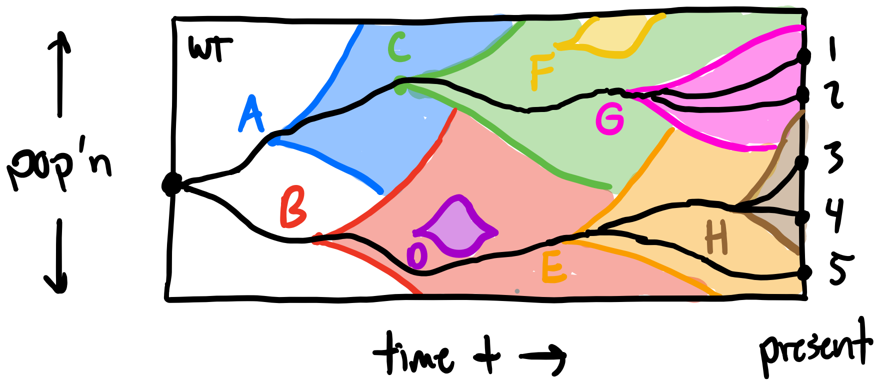
Chapter 15

Linked selection and clonal interference

Linked selection & clonal interference

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



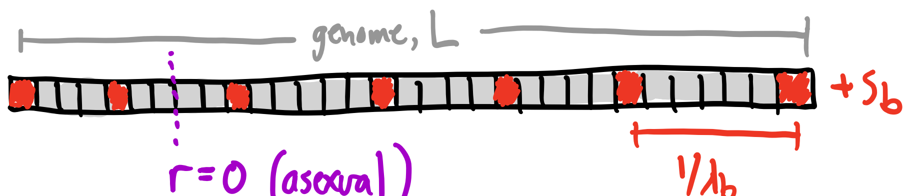
⇒ can't be reduced to L=1 or L=2 model (collective phase)

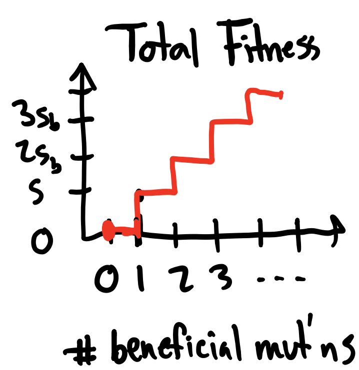
⇒ Most progress only recently, w/ big contribution from physicists

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

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

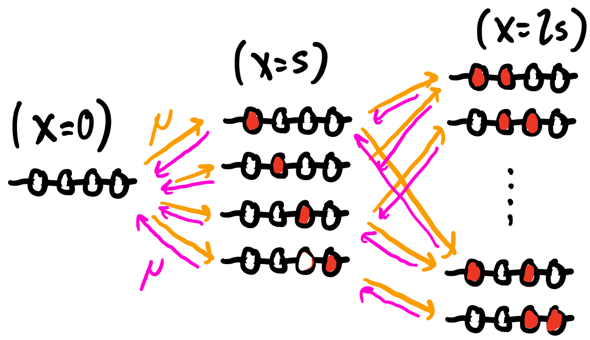
"Staircase" Model





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. $L\lambda_b \to \infty$, $N \to 0$)

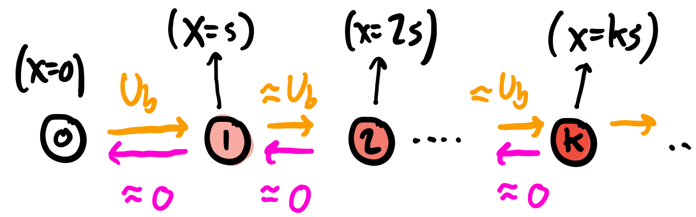
Genotype network:



+ $L \approx \infty$ levels

Key simplification:

"fitness class" $\quad f(k,t) \equiv \sum_{|\vec{g}|=k} f(\vec{g},t)$



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

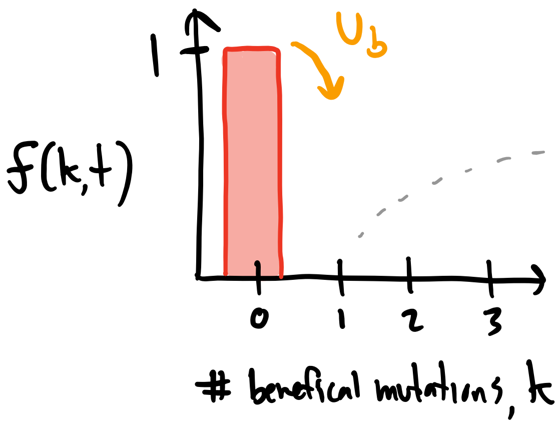
$$\frac{\partial f(k)}{\partial t} = \underbrace{s_b(k-\bar{k}(t))f(k)}_{\text{selection (nonlinear)}} + \underbrace{U_b\left[f(k-1)-f(k)\right]}_{\text{mutation}}$$

$$+ \underbrace{\sqrt{\frac{f(k)}{N}}\,\eta(k) - f(k)\sum_{k'}\sqrt{\frac{f(k')}{N}}\,\eta(k')}_{\text{genetic drift (stochastic)}}$$

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

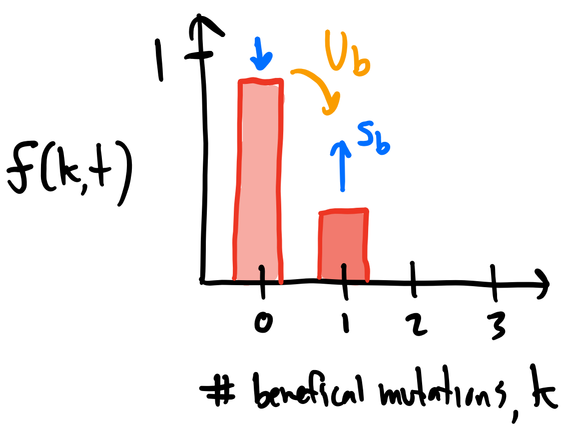
(e.g. yeast barcode experiment in HW 4 Problem #1)

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



$$\frac{df(1)}{dt} \approx s f(1) + U_b + \sqrt{\frac{f(1)}{N}}\,\eta_1(t)$$

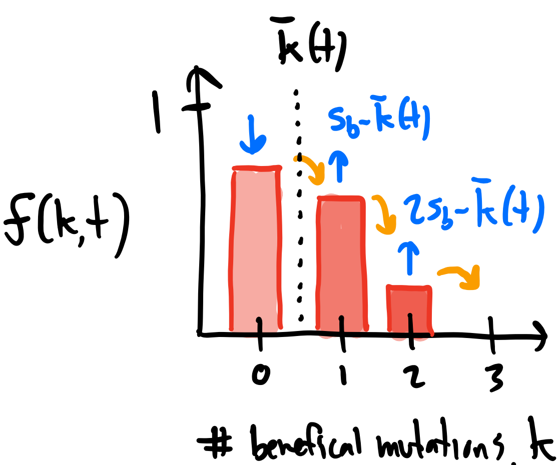
② First-step mutations ($k=1$) establish & grow exponentially



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

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

③ Double mutants establish before single mutants take over,



$\Rightarrow$ clonal interference!

$$\left(\text{since } \int_0^{\tau_{1/2}} N f(1,t)\cdot U_b \cdot s_b \, dt \sim NU_b \gg 1\right)$$

$\Rightarrow$ Is deterministic approx still useful?

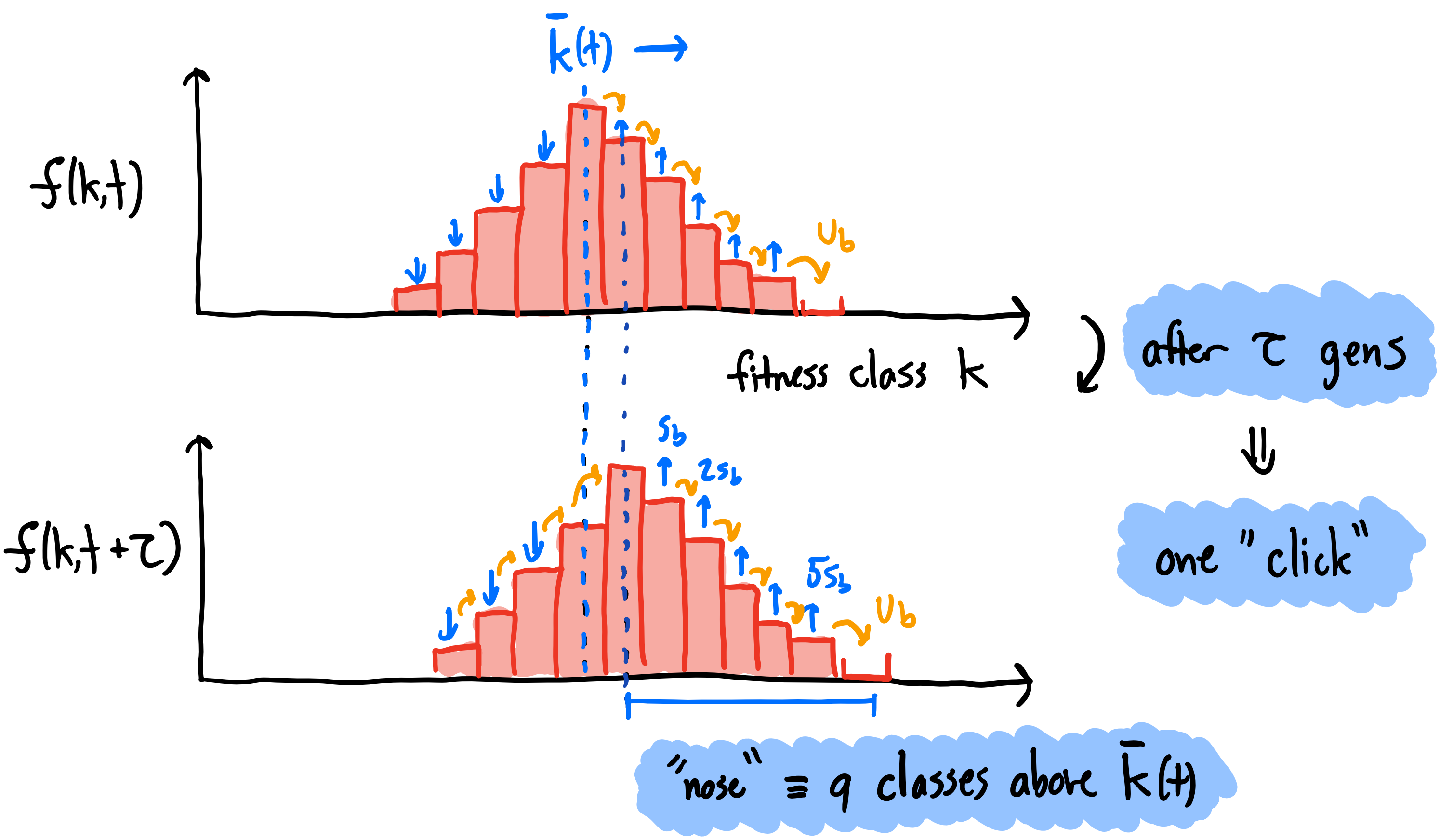
$$\boxed{\frac{\partial f(k)}{\partial t} = \underbrace{s_b(k-\bar{k}(t))f(k)}_{\text{selection (nonlinear)}} + \underbrace{U_b[f(k-1)-f(k)]}_{\text{mutation}} \underbrace{\cancel{+ \sqrt{\frac{f(k)}{N}}\eta(k) - f(k)\sum_{k'}\sqrt{\frac{f(k')}{N}}\eta(k')}}_{\text{genetic drift}} \rightarrow 0}$$

