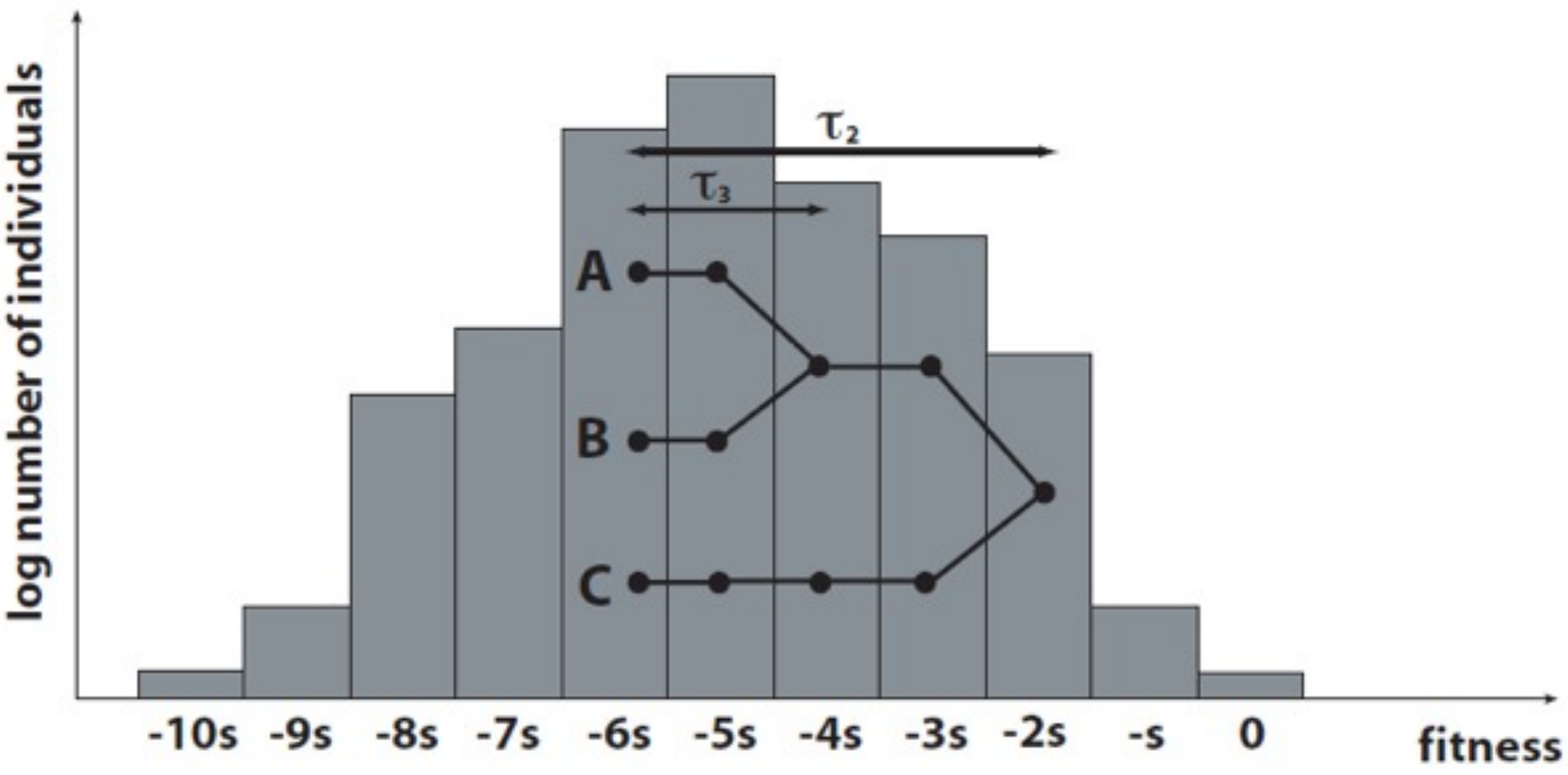


$$P_c^{k,k+m\to k-\ell} = \frac{1}{Nh_{k-\ell}\boxed{s(k-\ell)}}A_\ell^{k,m}$$

$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell}s_{k-\ell}}A_\ell^{k,m}$$

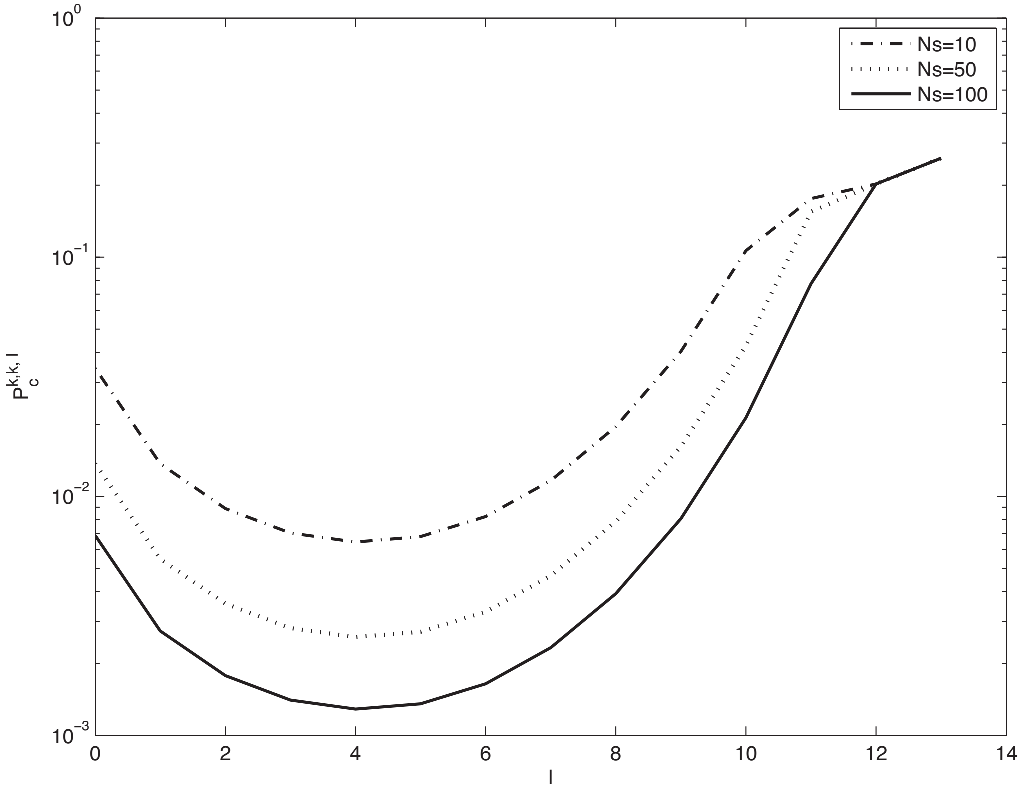


Comparison to variable population size

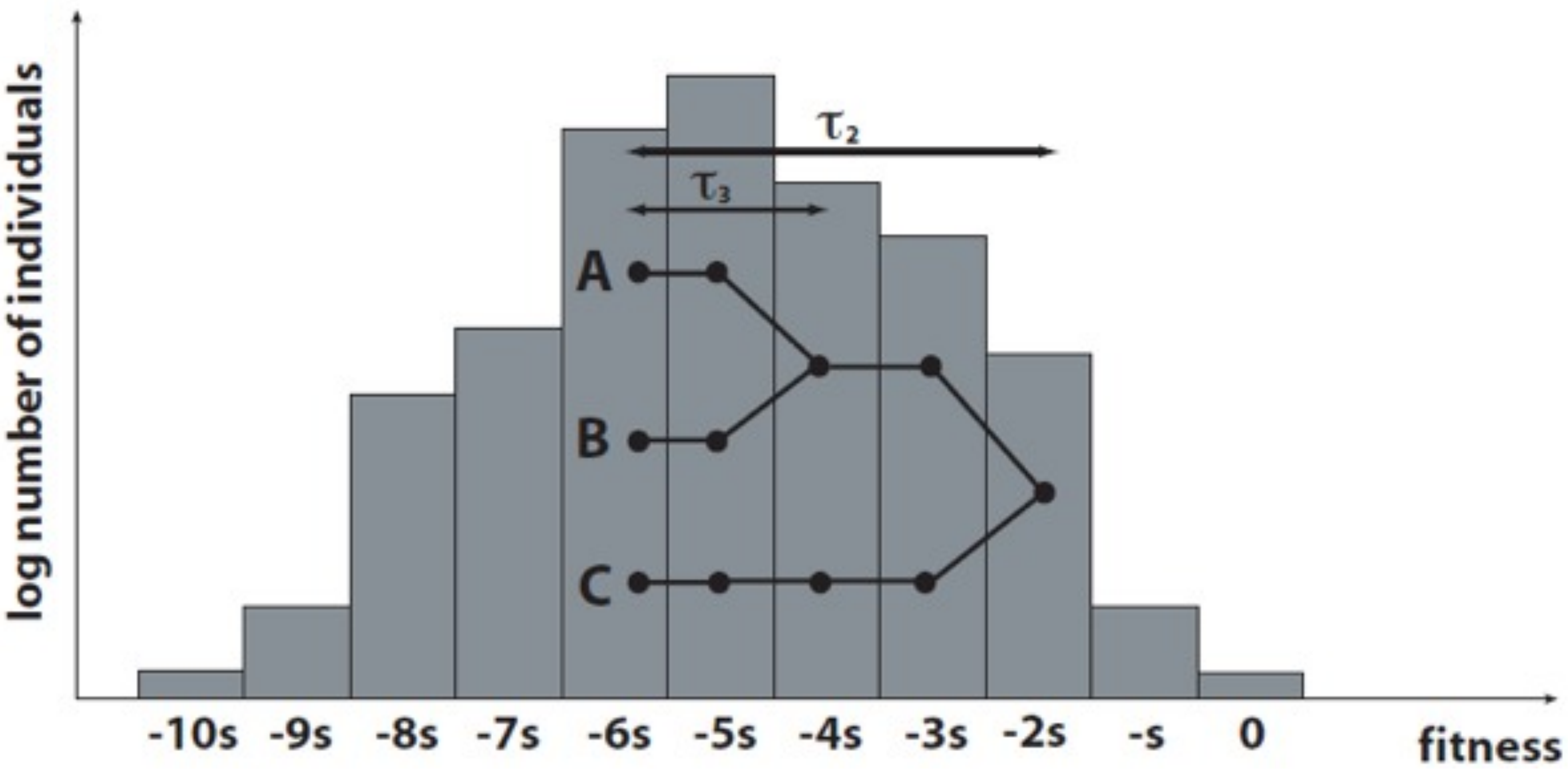


$$P_c^{k,k+m\to k-\ell} = \frac{1}{Nh_{k-\ell}s(k-\ell)}A_\ell^{k,m}$$

$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell}s_{k-\ell}}A_\ell^{k,m}$$



