# Coalescent models with linked selection

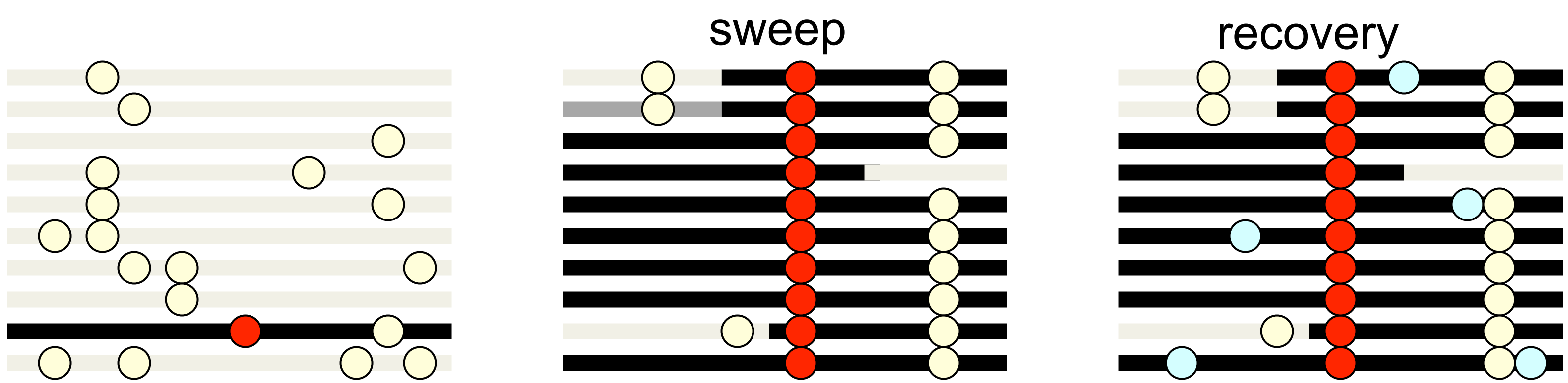
Graham Coop and Peter Ralph
Dept. of Evolution and Ecology
UC Davis

# Outline

- Genome-wide evidence for hitchhiking
- Multiple merger coalescent of full sweeps
- A multiple merger model of recurrent partial sweeps
- A simultaneous multiple merger model of recurrent soft sweeps

# The effect of selective sweeps on linked neutral variants

Maynard Smith and Haigh, Kaplan et al ‘89, etc




Reduced diversity high frequency derived alleles

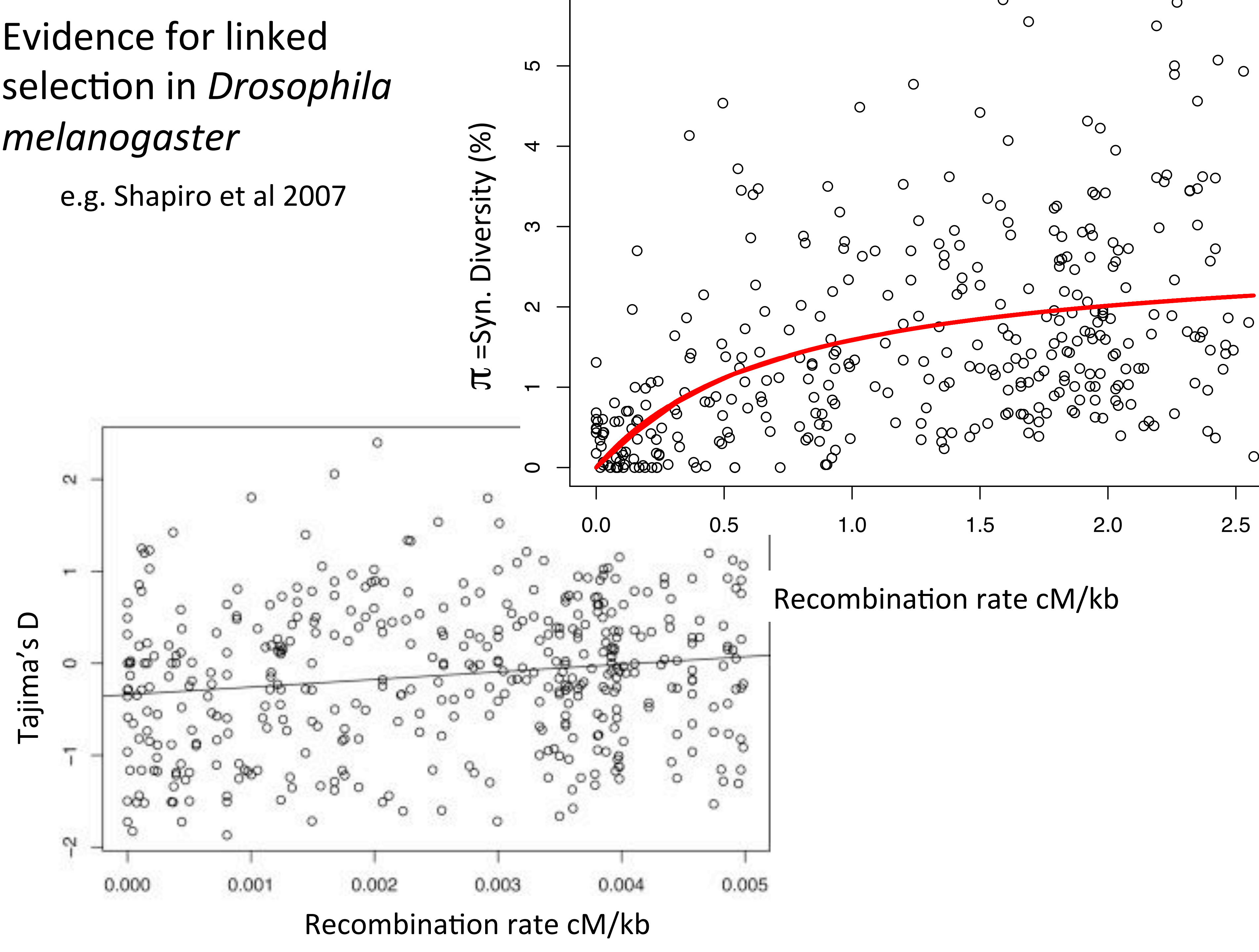
New mutations lead to a skew towards rare alleles

Selective sweep results in a characteristic reduction in coalescent time at linked neutral sites. Also a distortion in the genealogical tree towards external branches and away from internal branches.

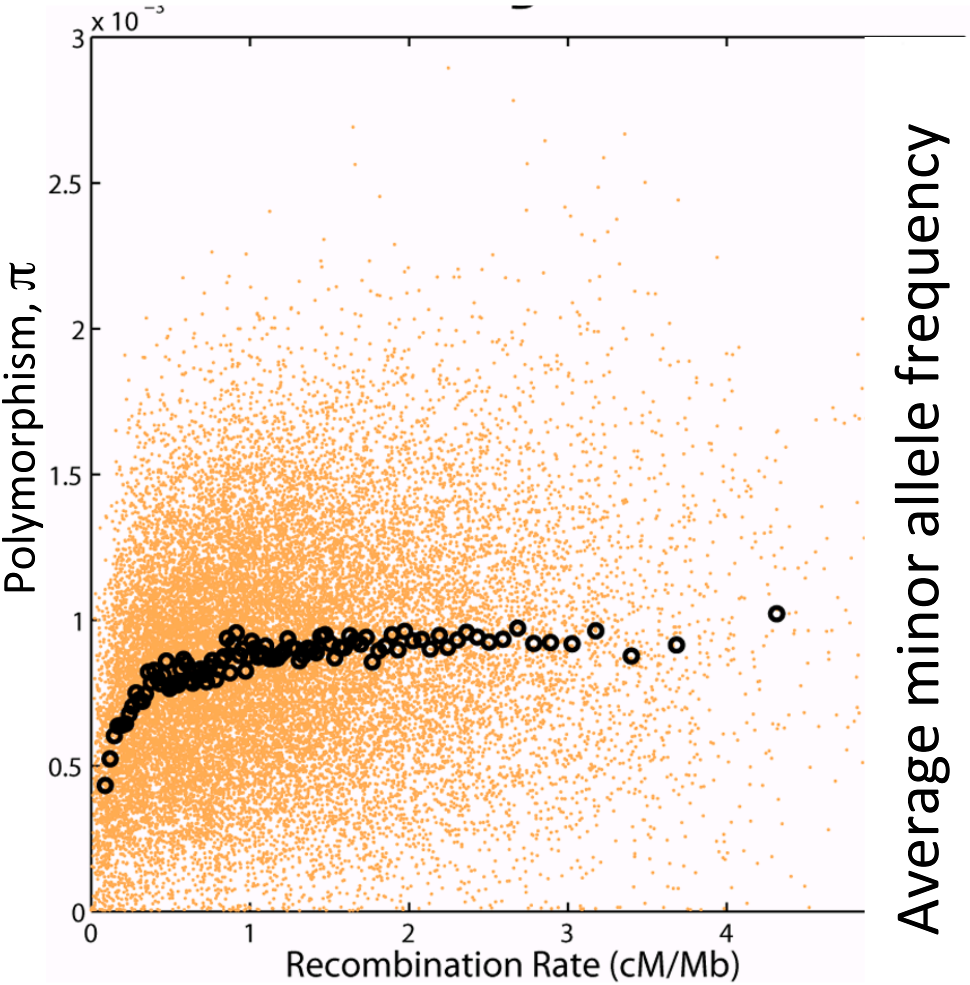
Background selection can also lead to a reduction in diversity, but lead to only a weak skew towards rare alleles

# Evidence for linked selection in *Drosophila melanogaster*

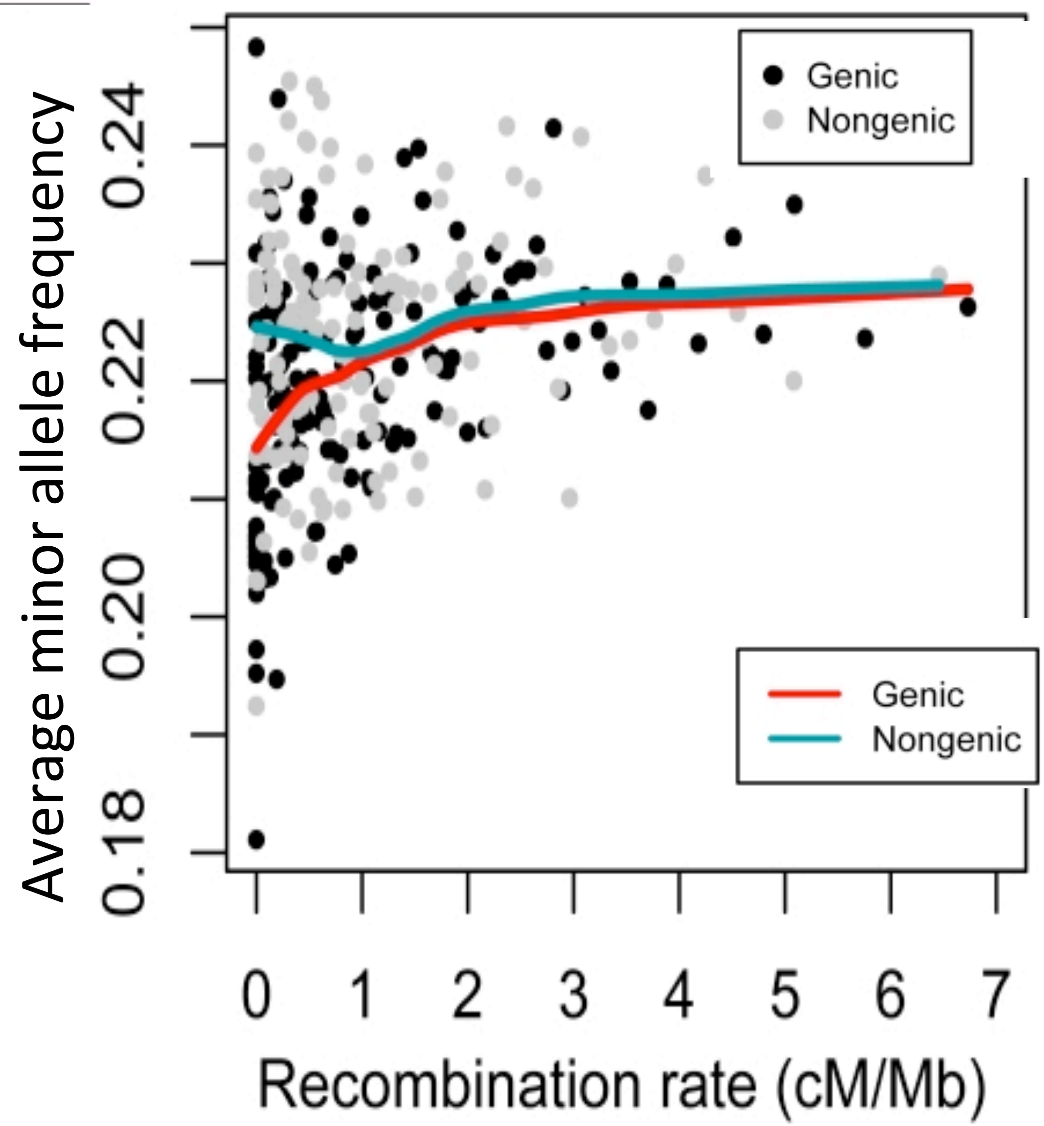
e.g. Shapiro et al 2007



Evidence for variation-reducing selection in humans
But not clear what mode of linked selection acts.




