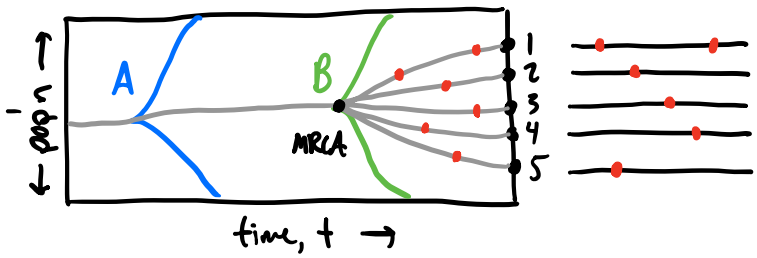
Chapter 13

The independent sites approximation

Last time: Extending coalescent approaches for:

① Selection

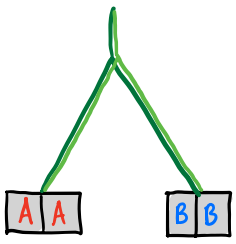
(successive mutations regime)



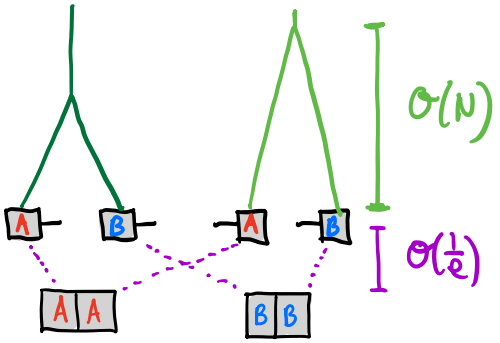
– OR –

② Recombination

$N_e \ll 1$ (effectively asexual)



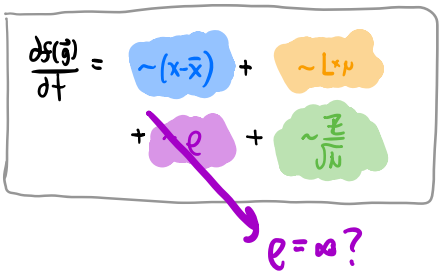
$N_e \gg 1$ (effectively independent)



Today: can we apply similar idea in forward-time picture?

(working directly w/ $f(\vec{g},t)$)

⇒ i.e., can we find a high recombination limit of our multi-locus model of evolution?



and therefore treat sel'n & recomb'n — and mut'n & drift — @ the same time?

Answer: Yes!

⇒ to start, consider 2-locus model

w/o selection or mutation (i.e. genotypes already exist)

⇒ 4 genotypes: $\vec{g} = (0,0), (1,0), (0,1), (1,1)$

⇒ 4 genotype freqs: $f_{00}, f_{10}, f_{01}, f_{11}$

Multi-locus SDEs reduce to:

(i) $\frac{df_{11}}{dt} = \underbrace{\rho\left[f_{10}f_{01} - f_{11}f_{00}\right]}_{\text{recombination}} + \underbrace{\sqrt{\frac{f_{11}}{N}}\eta_{11} - f_{11}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}}_{\text{genetic drift}}$

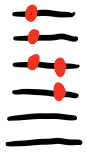
(ii) $\frac{df_{10}}{dt} = \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \sqrt{\frac{f_{10}}{N}}\eta_{10} - f_{10}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$

(ii) $\frac{df_{01}}{dt} = \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \sqrt{\frac{f_{01}}{N}}\eta_{01} - f_{01}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$

(iv) $\frac{df_{00}}{dt} = \rho\left[f_{10}f_{01} - f_{11}f_{00}\right] + \sqrt{\frac{f_{00}}{N}}\eta_{00} - f_{00}\sum_{\vec{g}}\sqrt{\frac{f_{\vec{g}}}{N}}\eta_{\vec{g}}$

$\Rightarrow$ Present day sample = Multinomial $(n, \vec{f})$

$(n_{11}, n_{10}, n_{01}, n_{00})$



$\Rightarrow$ Note: only 3 independent eqs (since $f_{11}+f_{10}+f_{01}+f_{00}=1$)

$\Rightarrow$ can eliminate $f_{00} = 1 - f_{11} - f_{10} - f_{01}$

& work w/ f_{11}, f_{10}, f_{01}

key idea: f_{11}, f_{10}, f_{01} is not only basis we can work with...