Comparison to variable population size

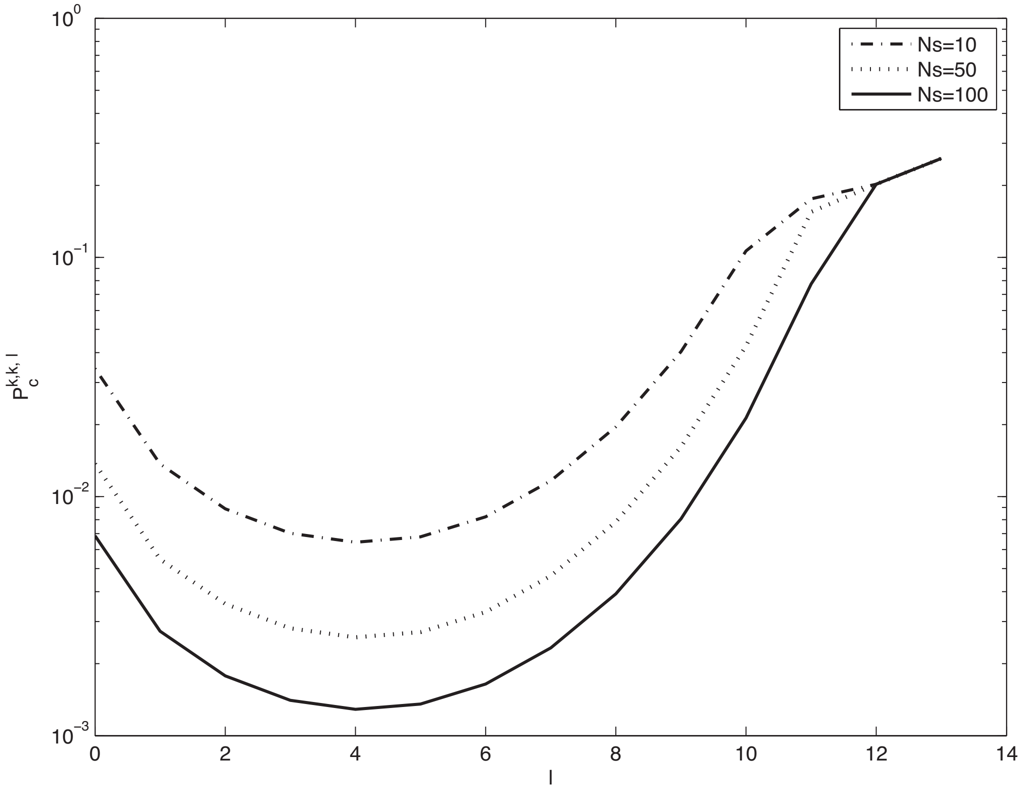


$$P_c^{k,k+m\to k-\ell} = \frac{1}{N h_{k-\ell} s(k-\ell)} A_\ell^{k,m}$$

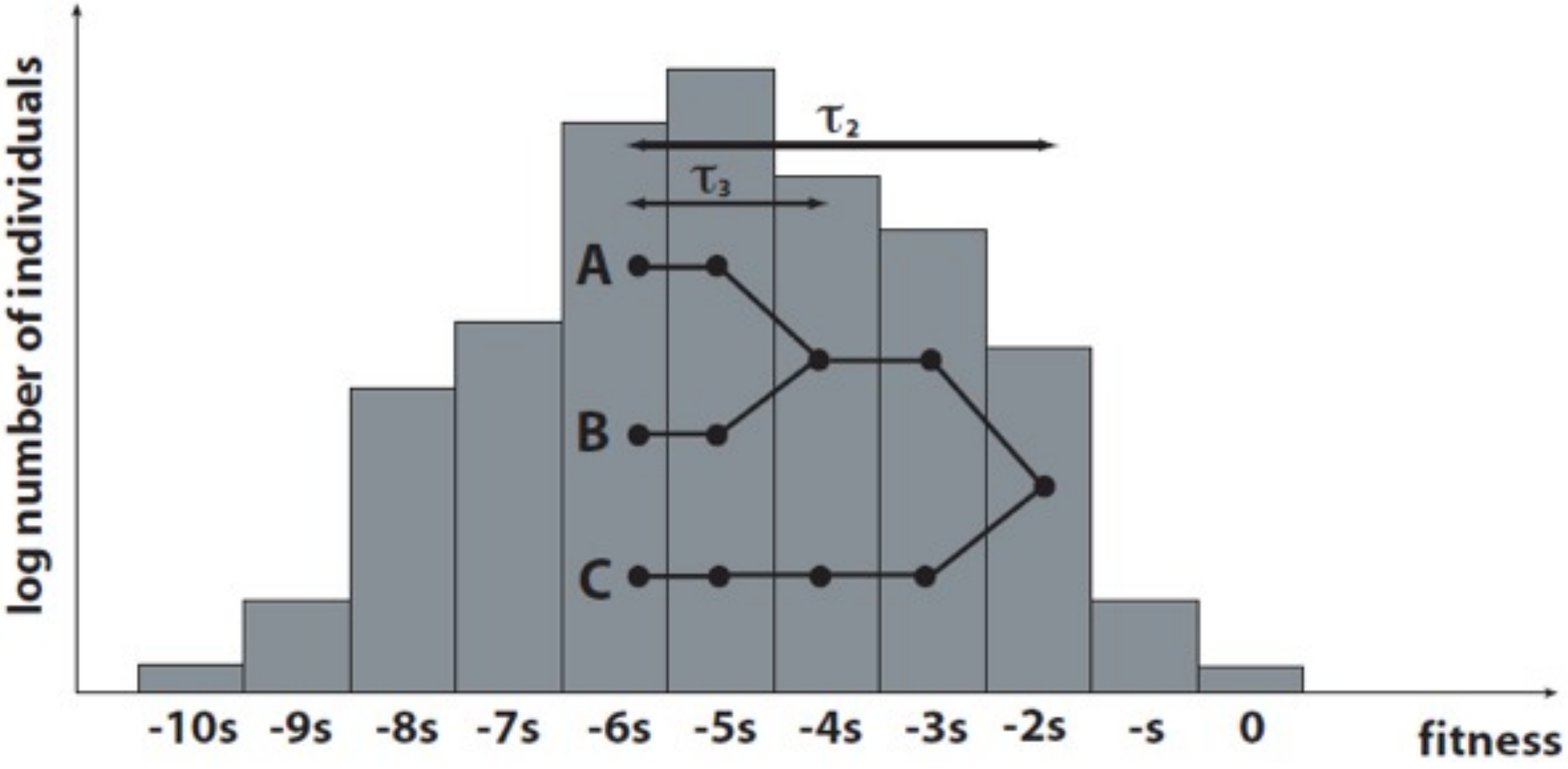
$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell} s_{k-\ell}} A_\ell^{k,m}$$

lineage spends ~1/s_k generations in each class

→ per generation coalescence probability in class k is 1/n_k



Comparison to variable population size



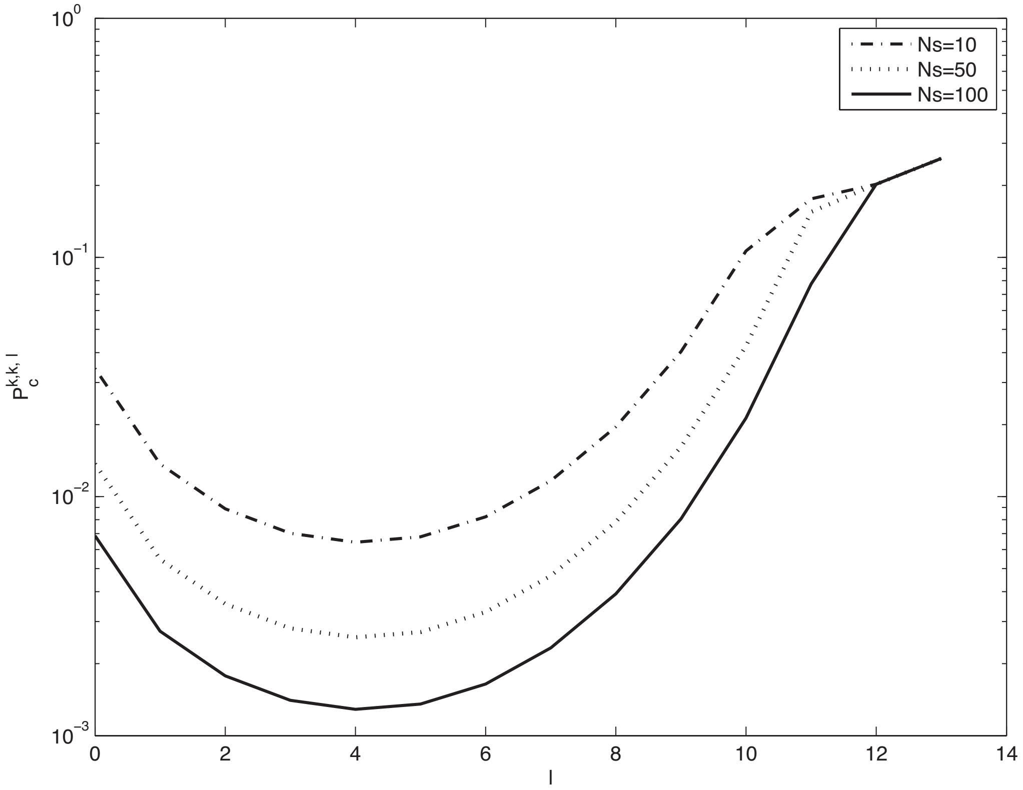
$$P_c^{k,k+m\to k-\ell} = \frac{1}{Nh_{k-\ell}\,s(k-\ell)} A_\ell^{k,m}$$

$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell}s_{k-\ell}} A_\ell^{k,m}$$

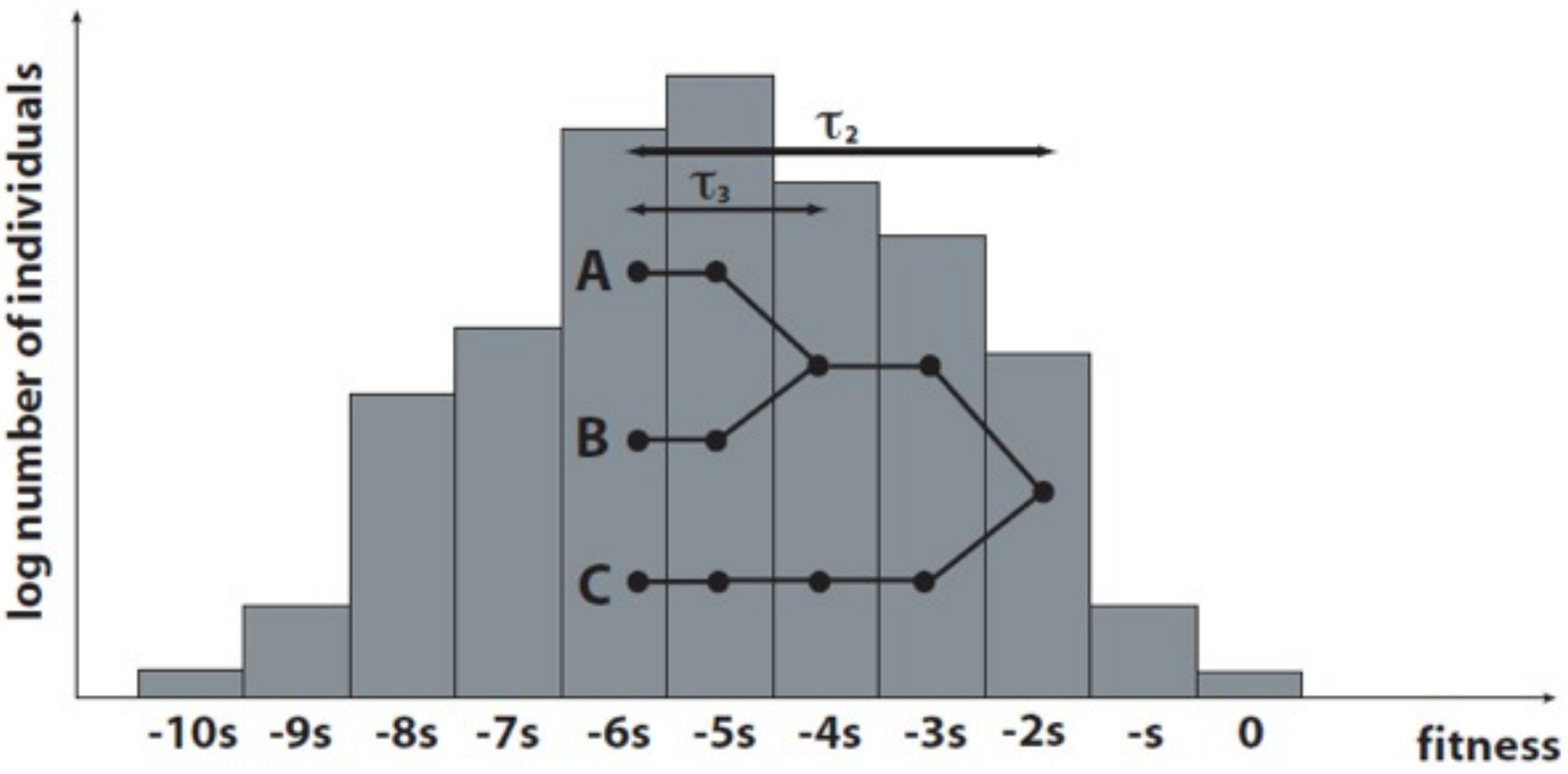
lineage spends ~1/s_k generations in each class