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

$\Rightarrow$ Not self-consistent! $\Rightarrow$ 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 $\tau(N, s_b, U_b)$ & $q(N, U_b, s_b)$?

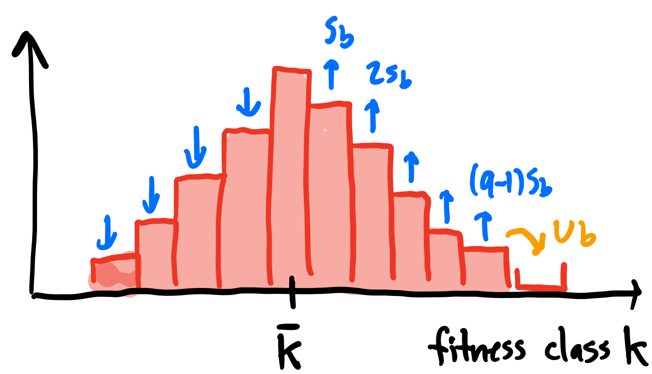
⇒ Today: heuristic analysis [~ Desai & Fisher 2007]

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"

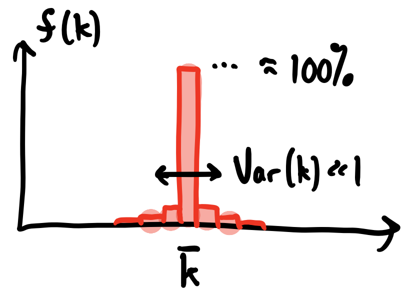
(since $s_b \gg U_b$)



② Genetic drift only important for establishing new nose

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

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

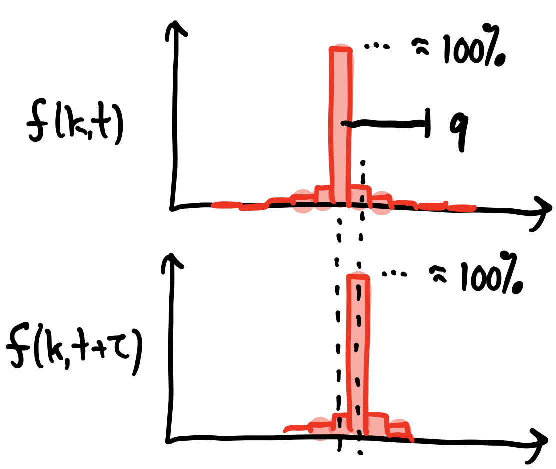


Problem 3 of HW 4:

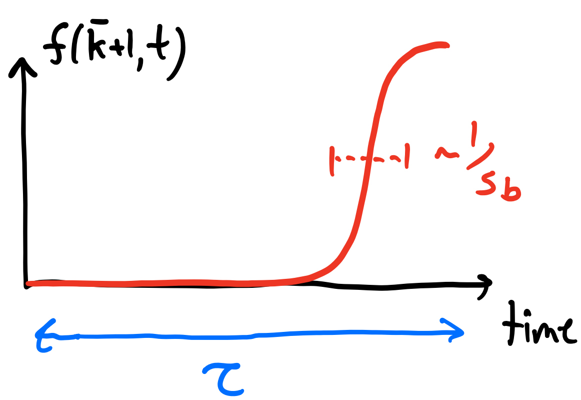
$$\underbrace{\frac{\partial \langle \bar{k} \rangle}{\partial t}}_{\equiv 1/\tau} = \left\langle \sum_k k \frac{\partial f(k)}{\partial t} \right\rangle = \left\langle \underbrace{\sum_k s_b (k-\bar{k})^2 f(k,t)}_{s_b \mathrm{Var}(k)} \right\rangle$$

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

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



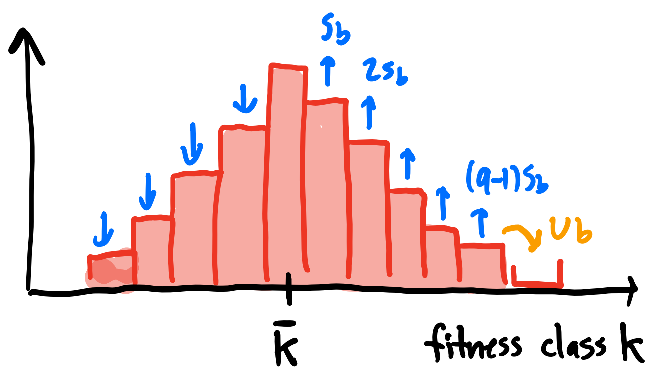
$\Rightarrow$



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

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

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

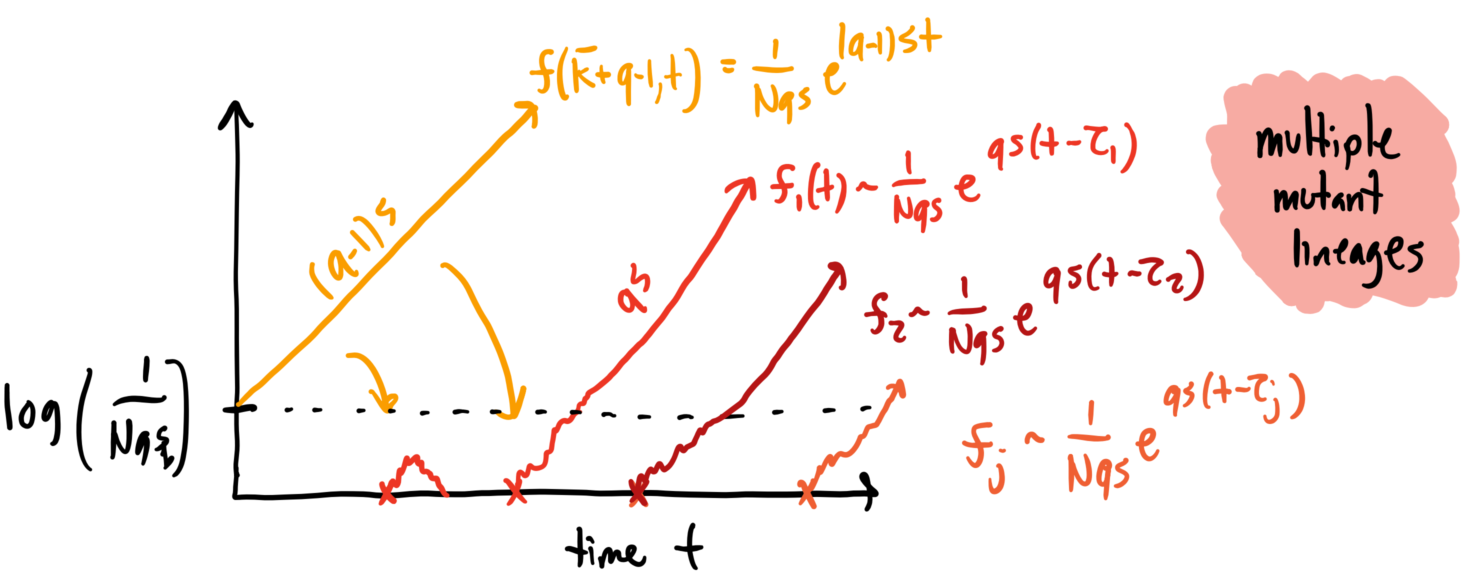


$\Rightarrow$ in one click (τ), must establish new nose

$\Downarrow$

$$f(\bar{k}+q, t) = \frac{1}{Nqs}\, e^{qs(t-\tau)}$$

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



$$\Rightarrow f(\bar{k}+q,t) = \sum_{j=0}^{J_{max}} f_j(t) \equiv \frac{1}{Nqs} e^{qs(t-\tau)}$$

τ ↓ establishment time for whole class.