$\Rightarrow$ one alternative that is often used:

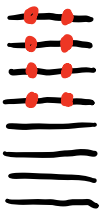
"allele freqs"

$$\begin{cases} f_1 \equiv f_{11} + f_{10} \Rightarrow \text{total freq of mutants @ site 1} \\ f_2 \equiv f_{11} + f_{01} \Rightarrow \text{" site 2} \end{cases}$$

$$D \equiv f_{11} - f_1 f_2 \equiv f_{11} f_{00} - f_{10} f_{01} \Rightarrow \text{"linkage disequilibrium" (LD)}$$

$\Rightarrow$ Why? LD is measure of how double mutant deviates from model where mut'ns are independent

e.g. one high-LD scenario:
(D large + positive)



$f_1 = 1/2, \; f_2 = 1/2$

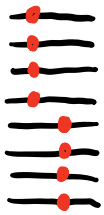
$f_{11} = 1/2$

$D = 1/2 - 1/4 = +1/4$

$r^2 = 1$

e.g. another high LD scenario:

(D large & negative)



$f_1 = 1/2$, $f_2 = 1/2$

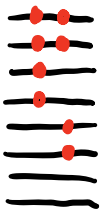
$f_{11} = 0$

$\Rightarrow D = 0 - 1/4 = -1/4$

$$r^2 = \frac{D^2}{f_1(1-f_1)f_2(1-f_2)} = \frac{1/16}{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}} = 1$$

e.g. a low LD scenario:

$(D \approx 0)\ (r^2 = 0)$



$f_1 = f_2 = 1/2$, $f_{11} = 1/4$

$D = 1/4 - 1/4 = 0$

$\Rightarrow$ sometimes write as correlation coefficient:

$$r \equiv \frac{D}{\sqrt{f_1(1-f_1)f_2(1-f_2)}} \Rightarrow r^2$$

(doesn't care about sign)

$\Rightarrow$ why is f_1, f_2, D a good basis?

$\Rightarrow$ let's rewrite our SDEs using def'ns:

$$f_1 \equiv f_{11} + f_{10}, \quad f_2 \equiv f_{11} + f_{01}, \quad D = f_{11} - f_1 f_2$$

$$\frac{\partial f_1}{\partial t} \equiv \frac{\partial f_{11}}{\partial t} + \frac{\partial f_{10}}{\partial t} = \rho\left[f_{10}f_{01} - f_{11}f_{00}\right] + \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \text{noise}$$

$\rho[\cdots] + \rho[\cdots] \to 0$

$$= 0 + \text{noise}$$

$$\Rightarrow \frac{\partial f_2}{\partial t} = 0 + \text{noise}$$

$$\Rightarrow \frac{\partial D}{\partial t} \equiv \frac{\partial}{\partial t}\left[f_{11} - f_1 f_2\right] = \frac{\partial f_{11}}{\partial t} - f_2 \frac{\partial f_1}{\partial t} - f_1 \frac{\partial f_2}{\partial t}$$

$\frac{\partial f_1}{\partial t} \to 0 + \text{noise}$, $\frac{\partial f_2}{\partial t} \to 0 + \text{noise}$

$$= -\rho D + \text{noise}$$

$\Rightarrow$ i.e. recombination cannot change mutation frequencies

$\Rightarrow$ can only change linkage disequilibrium!