Cai et al. 2009



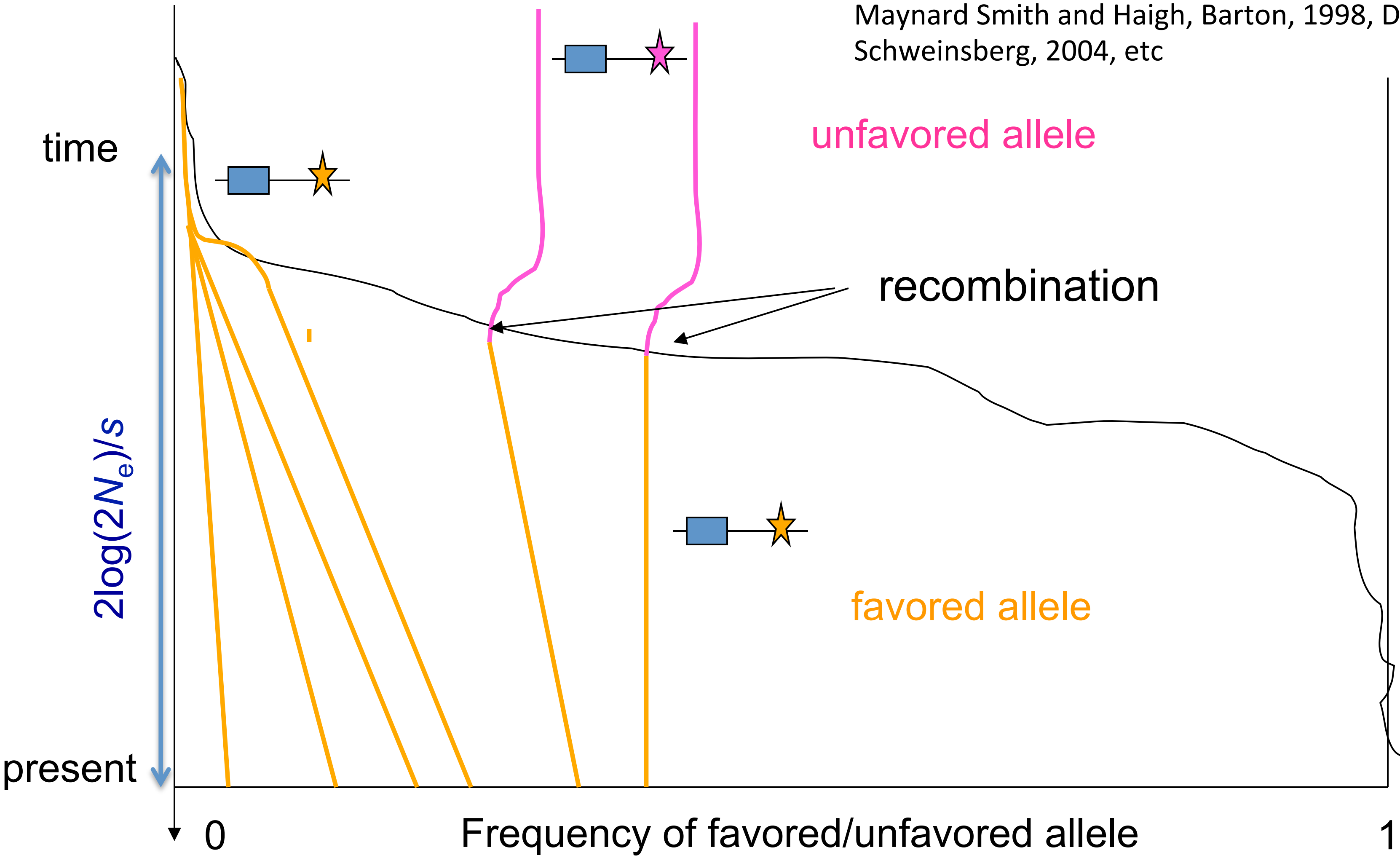

Lohmueller et al., 2011

Time-scale of selective sweep = t $=2\log(2N_e)/s$

Probability of failing to recombine off $q=\exp(-rt/2)$

Probability that i out of k lineages are forced to coalesce ~ Binom(k,q)

Maynard Smith and Haigh, Barton, 1998, Durrett and Schweinsberg, 2004, etc




Barton, 1998; Durrett and Schweinsberg, 2004; Etheridge et al., 2006; Pfaffelhuber et al., 2006,…

Sweeps occur at rate ν with
q ~ f(q) a iid r.v. across sweeps
i lineages out of k lineages
forced to coalesce at rate:

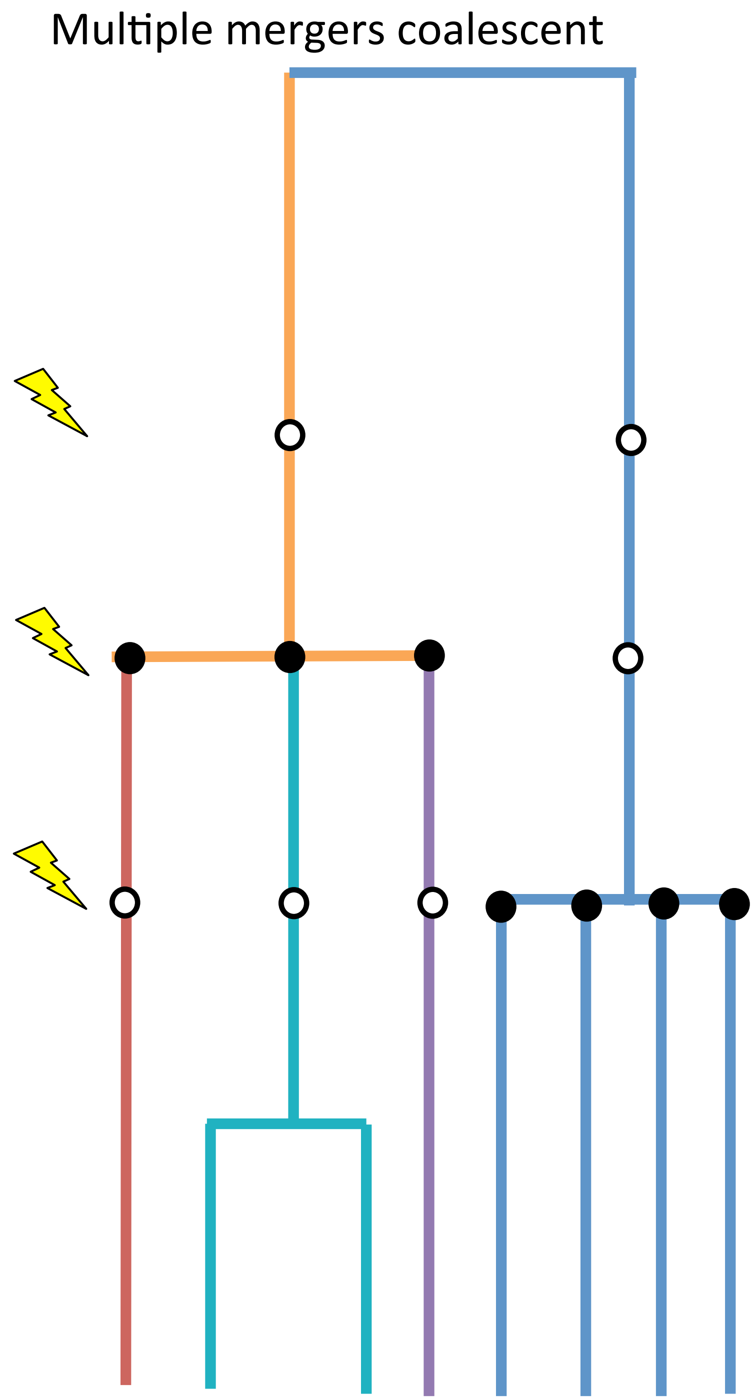
$$\lambda_{k,i} = \binom{k}{2}\frac{1}{2N}\delta_{i,2} + \nu I_{k,i} \quad \text{for } 2 \leq i \leq k,$$

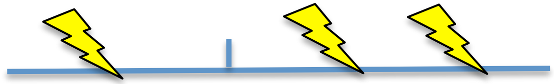
$$I_{k,i} = \binom{k}{i}\int_0^1 q^i(1-q)^{k-i} f(q)\,dq.$$

Gillespie '00, Durrett & Schweinsberg 05

Lambda coalescent:

$$\Lambda(dq) = q^2 \nu f(q)\,dq + \delta_0(dq)/2N$$





Homogeneous sweeps at rate $v_{BP}$,
recombination at rate $r_{BP}$.
Then i out of k lineages coalesce at rate:

$$\lambda_{k,i} = \frac{1}{2N}\binom{k}{2}\delta_{i,2} + \frac{\nu_{BP}}{r_{BP}} J_{k,i} \quad \text{for } 2 \le i \le k,$$

Kaplan et al 1989,

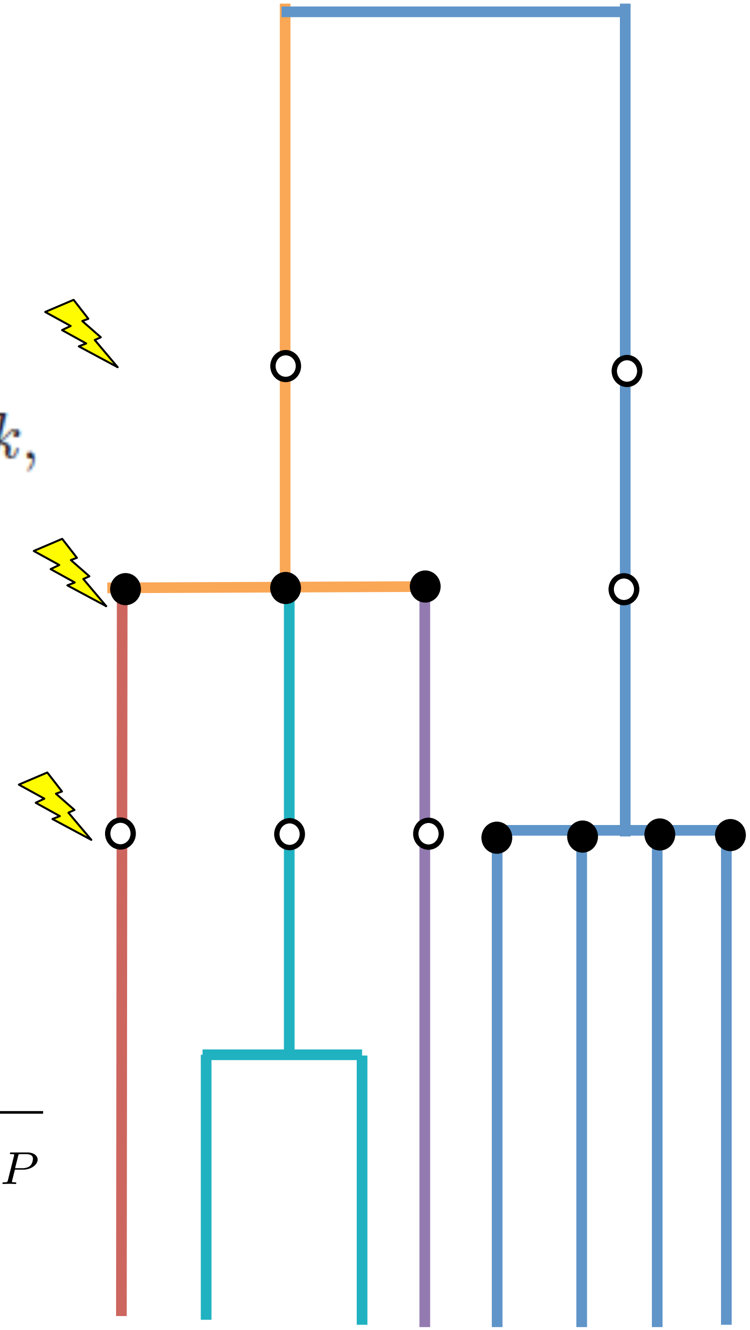
$$J_{k,i} = \binom{k}{i}\int_0^\infty q(r)^i (1-q(r))^{k-i}\,dr$$

