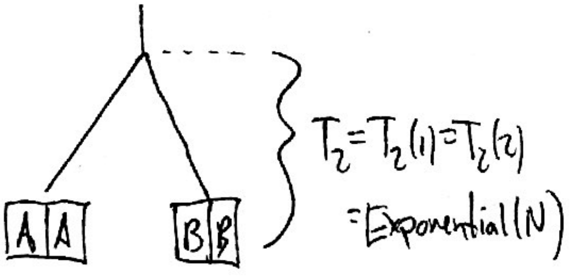
Quasi Linkage Equilibrium (QLE)

Last time, we talked about coalescent models for neutral recombining genomes, and saw that this led to 2 extreme limits:

$Ne \ll 1$

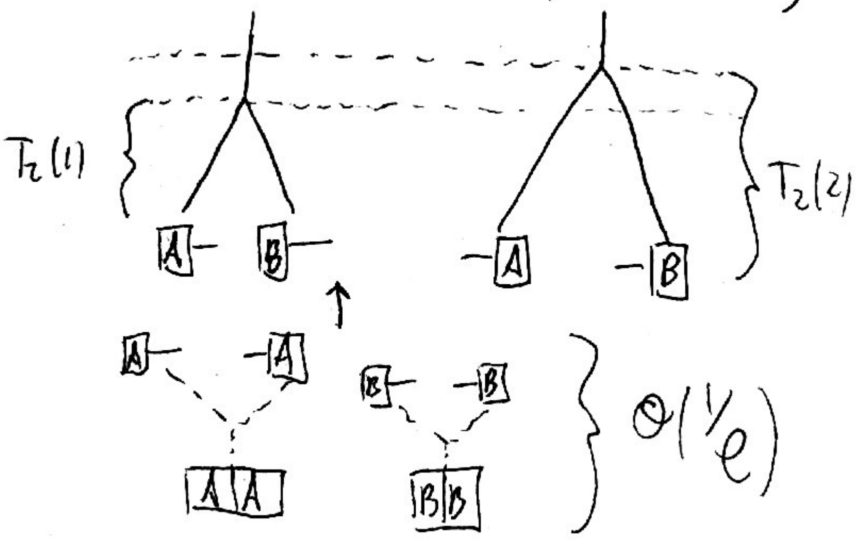
(effectively asexual)



⇑

not enough time for recombination to occur.

$Ne \gg 1$ (effectively independent)

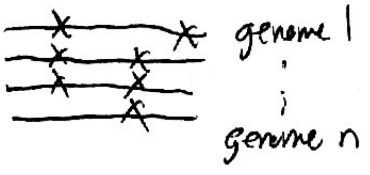


⇑

recombination will occur very fast, coarse grain over recomb. timescale.

⇒ in between, $Ne \sim \mathcal{O}(1)$ very complicated (ARG)

@ end of day, want <u>mutations</u> in sample, not trees...



⇒ also hard to get selection into this picture...

So today, want to talk about same concepts from a forward-time perspective, based on genotype frequencies, $f(\vec{g})$

$\Rightarrow$ to start, let's consider a simple 2-locus model for genotypes $\vec{g} = (0,0), (1,0), (0,1), (1,1)$, w/o selection

then genotype freqs satisfy the system of SDEs:

$$\frac{\partial f(11)}{\partial t} = \rho\left[f(10)f(01) - f(11)f(00)\right] + \sqrt{\frac{f(11)}{N}}\eta(11) - f(11)\sum_{\vec{g}'}\sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}')$$

$$\frac{\partial f(10)}{\partial t} = \rho\left[f(11)f(00) - f(10)f(01)\right] + \sqrt{\frac{f(10)}{N}}\eta(10) - f(10)\sum_{\vec{g}'}\sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}')$$

$$\frac{\partial f(01)}{\partial t} = \rho\left[f(11)f(00) - f(10)f(01)\right] + \sqrt{\frac{f(01)}{N}}\eta(01) - f(01)\sum_{\vec{g}'}\sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}')$$

$$\frac{\partial f(00)}{\partial t} = \rho\left[f(10)f(01) - f(11)f(00)\right] + \sqrt{\frac{f(00)}{N}}\eta(00) - f(00)\sum_{\vec{g}'}\sqrt{\frac{f(\vec{g}')}{N}}\eta(\vec{g}')$$

and sample comes from:

$$\left(f(11), f(10), f(01), f(00)\right) \xrightarrow{\text{Multinomial}(n,\vec{f})}$$

$n(11), n(10), n(01), n(00)$

this system really has only 3 independent eqs, since $f(11)+f(10)+f(01)+f(00)=1$

$\Rightarrow$ e.g. can eliminate $f(00)=1-f(11)-f(10)-f(01)$ & work w/ $f(11), f(10), f(01)$.