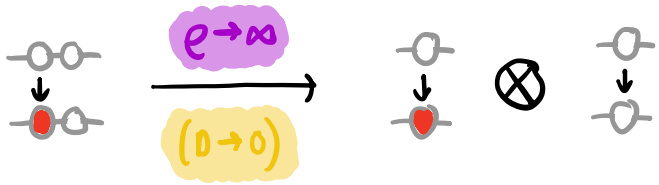
$\Rightarrow$ simple deterministic behavior: $D(t) = D(0)e^{-\rho t}$

$\Rightarrow$ suggests that if $e \to \infty$ $\Rightarrow$ $D(t) \approx 0$,

and maybe can treat 2-locus system as

direct product of independent single locus models:

e.g. pictorally:



Mathematically:

$$f_{11}(t) \approx f_1(t) f_2(t), \quad f_{10}(t) \approx f_1(t)[1 - f_2(t)], \ldots$$

$$\text{w/} \quad \frac{df_i}{dt} \approx \sqrt{\frac{f_i(1-f_i)}{N}}\,\eta_i(t) \qquad (i=1,2)$$

$\Rightarrow$ known as "linkage equilibrium", "free recombination",

"independent sites", "unlinked", etc.

$\Rightarrow$ can check validity using self-consistency argument:

$\Rightarrow$ assume $D \equiv f_{11} - f_{10} f_{01}$ is small &

calculate next order correction

$\Rightarrow$ correction is known as **Quasi-Linkage Equilibrium (QLE)**

$\Rightarrow$ easiest to see QLE for rare mutations ($f_1, f_2 \ll 1$)

$\Rightarrow$ then SDEs reduce to:

$$\frac{df_1}{dt} = \frac{df_{11}}{dt} + \frac{df_{10}}{dt} = \sqrt{\frac{f_{11}}{N}}\eta_{11}(t) + \sqrt{\frac{f_{10}}{N}}\eta_{10}(t)$$

$$= \sqrt{\frac{f_1 f_2 + D}{N}}\eta_{11}(t) + \sqrt{\frac{f_1 - f_1 f_2 - D}{N}}\eta_{10}(t) \equiv \sqrt{\frac{f_1}{N}}\eta_1(t)$$

$\Rightarrow$ i.e., define $\eta_1(t) \equiv \sqrt{\frac{N}{f_1}}\left[\sqrt{\frac{f_1 f_2 + D}{N}}\eta_{11} + \sqrt{\frac{f_1 - f_1 f_2 - D}{N}}\eta_{10}\right]$

(which satisfies $\langle \eta_1 \rangle = 0$, $\langle \eta_1^2 \rangle = 1$)

$\Rightarrow$ Similarly,

$$\frac{df_2}{dt} = \sqrt{\frac{f_2}{N}}\eta_2(t) \quad \text{w/} \quad \eta_2 \equiv \sqrt{\frac{N}{f_2}}\left[\sqrt{\frac{f_1 f_2 + D}{N}}\eta_{11} + \sqrt{\frac{f_2 - f_1 f_2 - D}{N}}\eta_{01}\right]$$

$$\Rightarrow \langle \eta_1 \eta_2 \rangle = \sqrt{\frac{N^2}{f_1 f_2} \cdot \frac{(f_1 f_2 + D)}{N}} = \sqrt{\frac{(f_1 f_2 + D)^2}{f_1 f_2}}$$

$\Rightarrow$ Finally, trickiest one:

$$\frac{\partial D}{\partial t} \equiv \frac{\partial f_{11}}{\partial t} - \frac{\partial}{\partial t}(f_1 f_2) \approx -\varepsilon D + \sqrt{\frac{f_{11}}{N}} \eta_{11} - \left\langle \left(\frac{\partial f_1}{\partial t}\right)_{diff} \left(\frac{\partial f_2}{\partial t}\right)_{diff} \right\rangle$$

$$= -\varepsilon D - \frac{f_1 f_2 + D}{N} + \sqrt{\frac{f_{11}}{N}} \eta_{11}(t)$$

$$\text{w/} \quad \langle \eta_{11} \eta_1 \rangle = \sqrt{\frac{f_1 f_2 + D}{f_1}}$$