→ per generation coalescence probability in class k is 1/n_k

historically varying population size - different effective population sizes depending on initial position in fitness distribution



Comparison to variable population size



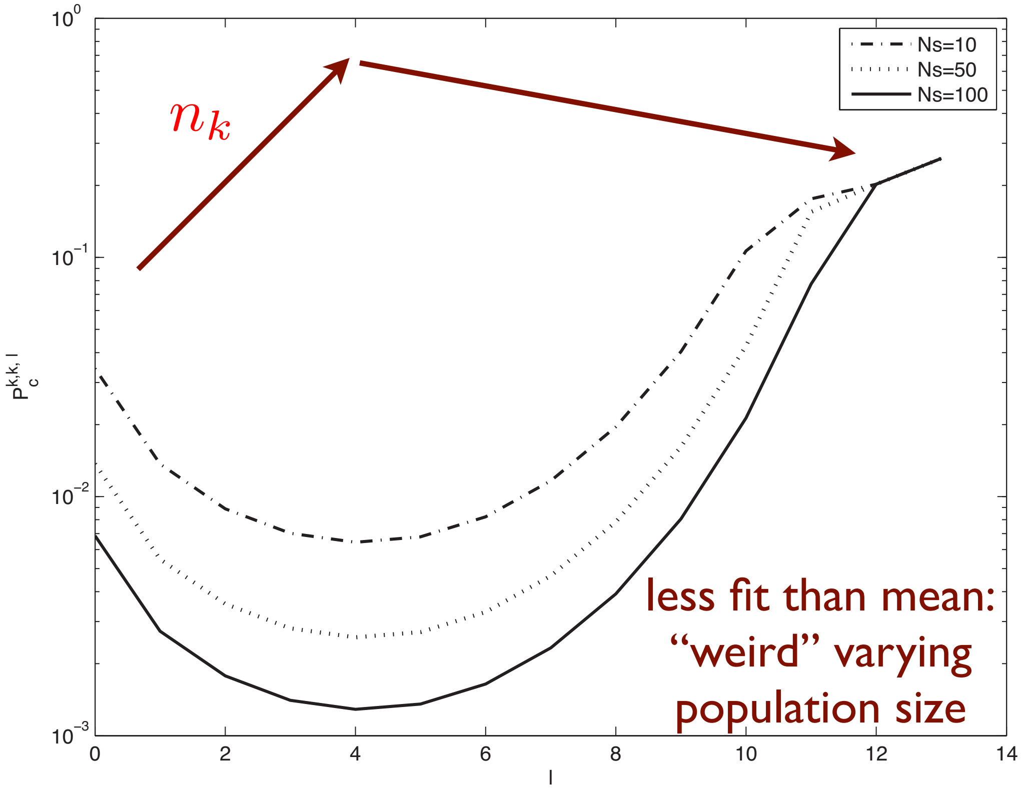
$$P_c^{k,k+m\to k-\ell} = \frac{1}{Nh_{k-\ell}\,s(k-\ell)} A_\ell^{k,m}$$

$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell}s_{k-\ell}} A_\ell^{k,m}$$

lineage spends ~1/s_k generations in each class

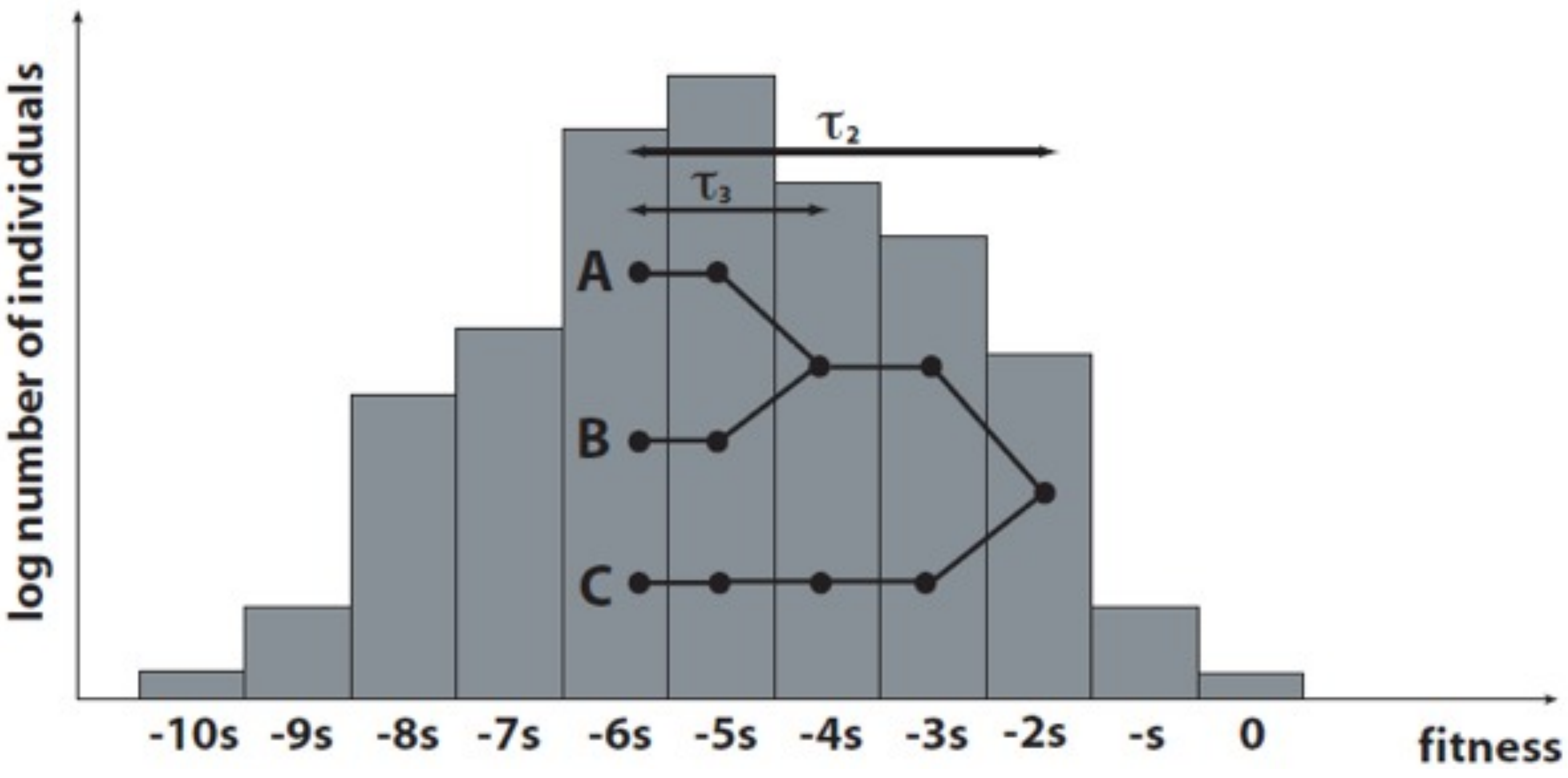
→ per generation coalescence probability in class k is 1/n_k

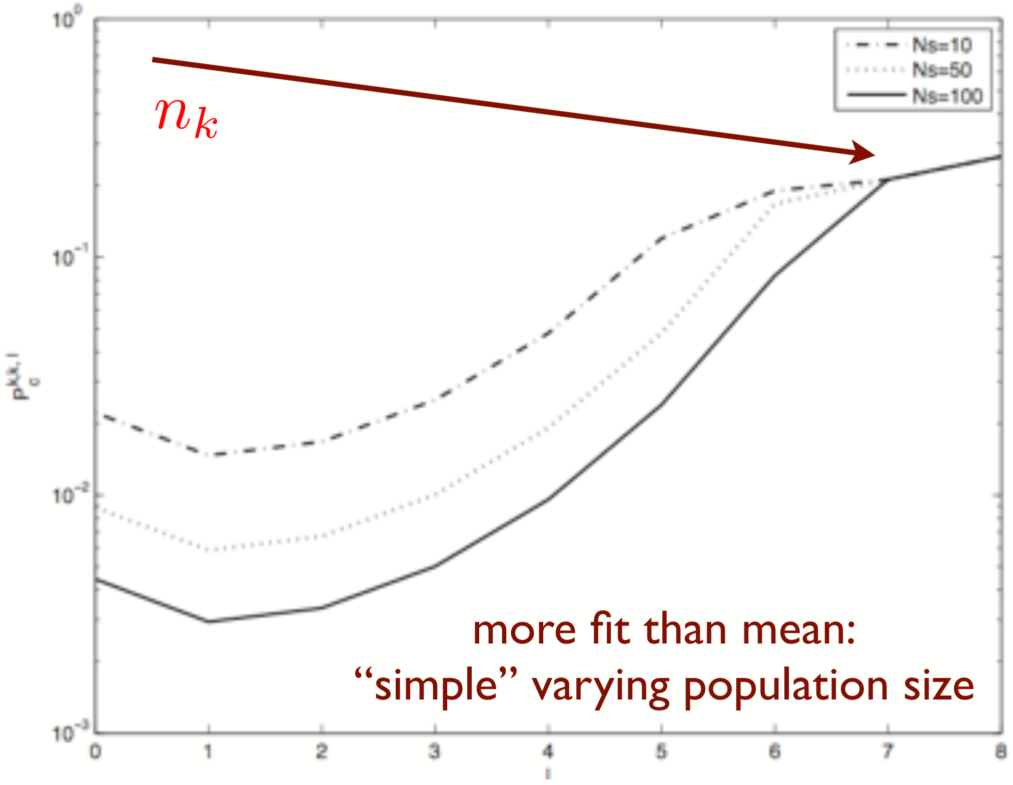
historically varying population size - different effective population sizes depending on initial position in fitness distribution