However, $(f(11), f(10), f(01))$ is not the only basis we could work with.

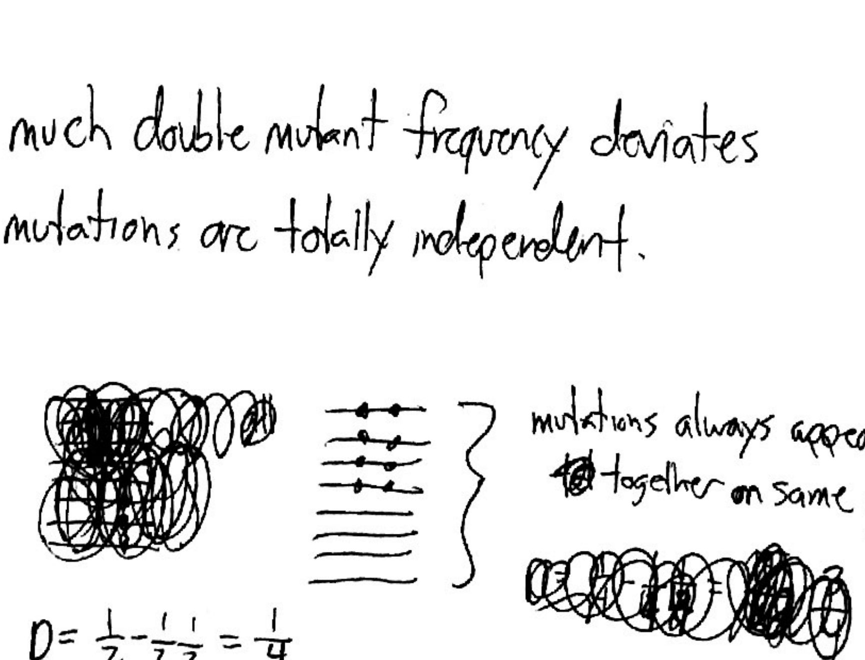
$\Rightarrow$ free to choose any other combination.

$\Rightarrow$ one combination that is often used:

$$\underbrace{f_1 \equiv f(11)+f(10)}_{\text{marginal allele freq of mutations @ site 1}}, \quad \underbrace{f_2 \equiv f(11)+f(01)}_{\text{allele freq @ site 2}}, \quad \underbrace{D \equiv f(11) - f_1 f_2 = f(11)f(00)-f(10)f(01)}_{\text{"Linkage disequilibrium" (LD)}}$$

$\Rightarrow$ LD is measure of how much double mutant frequency deviates from a model where mutations are totally independent.

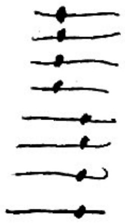
e.g. one high-LD scenario ($D=$ large & positive)



mutations always appear together on same genomes.

$$D = \frac{1}{2} - \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

e.g. another high-LD scenario (LD = large & negative)



} mutations always on separate chromosomes

$D = 0 - \frac{1}{2}\cdot\frac{1}{2} = -\frac{1}{4}$

⟹ Sometimes people measure as correlation coefficient:

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

e.g. in example 1, $r = +1$

example 2, $r = -1$

⟹ mutation @ site 1 gives you perfect info about site 2, & vice versa.

Why is f_1, f_2, D a good basis? Rewrite SDEs:

$$\frac{\partial f_1}{\partial t} \equiv \frac{\partial f(11)}{\partial t} + \frac{\partial f(10)}{\partial t} = \cancelto{0}{\rho\left(f(10)f(01) - f(00)f(11)\right]} + \rho\left(f(00)f(11) - f(10)f(01)\right)$$

$+$ noise (later)

$$\frac{\partial f_2}{\partial t} = 0 + \text{noise}, \quad \frac{\partial D}{\partial t} = \frac{\partial f_{11}}{\partial t} - \cancelto{0}{\frac{\partial f_1}{\partial t}(f_2)} - f_1 \cancelto{0}{\frac{\partial f_2}{\partial t}}$$

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

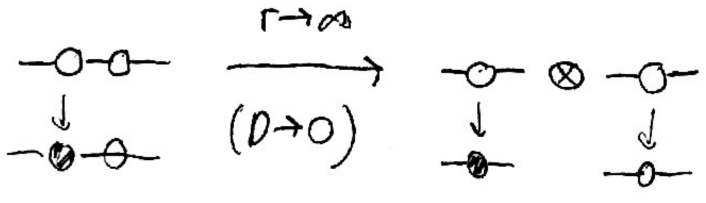
In other words, recombination cannot change allele frequencies, can only change linkage disequilibrium. change in linkage disequilibrium has very simple form (deterministically)

$$\Rightarrow \frac{\partial D}{\partial t} = -rD \Rightarrow D(t) = D(0)e^{-\rho t} \Rightarrow$$ decays to zero exponentially fast in ρ

$\Rightarrow$ suggests that if $\rho \to \infty$ (compared to what?)