$\Rightarrow$ in QLE ($f_{11} \approx f_1 f_2 +$ small correction) $\Rightarrow D \ll f_1 f_2$

$$\Rightarrow \langle \eta_1 \eta_2 \rangle = \sqrt{\frac{(f_1 f_2 + D)^2}{f_1 f_2}} \approx \sqrt{f_1 f_2} \ll 1 \quad (\text{since } f_1, f_2 \ll 1)$$

$$\Rightarrow \langle \eta_1 \eta_{11} \rangle = \sqrt{\frac{f_1 f_2 + D}{f_1}} \approx \sqrt{f_2} \ll 1$$

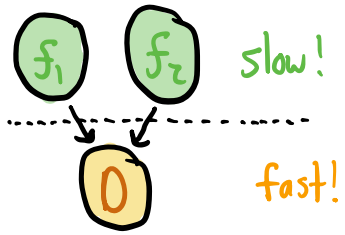
$\Rightarrow$ f_1 & f_2 independent of each other ($\&$ D)

$\Rightarrow$ $f_1 + f_2$ change on **drift timescale** $T_{drift} \sim N f_1, N f_2$

$\Rightarrow$ When $D \ll f_1 f_2$, LD equation reduces to:

$$\frac{\partial D}{\partial t} = -\frac{f_1 f_2}{N} - \varrho D + \sqrt{\frac{f_1 f_2}{N}}\,\eta_{\parallel}(t)$$

$\Rightarrow$ key idea: dynamics of D relax much faster than f_1, f_2 (since depends on ϱ)



$\Rightarrow$ Looks like Brownian particle in quadratic potential (Lecture 6)

w/ $\bar{x}_{eff} = -\frac{f_1 f_2}{N\varrho}$; $r_{eff} = \varrho$; $D_{eff} = \frac{f_1 f_2}{N}$

Solution: (i) $\langle D(t) \rangle = D(0)e^{-\varrho t} - \frac{f_1 f_2}{N\varrho}(1 - e^{-\varrho t}) \xrightarrow{t \gg 1/\varrho} -\frac{f_1 f_2}{N\varrho}$

(ii) $Var(D(t)) \xrightarrow{t \gg 1/\varrho} \frac{D_{eff}}{2 r_{eff}} = \frac{\frac{f_1 f_2}{N}}{2\varrho} = \frac{f_1 f_2}{2N\varrho}$

(iii) $Cov(D(t+\tau), D(t)) = Var(D(t))\, e^{-\varrho \tau}$

$\Rightarrow$ QLE ($f_{11} \approx f_1 f_2 +$ small correction) self-consistent if:

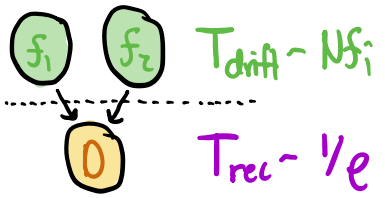
$\Rightarrow \left| \langle D \rangle \pm \sqrt{\mathrm{Var}(D)} \right| \ll f_1 f_2$

$\left| -\frac{f_1 f_2}{N\varrho} \pm \sqrt{\frac{f_1 f_2}{2N\varrho}} \right| \ll f_1 f_2$

$\Rightarrow N\varrho \gg \frac{1}{f_1 f_2} \gg 1$

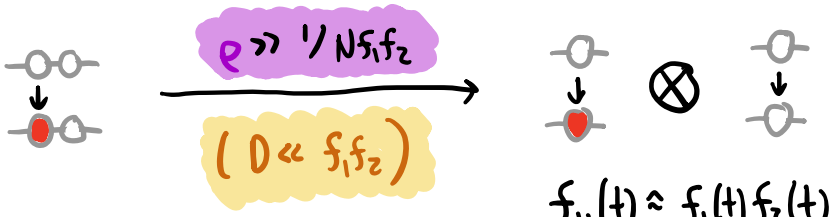
(like coalescent result, but now depends on f_1, f_2!)

