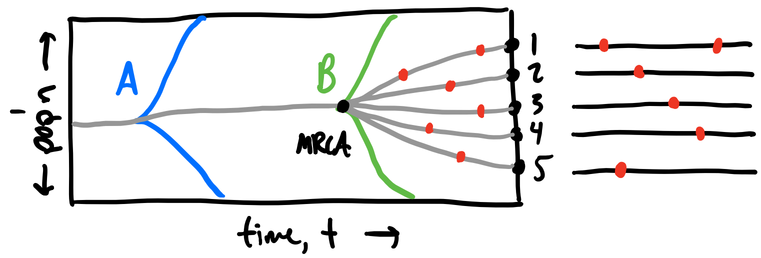
Chapter 13

The independent sites approximation

Last time: Extending coalescent approaches for:

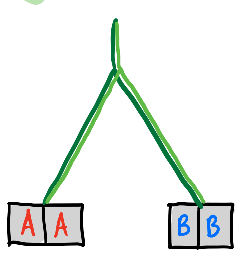
① Selection (successive mutations regime)



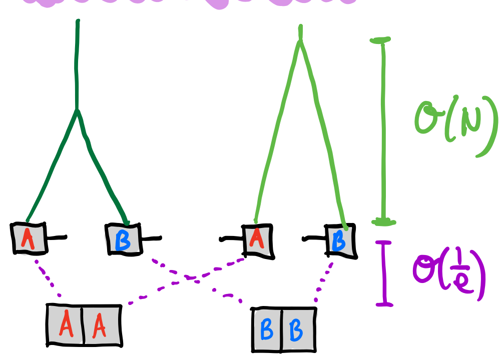
– OR –

② Recombination

$N\rho \ll 1$ (effectively asexual)



$N\rho \gg 1$ (effectively independent)



Today: can we apply similar idea in forward-time picture?
(working directly w/ $f(\vec{g},t)$)

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

$$\boxed{\frac{\partial f(\vec{g})}{\partial t} = \sim(x-\bar{x}) + \sim L^{*}\mu + \sim\rho + \sim\frac{Z}{\sqrt{N}}}$$

$\rho = \infty$?

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

Answer: Yes!

$\Rightarrow$ to start, consider 2-locus model w/o selection or mutation (i.e. genotypes already exist)

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

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

Multi-locus SDEs reduce to:

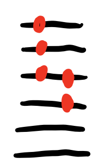
$$\text{(i)} \quad \frac{\partial f_{11}}{\partial t} = \underbrace{\rho\left[f_{10}f_{01} - f_{11}f_{00}\right]}_{\text{recombination}} + \underbrace{\sqrt{\frac{f_{11}}{N}}\,\eta_{11} - f_{11}\sum_{\vec{g}'}\sqrt{\frac{f_{\vec{g}}}{N}}\,\eta_{\vec{g}}}_{\text{genetic drift}}$$

$$\text{(ii)} \quad \frac{\partial f_{10}}{\partial t} = \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \sqrt{\frac{f_{10}}{N}}\,\eta_{10} - f_{10}\sum_{\vec{g}'}\sqrt{\frac{f_{\vec{g}}}{N}}\,\eta_{\vec{g}}$$

$$\text{(ii)} \quad \frac{\partial f_{01}}{\partial t} = \rho\left[f_{11}f_{00} - f_{10}f_{01}\right] + \sqrt{\frac{f_{01}}{N}}\,\eta_{01} - f_{01}\sum_{\vec{g}'}\sqrt{\frac{f_{\vec{g}}}{N}}\,\eta_{\vec{g}}$$

$$\text{(iv)} \quad \frac{\partial f_{00}}{\partial t} = \rho\left[f_{10}f_{01} - f_{11}f_{00}\right] + \sqrt{\frac{f_{00}}{N}}\,\eta_{00} - f_{00}\sum_{\vec{g}'}\sqrt{\frac{f_{\vec{g}}}{N}}\,\eta_{\vec{g}}$$

$\Rightarrow$ 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...