$D \to 0$ and maybe we can treat 2 locus system as direct product of single locus systems :

e.g. pictorally :



we call this limit "Linkage equilibrium", "free recombination", "independent sites", etc.

we can check this assumption using our now-familiar self consistency argument: assume that D is small & calculate next order correction $\Rightarrow$ this correction is called

"Quasi-linkage equilibrium"
(QLE)

Easiest to see QLE if we focus on rare mutations, $f_1, f_2 \ll 1$. Then SDEs reduce to (w/ noise now)

$$\frac{\partial f_1}{\partial t} = \sqrt{\frac{f(11)}{N}}\,\eta(11) + \sqrt{\frac{f(10)}{N}}\,\eta(10) = \sqrt{\frac{f_1 f_2 + D}{N}}\,\eta(11) + \sqrt{\frac{f_1(1-f_2)-D}{N}}\,\eta(10)$$

$$\equiv \sqrt{\frac{f_1}{N}}\,\tilde{\eta}_1$$

$$\frac{\partial f_2}{\partial t} = \sqrt{\frac{f(11)}{N}}\,\eta(11) + \sqrt{\frac{f(01)}{N}}\,\eta(01) = \sqrt{\frac{f_1 f_2 + D}{N}}\,\eta(11) + \sqrt{\frac{(1-f_1)f_2}{N}}\,\eta(01) \equiv \sqrt{\frac{f_2}{N}}\,\tilde{\eta}_2$$

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

w/ $\langle \eta_{11} \hat{\eta}_1 \rangle = \sqrt{\frac{f_{11}}{f_1}}$, $\langle \eta_{11} \hat{\eta}_2 \rangle = \sqrt{\frac{f_{11}}{f_2}}$, $\langle \tilde{\eta}_1 \hat{\eta}_2 \rangle = \frac{f_{11}}{\sqrt{f_1 f_2}}$

⟹ in QLE, we will assume that dynamics of D (f_{11}) will relax much faster than dynamics of f_1, f_2

⟹ since neutral, know that f_1, f_2 change on timescale

$$T_{drift} \sim Nf_1, Nf_2$$

⟹ on timescales $\ll T_{drift}$, f_1 & f_2 are effectively const.

⟹ equation for f_{11} looks like Branching process w/ mutation w/ effective params: $\mu_e = \rho f_1 f_2$, $s_e = -\rho$

⟹ solution from Lecture 8, p.4:

Gamma Dist'n w/ shape: $\alpha = 2N\rho f_1 f_2$

and $f_{11,max} = \dfrac{1-e^{-\rho t}}{2N\rho} \to \dfrac{1}{2N\rho}$

$$\Rightarrow \langle f_{11} \rangle = f_1 f_2, \quad Var(f_{11}) = \frac{f_1 f_2}{2N\rho}$$

$$\Rightarrow \langle D \rangle = 0 \quad Var(D) = \frac{f_1 f_2}{2N\rho}$$

we also know that relaxation timescale is ~ $1/e$

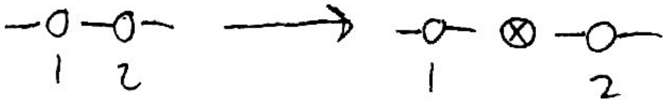
$\Rightarrow$ Now can check our assumptions : $\frac{1}{e} \ll T_{drift} \sim Nf_1, Nf_2$

$\Rightarrow$ QLE holds if $Nef_1, Nef_2 \gg 1$.

$\hookrightarrow$ similar to coalescent picture, except now explicit f dependence.

$\Rightarrow$ if this is true, then $\langle \hat{\eta}_2 \hat{\eta}_1 \rangle = \sqrt{\frac{f_{11}^2}{f_1 f_2}} = \sqrt{\frac{1}{(Nef_1)(Nef_2)}} \ll 1$

and $\frac{\partial f_1}{\partial t}, \frac{\partial f_2}{\partial t}$ equations really decouple! ✓.

so we showed that 

but only for sufficiently high freqs where can coarse-grain over recombination time.