$\Rightarrow$ separation of timescales also self-consistent since



✓

$\Rightarrow$ we're done! showed that:



i.e.

$\frac{\partial s(\vec{y})}{\partial t} = \sim \varrho + \sim \frac{\pi}{\sqrt{N}} \rightarrow \frac{\partial f_1}{\partial t} = \sqrt{\frac{f_1}{N}}\eta_1(t) \otimes \frac{\partial f_2}{\partial t} = \sqrt{\frac{f_2}{N}}\eta_2(t)$

⇒ can use same argument for selection too!

⇒ e.g. if $X(\vec{g}) \equiv s_1 g_1 + s_2 g_2$, can show:

(i) $\frac{df_1}{dt} = \frac{df_{11}}{dt} + \frac{df_{10}}{dt} \simeq s_1 f_1 + s_2 f_{11} + \text{noise}$

(ii) $\frac{df_2}{dt} = s_2 f_2 + s_1 f_{11} + \text{noise}$

(iii) $\frac{dD}{dt} = \frac{df_{11}}{dt} - f_1 \frac{df_2}{dt} - f_2 \frac{df_1}{dt} \approx (s_1 + s_2 - \rho) D + \text{noise}$

⇒ if $\rho \gg s_1 + s_2$ ⇒ $D(t) \to 0$

⇒ More generally, if recombination is faster than <u>all</u> other timescales ⇒ sites evolve independently

⇒ in practice, people often take this argument & run w/ it for <u>entire genome</u> (rarely check, since QLE is hard!)

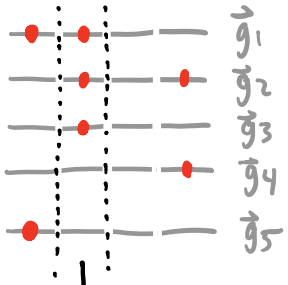
Linkage equilibrium approx ("independent sites")

L-site genome

recomb. is "fast"

collection of L single locus problems

$\Rightarrow$ Present-day sample:



mutations assigned independently based on $\{f_\ell(t)\}$

$n_\ell \sim \text{Binomial}(n, f_\ell(t))$

$\Rightarrow$ Data completely summarized by $\{n_\ell\}$: → "mutation counts"

$$\Pr[n_\ell = k] = \int \binom{n}{k} f_\ell^k (1-f_\ell)^{n-k} p_\ell(f_\ell)\, df_\ell$$

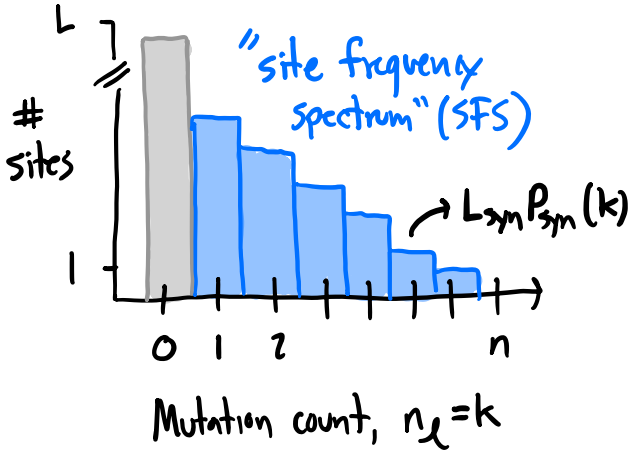
$$\frac{df_\ell}{dt} = s_\ell f_\ell(1-f_\ell) + \mu_\ell(1-2f_\ell) + \sqrt{\frac{f_\ell(1-f_\ell)}{N(t)}}\,\eta(t)$$