less fit than mean: “weird” varying population size

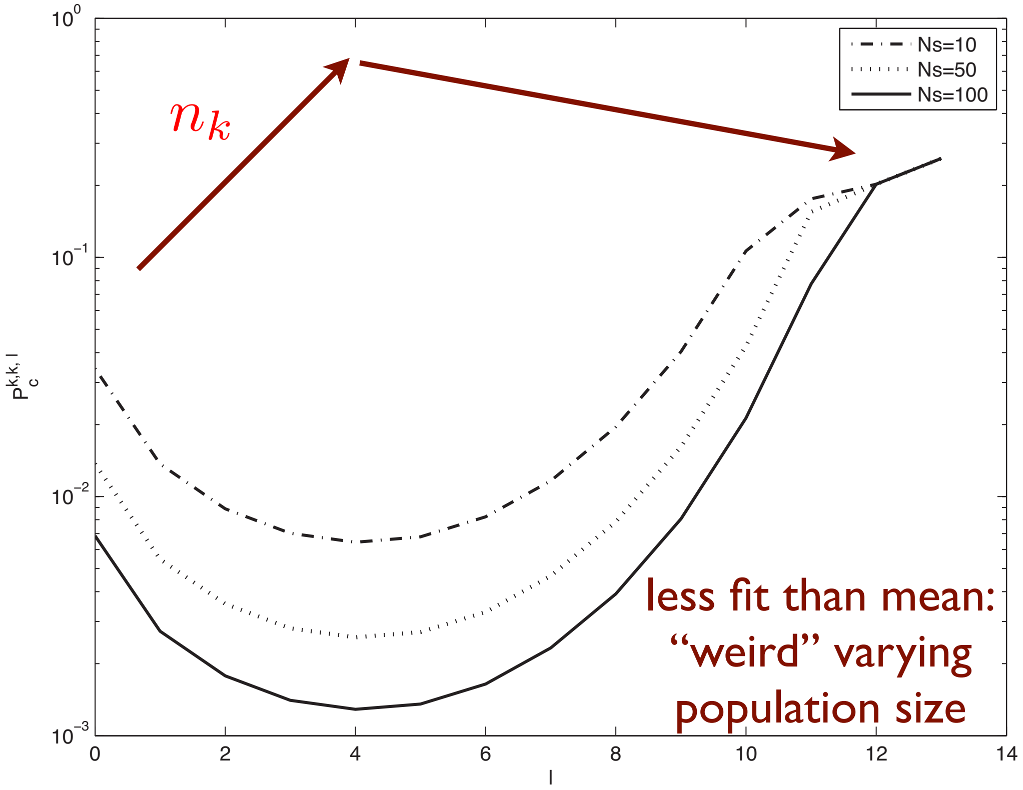
Comparison to variable population size





n_k

more fit than mean:
"simple" varying population size



n_k

less fit than mean:
"weird" varying
population size

$$P_c^{k,k+m\to k-\ell} = \frac{1}{N h_{k-\ell}\, s(k-\ell)} A_\ell^{k,m}$$

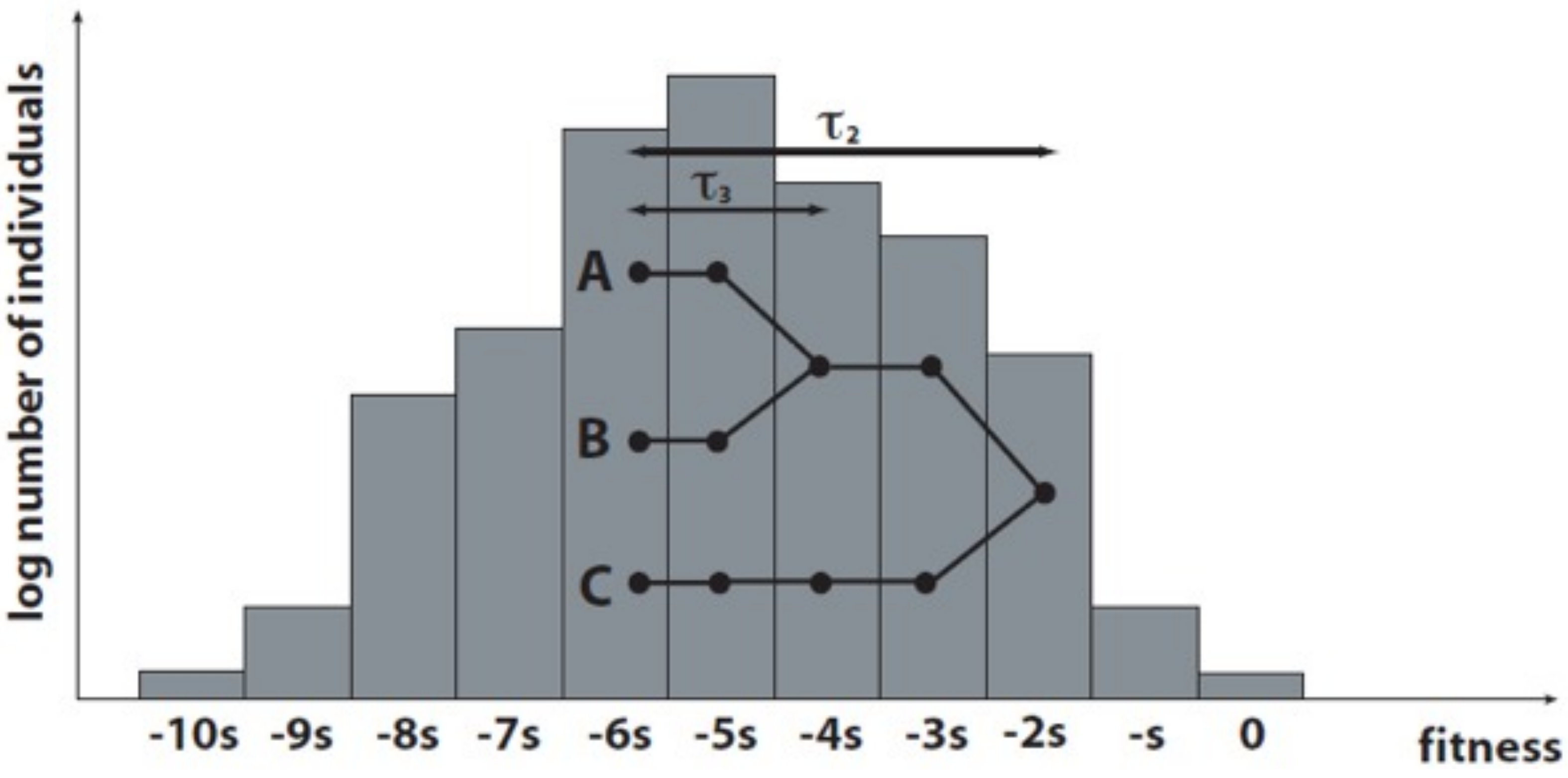
$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell} s_{k-\ell}} A_\ell^{k,m}$$

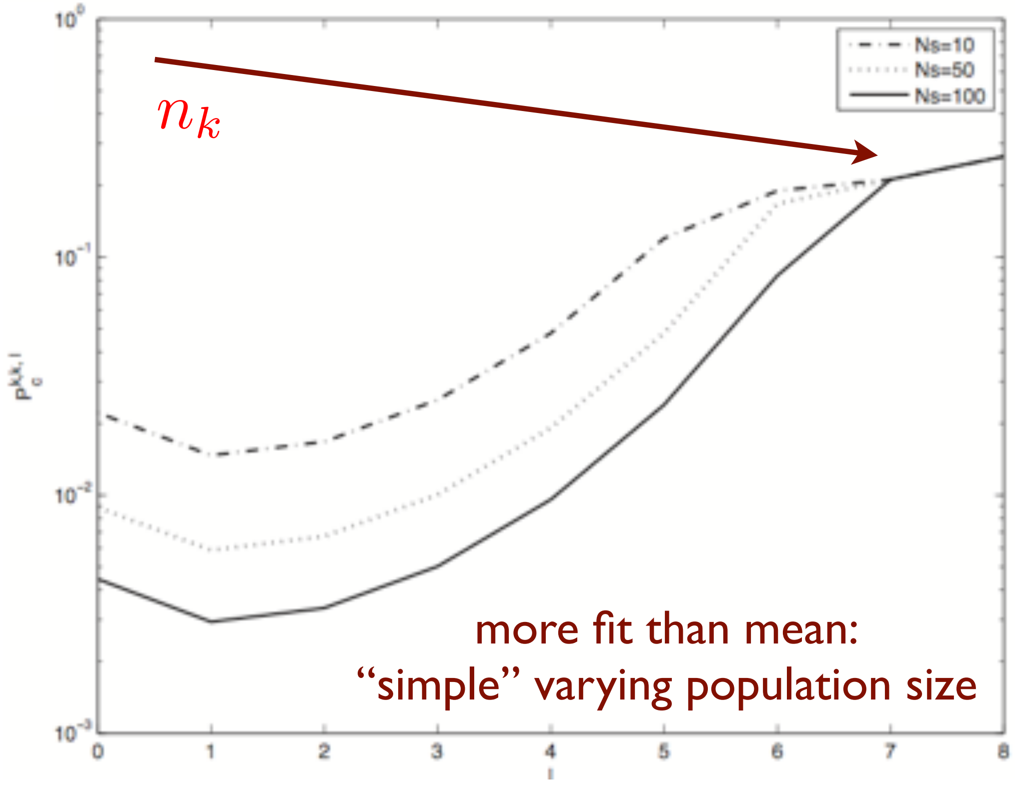
lineage spends ~1/s_k generations in each class

→ per generation coalescence probability in class k is 1/n_k

historically varying population size - different effective population sizes depending on initial position in fitness distribution

Comparison to variable population size





n_k

more fit than mean:
"simple" varying population size

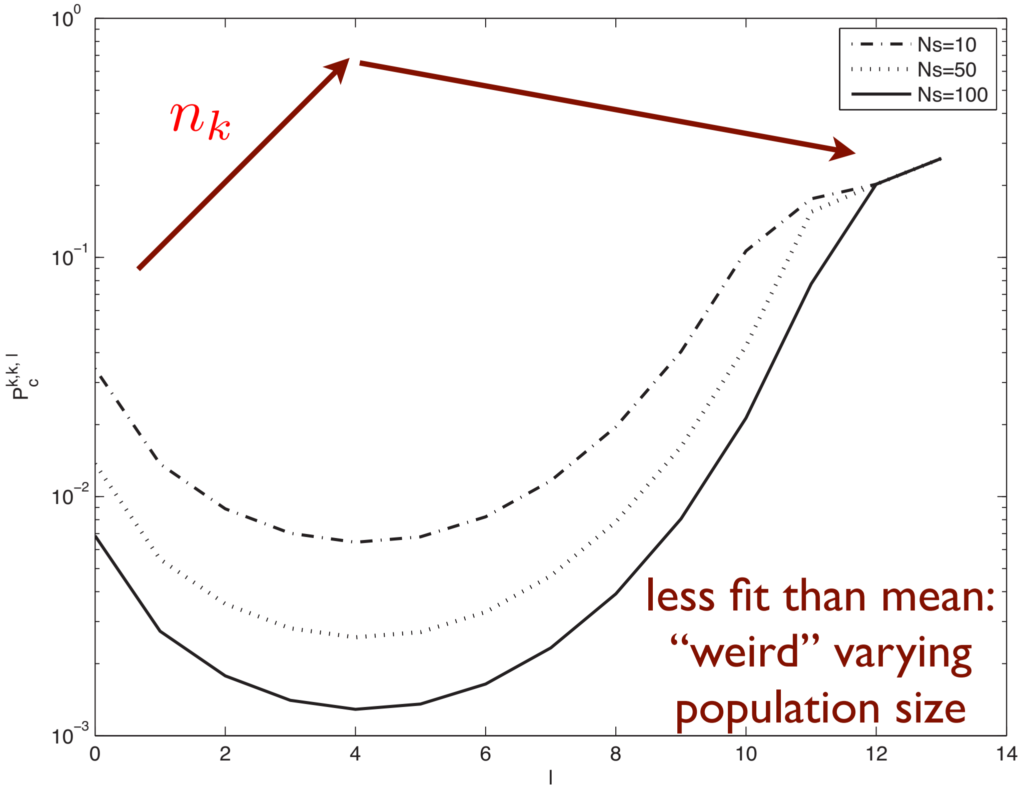
$$P_c^{k,k+m\to k-\ell} = \frac{1}{Nh_{k-\ell}s(k-\ell)}A_\ell^{k,m}$$

$$P_c^{k,k+m\to k-\ell} = \frac{1}{n_{k-\ell}s_{k-\ell}}A_\ell^{k,m}$$

lineage spends ~$1/s_k$ generations in each class

→ per generation coalescence probability in class k is $1/n_k$

historically varying population size - different effective population sizes depending on initial position in fitness distribution