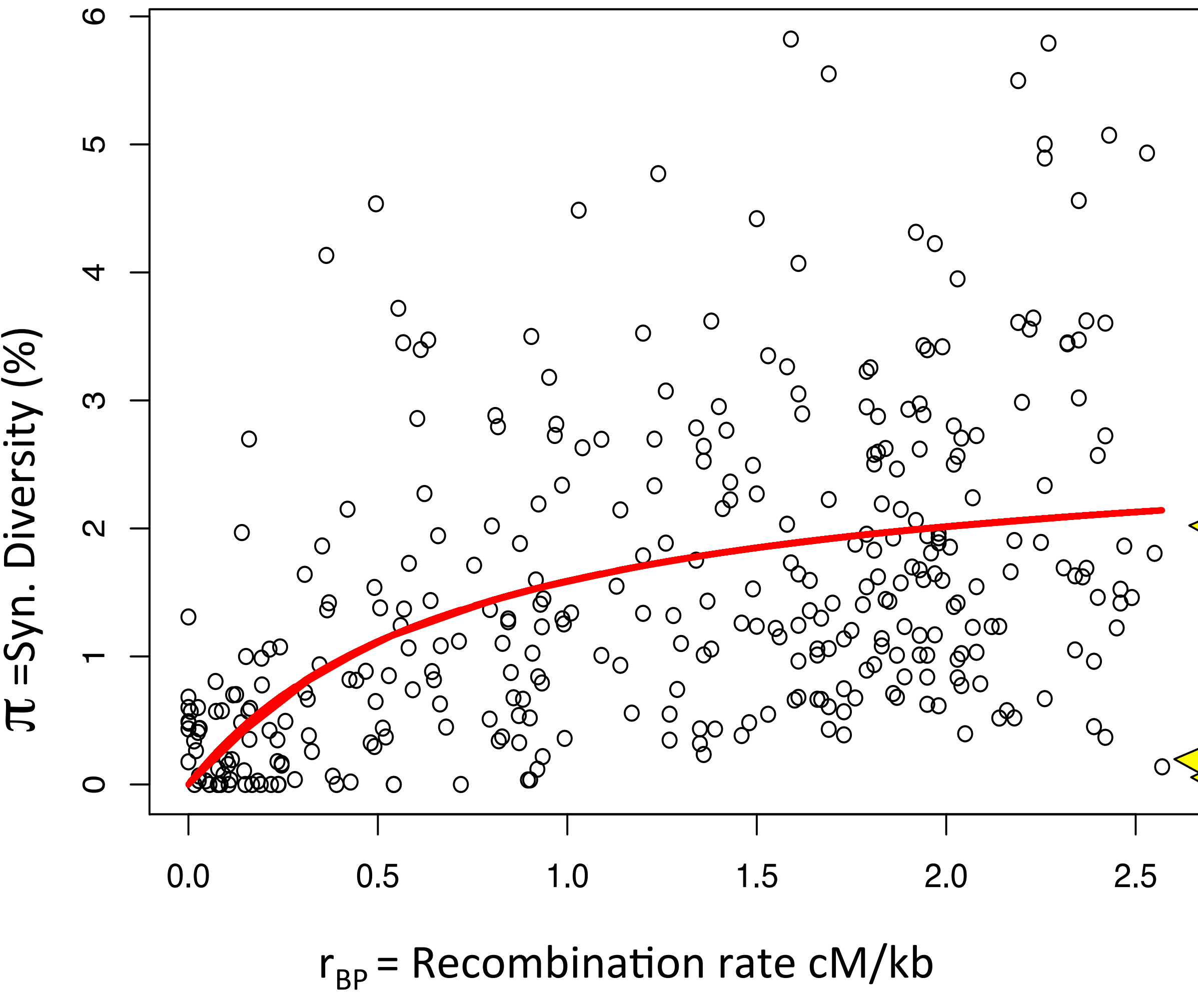
Durrett & Schweinsberg 05

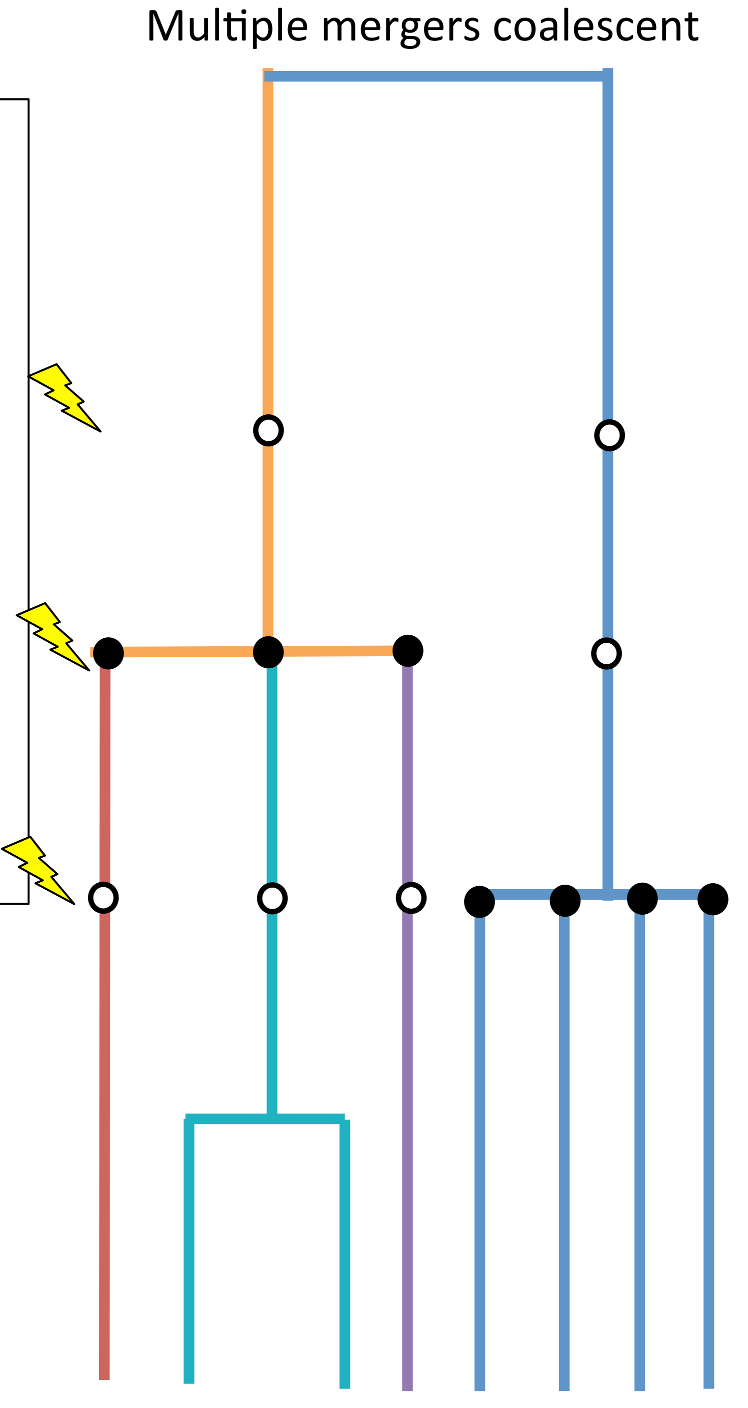
$$\mathbb{E}(\pi) = 2u\mathbb{E}(T_2) = \frac{4Nu}{1 + 2N\nu_{BP}J_{2,2}/r_{BP}}$$

Kaplan et al. (1989) and Stephan et al. (1992)

Multiple mergers coalescent



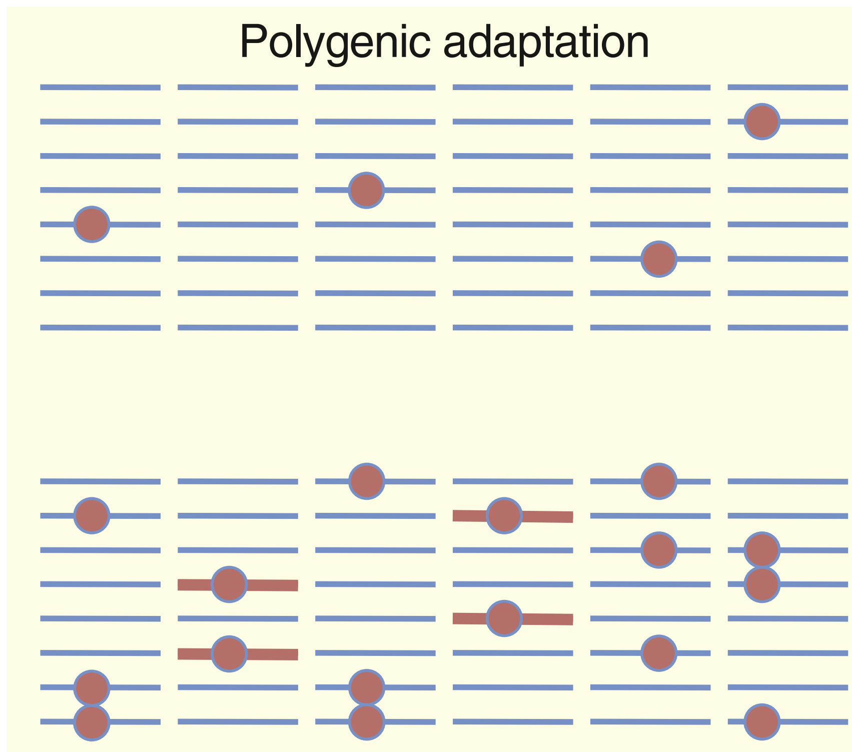



$$\mathbb{E}(\pi) = 2u\mathbb{E}(T_2) = \frac{4Nu}{1 + 2N\nu_{BP}J_{2,2}/r_{BP}}$$

Kaplan et al. (1989) and Stephan et al. (1992)

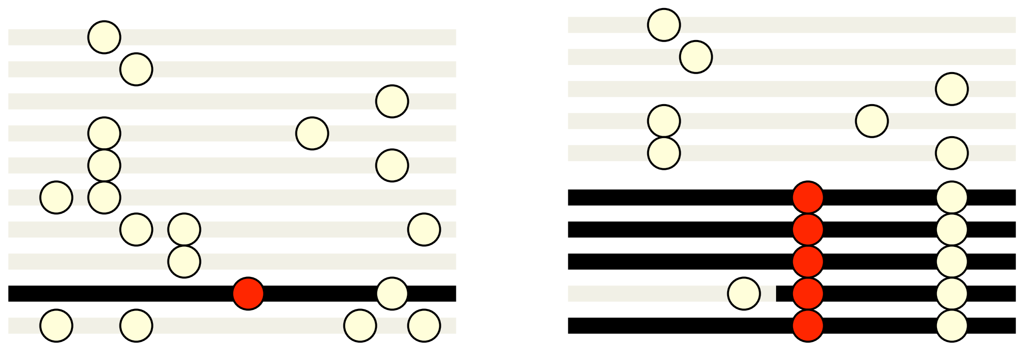
What if most newly arisen selected alleles do not sweep to rapidly fixation?
E.g. due to changing environment or genomic background
(Due to parallel mutation, other standing variation etc)

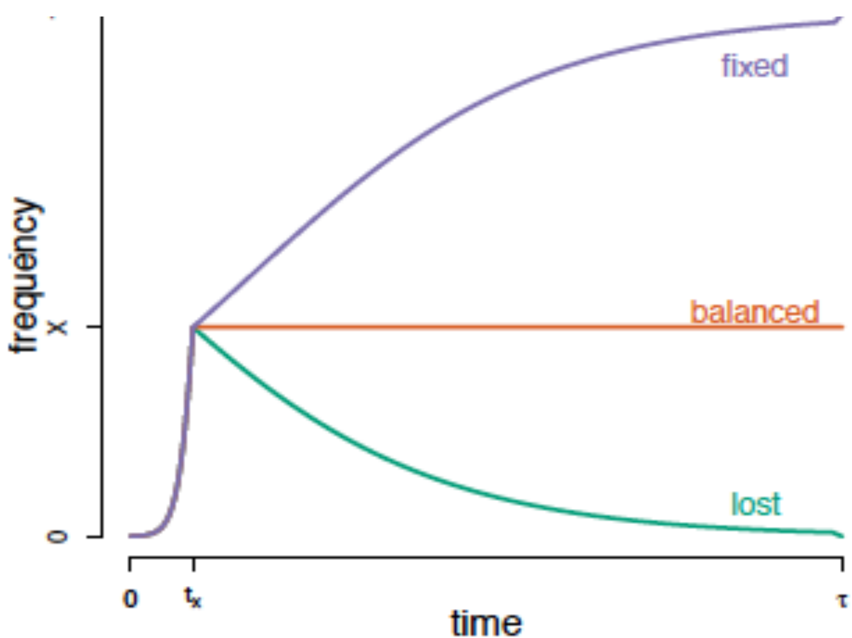



Pritchard, Pickrell, Coop 2010 Current Biology

Pennings and Hermisson, 2006a,b; Chevin and Hospital, 2008; Ralph and Coop, 2010, Innan and Kim, 2004; Hermisson and Pennings, 2005; Przeworski et al., 2005

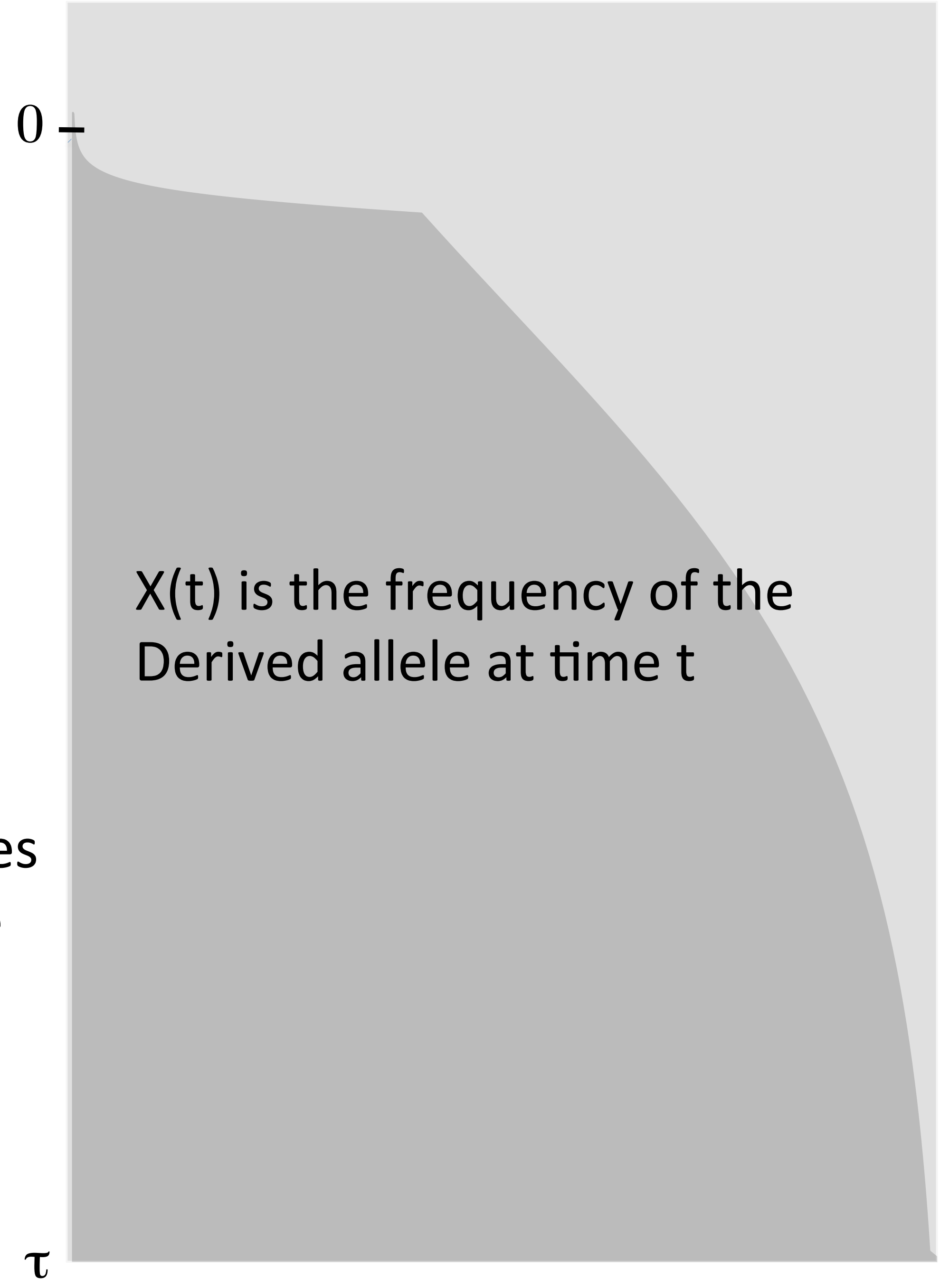
What if most newly arisen selected alleles do not sweep to rapidly fixation?
E.g. due to changing environment or genomic background
(Due to parallel mutation, other variation etc)