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

$$\underbrace{\int_0^{\tau_j} NU_b \cdot f_{q-1}(t) \cdot qs_b \, dt}_{\int_0^{\tau_j} \cancel{N}U_b \cdot \frac{1}{\cancel{Nqs}} e^{(q-1)st} \cdot \cancel{qs_b} \, dt \,=\, \frac{U_b}{qs_b} e^{(q-1)s_b\tau_j} \,\sim\, \mathcal{O}(k)} \sim \mathcal{O}(j)$$

Note: extra little bit will be important below!

$$\Rightarrow \boxed{\tau_j = \frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b} \cdot q \cdot j\right)}$$

$\Rightarrow$ Note:

$$\tau_j = \underbrace{\frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b} \cdot q\right)}_{\tau_1} + \underbrace{\frac{1}{(q-1)s_b} \cdot \log(j)}_{\tau_j - \tau_1}$$

$\tau_1 \gg \tau_j - \tau_1$

(most time spent waiting for <u>first</u> mut'n)

$\Rightarrow$ <u>many</u> mutations establish in quick succession $\left(\delta t \sim \frac{1}{q s_b} \ll \tau \right)$

$\Rightarrow$ Typical size of j^{th} lineage:

$$\Rightarrow \quad f_j(t) \sim \frac{1}{Nqs} e^{qs(t-\tau_j)} = \frac{e^{qst}}{Nqs} \left(\frac{s_b q j}{U_b} \right)^{-1-\frac{1}{q-1}}$$

extra bit will be important!

$\Rightarrow$ Size of entire nose class:

$$f(\bar{k}+q, t) = \sum_{j=1}^{J_{max}} f_j(t) = \frac{1}{Nqs} e^{qst} \left(\frac{s_b}{U_b} \right)^{-\frac{q}{q-1}} \xrightarrow{\quad} 1 \; \cancel{\sum_{j=1}^{J_{max}} \frac{1}{(q \cdot j)^{1+\frac{1}{q-1}}}}$$

$$\equiv \frac{1}{Nqs} e^{qs(t-\tau)}$$

set equal!

$\Rightarrow$ Time to establish <u>new</u> nose: $\tau = \frac{1}{(q-1)s} \log\left(\frac{s_b}{U_b} \right)$

vs $\tau_j \equiv \frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b} \cdot q \cdot j \right)$

[Note: $\tau < \tau_j$ b.c. <u>multiple</u> mutations contribute @ once]

One task remaining... how to determine $q(N, s_b, U_b)$?

$\Rightarrow$ follow new nose over time:

$$f(\bar{k}+q, \tau) \approx \frac{1}{Nqs} \xrightarrow{\tau} \frac{1}{Nqs} e^{(q-1)s\tau} \xrightarrow{\tau} \frac{1}{Nqs} e^{(q-1)s\tau + (q-2)s\tau} \longrightarrow \cdots$$

(right after est.)

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

$$f(q\tau) \sim \frac{1}{Nqs_b} e^{(q-1)s_b\tau + (q-2)s_b\tau + \ldots + s_b\tau} \sim \frac{1}{Nqs_b} e^{\frac{q^2 s_b\tau}{2}} \sim \mathcal{O}(1)$$

$\Rightarrow$ system of 2 eqs for τ & q!

$$\boxed{\frac{q^2 s_b \tau}{2} \approx \log(Ns_b)} \quad + \quad \boxed{\tau = \frac{1}{qs_b} \log\left(\frac{s_b}{U_b}\right)}$$

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

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

(compare to $\sim NU_b s_b^2$ in successive mutations regime)

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

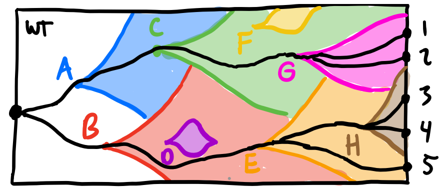
$$\Rightarrow \log\left(\frac{s_b}{U_b}\right) \ll \log(Ns_b) \ll \log^2\left(\frac{s_b}{U_b}\right)$$

Note: used heuristic derivation here...

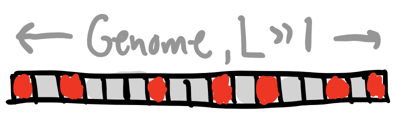
for formal analysis (using branching processes)

see Appendix A and B below

Recap : clonal interference



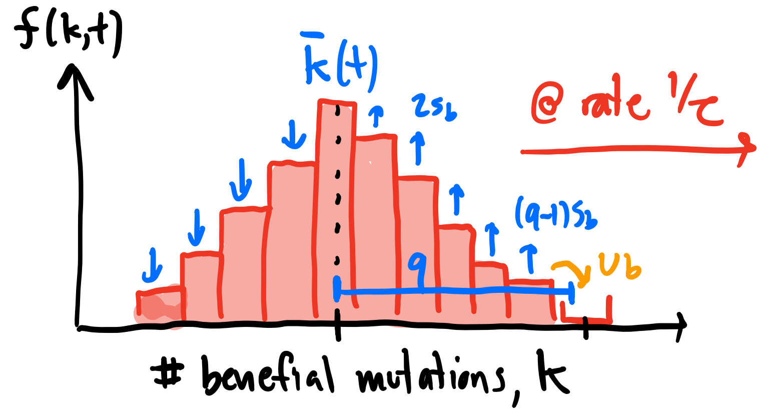
"Staircase" model



↳ selected mut'ns, $+s_b$

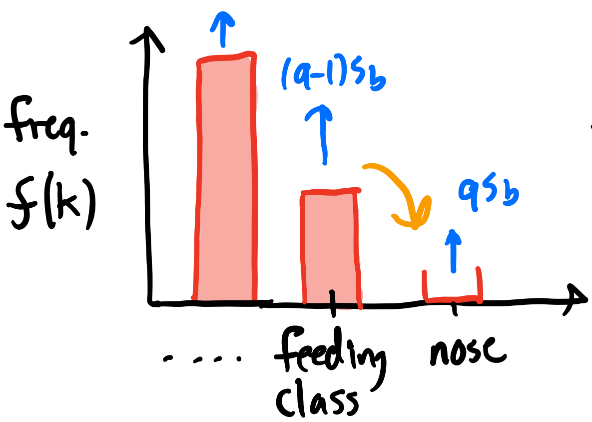
total rate $U_b \equiv L\lambda_b N$

$\Rightarrow$

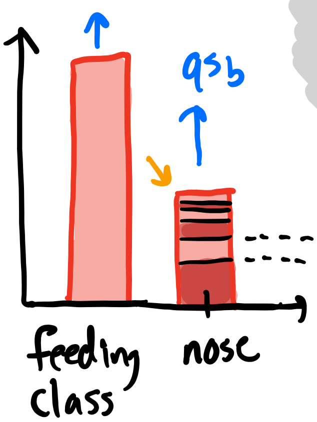


key behavior occurs @ "nose":

Multiple mutations contribute to nose!



$t+\tau$ $\Rightarrow$



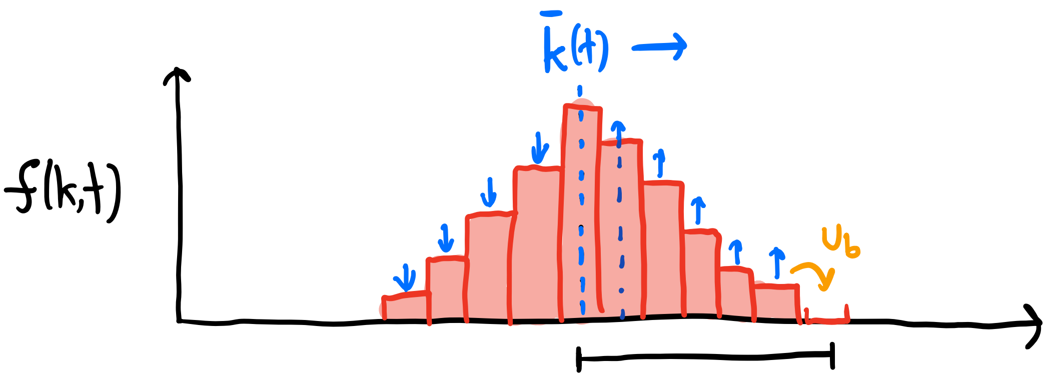
$$f_j(t) \sim \frac{1}{Nqs} e^{qs(t-\tau_j)}$$

$$\tau_j \sim \frac{1}{(q-1)s_b} \log\left(\frac{s_b q j}{U_b}\right)$$

Total contribution:

$$f_{nose}(t) \equiv \sum_{j=1}^{\infty} f_j(t) \equiv \frac{1}{Nqs} e^{qs(t-\tau)} \Rightarrow \tau \sim \frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b}\right)$$

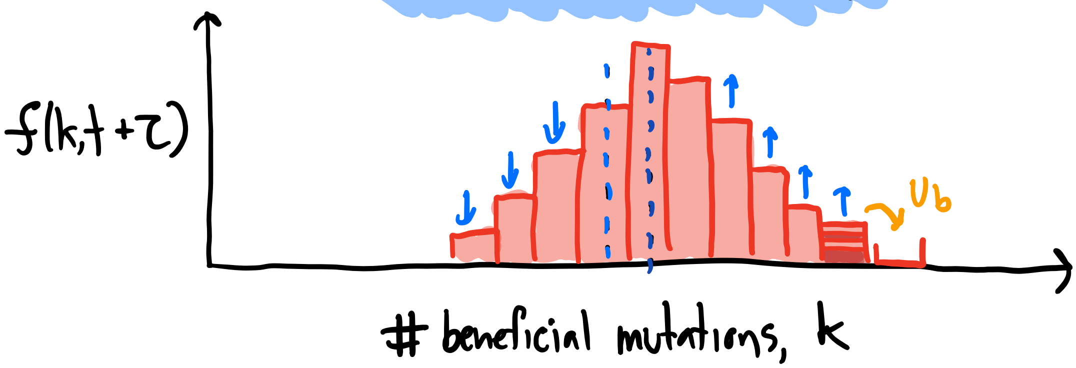
⇒ Complete picture of dynamics of fitness dist'n:



Width: $q \sim \dfrac{2\log(Ns_b)}{\log(s_b/U_b)}$

"click time"

$$\tau \sim \frac{\log^2(s_b/U_b)}{2s_b \log(Ns_b)}$$



Rate of adaptation

$$\left\langle \frac{\partial \bar{x}(t)}{\partial t} \right\rangle \approx \frac{s_b}{\tau}$$

⇒ early tests for clonal interference in lab evolution experiments:

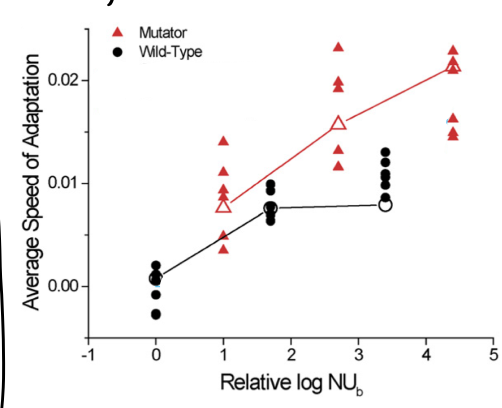
successive mutations:

$$\left\langle \frac{d\bar{x}}{dt} \right\rangle \sim s_b^2 \cdot NU_b$$

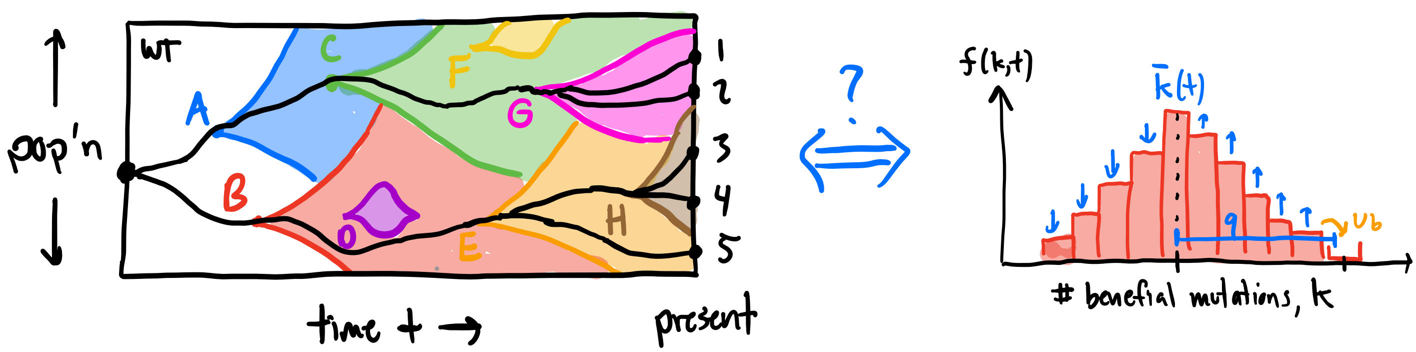
clonal interference:

$$\left\langle \frac{d\bar{x}}{dt} \right\rangle \sim s_b^2 \cdot \frac{\log(Ns_b)}{\log^2(s_b/U_b)}$$

e.g. Desai et al, Curr Bio 2007

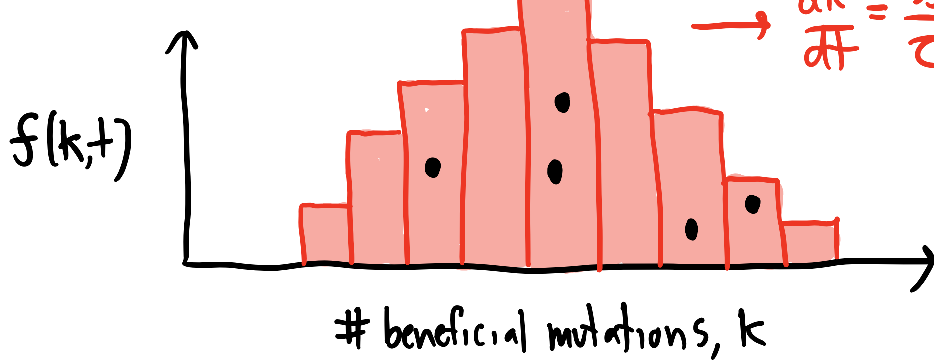


Next: Can we use this picture to understand genetic diversity backwards in time?



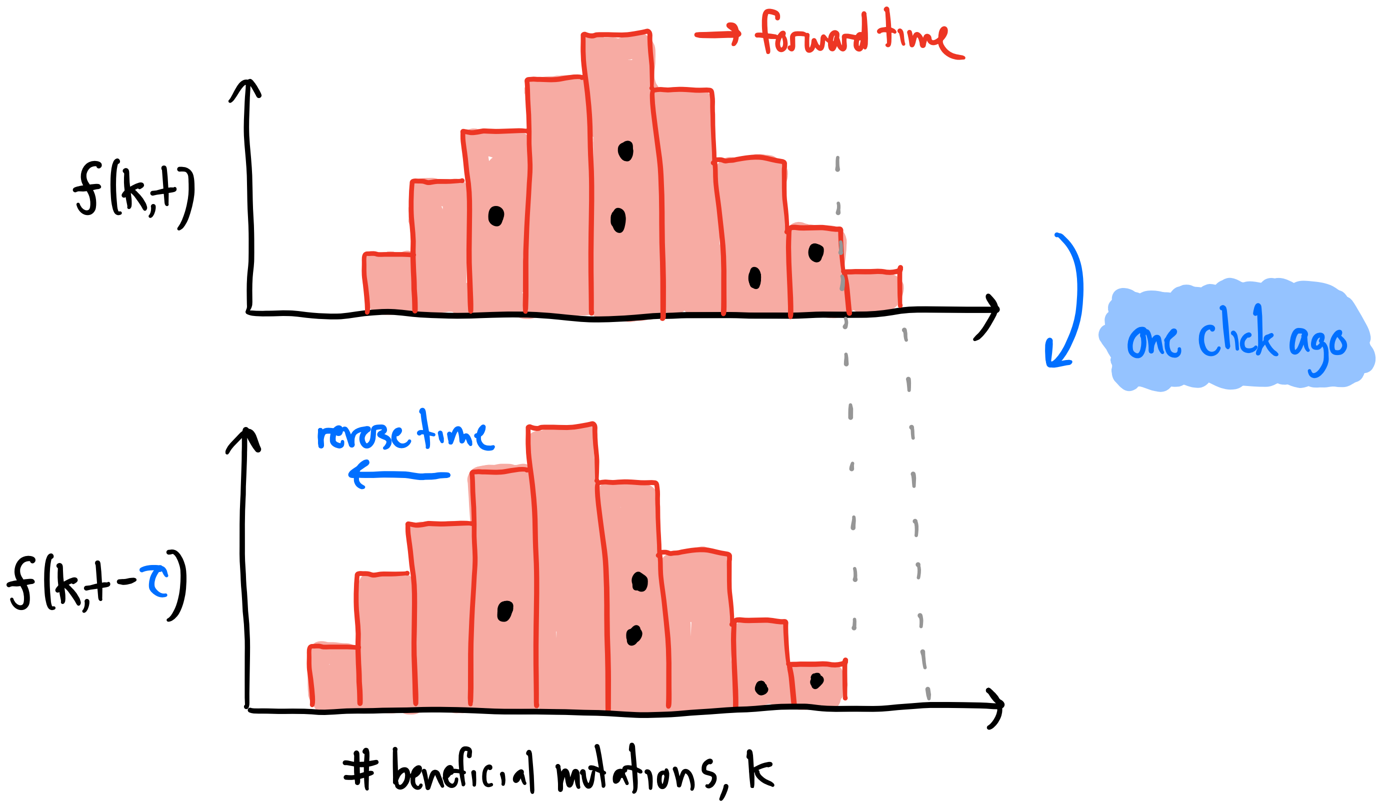
Answer: Yes we can! Let's start w/ some cartoons...

Step 1: draw sample of individuals from pop'n (present day)