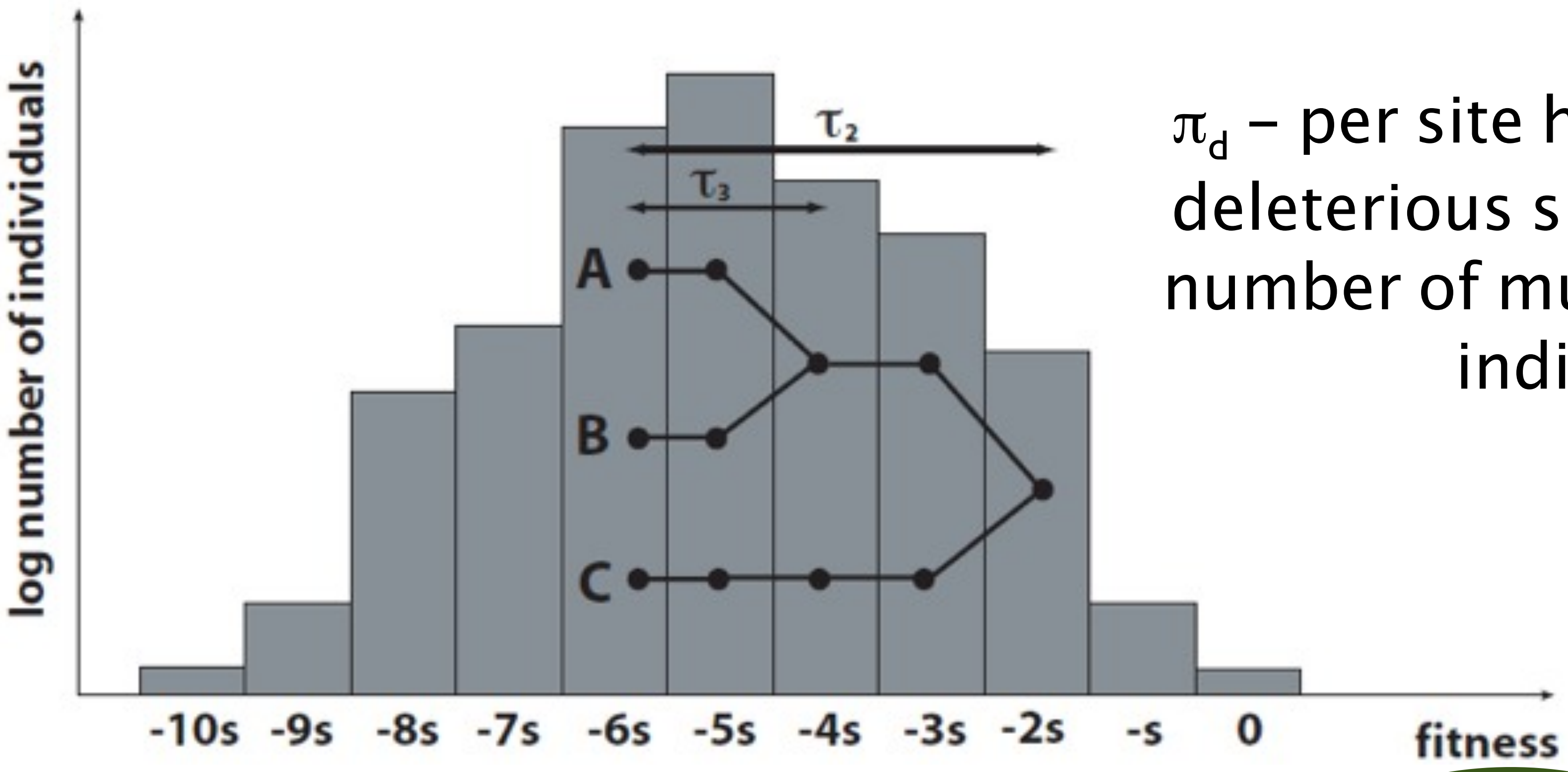


n_k

less fit than mean:
"weird" varying
population size

→ really strange variation in population size for two individuals from different classes

From coalescence probabilities to selected diversity



π_d – per site heterozygosity at deleterious sites – distance in number of mutations between individuals

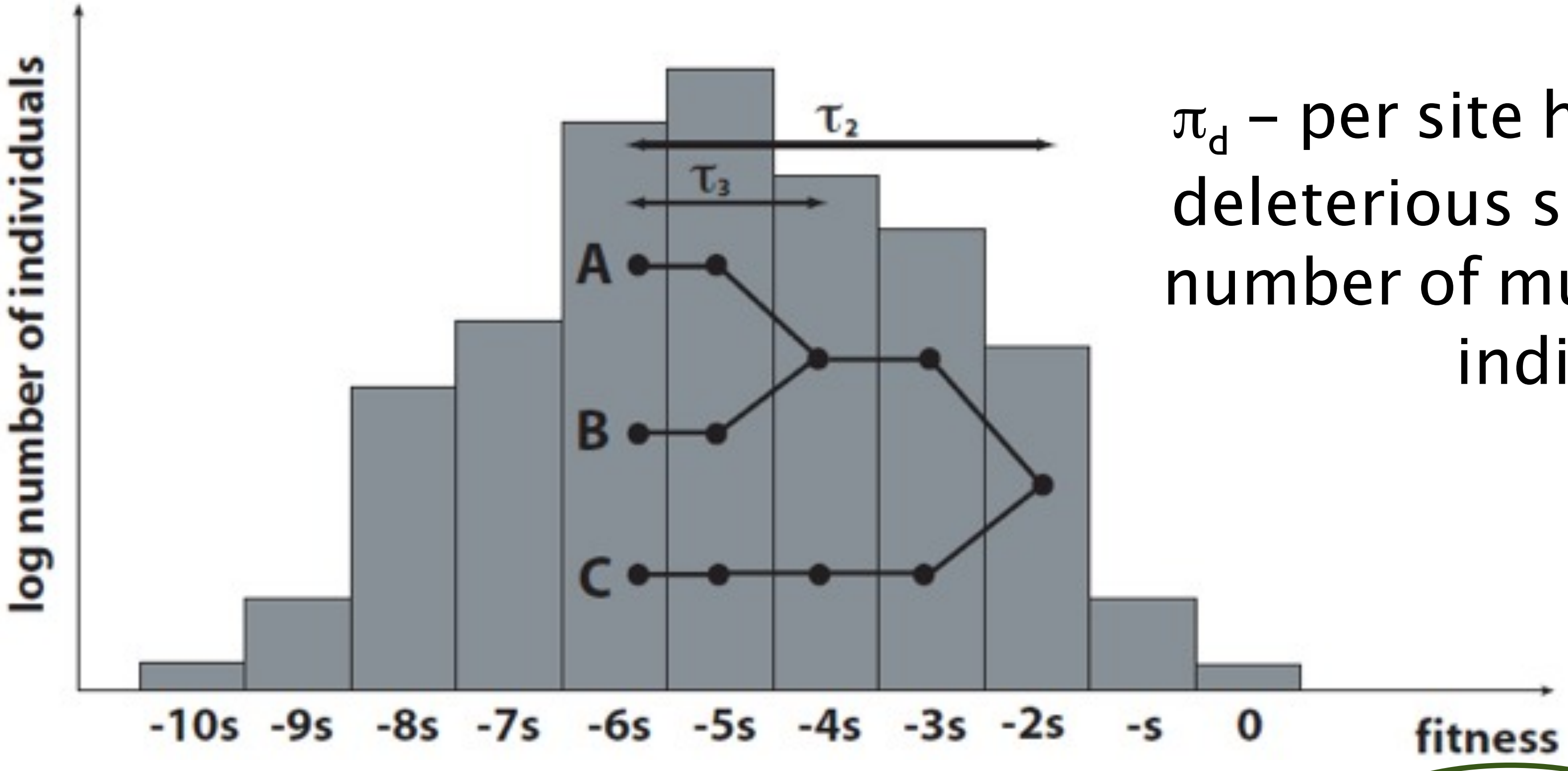
$$P(\pi_{AB}=4)=P(\tau_3=2)=P_c^{k,k\to k-2}(1-P_c^{k,k\to k-1})(1-P_c^{k,k\to k})$$

coalesced k-2 classes ago

did not coalesce k-1 class ago

did not coalesce in k class

From coalescence probabilities to selected diversity



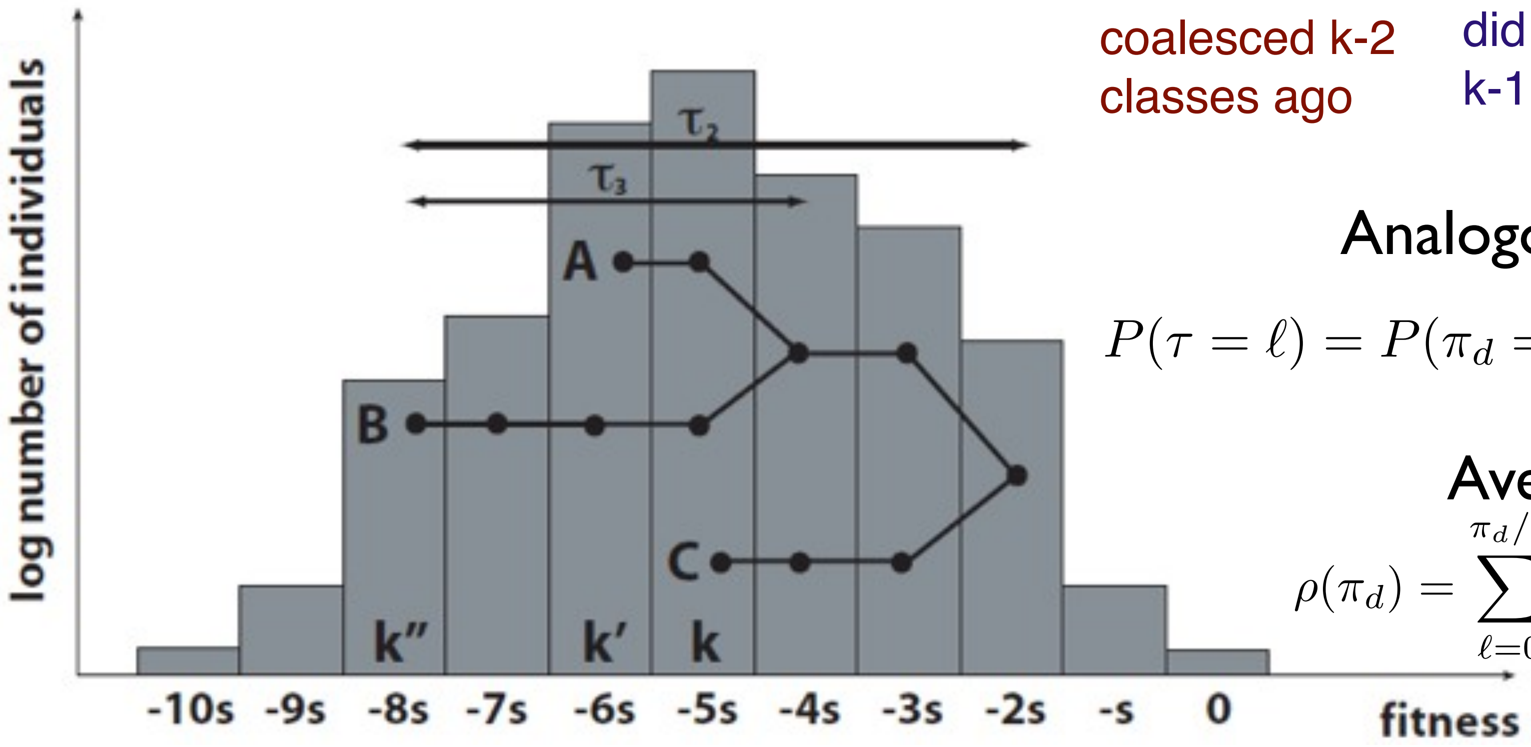
π_d – per site heterozygosity at deleterious sites – distance in number of mutations between individuals

$$P(\pi_{AB}=4)=P(\tau_3=2)=P_c^{k,k\to k-2}(1-P_c^{k,k\to k-1})(1-P_c^{k,k\to k})$$