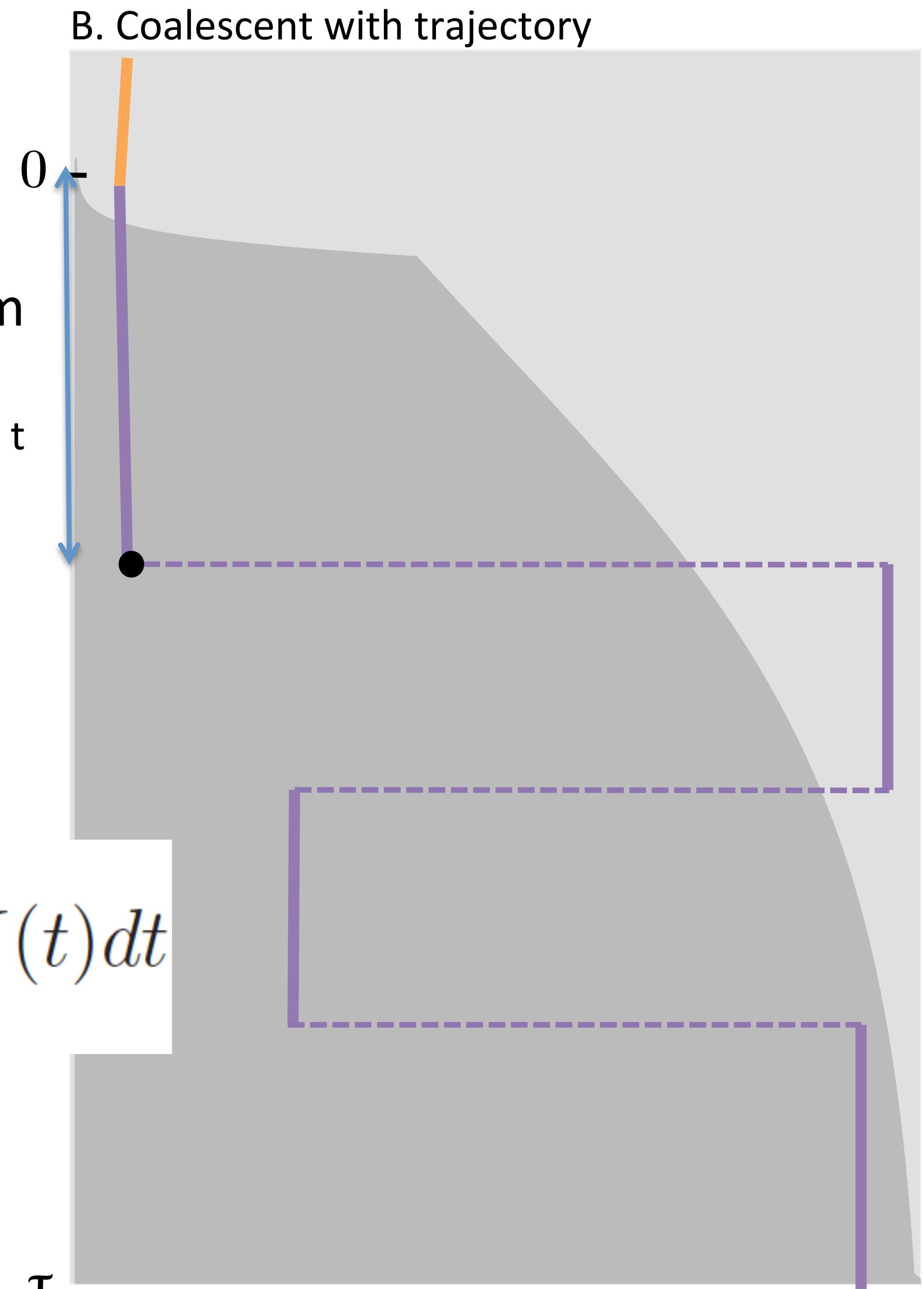
B. Coalescent with trajectory

The derived allele arose τ Generations ago




Conditions on trajectory:
Selected allele initially quickly increases in frequency. If it approaches 0 or 1 it does not renter the Population.

Imagine a neutral site
a genetic distance r away from
the selected locus




Probability that the
lineage is of the derived
type at time $0$ =

$$q(r, X) = r \int_0^{\tau} e^{-rt} X(t) dt$$

For r τ >>1

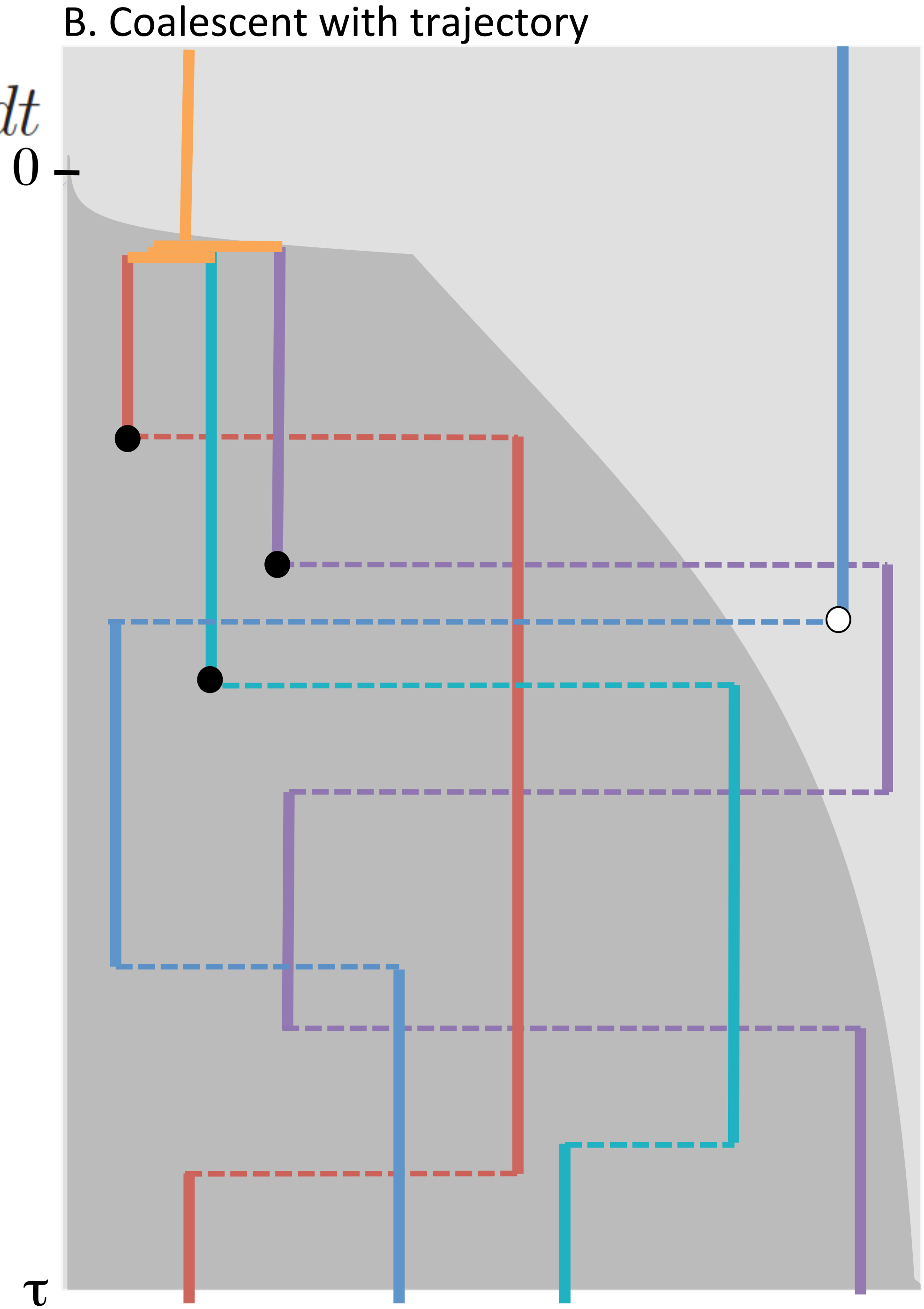
$$q(r, X) = r \int_0^{\tau} e^{-rt} X(t)\,dt$$

Probability that i out of k lineages are force to coalesce is binomial:

$$\binom{k}{i} q^i (1-q)^{k-i},$$

$$\text{for} \quad 2 \le i \le k,$$

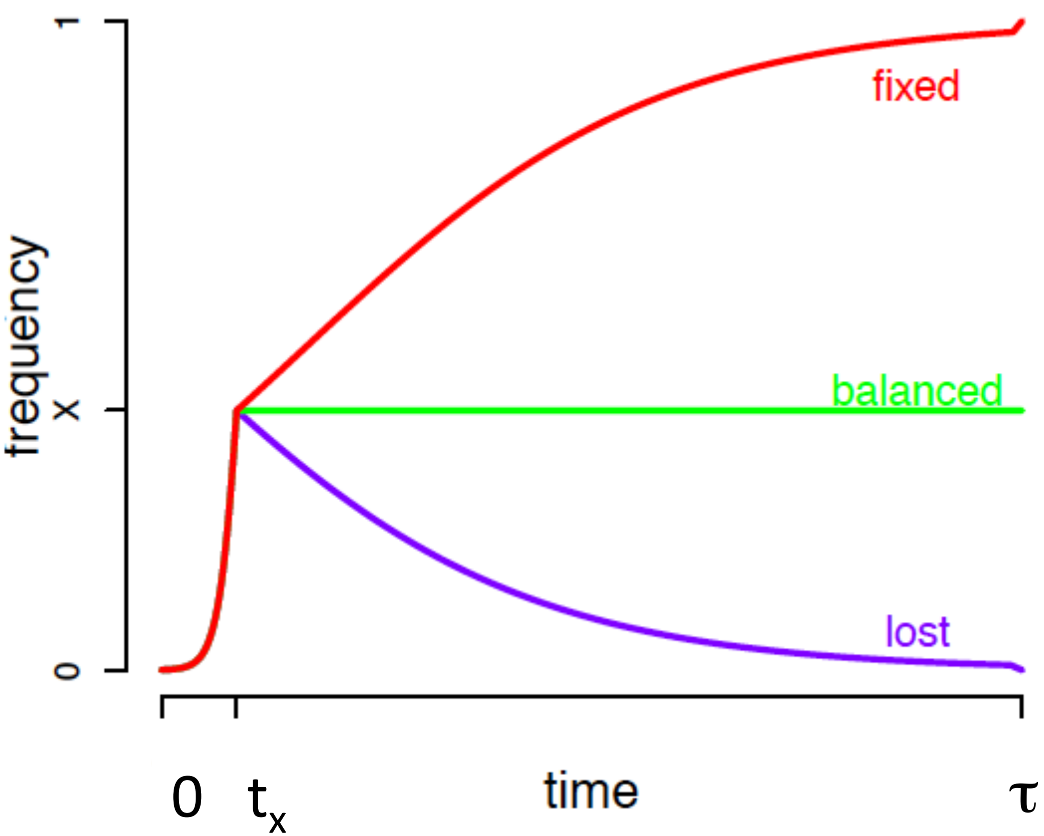
Assuming that the all coalescence happens close to time 0, rN >> 1



# Simple trajectories

Selected allele moves quickly from 1/2N to x in time $t_x$

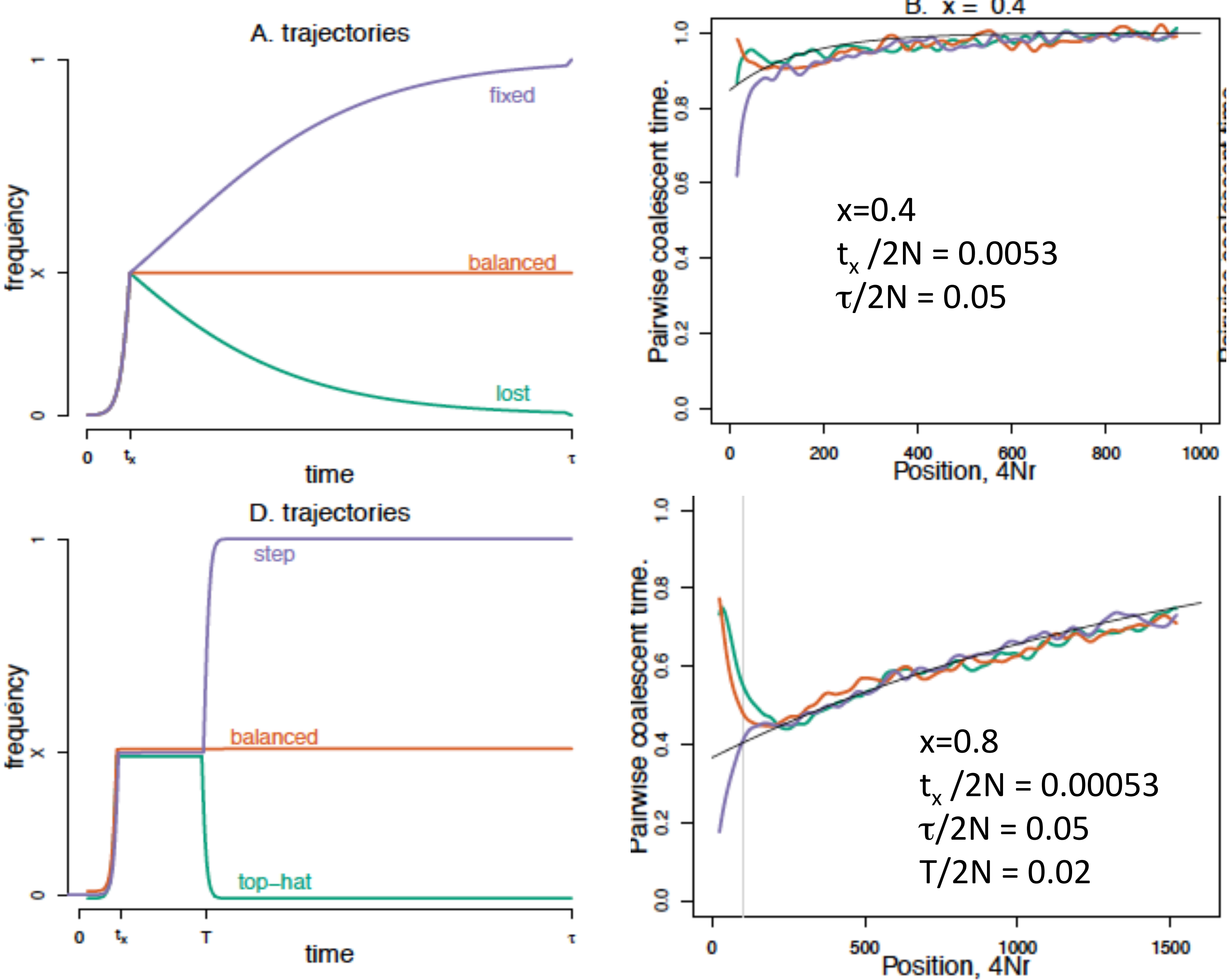
Then stays at x, or goes to fixation, or loss on a slower time-scale (e.g. with selection coefficient $s_2$, $-s_2$, or 0 respectively)