$$\frac{d\bar{k}}{dt} = \frac{s_b}{\tau}$$

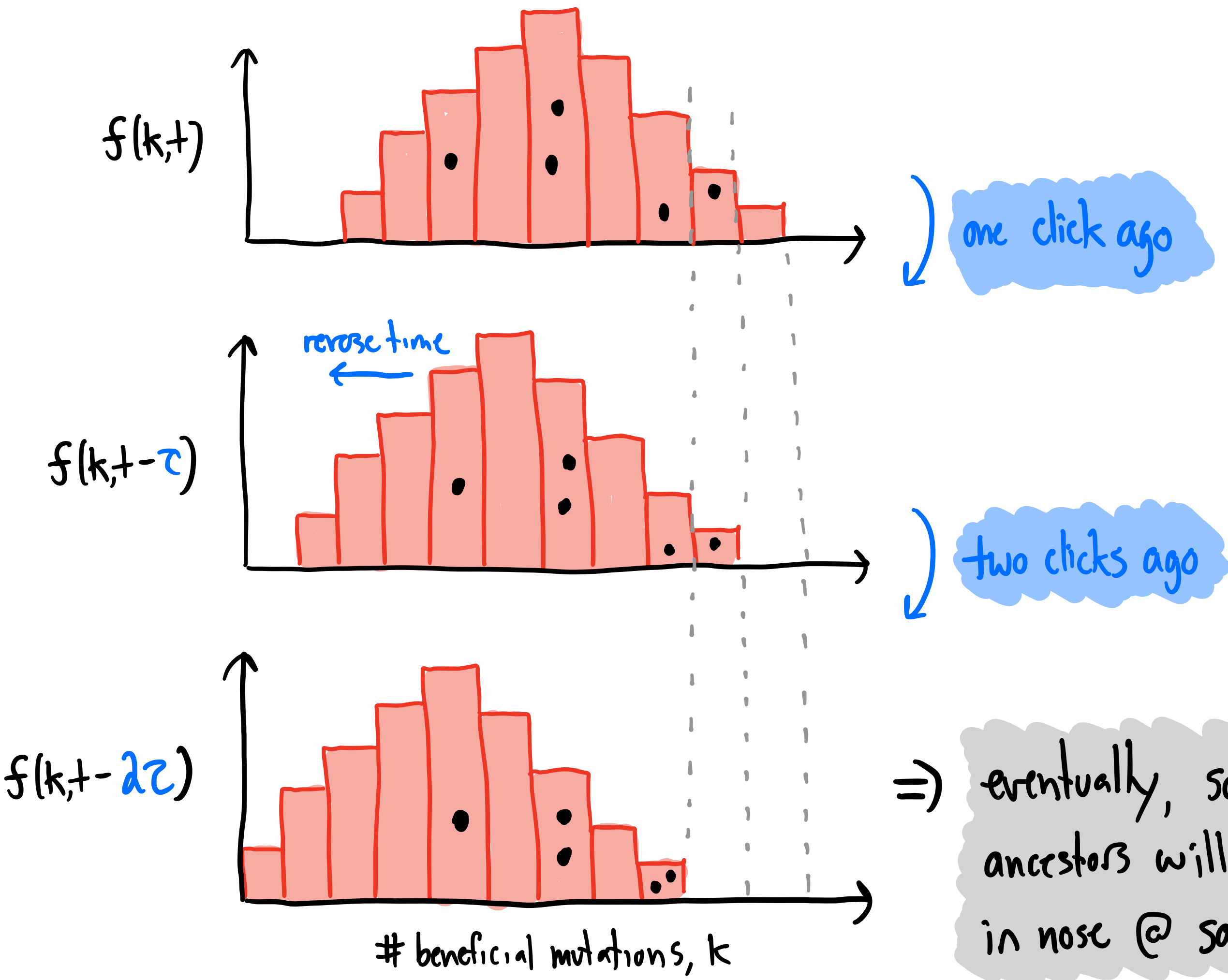
Step 2: where was everyone one click ago?



① can only coalesce if in same fitness class

② But little chance of coalescing in "bulk" of dist'n

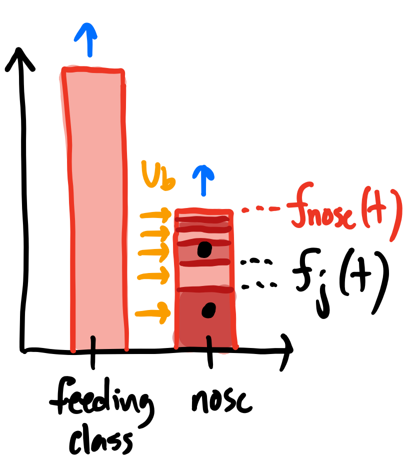
(since $\tau \ll Nf_{q-1}(\tau)$, $Nf_{q-2}(\tau)$, etc.)



⇒ eventually, some ancestors will exist in nose @ same time!

Two possible scenarios:

① Individuals are from separate lineages in the nose

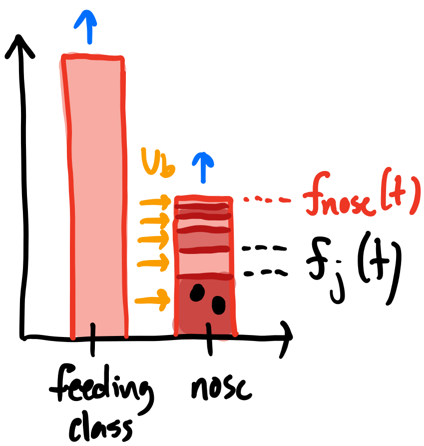


⇓

separate ancestors in feeding class

(distinct mut'n events)

② Individuals from <u>same</u> lineage in nose



$\Downarrow$

common ancestor in feeding class

(coalescence w/in τ gens)

$$\Rightarrow \text{Probability: } p_c(2) = \sum_{j=1}^{\infty} \left(\frac{f_j(t)}{f_{nose}(t)} \right)^2 = \sum_{=1}^{\infty} \left[\frac{\frac{1}{Nqs_b} e^{qs_b(t-\tau_j)}}{\frac{1}{Nqs_b} e^{qs_b(t-\tau)}} \right]^2$$

$$= \sum_{j=1}^{\infty} e^{-2qs_b(\tau_j - \tau)}$$

$\Uparrow$

only depends on establishment times τ_j !

⇒ if we plug-in typical values of τ_j & τ from heuristics:

$$\tau_j \sim \frac{1}{(q-1)s_b} \log\left(\frac{s_b q j}{U_b}\right) ; \qquad \tau \sim \frac{1}{(q-1)s_b} \log\left(\frac{s_b}{U_b}\right)$$

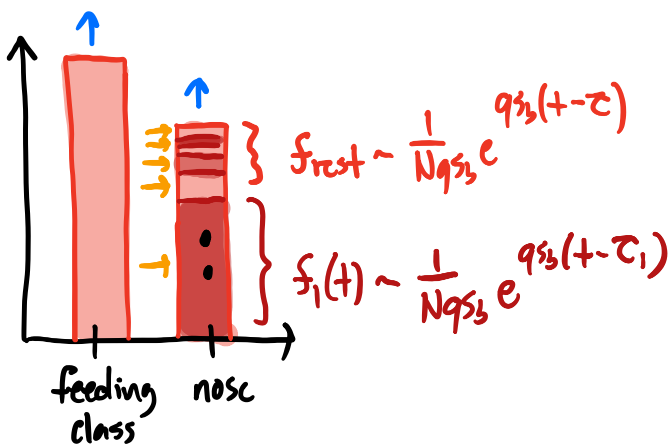
$$\Rightarrow \quad p_c(2) = \sum_{j=1}^{\infty} e^{-2qs_b(\tau_j - \tau)} = \sum_{j=1}^{\infty} (qj)^{-\frac{2q}{q-1}} \approx 1/q^2$$

~~⇒ suggests coalescence after $\sim q^2$ clicks ($T_{MRCA} \sim q^2 \tau$)~~

⇒ missing key part of puzzle: fluctuations

⇒ coalescence rare for typical lineage sizes, but small chance of having anomalously early mutant where coalescence is much more likely!

e.g. if <u>first</u> successful mutation occurs when $\tau_1 \lesssim$ <u>typical</u> τ ...



$$\Downarrow$$

$$\frac{f_1(t)}{f_{nosc}(t)} \gtrsim O(1)$$

$$\Downarrow$$

$$p_c(2) \sim O(1) \,!$$

$\Rightarrow$ not a huge shift in <u>time</u>: typically, $\tau_1 - \tau \sim \frac{\log(q)}{qs_b}$

so $\Delta\tau_1$ is $\ll 1/s_b \ll \tau$ (i.e. $\ll$ click time)

$\Rightarrow$ occurs w/ total probability:

$$p_{jackpot} \sim \int_0^{\tau} d\tau_1 \, NU_b f_{q-1}(t) \cdot qs_b$$

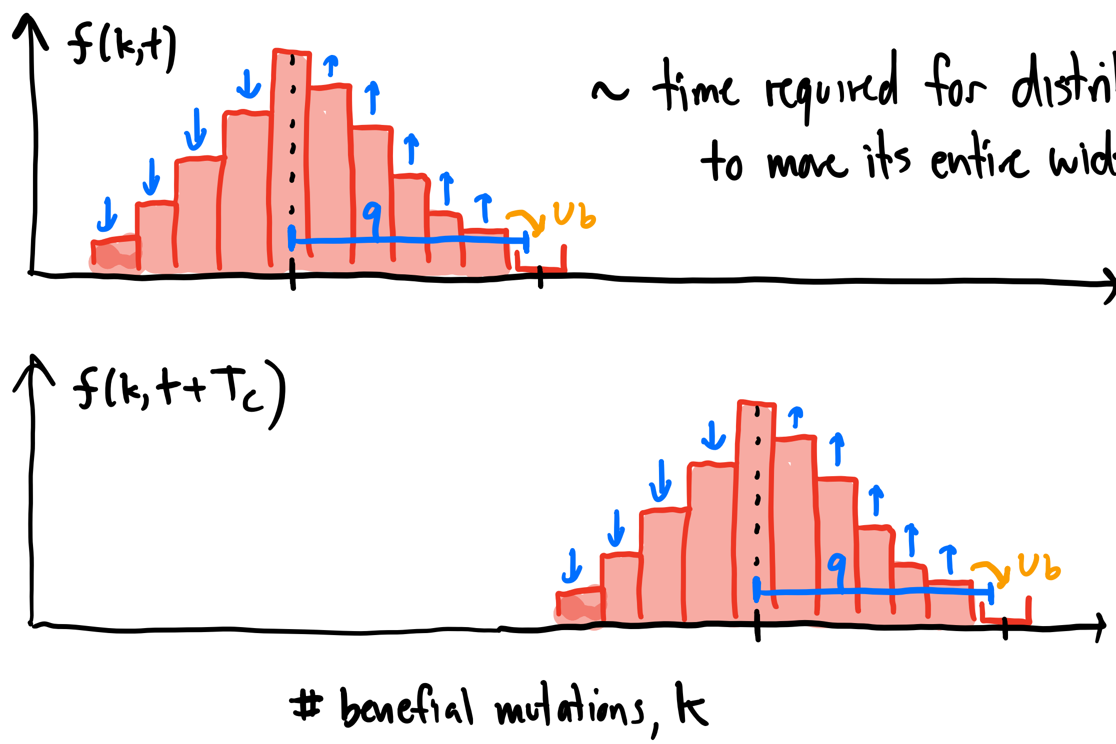
$$\sim \int_0^{\tau} d\tau_1 \, \cancel{N}U_b \cdot \frac{e^{(q-1)s_b t}}{\cancel{Nqs_b}} \cdot \cancel{qs_b} \sim \frac{U_b}{(q-1)s_b} e^{(q-1)s_b \tau}$$

$$\sim 1/q$$

$\Rightarrow$ $\boxed{p_{jackpot} \sim 1/q}$ $\left(\gg 1/q^2 \Rightarrow \right.$ more likely to coalesce via rare jackpot than normal establishment process $\left. \right)$

$\Rightarrow$ typical coalescence after $1/p_{jackpot} \sim q$ clicks

$\Rightarrow$ $\boxed{\text{coalescent timescale} \quad T_c \equiv q\tau \sim \frac{1}{s_b} \log\left(\frac{s_b}{U_b}\right)}$

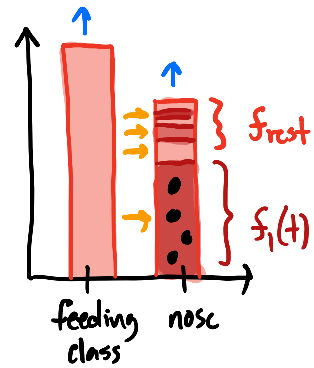


~ time required for distribution to move its entire width

$\Rightarrow$ fluctuations were crucial for determining T_c !