coalesced k-2 classes ago

did not coalesce k-1 class ago

did not coalesce in k class



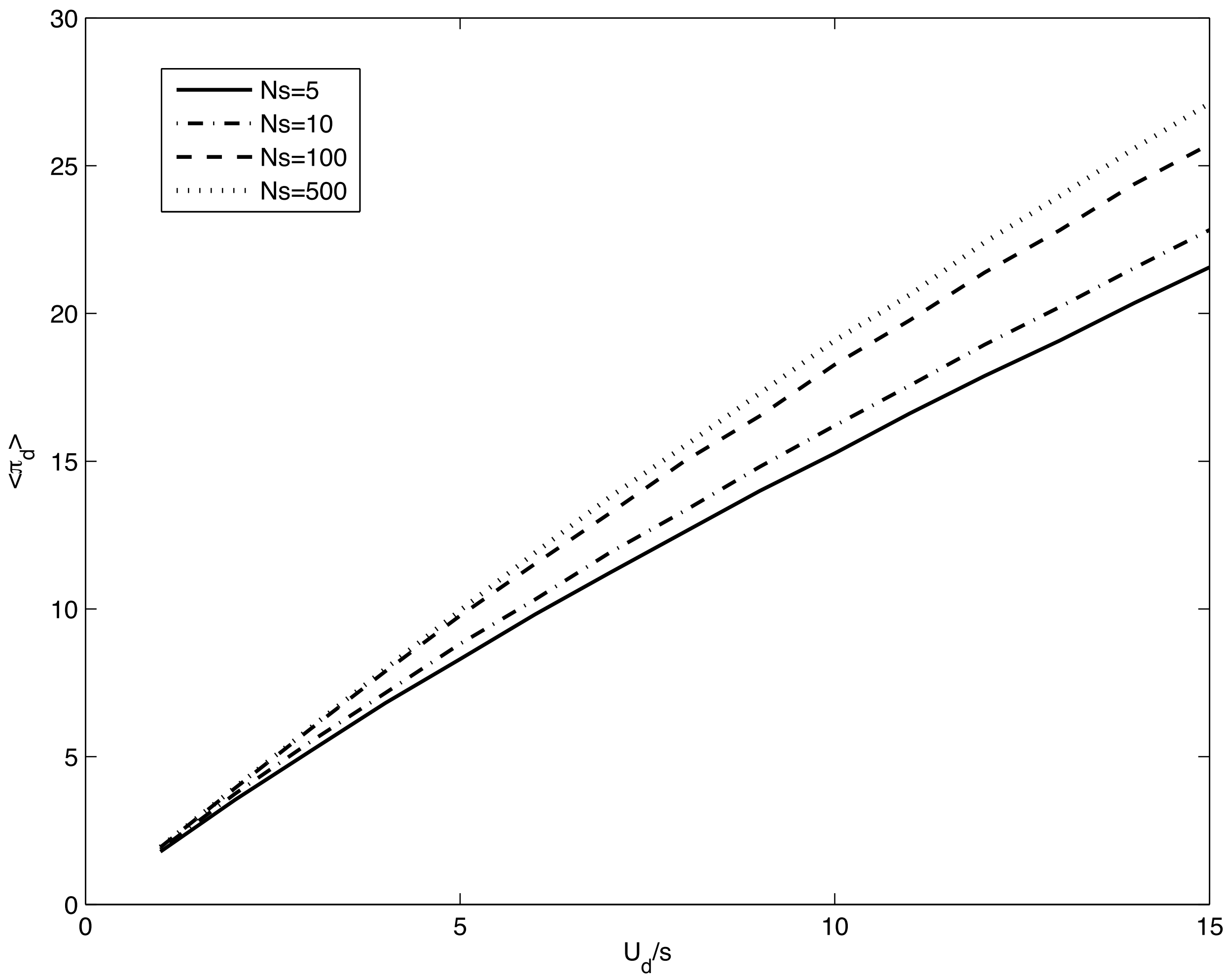
Analogous expressions apply for k, k’, k”

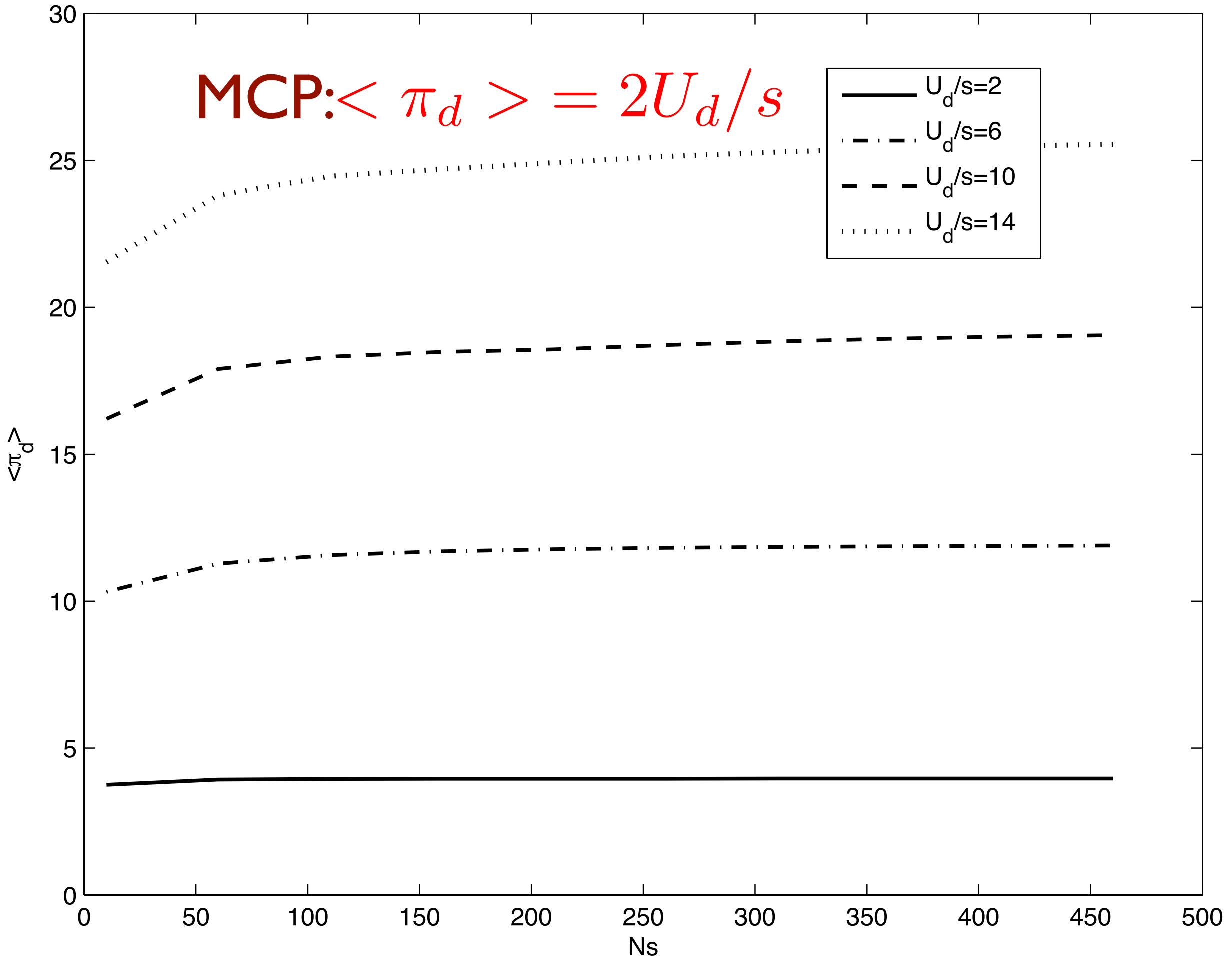
$$P(\tau=\ell)=P(\pi_d=2\ell+m)=P_c^{k,k+m\to k-\ell}\prod_{j=0}^{\ell-1}(1-P_c^{k,k+m\to k-j})$$

Average over distribution of k, k’, k”:

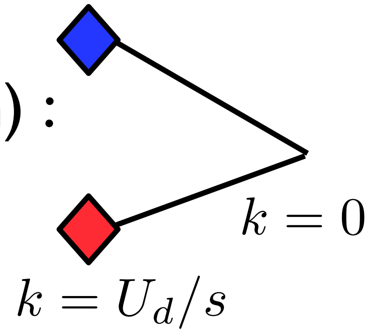
$$\rho(\pi_d)=\sum_{\ell=0}^{\pi_d/2}\sum_{k=0}^{\infty}H(k,k+m=k+\pi_d-2\ell)P_k^{k+m=k+\pi_d-2\ell}(\tau=\ell)$$

Scaling of $<\pi_d>$



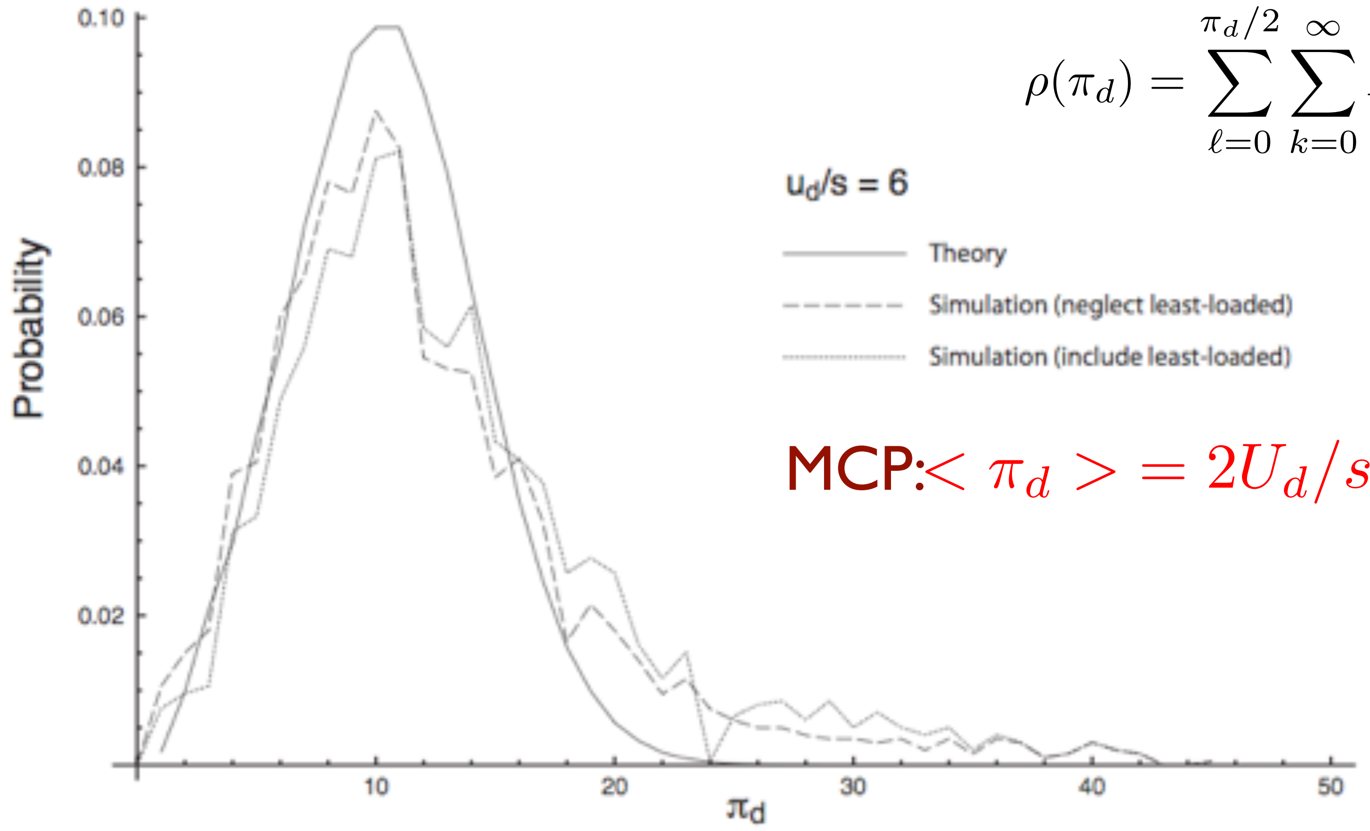

MCP: $< \pi_d > = 2U_d/s$

- large selection - weak N dependence
- mean coalescence path approximation for large N and large U_d/s (weaker selection) :
  - large number of lineages in each fitness class - coalescence events unlikely
  - all coalescence happens in zeroth class (like in EPS)
  - coalescence time is dominated by time it takes to get to zeroth class (unlike EPS)
- for small N - larger probability to coalesce in bulk - smaller $<\pi_d>$



