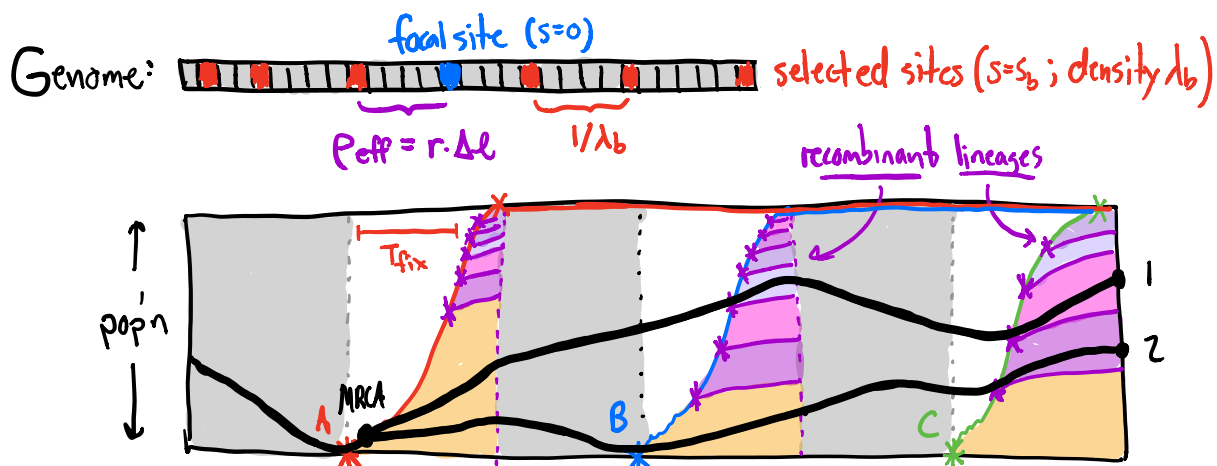


- Announcements:
- (1) No weekly feedback next week (thanks for effort!)
  - (2) No office hrs this week (always happy to chat!)

Last time: Linked selection via "classic selective sweeps"



Coalescence Prob Per Sweep:

$$P_c(n, \Delta \ell) = \exp[-n \cdot \Delta \ell \cdot \frac{r}{s_b} \cdot \log(Ns_b)]$$

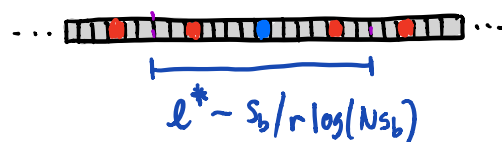
Total coalescence Rate from sweeps

$$R = \int_{-\infty}^{\infty} P_c(2, |\Delta \ell|) \cdot 2N\mu\lambda_b s_b \cdot d\Delta \ell$$

When  $N \gg \frac{1}{R} \gg T_{fix}$ :

$$\langle T_{MRCB} \rangle \approx \frac{1}{R} = \frac{r \log(Ns_b)}{2N\mu\lambda_b s_b^2}$$

$$\approx \int_{-l^*/2}^{l^*/2} \Theta(1) \cdot 2N\mu\lambda_b s_b \cdot d\Delta \ell$$



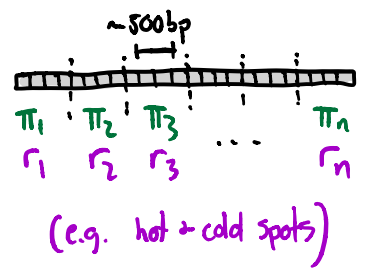
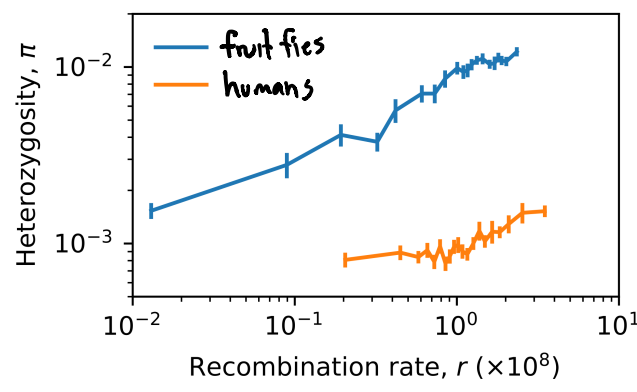
Today: (1) wrap up this case (2) clonal interference

⇒ Key prediction: genetic diversity (e.g.  $\pi$ ) @ neutral (e.g. syn) sites

depends on local recombination rate  $r$ !  
 (since controls linkage to selected sites)

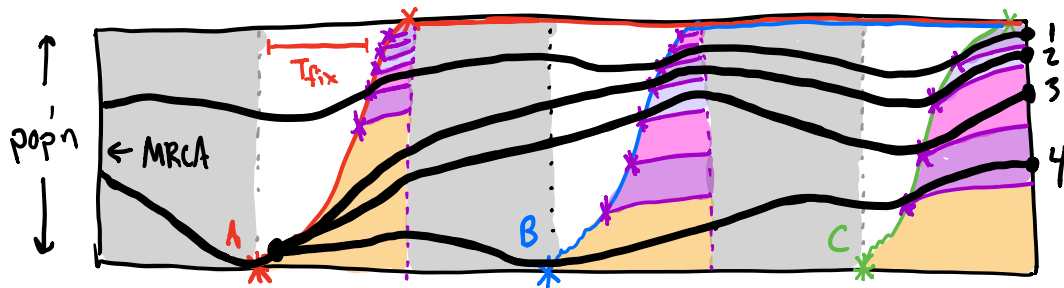
$$\langle \pi \rangle \approx \frac{r \log(Ns_b)}{S_b \cdot Ns_b \cdot \lambda_b}$$