$\Rightarrow$ coalescence is "bursty" :

e.g. in larger sample size n :

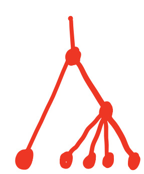


$$p_c(n\to 1) = \left(\frac{f_1(t)}{f_1(t)+f_{rest}(t)}\right)^n$$

$$\approx \begin{cases} \sim 1 & \text{if } f_1(t) \gtrsim n\cdot f_{rest}(t) \\ \ll 1 & \text{else} \end{cases}$$

$$\Rightarrow \quad P_{jackpot}(n) = \int_0^{\tau - \log(n)/qs_b} d\tau_1 \; NU_b f_{q-1}(t)\cdot q s_b \sim 1/qn$$

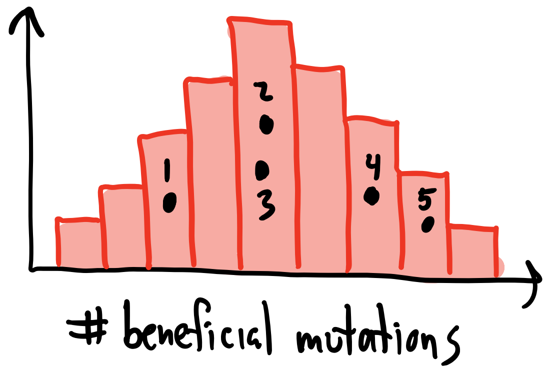
$\Rightarrow$ i.e. multiple mergers likely!



* For "formal" treatment, see Appendix C ...

Another interesting feature of genealogies + travelling wave:

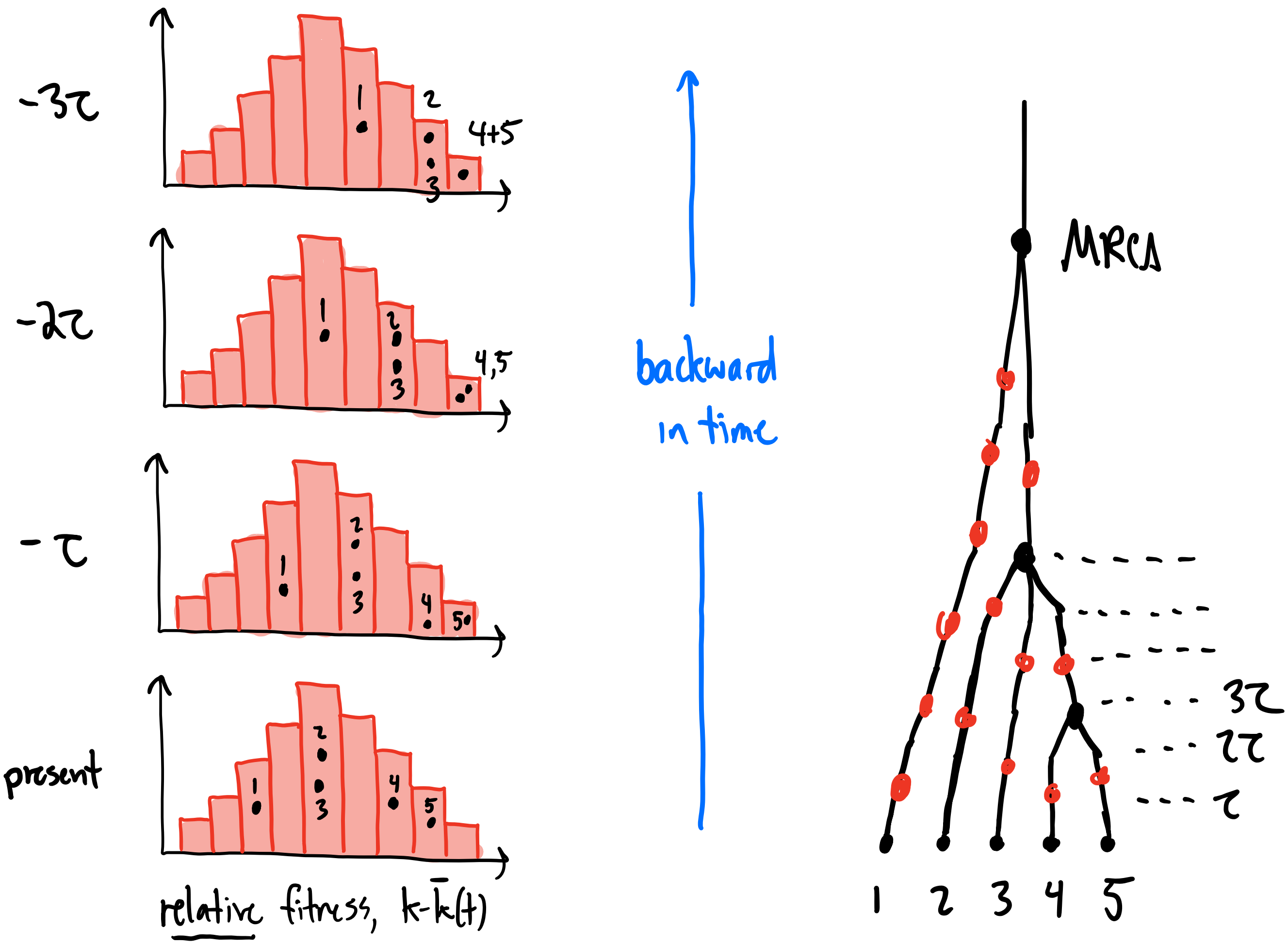
⇒ consider same example:



⇒ which individual's <u>descendents</u> are more likely to take over pop'n in <u>future</u>?

⇒ e.g. $5 \to 4 \to 2,3 \to 1$

⇒ now let's try to "simulate" genealogy...



$\Rightarrow$ time (+ burstiness) of coalescence in past

$\Rightarrow$ info about fitness in present

$\Rightarrow$ forecasts about who takes over in future!

eLIFE
elifesciences.org

RESEARCH ARTICLE

Predicting evolution from the shape of genealogical trees

Richard A Neher[1]*, Colin A Russell[2], Boris I Shraiman[3]*

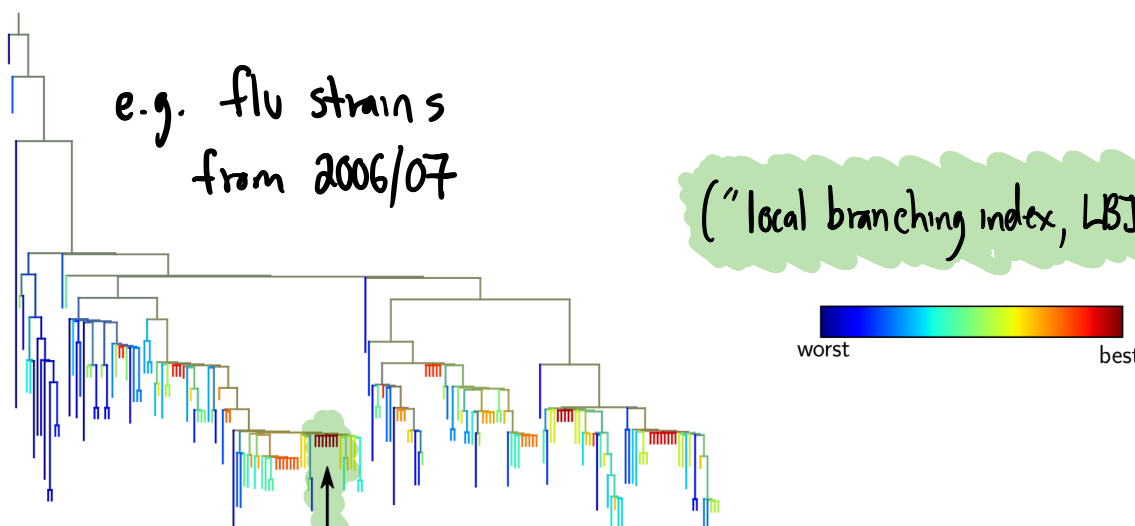
[1]Evolutionary Dynamics and Biophysics, Max Planck Institute for Developmental Biology, Tübingen, Germany; [2]Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom; [3]Kavli Institute for Theoretical Physics, University of California, Santa Barbara, Santa Barbara, United States