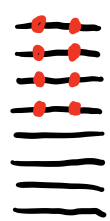
$\Rightarrow$ one alternative that is often used:

"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)



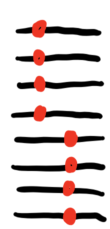
$f_1 = 1/2, \; f_2 = 1/2$

$f_{11} = 1/2$

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

---

e.g. another high LD scenario: (D large + negative)

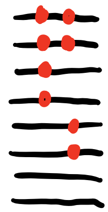


$f_1 = 1/2, \; f_2 = 1/2$

$f_{11} = 0$

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

e.g. a low LD scenario:
($D \approx 0$)



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

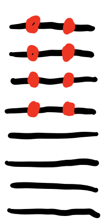
$D = 1/4 - 1/4 = 0$

$\Rightarrow$ sometimes write as <u>correlation</u> <u>coefficient</u>:

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

(doesn't care about sign)

Example 1:



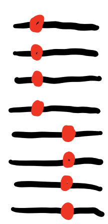
$f_1 = f_2 = 1/2$

$D = +1/4$

$$r = \frac{+1/4}{\sqrt{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}}} = +1$$

$\Rightarrow r^2 = 1$

Example 2:



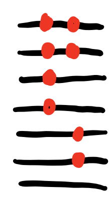
$f_1 = f_2 = 1/2$

$D = -1/4$

$$r = \frac{-1/4}{\sqrt{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}}} = -1$$

$\Rightarrow r^2 = 1$

Example 3:



$f_1 = f_2 = 1/2$

$D = 0$

$$r = \frac{0}{\sqrt{\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}}} = 0$$

$\Rightarrow r^2 = 0$

Question: Why is f_1, f_2, D a useful basis?

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

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

$$\frac{\partial f_1}{\partial t} \equiv \frac{\partial f_{11}}{\partial t} + \frac{\partial f_{10}}{\partial t} = \cancelto{0}{\rho\left[f_{10}f_{01} - f_{11}f_{00}\right] + \rho\left[f_{11}f_{00} - f_{10}f_{01}\right]} + \text{noise}$$

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

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

$$\Rightarrow \quad \frac{\partial D}{\partial t} \equiv \frac{\partial}{\partial t}\left[f_{11} - f_1 f_2\right] = \frac{\partial f_{11}}{\partial t} - f_2 \cancelto{0+\text{noise}}{\frac{\partial f_1}{\partial t}} - f_1 \cancelto{0+\text{noise}}{\frac{\partial f_2}{\partial t}}$$

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

Answer: recombination cannot change mutation frequencies

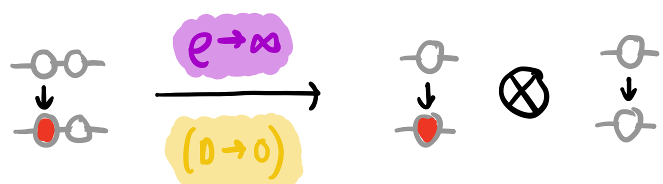
$\Rightarrow$ can only change linkage disequilibrium!

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

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

and maybe can treat 2-locus system as

<u>direct product</u> of <u>independent</u> single locus models:

e.g. pictorally: 

Mathematically:

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

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

$\Rightarrow$ known as "<u>linkage equilibrium</u>", "<u>free recombination</u>", "<u>independent sites</u>", "<u>unlinked</u>", etc.

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

$\Rightarrow$ assume $D \equiv f_{11} - f_{10} f_{01}$ is <u>small</u> & calculate <u>next order correction</u>

$\Rightarrow$ 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:

$$\frac{\partial f_1}{\partial t} = \frac{\partial f_{11}}{\partial t} + \frac{\partial f_{10}}{\partial t} = \sqrt{\frac{f_{11}}{N}}\,\eta_{11}(t) + \sqrt{\frac{f_{10}}{N}}\,\eta_{10}(t)$$

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

$\Rightarrow$ 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,

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

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

$\Rightarrow$ Finally, trickiest one:

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

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

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

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

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

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

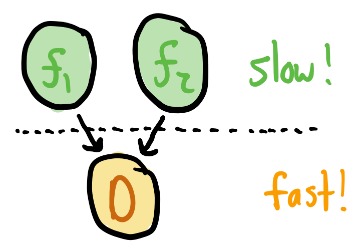
$\Rightarrow$ f_1 & f_2 independent of each other (& D)

⇒ f_1 + f_2 change on drift timescale $T_{drift} \sim Nf_1 , Nf_2$

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

$$\frac{\partial D}{\partial t} = -\frac{f_1 f_2}{N} - \rho D + \sqrt{\frac{f_1 f_2}{N}}\,\eta_{11}(t)$$

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



⇒ Looks like Brownian particle in quadratic potential (Lecture 6)

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

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

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

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

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

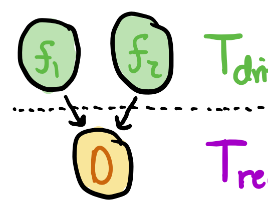
⇒ $\left|\langle D \rangle \pm \sqrt{\mathrm{Var}(D)}\right| \ll f_1 f_2$

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

⇒ $N\epsilon \gg \frac{1}{f_1 f_2} \gg 1$

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

⇒ separation of timescales also self-consistent since

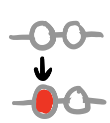


$T_{drift} \sim N f_i$

$T_{rec} \sim 1/\epsilon$

✓

⇒ we're done! showed that:



$$\xrightarrow[(D \ll f_1 f_2)]{\epsilon \gg 1/N f_1 f_2}$$

$f_{11}(t) \approx f_1(t) f_2(t)$

i.e.

$$\boxed{\frac{\partial f(\vec g)}{\partial t} = \sim\epsilon + \sim\frac{Z}{\sqrt{N}}} \rightarrow \boxed{\frac{\partial f_1}{\partial t} = \sqrt{\frac{f_1}{N}}\,\eta_1(t)} \otimes \boxed{\frac{\partial f_2}{\partial t} = \sqrt{\frac{f_2}{N}}\,\eta_2(t)}$$

$\Rightarrow$ can use same argument for selection too!

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

(i) $\dfrac{\partial f_1}{\partial t} = \dfrac{\partial f_{11}}{\partial t} + \dfrac{\partial f_{10}}{\partial t} \approx s_1 f_1 + s_2 f_{11} + \text{noise}$

(ii) $\dfrac{\partial f_2}{\partial t} = s_2 f_2 + s_1 f_{11} + \text{noise}$

(iii) $\dfrac{\partial D}{\partial t} = \dfrac{\partial f_{11}}{\partial t} - f_1 \dfrac{\partial f_2}{\partial t} - f_2 \dfrac{\partial f_1}{\partial t} \approx (s_1 + s_2 - \rho) D + \text{noise}$

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

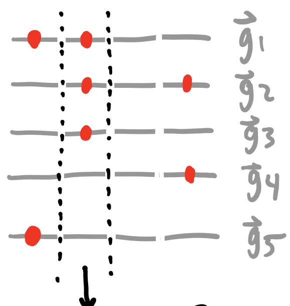
$\Rightarrow$ More generally, if recombination is faster than <u>all</u> other timescales $\Rightarrow$ sites evolve independently

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

Linkage equilibrium approx ("independent sites")

L-site genome $\xrightarrow{\text{recomb. is "fast"}}$ collection of L single locus problems

$\Rightarrow$ Present-day sample:



mutations assigned <u>independently</u> based on $\{f_\ell(t)\}$

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

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

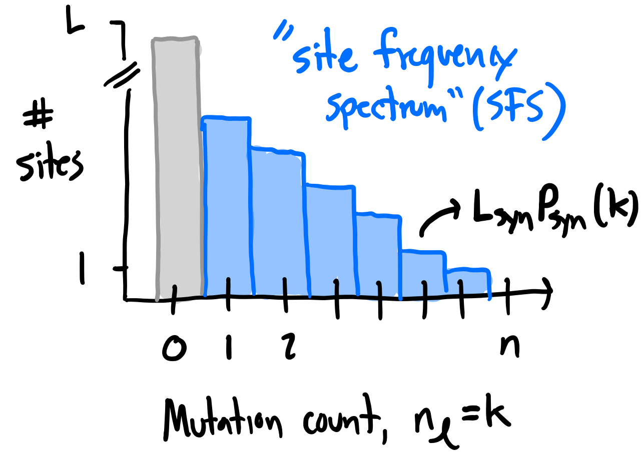
$$\Pr[n_\ell = k] = \int \binom{n}{k} f_\ell^k (1-f_\ell)^{n-k} \; p_\ell(f_\ell)\, df_\ell$$

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

$\Rightarrow$ common to group "similar" sites together e.g. all <u>synonymous</u> sites

$\downarrow s_\ell \approx 0$

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



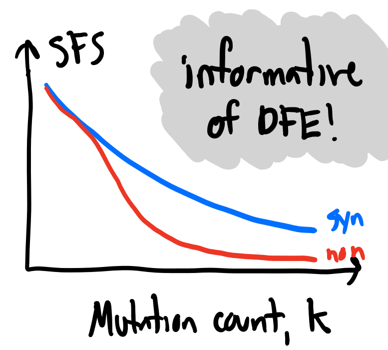
$\Rightarrow$ e.g. if $N(t) \approx N$ $\Rightarrow$ $P_{syn}(k) = \dfrac{2N\mu}{k}$

$\Rightarrow$ can do same thing for <u>nonsynonymous</u> mut'ns:

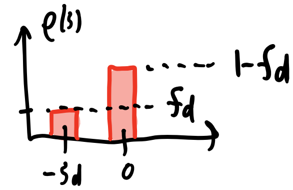
$$P_{non}(k) = \iint \binom{n}{k} f^k (1-f)^{n-k} p(f|s)\, \rho(s)\, df\, ds$$

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

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

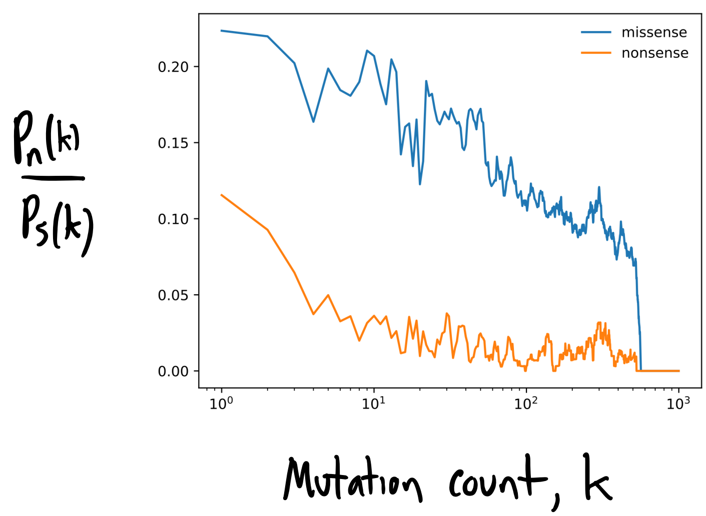


e.g. if $\rho(s) = (1-f_d)\,\delta(s) + f_d\,\delta(s+s_d)$



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

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



$\Rightarrow$ informative of "constraint" (strong negative sel'n)

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

$\Rightarrow$ why? strongly constrained $\approx$ important for organism

$\Rightarrow$ when do we expect independent sites approx to work?

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

recombination rate per site

distance between SNVs ($\approx 1/\pi \approx 1/2N\mu$)

$\Rightarrow$ $N\ell_{eff} \gg 1$ $\Rightarrow$ $\frac{r}{\mu} \gg 1$

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

$\Rightarrow$ $\ell_{eff} \gg s$ $\Rightarrow$ $Ns \ll \frac{r}{\mu} \sim \mathcal{O}(1)$

$\Rightarrow$ bad approximation for strong selection!

$\Rightarrow$ need to turn to other approaches...