⇒ can test using natural variation in  $r$  along genome:



⇒ sometimes spun as "~~linked selection~~ ⇒ local reductions in  $N_e$ "  
 e.g.  ~~$\pi_e \approx 2N_e(e)\mu$~~   
 ⇒ **WRONG!**

⇒ can see by examining larger sample sizes:



Recall: Coalescence Prob Per Sweep:  $p_c(n, \Delta t) = \exp\left[-n \Delta t \cdot \frac{r}{s_b} \cdot \log(Ns_b)\right]$

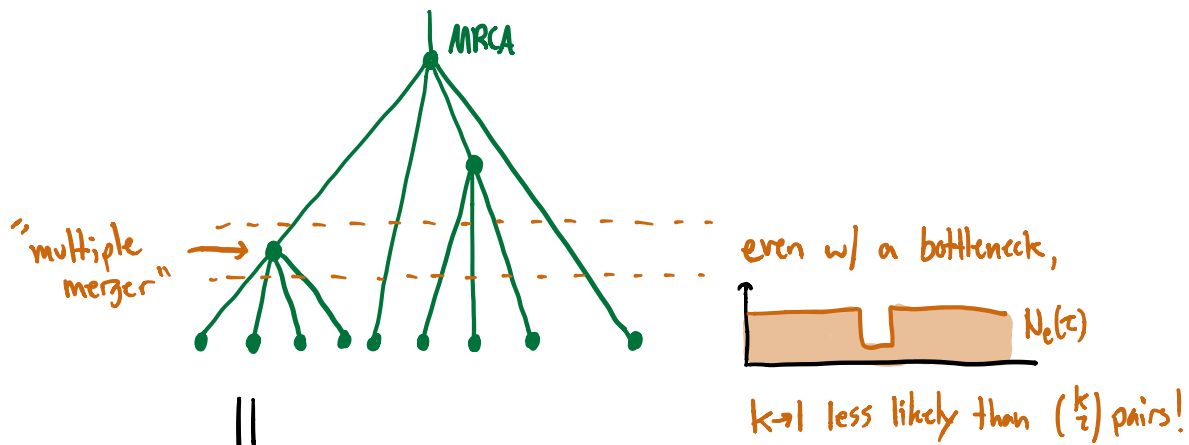
⇒ Total rate of sweeps w/  $n$  lineages coalescing:

$$R(n \rightarrow 1) = \int_0^{\infty} e^{-\frac{n \Delta t r \log(Ns_b)}{s_b}} \cdot 4N\mu s_b d\Delta t = \frac{4N\mu s_b}{\frac{nr}{s_b} \log(Ns_b)}$$

⇒  $R(n \rightarrow 1) = \frac{2}{n} R$  ⇒ Decays very slowly w/  $n$ !

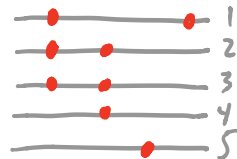
[compare to  $N \cdot \left(\frac{1}{N}\right)^n$  for neutral (Kingman) coalescent]

Upshot: if 2 lineages coalesce in a given timestep,  
 $\Rightarrow$  likely multiple lineages coalesce into same block!  
 $\Rightarrow$  can produce genealogies like:

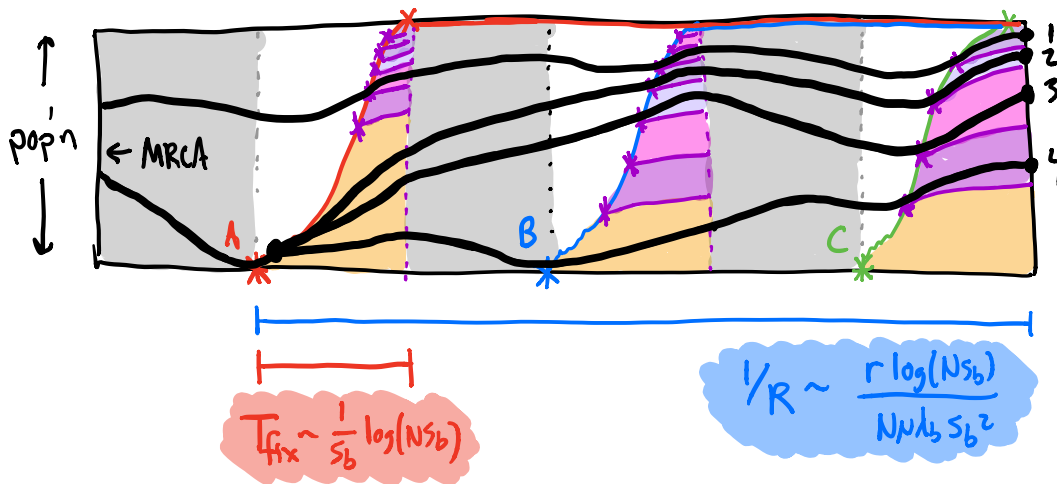


Important open question:

How can we detect these effects from mutation data?



⇒ when is this successive mutations - like picture a good approx?