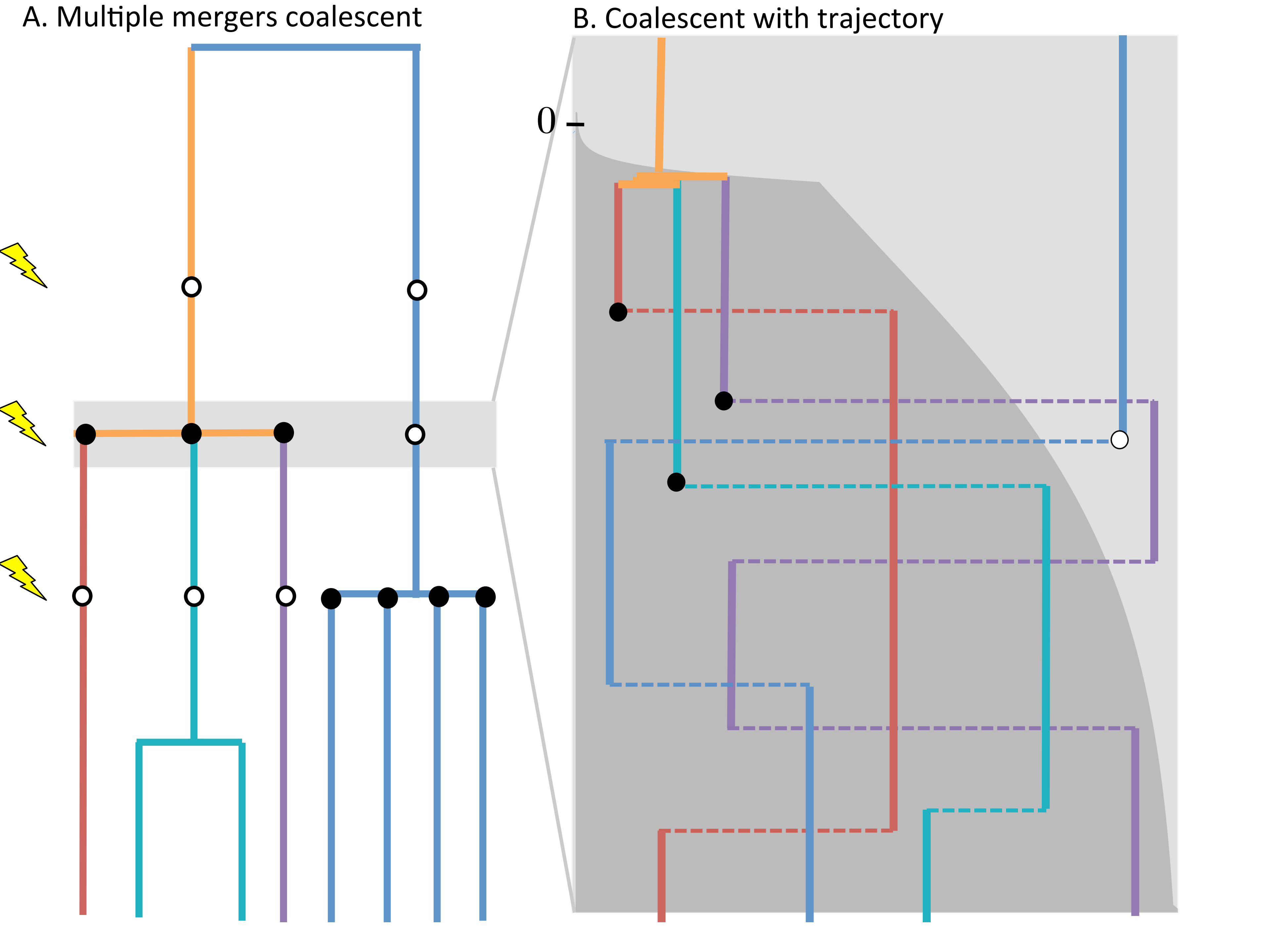

$$q \approx x e^{-r t_x}$$

Also holds for other trajectories when $r >> s_2$

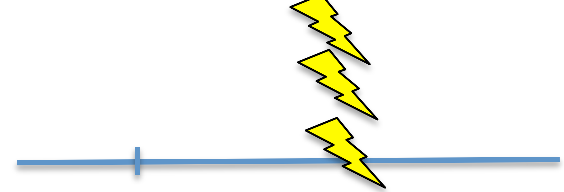
$$\mathbb{E}(T_2) = 2N(1 - q_x^2 e^{\tau/(2N)})$$



A. trajectories
frequency
fixed
balanced
lost
time
B. x = 0.4
Pairwise coalescent time.
Position, 4Nr
x=0.4
$t_x$ /2N = 0.0053
τ/2N = 0.05
D. trajectories
step
balanced
top–hat
time
x=0.8
$t_x$ /2N = 0.00053
τ/2N = 0.05
T/2N = 0.02


A. Multiple mergers coalescent
B. Coalescent with trajectory
0

# Recurrent sweep process



- Assume Neutral pairwise rate of coalescence: 1/(2N)
- Sweeps happen at rate ν
- At a fixed position, with constant q
- Total rate of coalescence of i out of k:

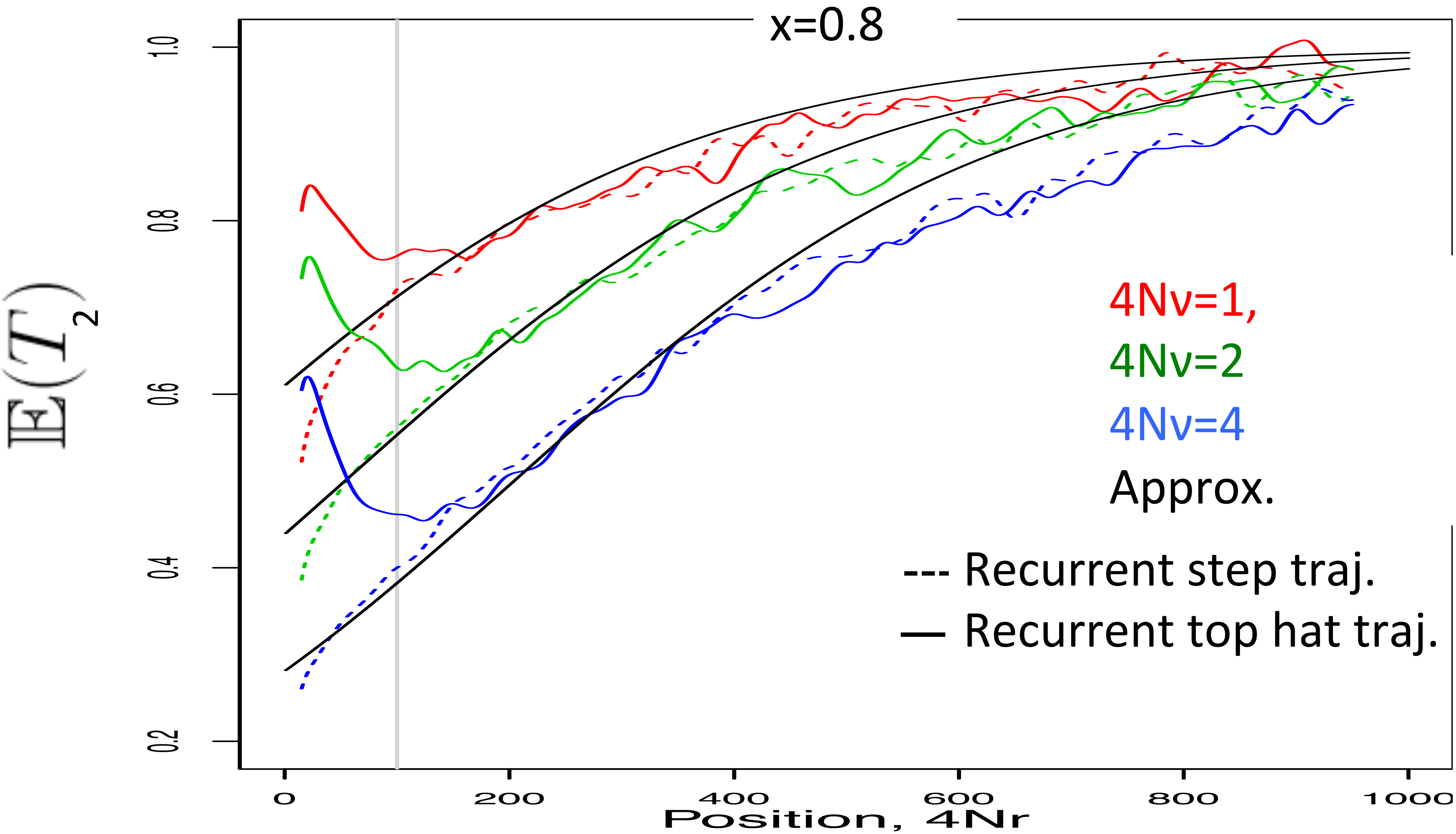
$$\lambda_{k,i} = \binom{k}{2}\frac{1}{2N}\delta_{i,2} + \nu I_{k,i} \quad \text{for } 2 \le i \le k,$$

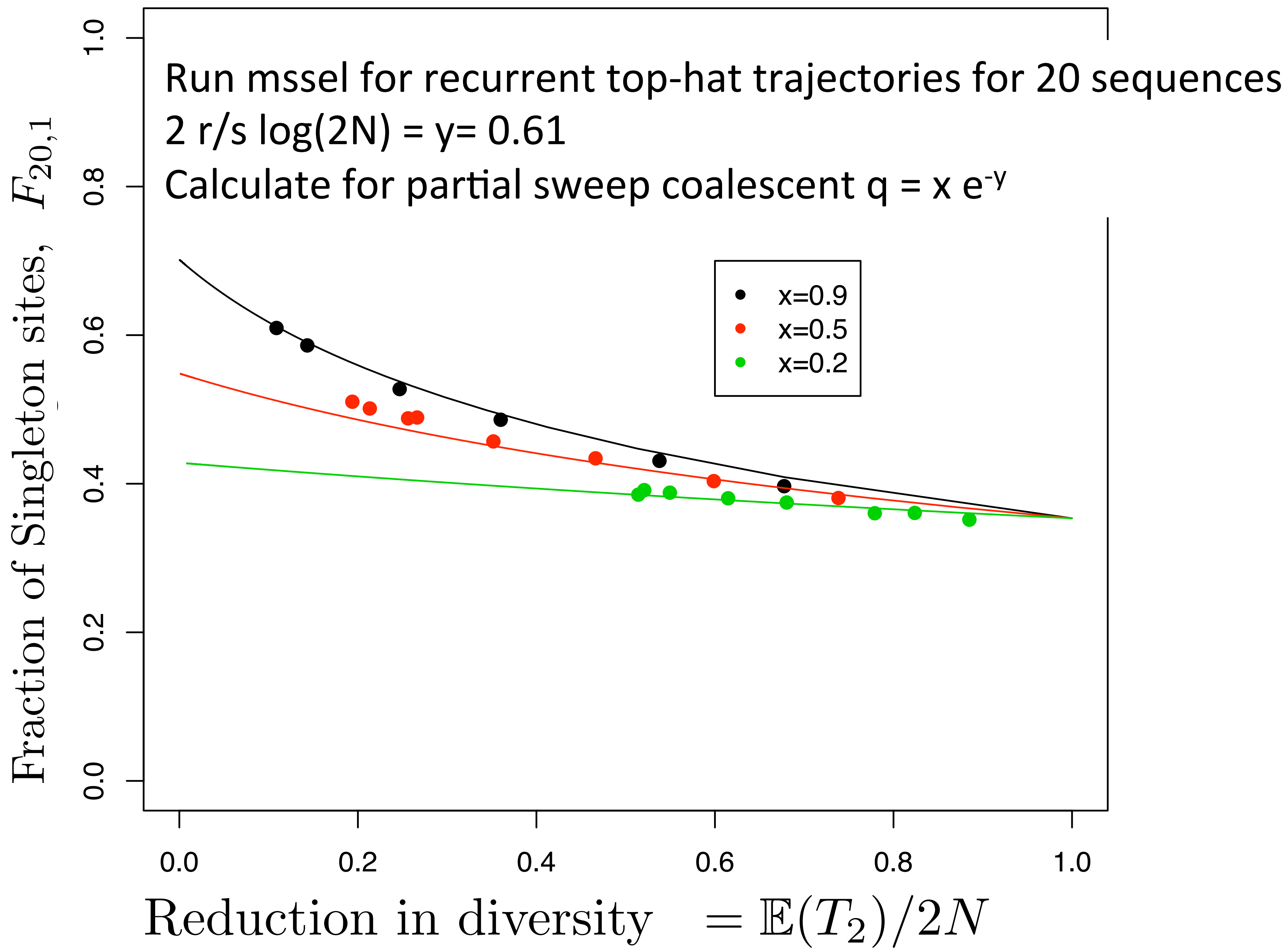
$$I_{k,i} = \binom{k}{i} q^i (1-q)^{k-i}.$$

Inspired by Gillespie '00, Durrett & Schweinsberg 05

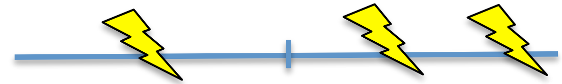
$$\mathbb{E}(T_2) = \frac{2N}{1 + 2N\nu q^2}$$

- For our simple approximation $q \approx xe^{-rt_x}$





Run mssel for recurrent top-hat trajectories for 20 sequences
2 r/s log(2N) = y= 0.61
Calculate for partial sweep coalescent q = x e^-y
x=0.9
x=0.5
x=0.2
Fraction of Singleton sites, $F_{20,1}$
Reduction in diversity $= \mathbb{E}(T_2)/2N$
0.0
0.2
0.4
0.6
0.8
1.0



Homogeneous sweeps at rate $\nu_{BP}$ ,recombination at rate $r_{BP}$.
Then i out of k lineages coalesce at rate:

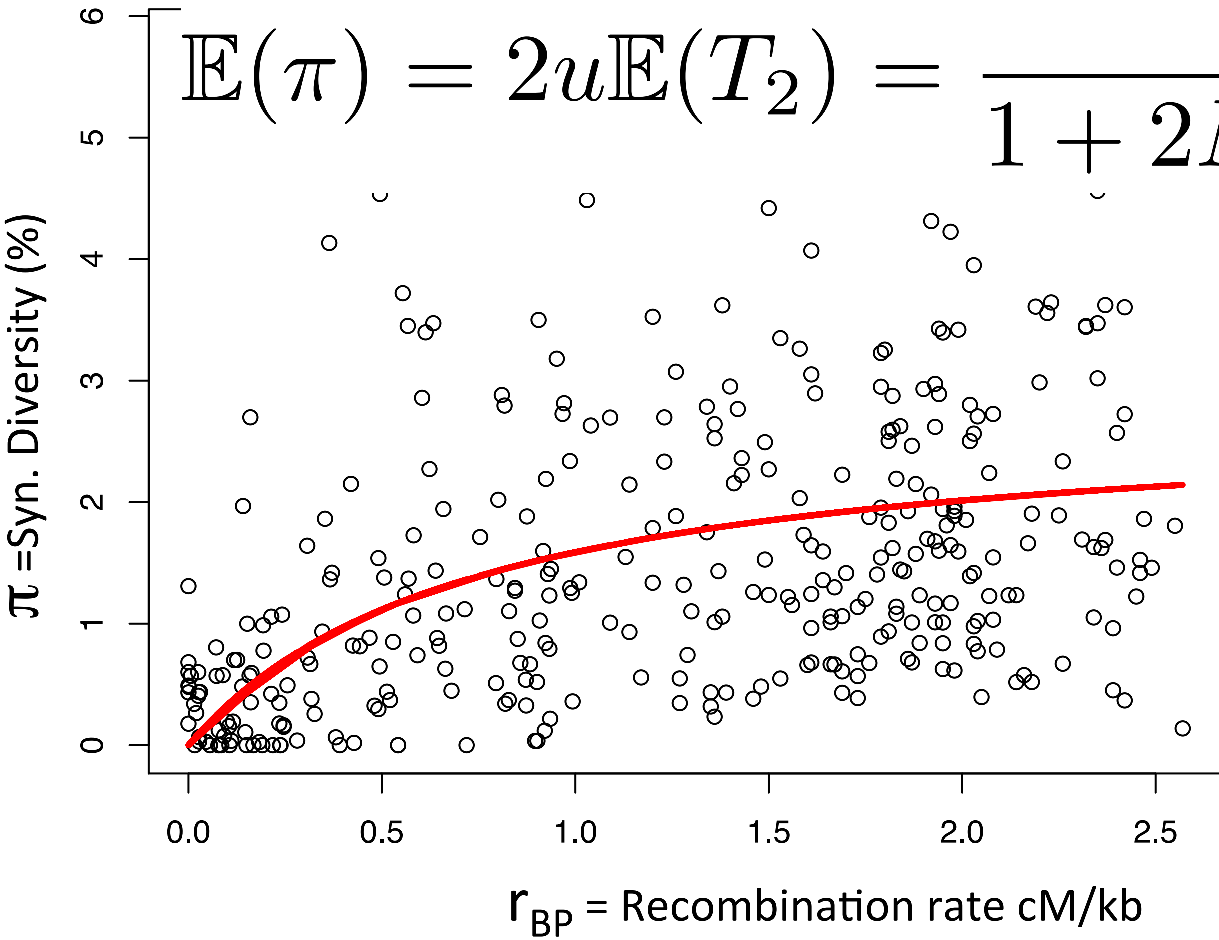
$$= \frac{\nu_{BP}}{r_{BP}} J_{k,i} \qquad \text{for } 2 \leq i \leq k,$$

$$J_{k,i} = \binom{k}{i} \mathbb{E}_X \left[ \int_0^\infty q(r, X)^i (1 - q(r, X))^{k-i} dr \right]$$