$\Rightarrow$ common to group "similar" sites together

e.g. all synonymous sites

$\downarrow s_\ell \approx 0$

$$P_{syn}(k) = \int \binom{n}{k} f^k (1-f)^{n-k} p(f|s=0)\, df$$



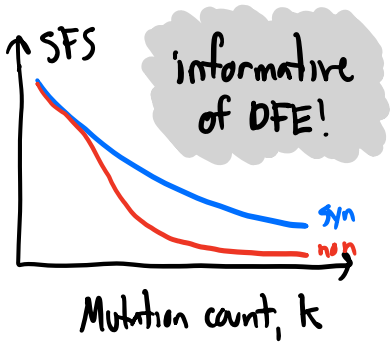
$\Rightarrow$ e.g. if $N(t) \approx N$ $\Rightarrow$ $P_{syn}(k) = \frac{2N\mu}{k}$

$\Rightarrow$ can do same thing for nonsynonymous mut'ns:

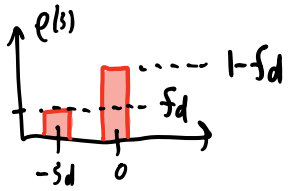
$$P_{non}(k) = \iint \binom{n}{k} f^k (1-f)^{n-k} p(f|s)\, \varrho(s)\, df\, ds$$

$$\approx \int_0^\infty \frac{2N\mu\, \varrho(-s)\, ds}{k(1 + 2Ns/n)^k} \qquad \text{if } N(t) \approx N \text{ \& } k \ll n$$

$$\Rightarrow \frac{P_{non}(k)}{P_{syn}(k)} = \int_0^\infty \frac{\varrho(-s)\, ds}{\left(1 + \frac{2Ns}{n}\right)^k}$$



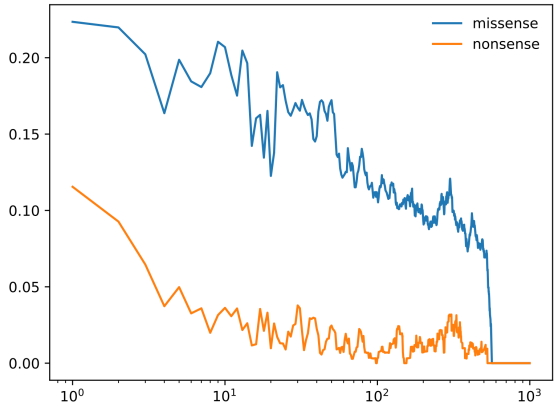
e.g. if $\rho(s) = (1-f_d)\,\delta(s) + f_d\,\delta(s+s_d)$



$$\Rightarrow \frac{P_{non}(k)}{P_{syn}(k)} = (1-f_d) + f_d\, e^{-k \log\left(1 + 2N s_y / n\right)}$$

e.g. real data from a human gut bacterium:
(Bacteroides fragilis)

$\frac{P_n(k)}{P_s(k)}$



Mutation count, k

⇒ informative of "constraint" (strong negative sel'n)

⇒ can coarse-grain over smaller subsets of sites to look for constraint on smaller regions (e.g. genes)

⇒ why? strongly constrained ≈ important for organism

⇒ when do we expect **independent sites approx** to work?

⇒ need $\rho_{eff} \approx r\Delta\ell$ large for all pairs of SNVs

- r: recombination rate per site
- $\Delta\ell$: distance between SNVs ($\approx 1/\pi \approx 1/2N\mu$)

⇒ $N\rho_{eff} \gg 1$ ⇒ $\frac{r}{\mu} \gg 1$

(in most organisms we've measured, $\frac{r}{\mu} \sim O(1)$!)

⇒ $\rho_{eff} \gg s$ ⇒ $Ns \ll \frac{r}{\mu} \sim O(1)$

⇒ bad approximation for strong selection!

⇒ need to turn to other approaches...