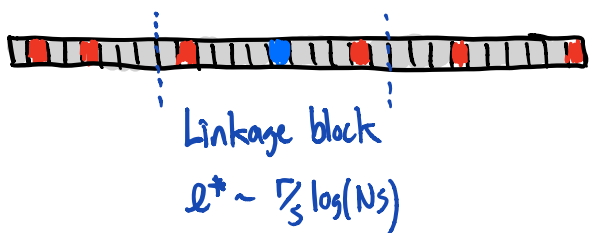


⇒ check self consistency:

Each coalescence-causing sweep should fix before next one occurs

$$\Rightarrow RT_{\text{fix}} \ll 1 \Rightarrow \frac{N \mu \lambda_b s_b^2}{r \log(Ns_b)} \cdot \frac{1}{s_b} \cdot \log(Ns_b) = \frac{N}{r} \cdot \lambda_b \cdot N s_b \ll 1$$

Alternative interpretation: multiple sweeps cannot occur w/in  $l^*$  of each other in a single fixation time:



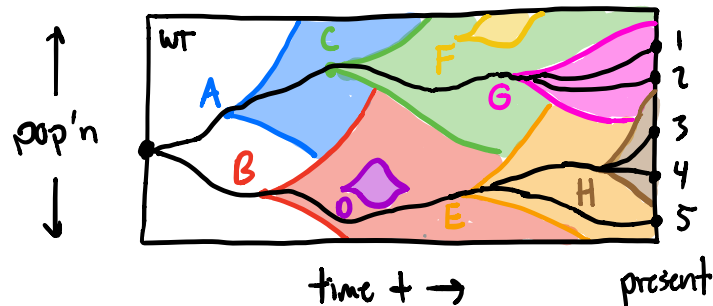
$$N \cdot \mu \lambda_b l^* \cdot s_b \cdot T_{\text{fix}} \ll 1$$

$$\Downarrow \frac{N}{r} \cdot \lambda_b \cdot N s \ll 1$$

$\Rightarrow$  if  $\frac{\mu}{r} \sim \mathcal{O}(1) \Rightarrow$  need  $\lambda_b \ll \frac{1}{N S_b} \ll 1$

$\Rightarrow$  will always break down in sufficiently large pop'ns!

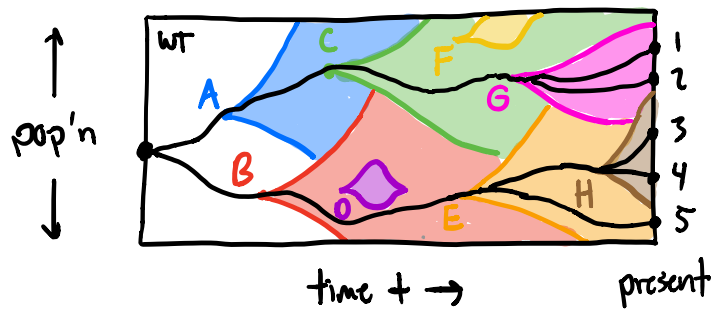
$\Rightarrow$  Back to clonal interference regime!



$\Rightarrow$  Finally time to consider in detail...

# Linked selection + clonal interference

(a.k.a. "Hill-Robertson Interference")



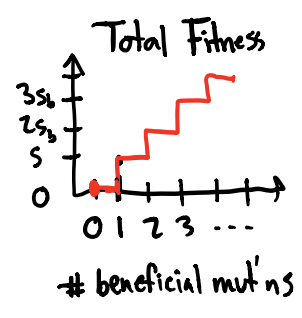
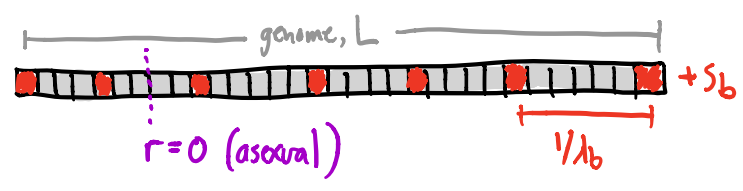
⇒ can't be reduced to  $L=1$  or  $L=2$  model (collective phase)

⇒ Most progress only recently, w/ big contribution from physicists

[e.g. Tsimring et al PRL '96, Rouzine et al '03, Desai + Fisher '07, ...]

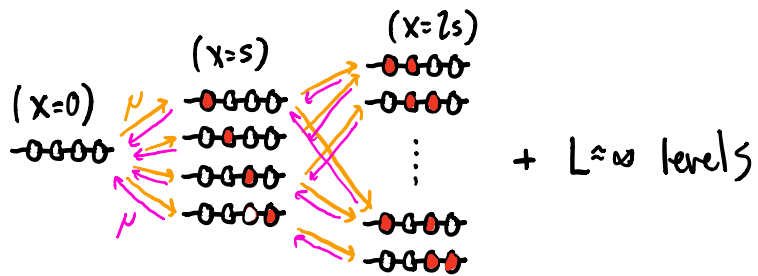
⇒ Analytical progress enabled by starting w/ very simple model:

## "Staircase" Model



- ① All mutations provide same benefit ( $S_b$ )
- ② Occur @ total rate  $U_b \equiv L\lambda_b N$
- ③ Never run out (e.g.  $L\lambda_b \rightarrow \infty, N \rightarrow 0$ )

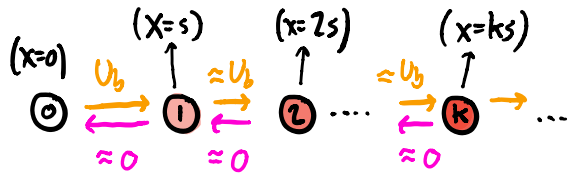
Genotype network:



Key simplification:

"fitness class"

$$f_k \equiv \sum_{|j|=k} f(j)$$



$\Rightarrow$  coarse-grained SDE (1+1 dimensional vs  $2^L+1$  dim.)

$$\frac{df(k)}{dt} = \underbrace{s_b(k - \bar{k}(t))f(k)}_{\text{selection (nonlinear)}} + \underbrace{U_b [f(k-1) - f(k)]}_{\text{mutation}}$$

$$+ \sqrt{\frac{f(k)}{N}} \eta(k) - f(k) \sum_{k'} \sqrt{\frac{f(k')}{N}} \eta(k')$$