Where $J_{k,i}$ depend only on the form taken by trajectories
So rate of coalescence controlled by $\frac{\nu_{BP}}{r_{BP}}$

E.g. for our simple trajectory $J_{k,i}$ is a function of x (freq. sweeps achieve) and so number of lineages forced to coalesce by x (or distribution on x).

$$\mathbb{E}(\pi) = 2u\mathbb{E}(T_2) = \frac{4Nu}{1 + 2N\nu_{BP}J_{2,2}/r_{BP}}$$




Data from
*Drosophila melanogaster*
(Shapiro et al 2007)

$2Nv_{BP}\ J_{2,2}=7x10^{-9}$

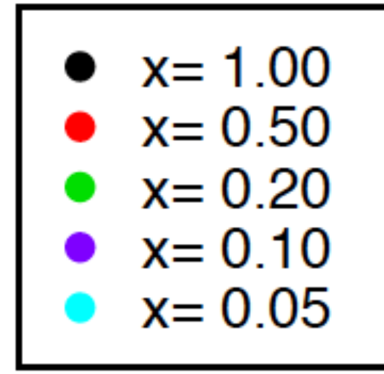
Assuming none of the reduction is due to Background Selection

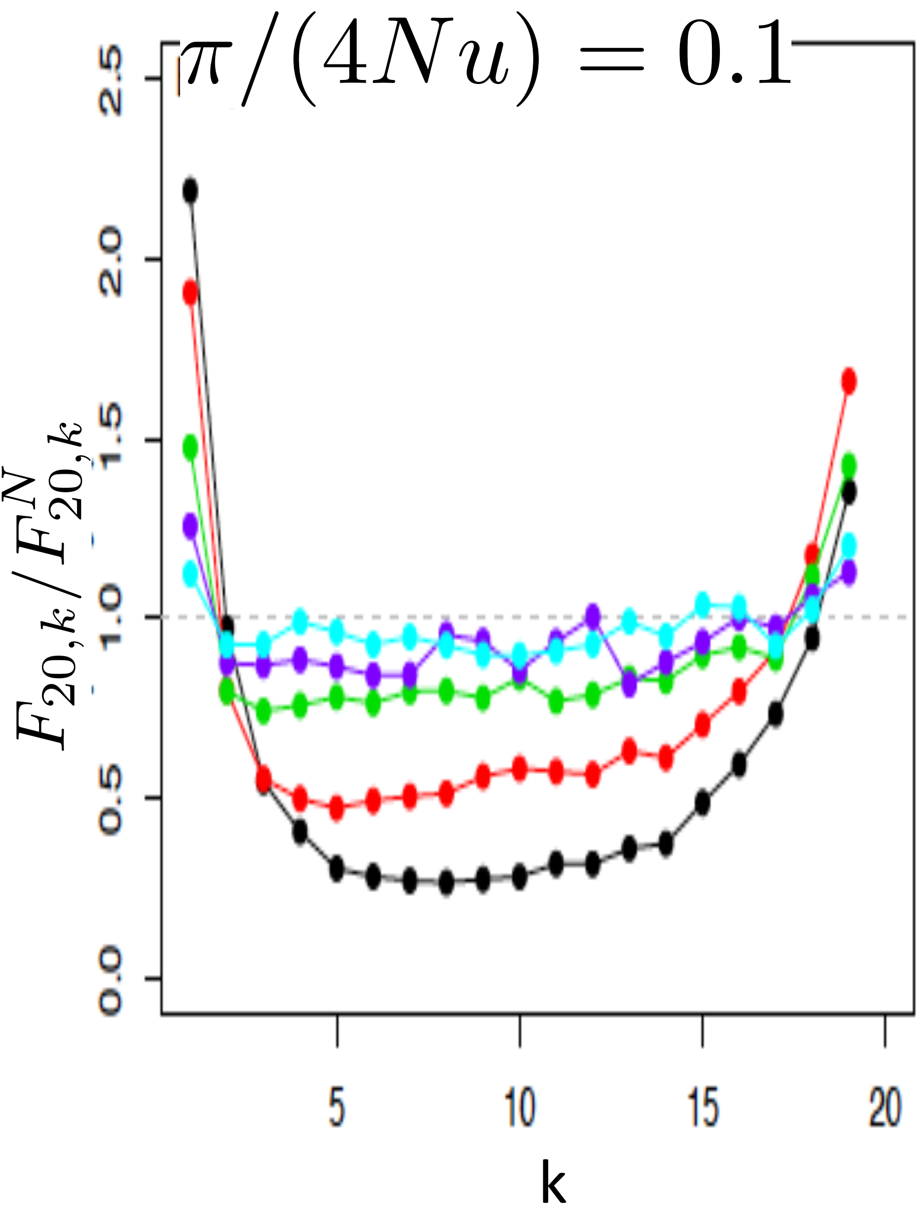
Under our simple partial sweep model: $J_{2,2} = x^2/t_x$

$t_x$ = 1000 gens (s~0.1%), $N=10^6$, $v_{BP}\ x^2 = 3x10^{-13}$

| | | | |
|---|---|---|---|
| x = | 100% | 20% | 5% |
| $v_{BP}$= | 3e-13 | 8e-12 | 1e-10 per generation |

For same reduction in diversity we can get very different distortions to frequency spectrum



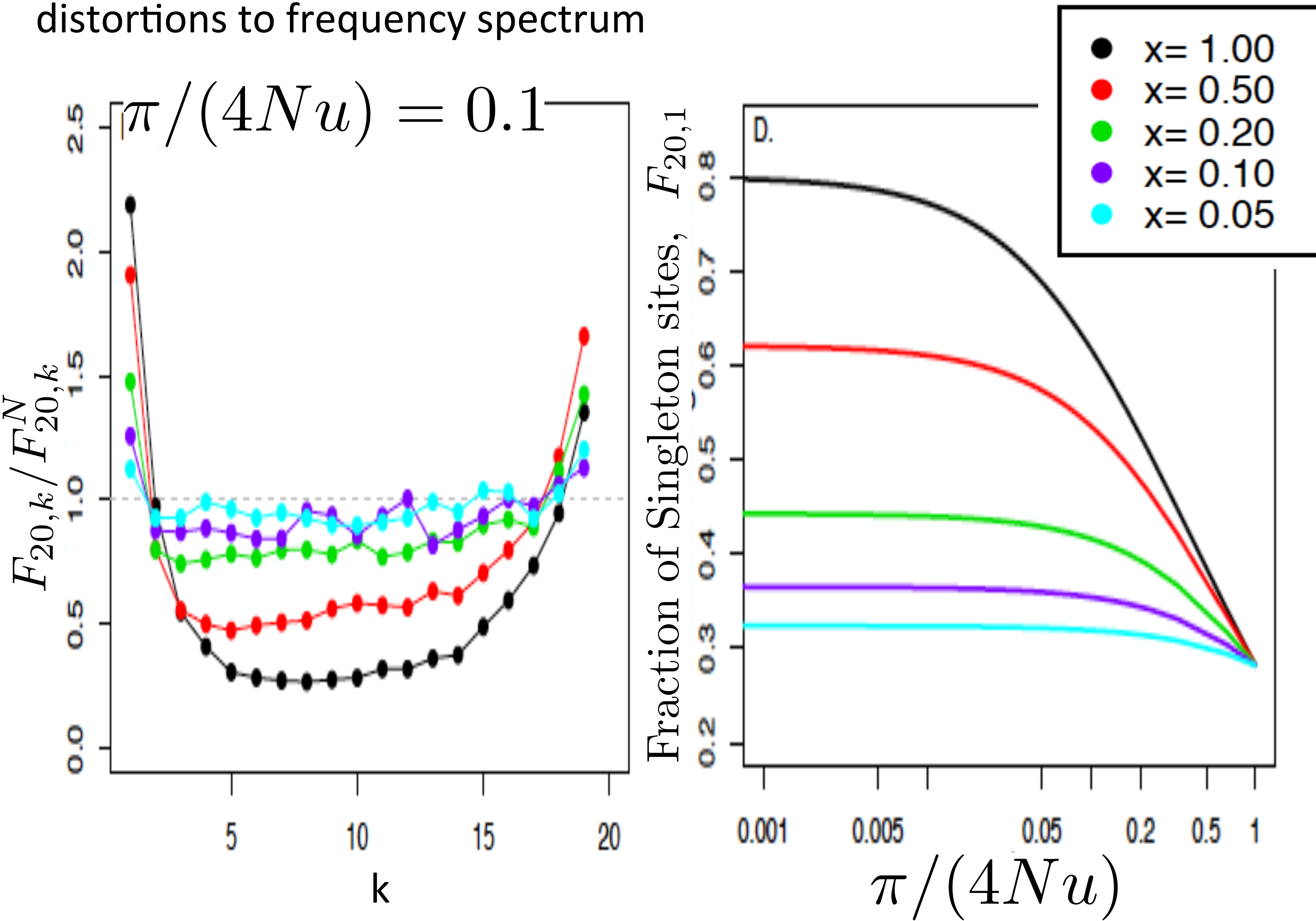




$$F^N_{n,k} = \mathbb{E}(\text{Fraction of sites seen in k out of n})$$

Under Kingman coalescent

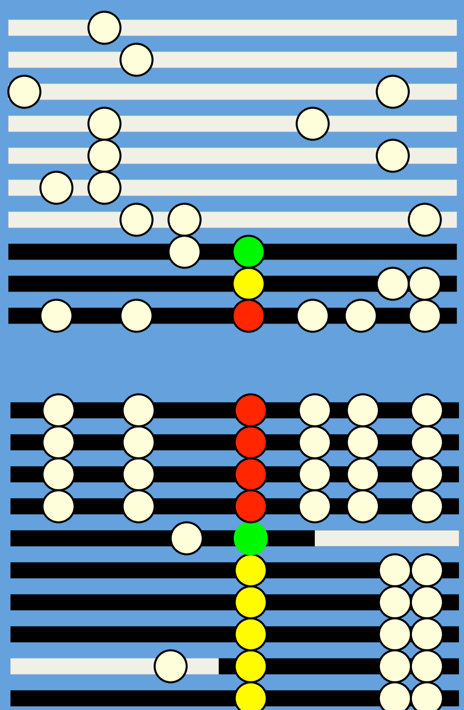
$$F^N_{n,k} = (1/k)/\sum_{j=1}^{n-1}(1/j)$$

For same reduction in diversity we can get very different distortions to frequency spectrum



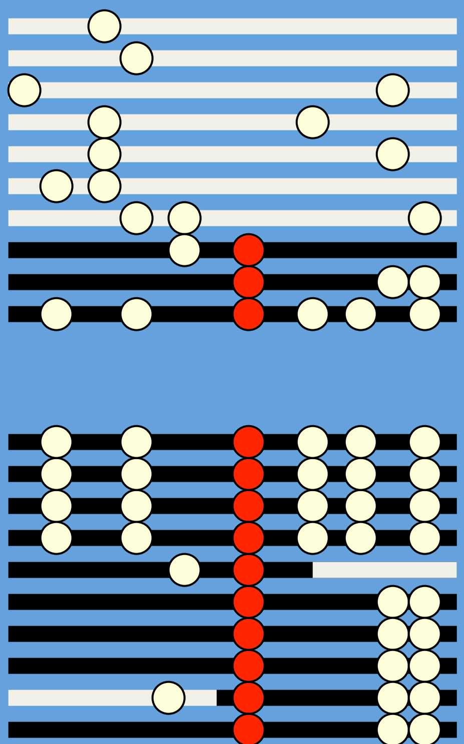
# Soft Sweeps

Selection on
multiple mutations
either standing or new



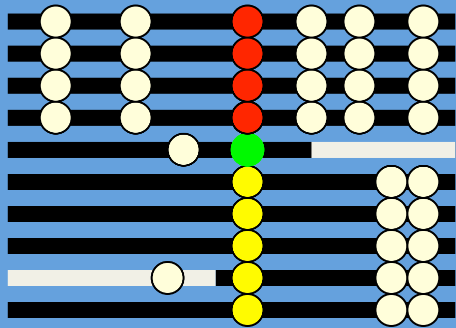
Hermisson and Pennings 05,
Pennings and Hermisson 06

Selection on
standing variation



Przeworski, Coop and Wall 2005

Kim and Innan 05

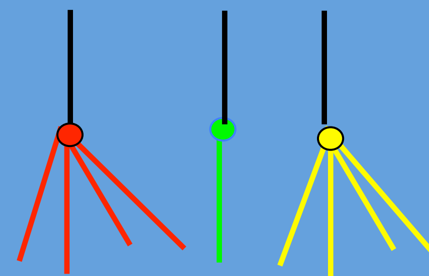


# Soft Sweeps

Pennings and Hermisson showed:

Mutation rate at selected site =$\rho$

At selected site: Lineages assigned to coalescent families (tables) following infinite alleles model with param. 4N$\rho$

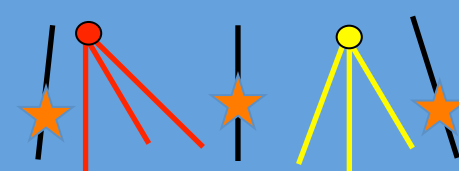


At distance r away lineages recombine off, with probability q, and so escape coalescence.