⇒ implemented this idea for HA gene in influenza

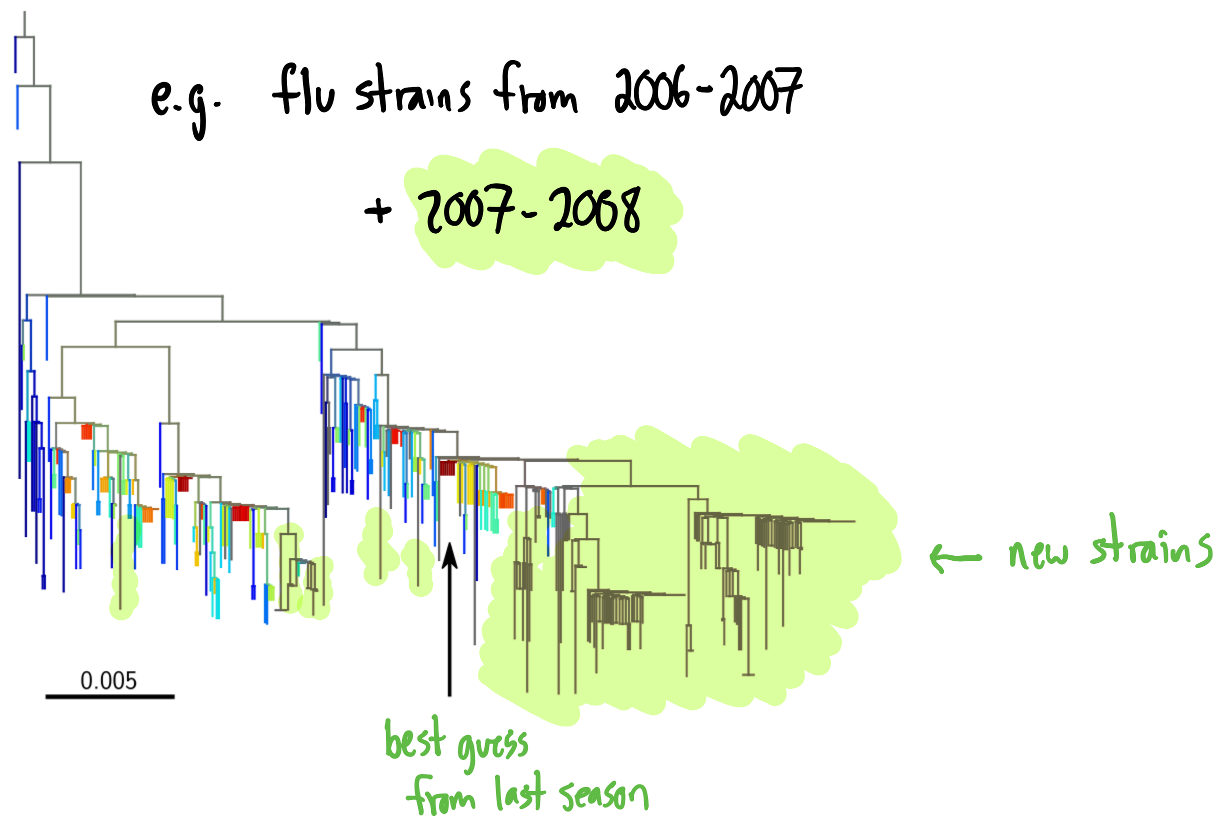
(data from Problem #1 in HW 1)

e.g. flu strains from 2006/07

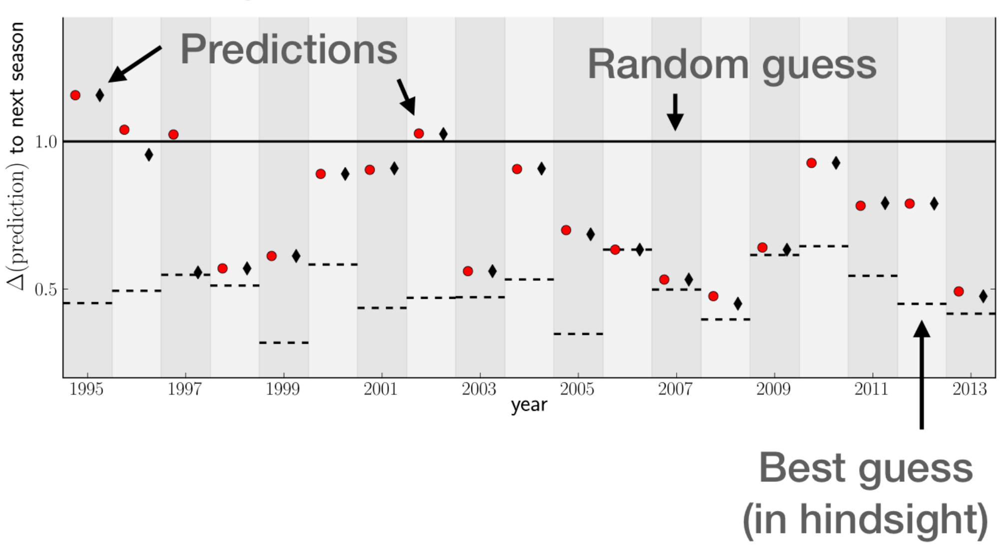
("local branching index, LBI")



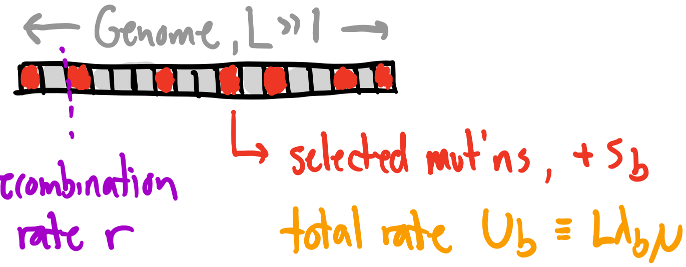
highest inferred "fitness" (LBI)



Evaluating performance over multiple flu seasons:

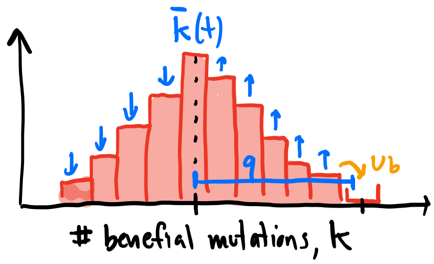


What about recombination?



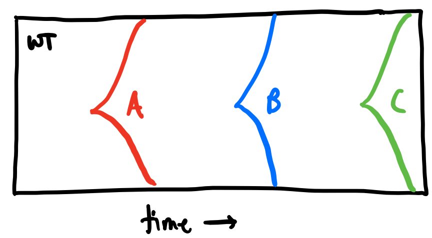
$r = 0$

$\Rightarrow$

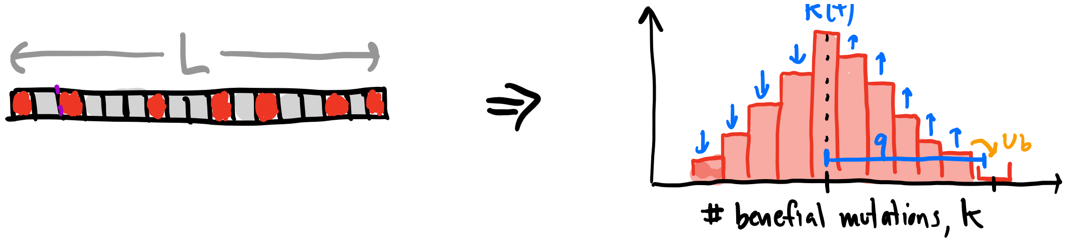


$r \gg N \cdot \mu \lambda_b \cdot s$

$\Rightarrow$

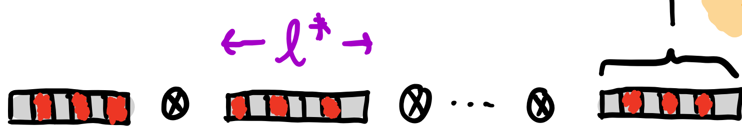


$r=0$



$U_{b,\mathrm{eff}} = \ell^* \lambda_b \mu$?

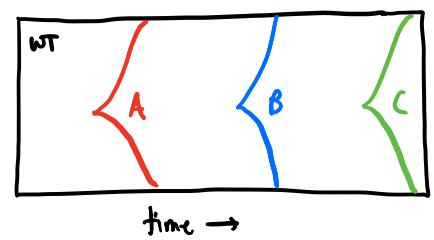
In between?



Blocks are ~independent

... but multiple mut'ns / block!

$r \gg N \cdot \mu \lambda_b \cdot s$



If true, need:

1. w/in blocks, recombination should be rare! ($r \approx 0$)

$$\Rightarrow \quad r\ell^* \cdot T_c(N, s_b, U_{\mathrm{eff}}(\ell^*)) \ll 1$$

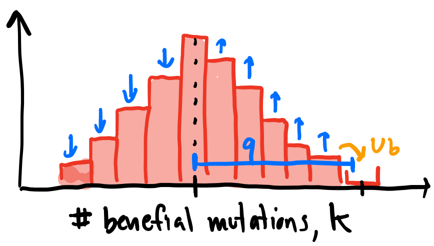
2. between blocks, recombination should be frequent!

$$\Rightarrow \quad r\ell^* \cdot T_c \gg 1$$

$\Rightarrow$ can we (almost) satisfy both w/ $r\ell^* \cdot T_c \sim \mathcal{O}(1)$?