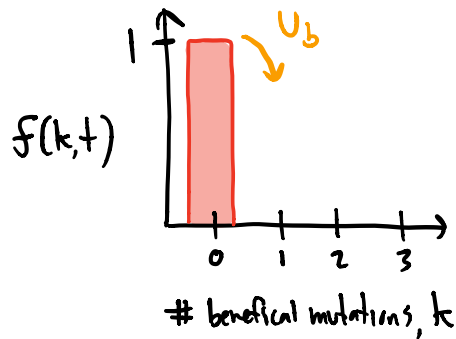
genetic drift (stochastic)

$\Rightarrow$  let's consider behavior when  $Ns_b \gg NU_b \gg 1$

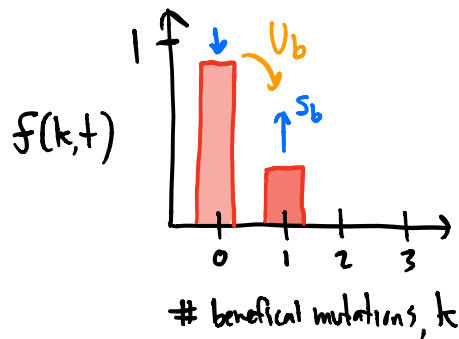
(e.g. yeast barcode experiment in HW 4 Problem #1)



① Start w/ wildtype population @  $t=0$



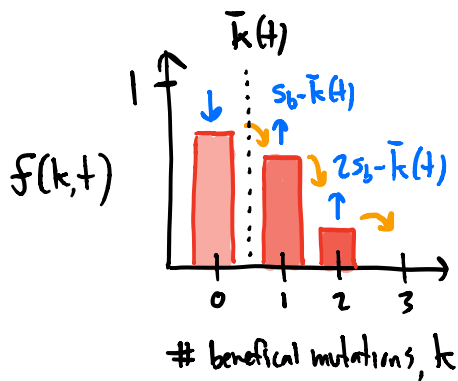
② First-step mutations ( $k=1$ ) establish & grow exponentially



$$\Rightarrow f(1,t) \approx \frac{U_b}{s_b} (e^{s_b t} - 1)$$

(deterministic approx good @ first, since  $NU_b \gg 1$ )

③ Double mutants establish before single mutants take over,



$\Rightarrow$  clonal interference!

(since  $\int_0^{z_{1/2}} NS(1,t) \cdot U_b \cdot s_b dt \sim NU_b \gg 1$ )

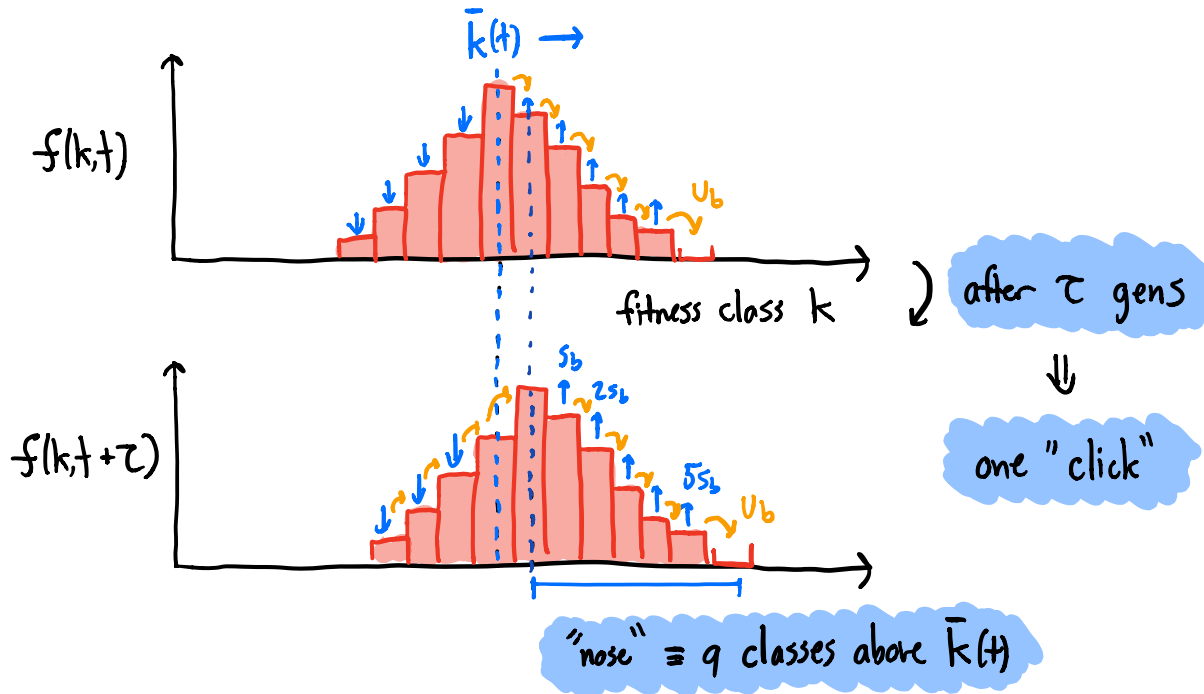
⇒ Is deterministic approx still useful?

$$\frac{df(k)}{dt} = \underbrace{s_b(k - \bar{k}(t))f(k)}_{\text{selection (nonlinear)}} + \underbrace{U_b[f(k-1) - f(k)]}_{\text{mutation}} + \underbrace{\int \frac{f(k')}{N} \eta(k') - f(k) \sum_k \frac{f(k')}{N} \eta(k')}_{\text{genetic drift}} \rightarrow 0$$

⇒ can show:  $f_{\text{det}}(k,t) = \frac{1}{k!} \cdot \left[ \frac{U_b}{s_b} (e^{s_b t} - 1) \right]^k \cdot e^{-\frac{U_b}{s_b} (e^{s_b t} - 1)}$  X

⇒ **Not self-consistent!** ⇒ Predicts  $s_b \bar{k}(t) \approx U_b e^{s_b t}$   
 (eventually all  $f(k,t) \ll 1/N!$ )

⇒ Instead, if we simulate model, observe "travelling wave":



⇒ What determines  $\tau(N, s_b, U_b) + q(N, U_b, s_b)$ ?