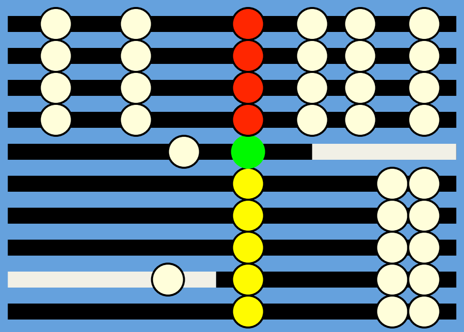
Remaining lineages assigned to coalescent families

$q=e^{-rt}$

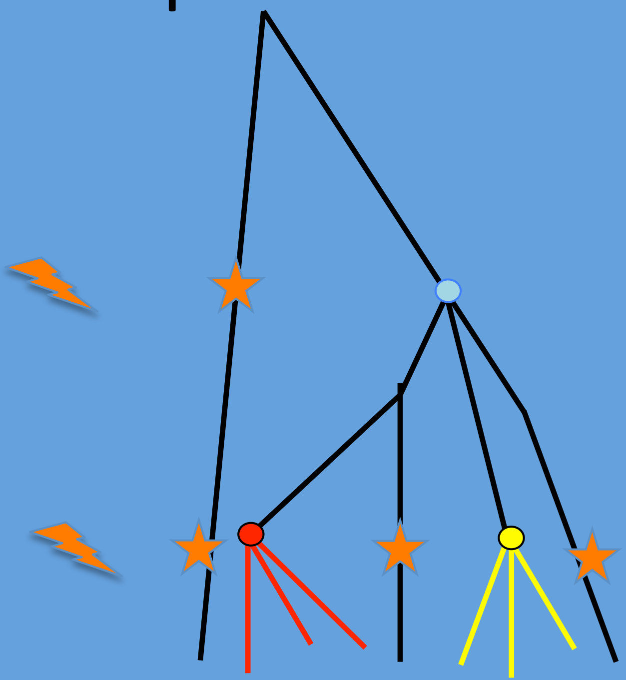
Where t= time of sweep



Hermisson and Pennings 05,
Pennings and Hermisson 06



# Recurrent Soft Sweeps



Neutral coalescence at rate 1/(2N)

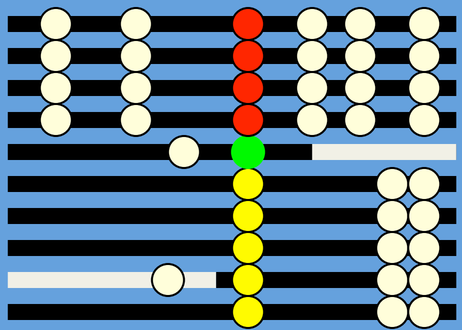
Sweeps occur at rate $\nu_{BP}$ homogeneously along sequence recombining at rate $r_{BP}$

i out of k lineages caught in sweep at rate:

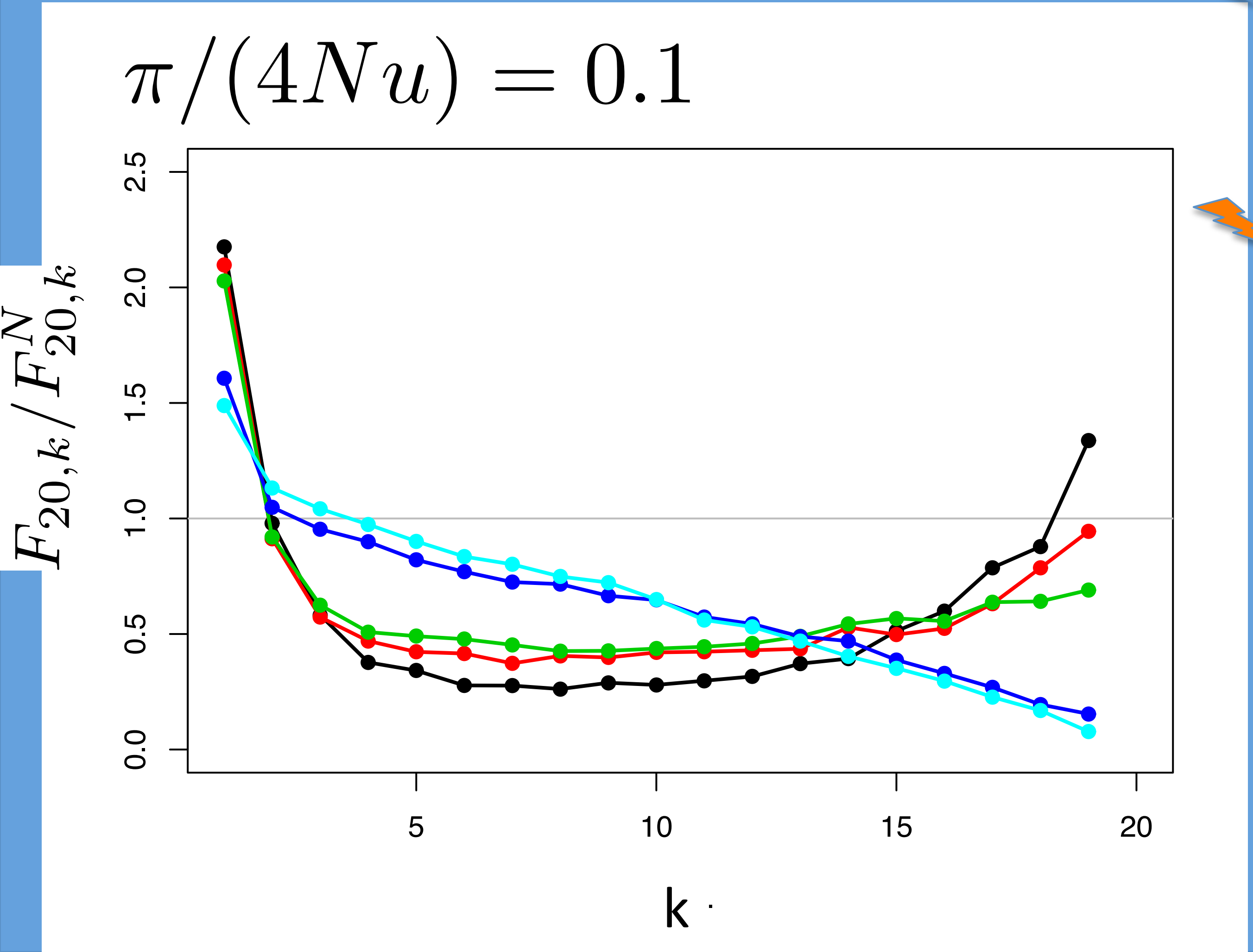
$$\binom{k}{i} \frac{\nu_{BP}}{t\, r_{BP}} \int_0^\infty \left(e^{-r}\right)^i \left(1 - e^{-r}\right)^{k-i} dr$$

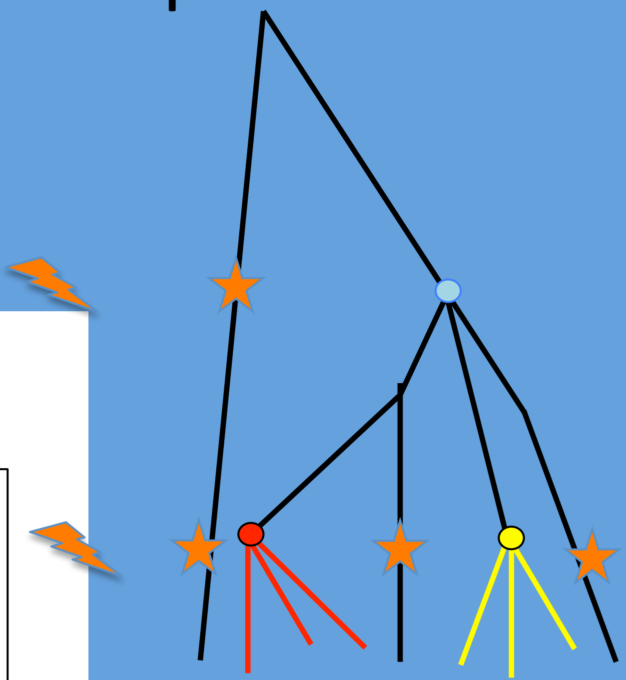
The i lineages are then forced into coalescence families according to infinite alleles model with parameter 4Nρ

$$\mathbb{E}[\pi] = \frac{\theta}{1 + 2N\nu_{BP} J_{2,2} / \left(r_{BP}(1 + 4N\rho)\right)}$$



# Recurrent Soft Sweeps



# Conclusions

- A broad range of linked selection models can be approximated by coalescent models with multiple mergers
- Range of biological models of linked selection depressingly large and predictions overlap.
- Idea: Rather than estimating one model why not estimate rates of different types of coalescence across genome.

# What we need

- Given that the rate of sweeps differs across the genome, what can we hope to learn about the multiple merger process?
- We need theory to predict frequency spectra and haplotype patterns under these models.
- What set of statistics are most informative?
- What set of coalescent processes can we hope to distinguish?

# Thanks

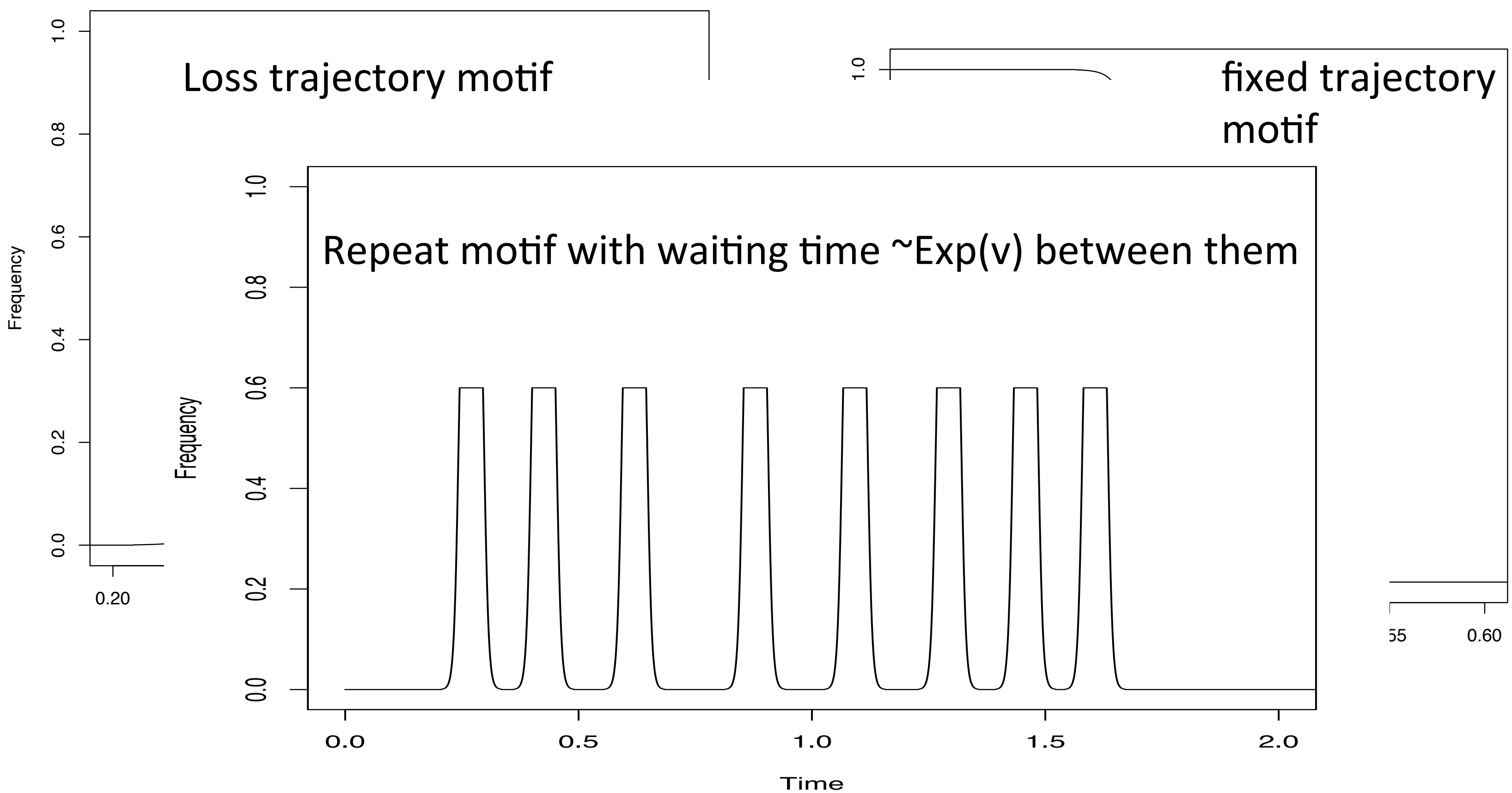
Peter Ralph

Thanks to Yaniv Brandvain, Chuck Langley, Molly Przeworski,
Alisa Sedghifar, and Guy Sella for helpful conversations