$\Rightarrow$ can do same argument for selection, e.g. $X(\vec{g}) = s_1 g_1 + s_2 g_2$

then can show:

$$\frac{\partial f_1}{\partial t} = s_1 f_1 + s_2 D + \text{noise}$$

$$\frac{\partial f_2}{\partial t} = s_2 f_2 + s_1 D + \text{noise}$$

$$\frac{\partial D}{\partial t} = \frac{\partial f_{11}}{\partial t} - f_2 \frac{\partial f_1}{\partial t} - f_1 \frac{\partial f_2}{\partial t} =$$

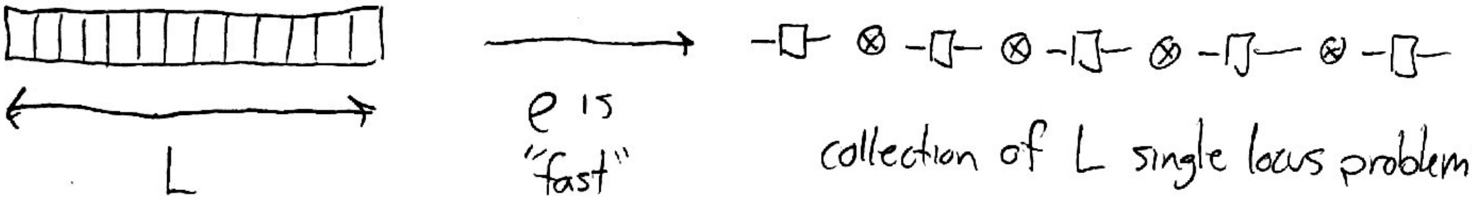
$$= (s_1 + s_2 - r) D + \text{noise.}$$

$\Rightarrow$ so if $r > s_1 + s_2 \Rightarrow D(t) \rightarrow 0$ w/ time.

$\Rightarrow$ General conclusion: if r is faster than all other timescales in system then loci evolve independently!

$\Rightarrow$ in practice, people take this argument & run w/ it for entire genome: (though people rarely check it like we just did because QLE is hard...)

Linkage equilibrium approx ("independent sites")

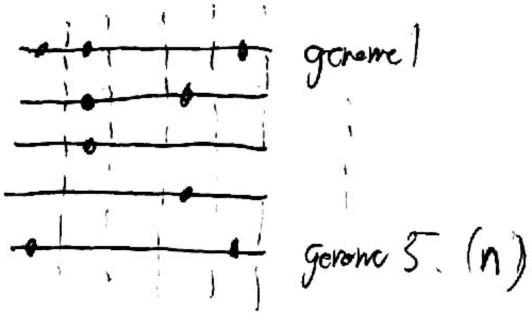


collection of L single locus problems.

⇒ when it works, one of the most powerful approximations in pop gen. Since it lets us use single locus results to look @ real data

↳ crazy if you think about it.

now when we draw a sample of individuals, we can assign mutations indepedtly given current allele frequencies, $\{f_\ell(t)\}$



⇒ by definition, $D \approx 0$ & no information in haplotype structure.

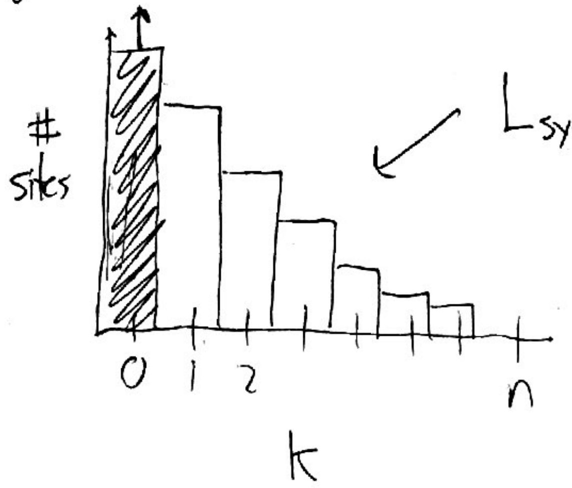
⇒ instead, data can be completely sumarized by n_ℓ = # of individuals w/ mutation @ site ℓ.