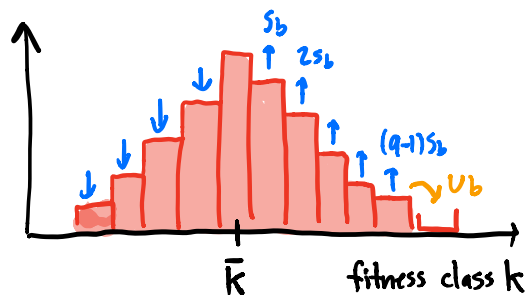
⇒ Today: heuristic analysis [ $\sim$  Desai + Fisher 2007]

applies when:  $Ns_b \gg NU_b \gg 1$  +  $s_b \tau \gg 1$  &  $q \gg 1$

Leads to simplifications:

① mutations only important for establishing new "nose"

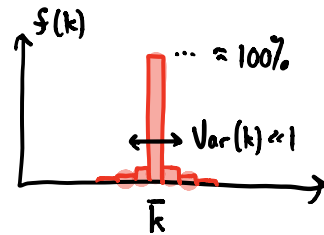
(since  $s_b \gg U_b$ )



② Genetic drift only important for establishing new nose

(since  $\tau \gg 1/s_b$ , individual mutations establish before next click.)

③ most of pop'n is near  $k \approx \bar{k}(t)$

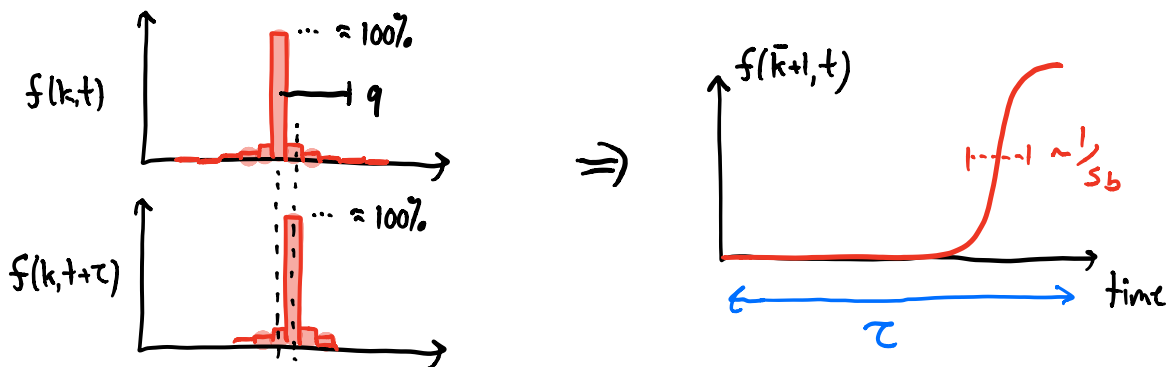


Problem 3 of HW 4:

$$\underbrace{\frac{d\langle \bar{k} \rangle}{dt}}_{\equiv 1/\tau} = \left\langle \sum_k k \frac{df(k)}{dt} \right\rangle = \left\langle \underbrace{\sum_k s_b (k - \bar{k})^2 f(k, t)}_{s_b \text{Var}(k)} \right\rangle$$

$$\Rightarrow \text{Var}(k) = \frac{1}{s_b \tau} \ll 1 \quad (\text{by assumption})$$

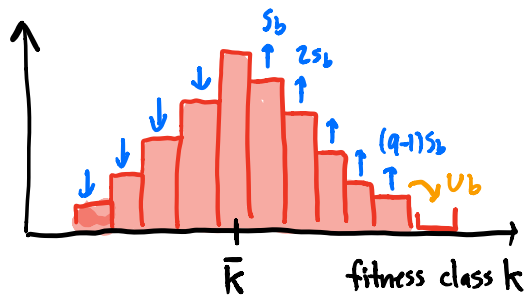
④ Also implies that  $\bar{k}(t)$  clicks suddenly:



$\Rightarrow$  i.e. for most  $t \in [0, \tau] \Rightarrow \bar{k}(t) = \bar{k}(0)$

$\Rightarrow$  everyone grows as  $f(k, t) \sim f(k, 0) e^{(k - \bar{k}(0))st}$

$\Rightarrow$  Now we have all ingredients to understand wave:

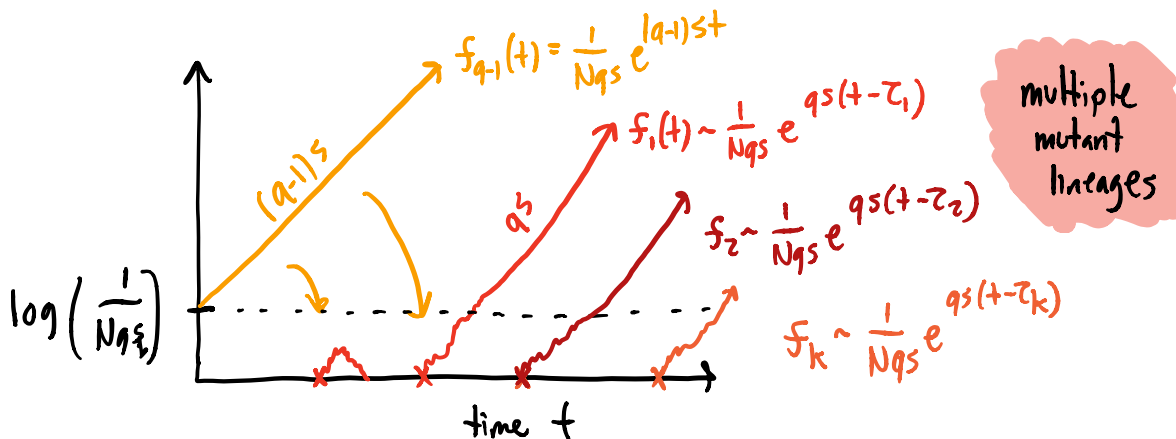


$\Rightarrow$  in one click ( $\tau$ ), must establish new nose

$$\Downarrow$$

$$f_q(t) = \frac{1}{Nq_s} e^{qs(t-\tau)}$$

$\Rightarrow \tau \approx$  establishment time of nose class!



$$\Rightarrow f_q(t) = \sum_{k=0}^{K_{\max}} f_k(t) \equiv \frac{1}{Nq_s} e^{qs(t-\tau)}$$

$\downarrow$   
 establishment time  
 for whole class.

$\Rightarrow k^{\text{th}}$  successful mutant establishes when:

$$\int_0^{\tau_k} NU_b \cdot f_{q-1}(t) \cdot qs_b dt \sim \mathcal{O}(k)$$

$$\int_0^{\tau_k} NU_b \cdot \frac{1}{Nq_s} e^{(q-1)st} \cdot qs_b dt = \frac{U_b}{q_s} e^{(q-1)s_b \tau_k} \sim \mathcal{O}(k)$$

Note: extra  
little bit  
will be  
important  
below!

$$\Rightarrow \tau_k = \frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b} \cdot q \cdot k\right)$$

$$\Rightarrow \text{Note: } \tau_k = \underbrace{\frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b} \cdot q\right)}_{\tau_1} + \underbrace{\frac{1}{(q-1)s_b} \cdot \log(k)}_{\tau_k - \tau_1}$$

$\gg$

(most time spent waiting for first mut'n)

⇒ many mutations establish in quick succession ( $\delta t \sim \frac{1}{q s_b} \ll \tau$ )

⇒ Typical size of  $k^{\text{th}}$  lineage:

$$\Rightarrow f_k(t) = \frac{1}{N q s} e^{q s (t - \tau_k)} = \frac{e^{q s t}}{N q s} \left( \frac{s_b q k}{U_b} \right)^{-1 - \frac{1}{q}}$$

important!

⇒ Size of entire nose class:

$$\Rightarrow f_q(t) = \sum_{k=1}^{K_{\text{nose}}} f_k(t) = \frac{1}{N q s} e^{q s t} \left( \frac{s_b}{U_b} \right)^{-1 - \frac{1}{q}} \sum_{k=1}^{K_{\text{nose}}} \frac{1}{q} \frac{1}{k^{1 + \frac{1}{q}}}$$

$$\equiv \frac{1}{N q s} e^{q s (t - \tau)}$$

↙ set equal! ↘

$$\Rightarrow \tau = \frac{(1 + \frac{1}{q})}{q s_b} \log \left( \frac{s_b}{U_b} \right) \approx \frac{1}{q s_b} \log \left( \frac{s_b}{U_b} \right)$$

$$\text{vs } \tau_k = \frac{1}{q s_b} \log \left( \frac{s_b}{U_b} \cdot q \cdot k \right) \quad \left[ \tau < \tau_k \text{ because multiple mutations contribute } \rho \text{ once} \right]$$

⇒ how to determine  $q$ ?

⇒ follow new nose over time:

$$f(\bar{k}+q, \tau) \approx \frac{1}{Nq_s} \xrightarrow{\tau} \frac{1}{Nq_s} e^{(q-1)s\tau} \xrightarrow{\tau} \frac{1}{Nq_s} e^{(q-1)s\tau + (q-2)s\tau} \rightarrow \dots$$

(right after est.)

⇒ After  $q$  clicks, old nose is new mean! (majority of pop'n)

$$f(q\tau) \sim \frac{1}{Nq_s b} e^{(q-1)s\tau + (q-2)s\tau + \dots + s\tau} \sim \frac{1}{Nq_s b} e^{\frac{q^2 s \tau}{2}} \sim \mathcal{O}(1)$$

⇒ system of 2 eqs for  $\tau$  &  $q$ :

$$\frac{q^2 s_b \tau}{2} \approx \log(N s_b) \quad + \quad \tau = \frac{1}{q s_b} \log\left(\frac{s_b}{v_b}\right)$$



$$\Rightarrow \text{solution: } q = \frac{2 \log(Ns_b)}{\log\left(\frac{s_b}{u_b}\right)} ; \tau = \frac{1}{2s_b} \frac{\log^2\left(\frac{s_b}{u_b}\right)}{\log(Ns_b)}$$

$$\Rightarrow \left\langle \frac{d\bar{x}}{dt} \right\rangle = \frac{s_b}{\tau} = \frac{2s_b^2 \log(Ns_b)}{\log^2(s_b/u_b)}$$

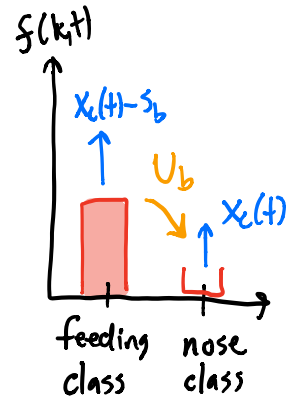
(compare to  $\sim N u_b s_b^2$  in *successive mutations* regime)