Linkage block ansatz

$\ell^* \sim 1/rT_c$ $\quad$ $U_{b,\mathrm{eff}} = \ell^* \lambda_b \mu$ $\rightarrow$



$$T_c \sim \frac{1}{s_b}\log\left(\frac{s_b}{U_{b,\mathrm{eff}}}\right)$$

$\Rightarrow$ Self consistency: $$T_c \sim \frac{1}{s_b}\log\left(\frac{s_b}{\mu\lambda_b}\cdot rT_c\right)$$

$\Rightarrow$ solution: $$T_c \sim \frac{1}{s_b}\log\left(\frac{r}{\mu\lambda_b}\right)$$

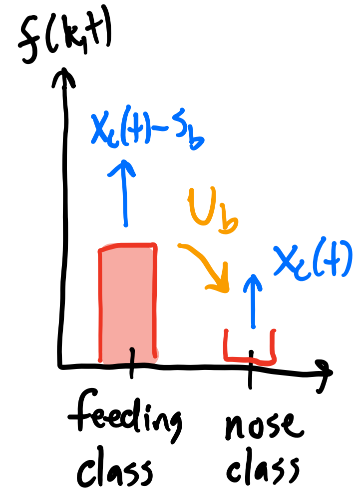
$$\Rightarrow \ell^* \sim \frac{s_b}{r}\log^{-1}\left(\frac{r}{\mu\lambda_b}\right)$$

$\Rightarrow$ self consistent if $NU_{b,\mathrm{eff}}\log(Ns_b) \gg 1$ & $U_{\mathrm{eff},b} \ll s_b$

$$\Rightarrow N\cdot\mu\lambda_b\cdot s_b \gg r \gg \mu\lambda_b$$

Appendix A: 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 1st 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:

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

$+$

mutation: $U_b f_{q-1}(t) = \frac{U_b}{2Nqs_b} e^{\int_0^t (X_c(t')-s_b)dt}$

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

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

$\Rightarrow$ will be harder because time-varying fitness

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

$\Rightarrow$ 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 τ...)

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

$$\frac{\partial H_f}{\partial t} = \left[x_c(t) z - \frac{z^2}{2N} \right] \frac{\partial H_f}{\partial z} - z \, v_b f_{q-1}(t) \, H_f$$

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

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

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

w/ $\psi(t) = 0$, $z(0) = z$,

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

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

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

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

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

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

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

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

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

(can plug in & check...)

so $$H_f(z,t) = \exp\left[-\int_0^t \frac{z\, U_b f_{q-1}(t-\tau)\, e^{\int_0^\tau x_c(t-\tau')d\tau'}}{1+\frac{z}{2N}\int_0^\tau e^{\int_0^{\tau'} x_c(t-\tau'')d\tau''}d\tau'}\, d\tau\right]$$

$$= \exp\left[-\int_0^t \frac{z \cdot U_b f_{q-1}(u)\, e^{\int_u^t x_c(u')du'}}{1+\frac{z}{2N}\int_u^t e^{\int_{u'}^t x_c(u'')du''}du'}\, du\right]$$

$\Rightarrow$ again, helpful to define $v(t)$ s.t. $f_q(t) \equiv \frac{v(t)}{2Nqs_b} e^{\int_0^t x_c(t')dt'}$

$$\Rightarrow \quad H_v(z,t) \equiv \left\langle e^{-z \cdot v(t)} \right\rangle \equiv H_f\left(2Nqs_b\, e^{-\int_0^t x_c(t')dt'} z, t\right)$$

$$\Rightarrow \quad H_v(z,t) = \exp\left[-\int_0^t \frac{z\, U_b f_{q-1}(u)\, 2Nqs_b\, e^{-\int_0^u x_c(u')du'}}{1+qsz \cdot \int_u^t du'\, e^{-\int_0^{u'} x_c(u'')du''}}\, du\right]$$

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

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

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

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

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

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

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

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

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 $O\left(\frac{1}{s_b}\right)$ of τ.

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

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

$$\log H_v(z) \approx -\int_{-\infty}^{\infty} \frac{z \cdot U_b e^{-s_b t}\, dt}{1 + 2 \cdot z \cdot e^{-q s_b t}}$$

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

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

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

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

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

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

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

$\Rightarrow$ typical value of establishment time:

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

$\Rightarrow$ consistent w/ results from simpler heuristic argument!

Appendix B: How many lineages contribute to new nose?

Recall in heuristic argument, we had:

$$f_{nose}(t) \equiv \sum_{j=1}^{J_{max}} f(t) = \frac{1}{Nqs} e^{qs_b(t-\tau)} \cdot \sum_{j=1}^{J_{max}} \frac{1}{qj^{1+1/q}}$$

& argued that sum over k converged to ≈ 1.

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

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

$$\Rightarrow \sum_{j=1}^{J_{max}} \frac{1}{qj} \frac{1}{^{1+1/q}} \approx \int_1^{max} \frac{dk}{qj^{1+1/q}} = 1 - e^{-\frac{1}{q}\log J_{max}}$$

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

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

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

$$\Rightarrow \quad \tau_{j_{max}} - \tau_1 \sim \frac{1}{q s_b} \log(J_{max}) \gg \frac{1}{s_b}$$

Thus, mutations that establish $\gg 1/s_b$ after τ_1 have <u>negligible</u> contribution to $f_{nose}(t)$, τ, etc

$\Rightarrow$ since $1/s_b \ll \tau$, this happens <u>long</u> before next click.

$\Rightarrow$ can take $J_{max} \simeq \infty$ w/o losing any accuracy

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

Appendix C: formal analysis of coalescence in the nose

Recall: main result for stochastic size of nose:

$$f_{nose}(t) \equiv \frac{v}{2Nqs_b} e^{qs_b t} \Rightarrow H_v(z) \approx e^{-\frac{U_b}{s_b} z^{1-1/q}}$$

(supplement of lecture 19)

Let's fine-grain this further:

$\Rightarrow$ let $f_\ell(t) \equiv$ freq of lineage in nose founded by beneficial mutation @ site ℓ

$\Rightarrow$ then $H_{v_\ell}(z) \approx e^{-\frac{\mu}{s_b} z^{1-1/q}}$

$\Rightarrow$ Probability that 2 individuals coalesce = probability that they came from same lineage:

$$\Rightarrow \quad p_c(2) = \left\langle \sum_{\ell=1}^{L/\Lambda_b} \left(\frac{f_\ell(t)}{\sum_{\ell'} f_{\ell'}(t)} \right)^2 \right\rangle = \left\langle \sum_{\ell=1}^{L/\Lambda_b} \left(\frac{v_\ell}{\sum_{\ell'} v_{\ell'}} \right)^2 \right\rangle$$

$\Rightarrow$ Trick: using $\displaystyle\int_0^\infty \frac{\lambda^\alpha}{\Gamma(\alpha)} z^{\alpha-1} e^{-\lambda z}\, dz = 1$, can write as

$$p_c(2) = \left\langle \sum_\ell \left(\frac{v_\ell}{\sum_{\ell'} v_{\ell'}} \right)^2 \right\rangle = \left\langle \sum_\ell v_\ell^2 \int_0^\infty dz\, z\, e^{-\left(\sum_{\ell'} v_{\ell'}\right) z} \right\rangle$$

$$= \sum_\ell \left\langle \int_0^\infty dz \cdot z \cdot \left(v_\ell^2 e^{-z v_\ell} \right) \cdot \prod_{\ell' \neq \ell} e^{-v_{\ell'} z} \right\rangle$$

$$= \sum_\ell \int_0^\infty dz \cdot z \cdot \frac{\partial^2 H_{v_\ell}(z)}{\partial z^2} \cdot \prod_{\ell' \neq \ell} H_{v_{\ell'}}(z)$$

$\Rightarrow$ using results above for $H_{v_\ell}(z)$ and $H_v(z)$,

$$\Rightarrow \quad \frac{\partial H_{v_\ell}}{\partial z} = -\frac{\mu_\ell}{s_b}\left(1 - \frac{1}{q}\right) z^{-1/q} H_{v_\ell}(z)$$

$$\Rightarrow \quad \frac{\partial^2 H_{v_\ell}}{\partial z^2} = \frac{1}{q}\,\frac{\nu_\ell}{s_b}\left(1-\frac{1}{q}\right) z^{-1-1/q}\, H_{v_\ell}(z) + \mathcal{O}(\nu_\ell^2)$$

so that:

$$p_c(2) = \sum_\ell \int_0^\infty dz \cdot z \cdot \frac{\partial^2 H_{v_\ell}(z)}{\partial z^2} \cdot \prod_{\ell' \neq \ell} H_{v_{\ell'}}(z)$$

$$= \frac{1}{q} \sum_\ell \frac{\nu_\ell}{U_b} \int_0^\infty dz\, \frac{U_b}{s_b}\left(1-\frac{1}{q}\right) z^{-1/q}\, \underbrace{H_v(z)}_{=\prod_\ell H_{v_\ell}(z)}$$

$$= \frac{1}{q} \int_0^\infty -\frac{\partial H_v(z)}{\partial z} = \frac{1}{q}\left[\cancelto{1}{H(0)} - \cancelto{0}{H(\infty)}\right]$$

$$= \frac{1}{q} \quad \text{as desired}$$

Can do same thing for larger samples:

$$p_c(n) = \left\langle \sum_\ell \left(\frac{\nu_\ell}{\sum_{\ell'} \nu_{\ell'}}\right)^n \right\rangle = \sum_\ell \int_0^\infty dz \cdot \frac{(-1)^n z}{\Gamma(n)}\, \frac{\partial^n H_{v_\ell}(z)}{\partial z^n} \prod_{\ell' \neq \ell} H_{v_{\ell'}}(z)$$

$$\Rightarrow \quad \frac{\partial^n H_{v_e}(z)}{\partial z^n} = \frac{N_e}{S_b} \frac{(-1)^n}{q}\left(1-\frac{1}{q}\right) \prod_{k=1}^{n-2}\left(k+\frac{1}{q}\right) z^{-n+1-1/q} H_{v_e}(z) + \mathcal{O}(N_e^2)$$

$$\approx \frac{N_e}{S_b} \frac{(-1)^n (n-2)!}{q} z^{-n+1-1/q} H_{v_e}(z)$$

$$\Rightarrow \quad p_c(n) = \frac{1}{q} \frac{\Gamma(n-1)}{\Gamma(n)} \cdot \sum_e \frac{N_e}{U_b} \cdot \int_0^\infty dz \cdot \frac{-\partial H_v(z)}{\partial z}$$

$$\approx \frac{1}{q(n-1)}$$

$$\Rightarrow \quad \boxed{p_c(n) = \frac{p_c(2)}{n-1}} \leftarrow$$

also known as Bolthausen-Sznitman coalescent (BSC)