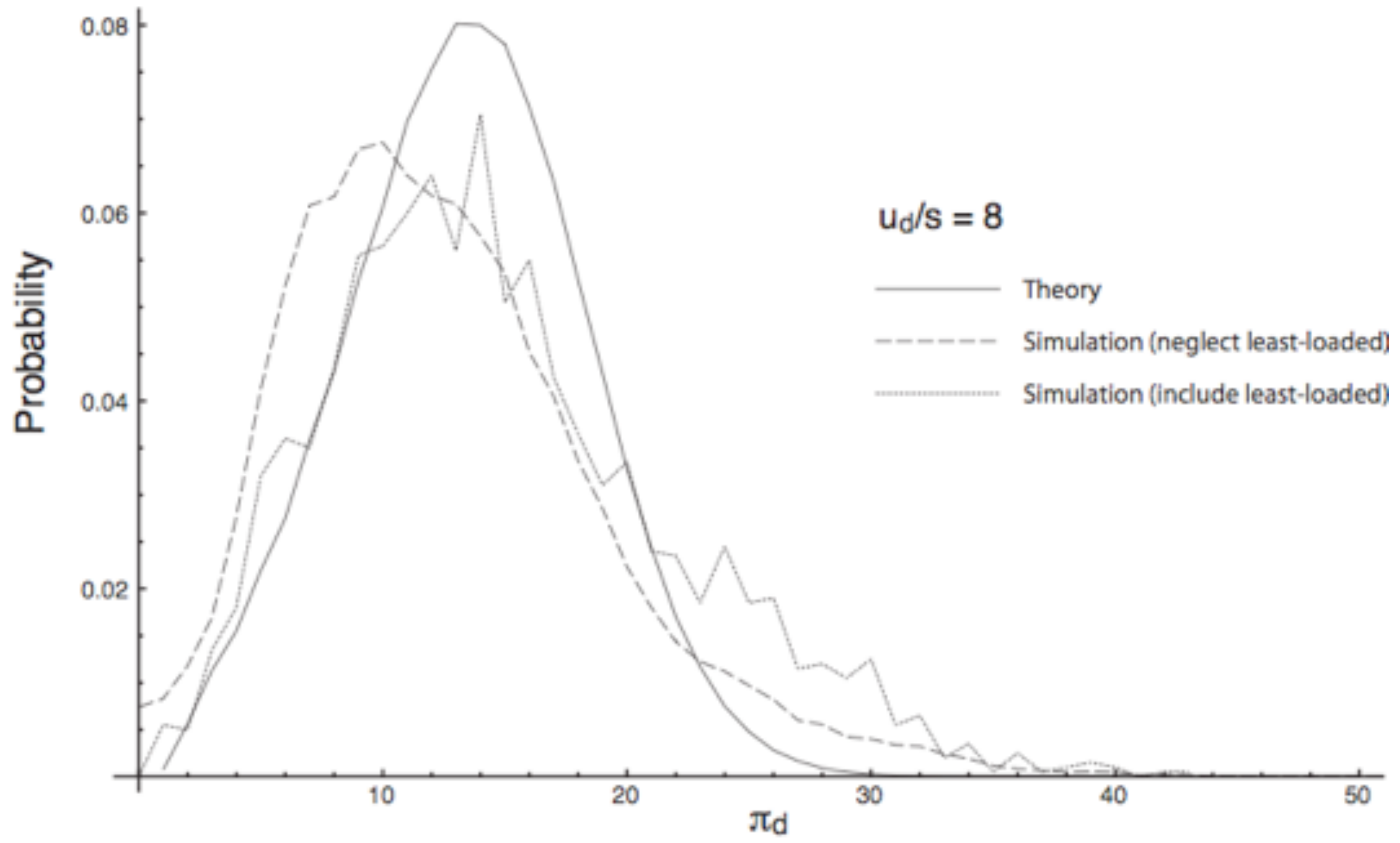
Distribution of per site heterozygosity π_d

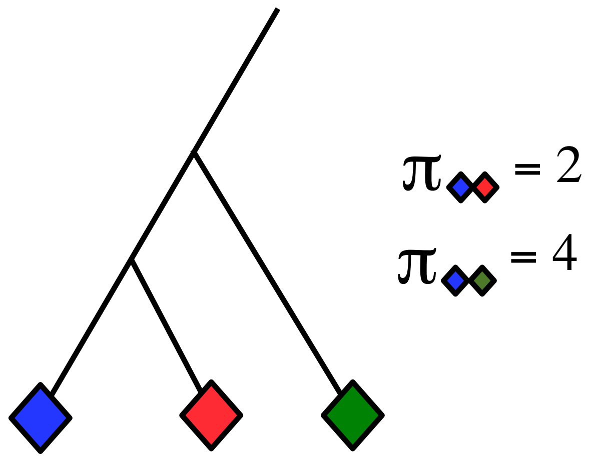
$$\rho(\pi_d) = \sum_{\ell=0}^{\pi_d/2} \sum_{k=0}^{\infty} H(k, k+m = k+\pi_d - 2\ell) P_k^{k+m=k+\pi_d-2\ell}(\tau = \ell)$$



MCP: $< \pi_d > = 2U_d/s$

π_d – distance in number of mutations between individuals





FGS:
$$\rho(\pi_d = r) = \sum_{k=r-k-m} H(k, k+m)$$
$$= e^{-2U_d/s} \frac{1}{r!} (\frac{2U_d}{s})^r$$

Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

$$\Psi(t) = \sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\sum_{\ell=0}^{k} \Psi(t|, k+m, \ell)\phi_k^{k+m}(\tau = \ell)H(k, k+m)$$

Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

$$\Psi(t) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \sum_{\ell=0}^{k} \Psi(t|, k+m, \ell) \phi_k^{k+m}(\tau = \ell) H(k, k+m)$$

distribution of actual coalescence time conditional on them coalescing

probability to coalesce ℓ steps ago

average over class frequencies

Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

$$\Psi(t) = \sum_{k=0}^{\infty} \sum_{m=0}^{\infty} \sum_{\ell=0}^{k} \Psi(t|, k+m, \ell) \phi_k^{k+m}(\tau = \ell) H(k, k+m)$$

(longer of the actual mutation times+time for coalescence in class k-ℓ) ~ distribution of actual coalescence time conditional on them coalescing

probability to coalesce ℓ steps ago

average over class frequencies

Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

$$\Psi(t)=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\sum_{\ell=0}^{k}\Psi(t|,k+m,\ell)\phi_k^{k+m}(\tau=\ell)H(k,k+m)$$

(longer of the actual mutation times+time for coalescence in class k-ℓ) ~ distribution of actual coalescence time conditional on them coalescing

probability to coalesce ℓ steps ago

average over class frequencies

- as in the traditional coalescent - neutral mutations distributed according to a Poisson process where time is drawn from distribution of coalescence times (branch lengths)

$$\rho(\pi_n)=\int\frac{[2U_nt]^{\pi_n}}{\pi_n!}e^{-2U_nt}\Psi(t)dt$$

Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

$$\Psi(t)=\sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\sum_{\ell=0}^{k}\Psi(t|,k+m,\ell)\phi_k^{k+m}(\tau=\ell)H(k,k+m)$$

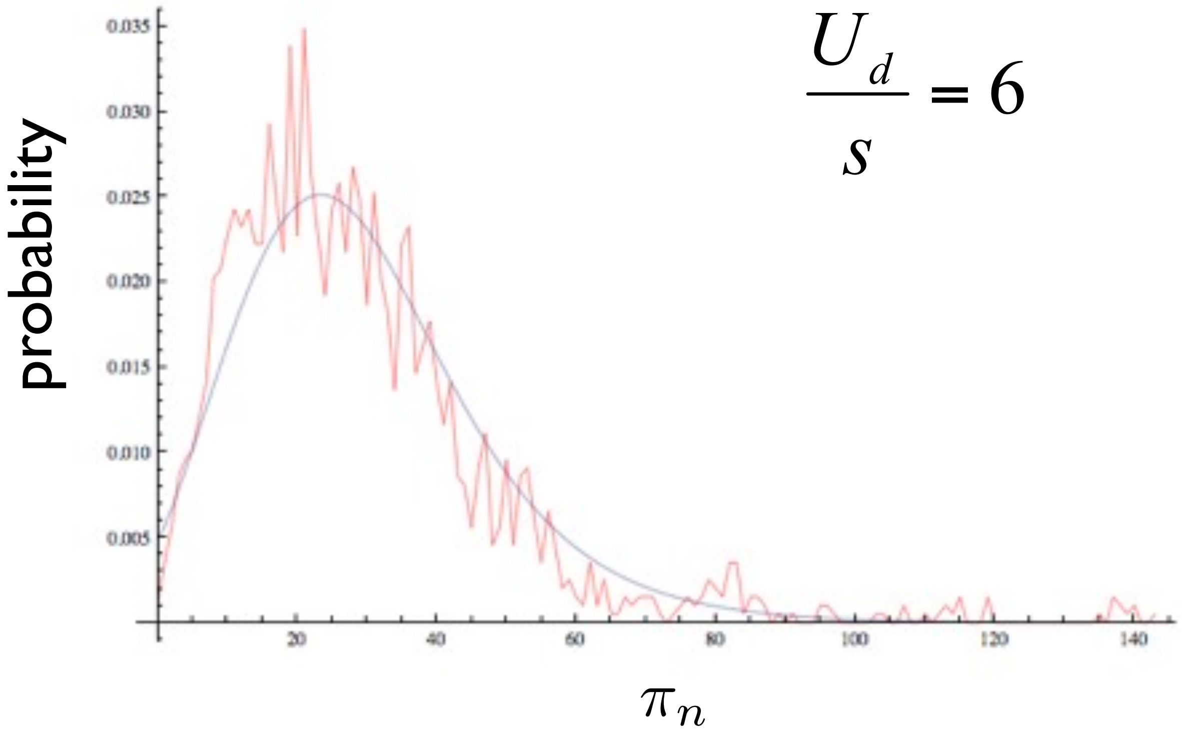
$\Big($longer of the actual mutation times+time for coalescence in class k-$\ell\Big)\sim$ distribution of actual coalescence time conditional on them coalescing

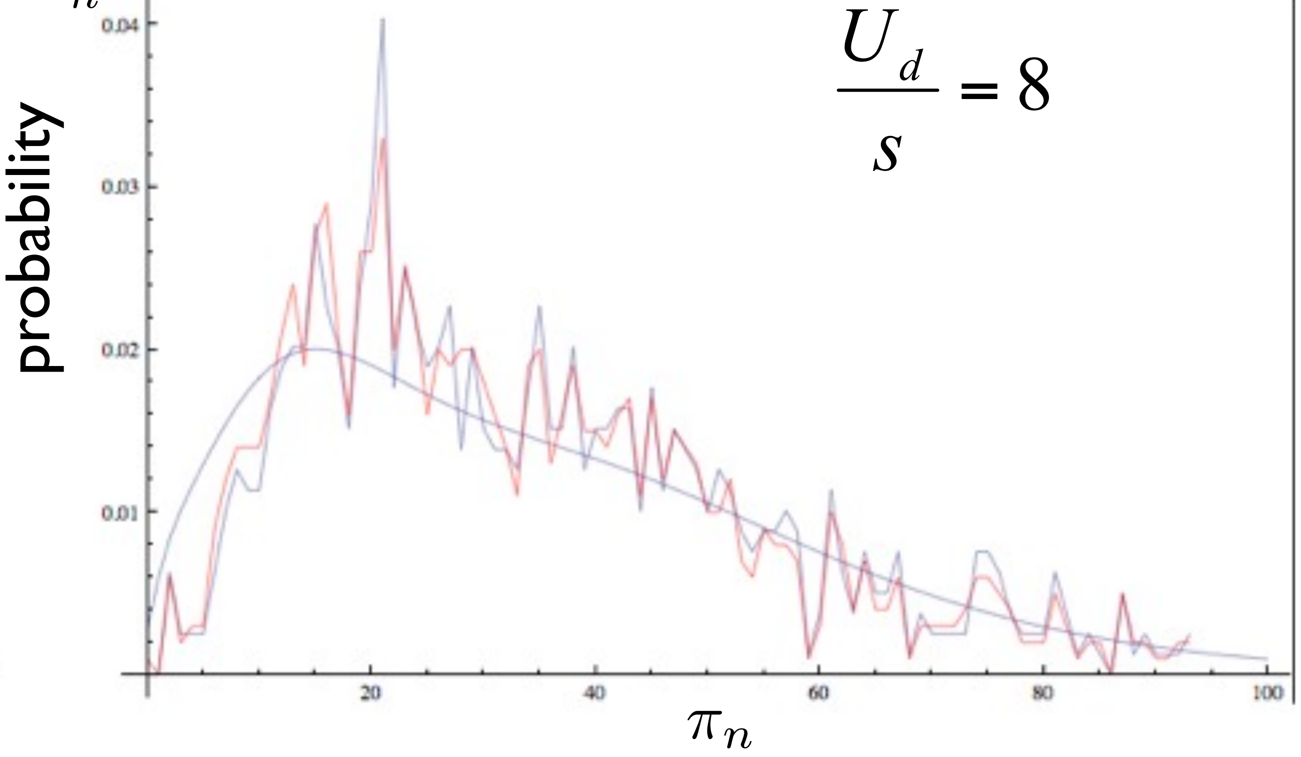
probability to coalesce ℓ steps ago

average over class frequencies

- as in the traditional coalescent - neutral mutations distributed according to a Poisson process where time is drawn from distribution of coalescence times (branch lengths)

$$\rho(\pi_n)=\int\frac{[2U_nt]^{\pi_n}}{\pi_n!}e^{-2U_nt}\Psi(t)dt$$





Effective time to real times and neutral diversity

- need to translate step-times into real times to get the distribution of actual coalescence time between two randomly chosen individuals $\Psi(t)$

$$\Psi(t) = \sum_{k=0}^{\infty}\sum_{m=0}^{\infty}\sum_{\ell=0}^{k} \Psi(t|, k+m, \ell)\phi_k^{k+m}(\tau=\ell)H(k,k+m)$$