$\Rightarrow$  Self consistency:  $s\tau \gg 1$  +  $q \gg 1$

$$\Rightarrow \log\left(\frac{s_b}{u_b}\right) \ll \log(Ns_b) \ll \log^2\left(\frac{s_b}{u_b}\right)$$

## Supplement: Formal analysis of the nose class

⇒ we can understand the establishment of the nose class more formally using the branching process framework that we studied in the 1<sup>st</sup> half of the course



⇒ Under our assumptions, nose can be described by LBP model:

$$\frac{df_q}{dt} = X_c(t) f_q + U_b f_{q-1}(t) + \sqrt{\frac{f_q}{N}} \eta(t)$$

w/  $f_q(0) = 0$  & time-varying:

$$\text{selection: } X_c(t) = [q - \bar{k}(t)] s_b$$

$$\text{mutation: } U_b f_{q-1}(t) = \frac{U_b}{2Nq s_b} e^{\int_0^t (X_c(t) - s_b) dt}$$

⇒ In their analysis, Desai & Fisher (2007) assumed that  $\bar{k}(t) \approx 0$  throughout the establishment period, so that  $X_c(t) \equiv qs_b$  &  $f_{q-1}(t) = \frac{1}{Nqs} e^{(q-1)s_b t}$

⇒ Let's see how far we can get by relaxing this approx & explicitly modeling the "click" of  $\bar{k}(t)$ ...

⇒ will be harder because time-varying fitness

$$X_c(t) = qs_b - s_b \bar{k}(t)$$

⇒ From our discussion in class, can take

$$\bar{k}(t) \equiv \frac{e^{s_b(t-t_c)}}{1 + e^{s_b(t-t_c)}}$$

where  $t_c$  is the time that  $\bar{k}(t)$  clicks.

(later we will imagine that  $t_c$  is close to  $\tau$ ...)

$\Rightarrow$  From SDE, the generating function  $H_f(z, t) = \langle e^{-z \cdot f_1(t)} \rangle$

satisfies the PDE:

$$\frac{\partial H_f}{\partial t} = \left[ \chi_L(t) z - \frac{z^2}{2N} \right] \frac{\partial H_f}{\partial z} - z U_{bf_1}(t) H_f$$

w/ initial condition  $H_f(z, 0) = 1$

$\Rightarrow$  can solve w/ method of characteristics:

$$\text{define: } \psi(t_R) = \log \left[ H_f(z(t_R), t - t_R) \right]$$

$$\text{w/ } \psi(t) = 0, \quad z(0) = z,$$

$$\psi(0) = \log H_f(z, t)$$

$$\Rightarrow \psi \text{ satisfies: } \frac{d\psi}{dt_R} = - \frac{\partial H_f}{\partial t} + \frac{\partial H_f}{\partial z} \left( \frac{dz}{dt_R} \right)$$

$$\Rightarrow \frac{\partial \psi}{\partial t_R} = \left\{ \frac{\partial z}{\partial t_R} - \left[ \chi_2(t-t_R)z - \frac{z^2}{2N} \right] \right\} \frac{\partial \log H_f}{\partial z} + z(t_R) U_b f_{q-1}(t-t_R)$$

$$\Rightarrow \text{if } \frac{\partial z}{\partial t_R} = \chi_2(t-t_R)z - \frac{z^2}{2N} \quad \& \quad z(0) = z$$

$$\Rightarrow \psi(t_R) = \psi(0) + \int_0^{t_R} z(t'_R) U_b f_{q-1}(t-t'_R) dt'_R$$

$$\Rightarrow \log H_f(z, t) = - \int_0^t z(\tau) U_b f_{q-1}(t-\tau) d\tau$$

where  $\frac{\partial z}{\partial \tau} = \chi_2(t-\tau)z - \frac{z^2}{2N}$ ,  $z(0) = z$

$\Rightarrow$  solution for characteristic curve is given by:

$$z(\tau) = \frac{z e^{\int_0^\tau \chi_2(t-\tau') d\tau'}}{1 + \frac{z}{2N} \int_0^\tau e^{\int_0^{\tau'} \chi_2(t-\tau'') d\tau''} d\tau'}$$

(can plug in & check...)

$$\text{so } H_f(z, t) = \exp \left[ - \int_0^t \frac{z U_b f_{q-1}(t-\tau) e^{\int_0^\tau x_c(t-\tau') d\tau'}}{1 + \frac{z}{zN} \int_0^\tau e^{\int_0^{\tau'} x_c(t-\tau'') d\tau''} d\tau'} d\tau \right]$$

$$= \exp \left[ - \int_0^t \frac{z \cdot U_b f_{q-1}(u) e^{\int_u^t x_c(u') du'}}{1 + \frac{z}{zN} \int_u^t e^{\int_u^{\tau'} x_c(u'') du''} du'} du \right]$$

$\Rightarrow$  again, helpful to define  $v(t)$  s.t.  $f_q(t) \equiv \frac{v(t)}{zNq_s b} e^{\int_0^t x_c(t') dt'}$

$$\Rightarrow H_v(z, t) \equiv \langle e^{-z \cdot v(t)} \rangle \equiv H_f \left( zNq_s b e^{-\int_0^t x_c(t') dt'}, z, t \right)$$

$$\Rightarrow H_v(z, t) = \exp \left[ - \int_0^t \frac{z U_b f_{q-1}(u) zNq_s b e^{-\int_0^u x_c(u') du'}}{1 + q_s z \cdot \int_u^t du' e^{-\int_0^{u'} x_c(u'') du''}} du \right]$$