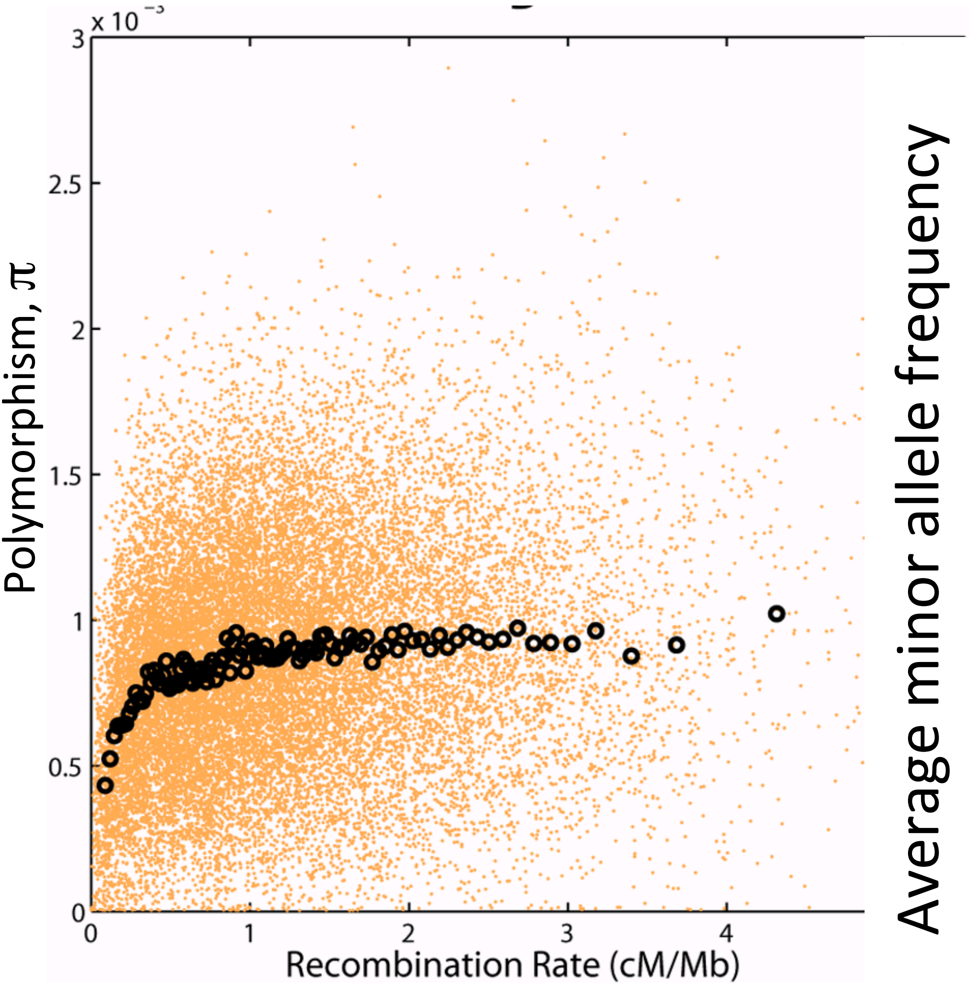
- For our simple approximation $q \approx x e^{-r t_x}$

$$E(T) = \frac{1}{1 + 2Nvq^2}$$

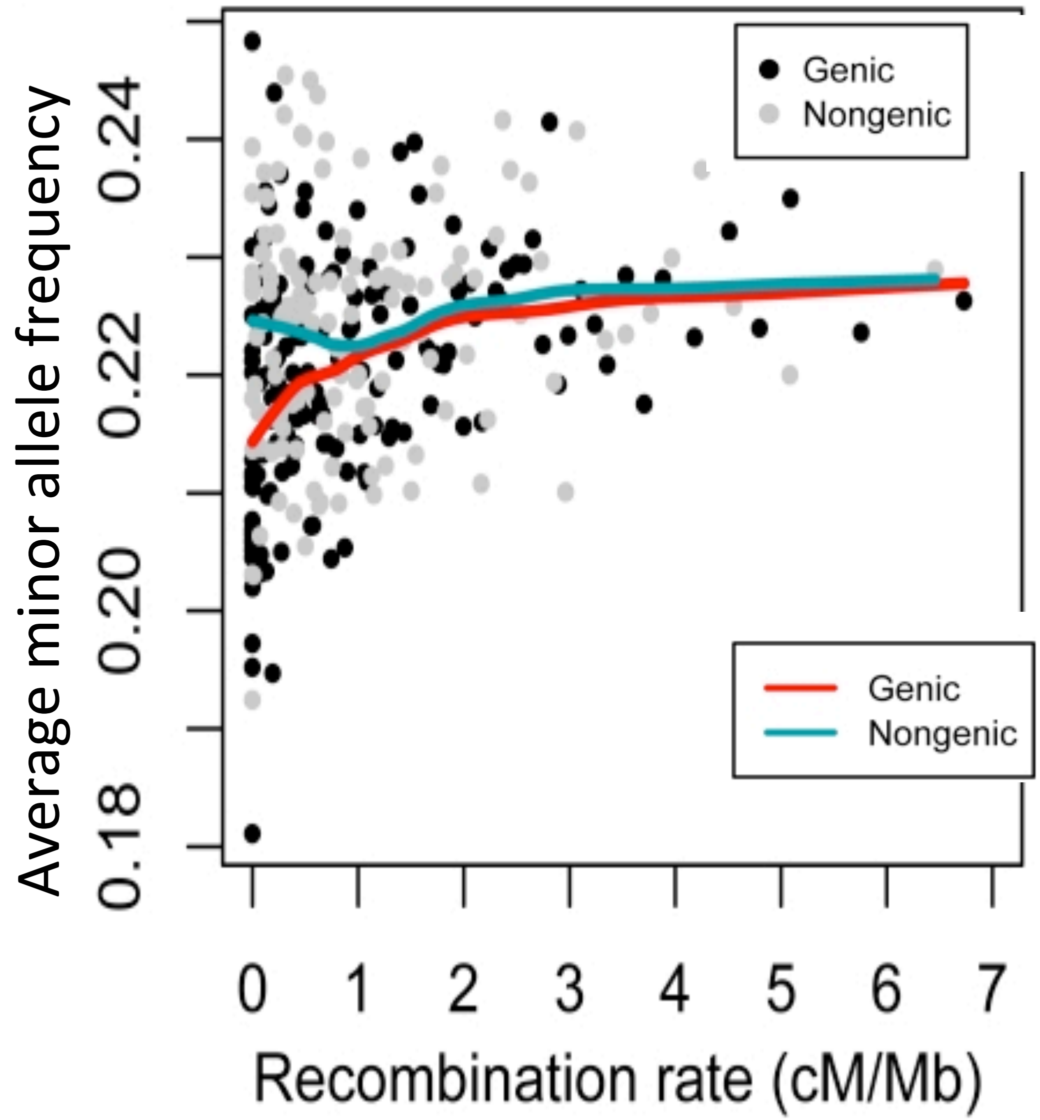
Simulate mssel with either



# Evidence for variation-reducing selection in humans
# But not clear what mode of linked selection acts.

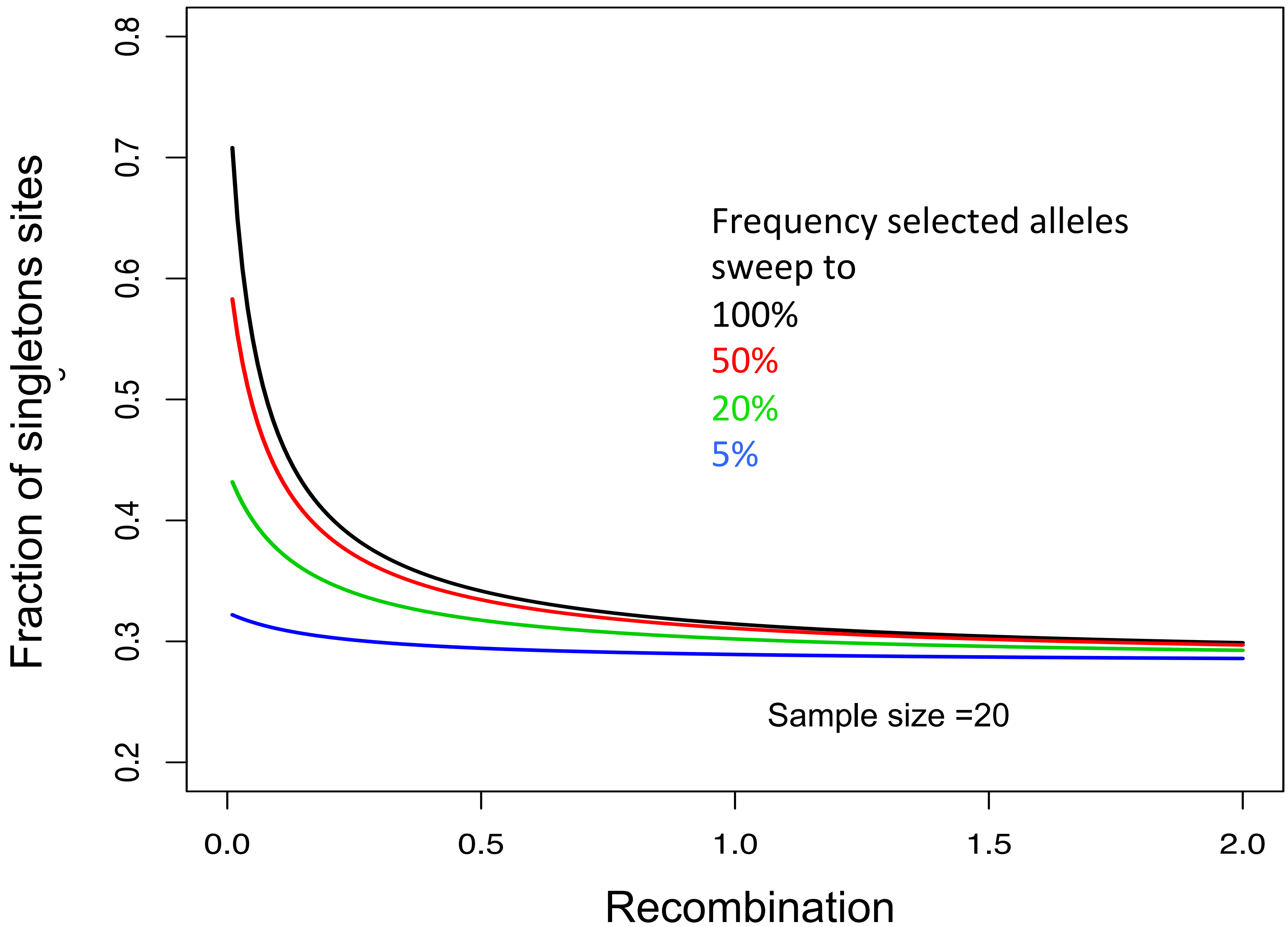


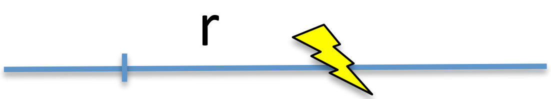

Cai et al. 2009



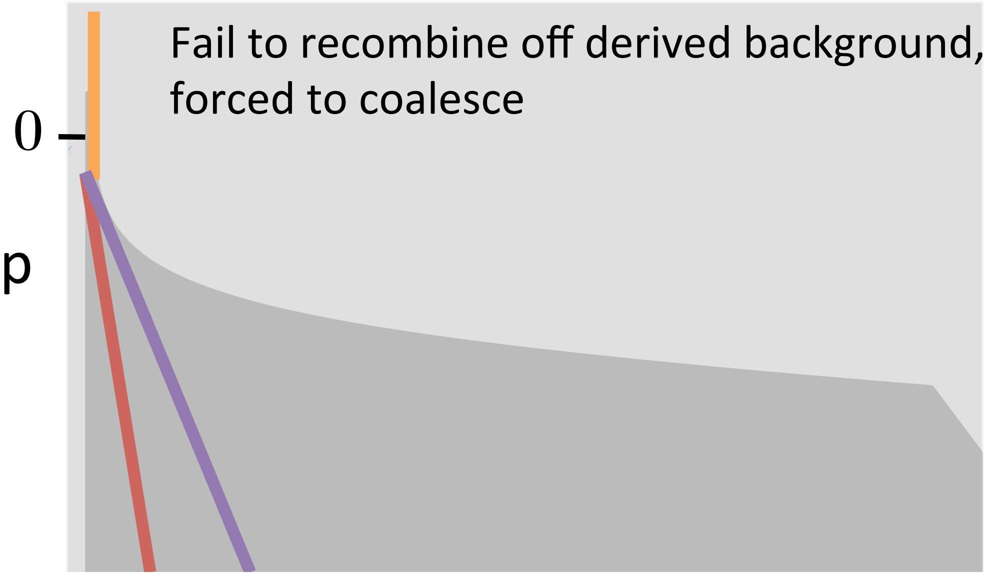

Lohmueller et al., 2011

# Matching the reduction in pi the distortion to the site frequency spectrum



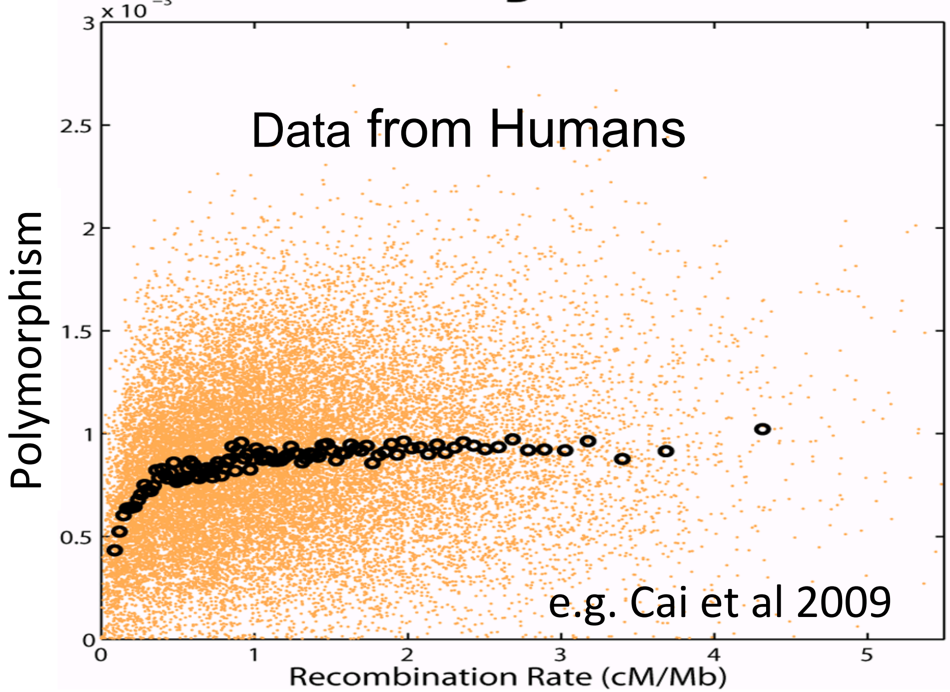



Soft sweep model due to
Parallel mutation during sweep




# Conclusions

- P




Rate of recombination (cM/Mb)

Hellmann et al using similar data

$$\pi \approx \frac{r_{BP}\pi_0}{r_{BP} + \alpha}$$

Estimated $\pi_0 = 1.6\times10^{-3}$ , $\alpha = 6 \times 10^{-11}$

Assuming none of the reduction is due to BS

$\alpha = 2Nv_{BP}\,(x^2/t_x)$

$t_x = 1000$ (s~1%)

N=10000

$v_{BP}\,x^2 = 3 \times 10^{-12}$

Note humans need a high sweep rate despite smaller effect of HH

| | | | | |
|---|---|---|---|---|
| x = | 100% | 50% | 20% | 5% |
| $v_{BP}$= | 3e-12 | 1e-11 | 8e-11 | 1e-09 !!! |

Solid coloured line recurrent loss trajectory.
Dashed coloured line recurrent fix trajectory

$t_x$ /2N = 0.0015
Pauses for 0.02 (2N generations)