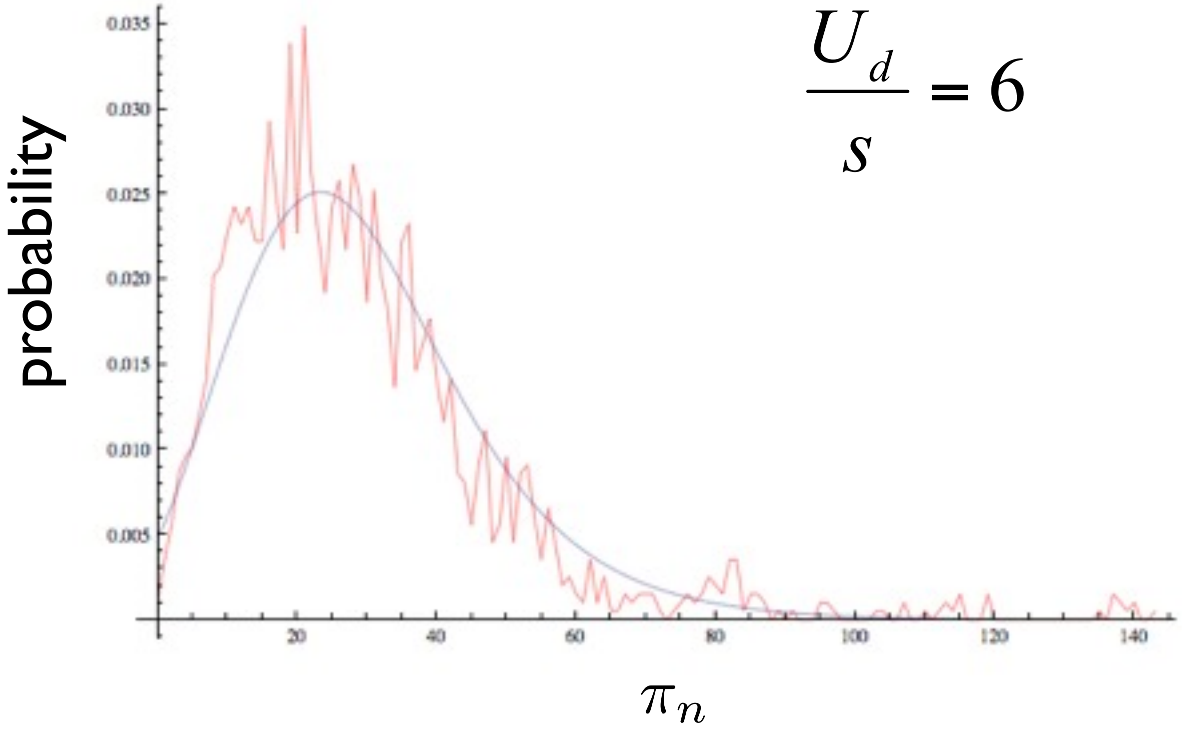
$\left(\text{longer of the actual mutation times+time for coalescence in class k-}\ell\right) \sim$ distribution of actual coalescence time conditional on them coalescing

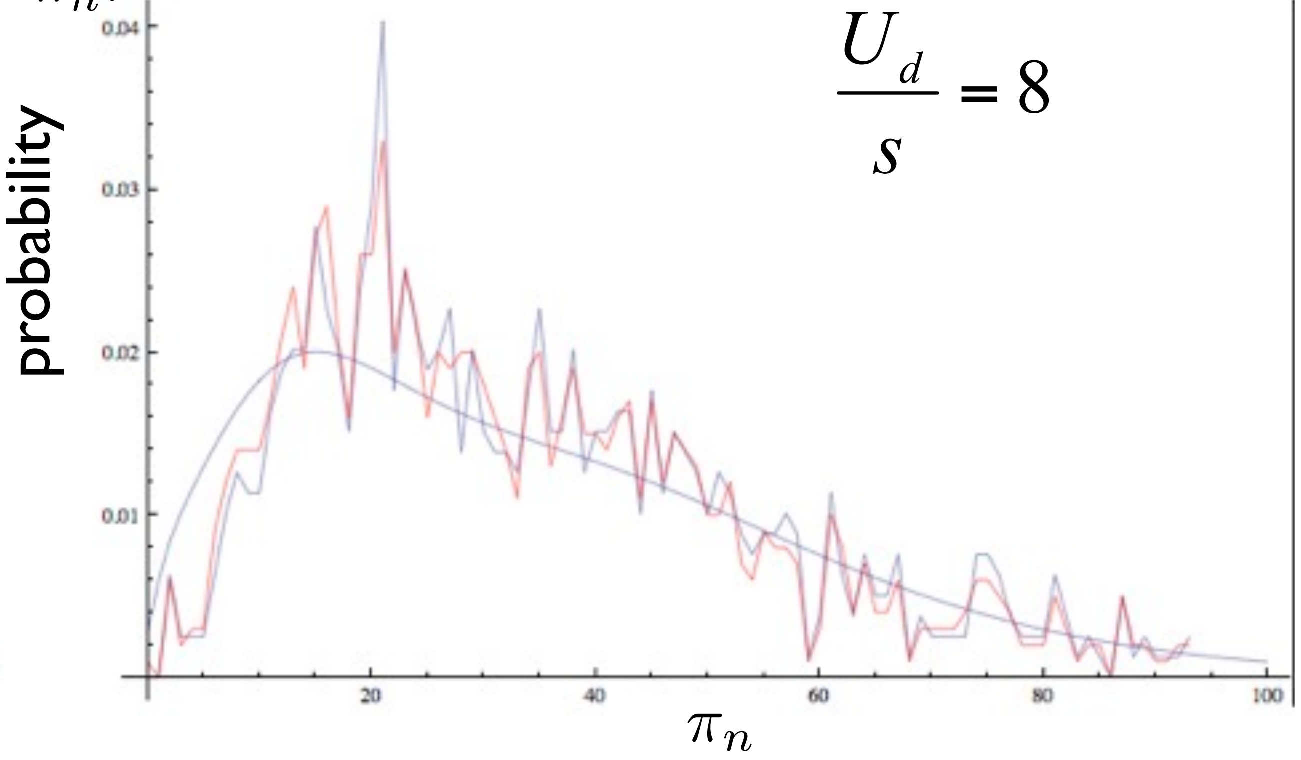
probability to coalesce ℓ steps ago

average over class frequencies

- as in the traditional coalescent - neutral mutations distributed according to a Poisson process where time is drawn from distribution of coalescence times (branch lengths)

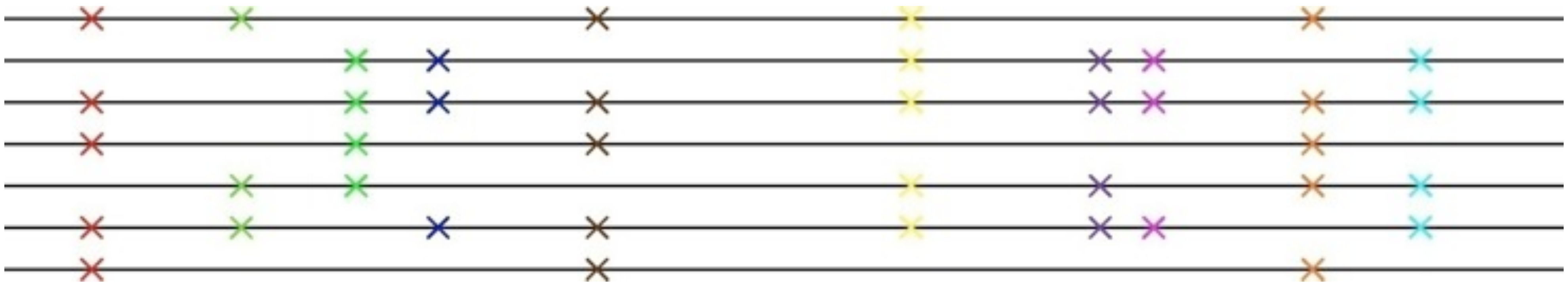
$$\rho(\pi_n) = \int \frac{[2U_n t]^{\pi_n}}{\pi_n!} e^{-2U_n t}\Psi(t)dt$$





- non-zero peak in distribution - unlikely for two individuals to be extremely closely related - from peak in fitness distribution
- non-exponential distribution - difference from neutral case

Connection to data



We can now calculate the expected distribution of any statistic describing variation when negative selection is operating.

We know a bit more about what we're looking for.

Summary

- expansion of coalescence framework to negative selection
  - idea: effectively see how individuals move through fitness distribution
  - do not follow individual ancestry
  - count time is steptimes

PRF

- the genetic variability cannot be mimicked by effective population size

~~N_e~~

- approach works for weak and strong selection
  - strong selection: reproduce results of background selection
  - weak selection: deviations from neutrality, background selection predictions
  - weak selection: heterozygosity signatures clearly distinct from neutral models

U_d/s

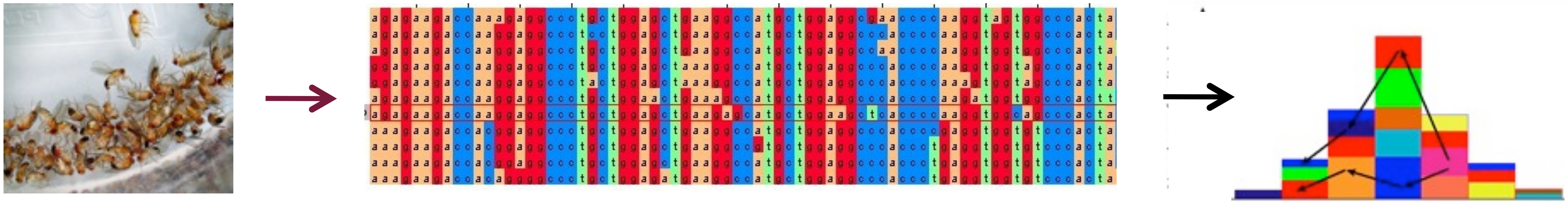
- coalescent probabilities depend on time varying ancestry dependent effective population size

n_k

- mean coalescence path approximation - weak selection, large N
  - coalescence in zeroth class determined by time to get there
  - no N dependence

$$< \pi_d > = 2U_d/s$$

- beneficial mutations
- positive and negative selection



MM Desai, AM Walczak, JB Plotkin, arXiv:1010.2478v1

MM Desai, AM Walczak, LE Nicolaisen, JB Plotkin, arXiv:1010.2479v1