$\Rightarrow$  similar to single-locus case, we expect  $v(t)$  to approach constant value  $v$  @ long times

$$\Rightarrow H_v(z) \equiv \lim_{t \rightarrow \infty} H_v(z, t)$$

$$\Rightarrow \log H_v(z) = - \int_0^{\infty} \frac{z \cdot U_b e^{-s_b t} dt}{1 + z \cdot q s_b \int_t^{\infty} dt' e^{-\int_0^{t'} x_c(t'') dt''}}$$

$\Rightarrow$  Now we have to plug in our expression for  $X_c(t)$ :

$$X_c(t) = q s_b - \frac{s_b e^{s_b(t-t_c)}}{1 + e^{s_b(t-t_c)}}$$

$$\Rightarrow \int_0^t x_c(t') dt' = q s_b t - \log \left[ \frac{1 + e^{s_b(t-t_c)}}{1 + e^{-s_b t_c}} \right]$$

$$\Rightarrow e^{-\int_0^t x_c(t') dt'} = e^{-q s_b t} \left[ \frac{1 + e^{s_b(t-t_c)}}{1 + e^{-s_b t_c}} \right]$$

$$\Rightarrow q s_b \int_t^{\infty} dt' e^{-\int_0^{t'} x_c(t'') dt''} = \frac{e^{-q s_b t}}{1 + e^{-s_b t_c}} + \left( \frac{q}{q-1} \right) \frac{e^{-q s_b t} e^{s_b(t-t_c)}}{1 + e^{-s_b t_c}}$$

and hence:

$$\log H_V(z) = - \int_0^{\infty} \frac{z \cdot U_b e^{-s_b t} dt}{1 + z \cdot e^{-q s_b t} \left[ 1 + e^{s_b(t-t_c)} \left( \frac{q}{q-1} \right) \right]}$$

(where we have assumed that the click time  $t_c$  is  $\gg 1/s_b$ )

$\Rightarrow$  for large  $q$  & relevant values of  $z$ , this integral will be dominated by times w/in  $o(1/s_b)$  of  $\tau$ .

$\Rightarrow$  can extend lower limit of integral to  $t = -\infty$   
w/o much error...

$\Rightarrow$  if  $t_c$  is also w/in  $o(1/s_b)$  of  $\tau$ , we can expand  $e^{s_b(t-t_c)}$  term in denominator, so that

$$\log H_V(z) \approx - \int_{-\infty}^{\infty} \frac{z \cdot U_b e^{-s_b t} dt}{1 + z \cdot z \cdot e^{-q s_b t}}$$



changing variables to  $\xi = (2z)^{1/q} e^{-s_b t}$ ,

$$\log H_v(z) = \exp \left[ -\frac{U_b}{S_b} \cdot z^{1-1/q} \cdot \left( 2^{-1/q} \int_0^\infty \frac{\xi d\xi}{1+\xi^q} \right) \right]$$

↑  $1 + \mathcal{O}(1/q)$

$$\Rightarrow H_v(z) = e^{-\frac{U_b}{S_b} z^{1-1/q}}$$

$\Rightarrow$  typical value of  $v$  occurs when  $H_v(z = \frac{1}{v^*}) = e^{-1}$

$$\Rightarrow v^* = \left( \frac{S_b}{U_b} \right)^{\frac{q}{q-1}}$$

$$\Rightarrow \text{substituting into } f_q(t) = \frac{v}{2NqS} e^{qst} = \frac{e^{qs(t-\tau)}}{2NqS}$$

$$\Rightarrow \text{typical value of } f_q^*(t) = \frac{e^{S_b t}}{NqS_b} \left( \frac{S_b}{U_b} \right)^{\frac{q}{q-1}}$$

⇒ typical value of establishment time:

$$\tau^* = \frac{1}{(q-1)s_b} \log\left(\frac{s_b}{u_b}\right)$$

⇒ consistent w/ results from  
simpler heuristic argument!

Supplement: How many lineages contribute to new nose?

Recall in heuristic argument, we had:

$$f_{\text{nose}}(t) \equiv \sum_{k=1}^{K_{\text{max}}} f_k(t) = \frac{1}{Nq_s} e^{qs_s(t-\tau)} \cdot \sum_{k=1}^{K_{\text{max}}} \frac{1}{q k^{1+1/q}}$$

+ argued that sum over  $k$  converged to  $\approx 1$ .

$\Rightarrow$  Let's look @ this more carefully...

$\Rightarrow$  if  $K_{\text{max}} \gg 1$  (will revisit below)

$$\Rightarrow \sum_{k=1}^{K_{\text{max}}} \frac{1}{q k^{1+1/q}} \approx \int_1^{K_{\text{max}}} \frac{dk}{q k^{1+1/q}} = 1 - e^{-\frac{1}{q} \log K_{\text{max}}}$$

Thus, sum converges to 1 provided that  
 $\log K_{\text{max}}$  is large compared to  $q$

$\Rightarrow$  how does this translate to establishment times  $\tau_k$ ?

recall that  $\tau_k - \tau_1 \sim \frac{1}{q s_b} \log(k)$ , so condition becomes:

$$\Rightarrow \tau_{k_{\max}} - \tau_1 \sim \frac{1}{q s_b} \log(K_{\max}) \gg \frac{1}{s_b}$$

Thus, mutations that establish  $\gg \frac{1}{s_b}$  after  $\tau_1$   
have negligible contribution to  $S_{\text{nose}}(t)$ ,  $\tau$ , etc

$\Rightarrow$  since  $\frac{1}{s_b} \ll \tau$ , this happens long before next click.

$\Rightarrow$  can take  $K_{\max} \approx \infty$  w/o losing any accuracy

i.e., can pretend that infinite # of muts  
contribute to establishment of new nose.