⇒

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

@ this point, people often gap "similar" sites together.

e.g. all synonymous sites (putatively neutral,) so $S_e = 0$



"site-frequency spectrum, SFS"

$\Rightarrow$ since there are lots of synonymous sites, & each one is an independent draw from $P_{syn}(n_e=k)$, then across genome, we get a self-averaged version of $P_{syn}(n_e=k)$, even from just 1 population! this means we can estimate demography, since

$$P_e(f_e) \Leftrightarrow \frac{\partial f_e}{\partial t} = \mu_e + \sqrt{\frac{f_e(1-f_e)}{N(t)}}\,\eta(t)$$

(mapping often not possible in closed form, but can do numerically)

$$\Rightarrow \text{if } N(t) = N \Rightarrow Pr(n_e=k) = \frac{2N\mu}{k} \rightarrow \frac{2N\mu}{f} \text{ when n large.}$$

$\searrow 2N\mu$ when $n=2$.

$\Rightarrow$ similarly, π_{syn} self averages to $2N\mu$, even w/ just 2 samples!

can do same thing for <u>non-synonymous mutations</u>

$$\Pr[n_e=k] = \iint \binom{n}{k} f_e^{k}(1-f_e)^{n-k}\, p(f_e|s_e)\,\rho(s_e)\, \underbrace{df_e}_{\text{random freq @ locus } e}\, \underbrace{ds_e}_{\text{selection coeff @ locus } e}$$

$$\Rightarrow \text{e.g. if } \rho(s) = (1-\lambda_d)\underbrace{\delta(s)}_{\text{neutral mutations}} + \lambda_d \underbrace{\delta(s+s_d)}_{\text{strongly deleterious mut. (e.g. LOF mutant)}}$$

$$\Rightarrow p(f) = \frac{2N\mu(1-\lambda_d)}{f} + \frac{2N\mu\lambda_d e^{-Nsf}}{f} \approx \begin{cases} \dfrac{2N\mu}{f} & \text{for } f \ll \dfrac{1}{Ns} \\ \text{(like synonymous)} & \\ \dfrac{2N\mu(1-\lambda_d)}{f} & \text{for } f \gg \dfrac{1}{Ns} \end{cases}$$

$\Rightarrow$ similarly for π:

$$\pi_{non} = (1-\lambda_d)2N\mu + \lambda_d \frac{2N\mu}{2Ns}$$

$\Rightarrow$ typically we don't know $2N\mu$, but can estimate it from π_{syn}

$$\Rightarrow \pi_{non}/\pi_{syn} = (1-\lambda_d) + \frac{\lambda_d}{Ns} \qquad (Ns \gg 1)$$

$\frac{\pi_n}{\pi_s}$ provides a measure of "constraint", how much negative selection is going on in population w/in T_{MRCA}.

e.g. in bacteria (E. coli in 2 different people's guts) often find $\pi_{non}/\pi_{syn} \approx 0.1$.

$\Rightarrow$ this is pretty crazy... suggests that $\lambda_d \gtrsim 0.9$

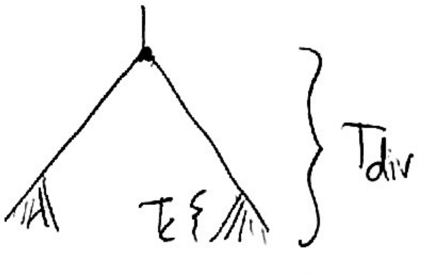
i.e., 90% of all amino acid changes in bactria are sufficiently strongly deleterious that they are strongly selected against.

$\Rightarrow$ in practice, people often coarse grain sites & look for constraint on even smaller portions of genome.

$\Rightarrow$ reason is that strongly constrained $\approx$ important for organism (interesting biology)

$\Rightarrow$ can also do same thing w/ substitutions between 2 species (d_N/d_S)

$\Rightarrow$ more time for mutations to occur, better signal



when do we expect QLE to work?

$\Rightarrow$ think about collection of 2 locus mini problems.

$$\rho_{eff} = r\Delta\ell \rightarrow \text{distance between sites}$$

↳ recombination rate per site.

if guestimate N from π_{syn}: $N = \frac{\pi_{syn}}{2\mu}$

$\Rightarrow$ $D_{max} \sim \frac{1}{N\rho} \sim \frac{\mu}{r(\pi\Delta\ell)} \approx \frac{\mu}{r}$ if $\Delta\ell \approx \frac{1}{\pi}$

(neighboring SNPs)

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

(weird right?) → so neutral muts right on boundary of ok.

$\Rightarrow$ selected mutations needs: $\frac{s}{r} \ll 1$

$\Rightarrow$ if $r \sim \mu \sim 10^{-8}-10^{-10}$ $\Rightarrow$ need $s \ll 10^{-8}-10^{-10}$

$\Rightarrow$ bad approximation most of time. need some other method to predict evolution in this case.