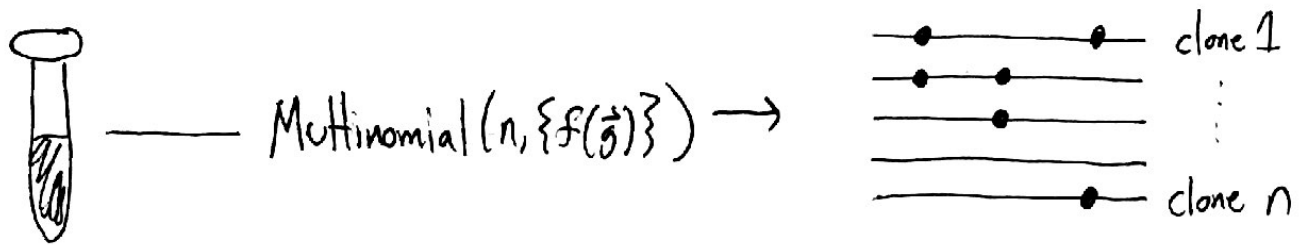


Introduction to multi-locus models of evolution

①

Now that we have introduced methods for measuring genomes (or amplicons) from different individuals in a population:



We need a corresponding set of models to predict the genotype frequencies, $f(\vec{g})$, that arise through evolution.

For genome of length $L \Rightarrow 2^L$ possible genotypes (^{WT+mutant} @ each site)

es. $L=1, g=0,1 \Rightarrow f(0)=1-f, f(1)=f$

$L=2, \vec{g} = \underbrace{(0,0)}_{\text{WT}}, \underbrace{(1,0), (0,1)}_{\text{single mutants}}, \underbrace{(1,1)}_{\text{double mutant}}$

$L=3: \vec{g} = (0,0,0), (1,0,0), \dots, (1,1,0), \dots, (1,1,1)$

\vdots
etc

\Rightarrow can we generalize our serial dilution (& diffusion models) to account for this case?

① Genetic drift: first consider case w/ no growth rate diffs
 & no additional mutations.

②

After one day of growth: $f(\vec{g}) \Rightarrow \frac{f(\vec{g}) e^{r\Delta t}}{\sum_{\vec{g}'} f(\vec{g}') e^{r\Delta t}} = \frac{f(\vec{g})}{\sum_{\vec{g}'} f(\vec{g}')} = f(\vec{g}) \checkmark$

After Poisson dilution:

$$n(\vec{g}, t+\Delta t) \sim \text{Poisson}(N_0 f(\vec{g})) \Rightarrow f(\vec{g}, t+\Delta t) = \frac{n(\vec{g}, t+\Delta t)}{\sum_{\vec{g}'} n(\vec{g}', t+\Delta t)}$$

\Rightarrow Repeating Taylor expansions from before...

$$\llcorner \Rightarrow f(\vec{g}, t+\delta t) = f(\vec{g}) + \sqrt{\frac{f(\vec{g})\delta t}{N_e}} z_{\vec{g}} - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')\delta t}{N_e}} z_{\vec{g}'}$$

where $z_{\vec{g}}$ are normal random variables w/

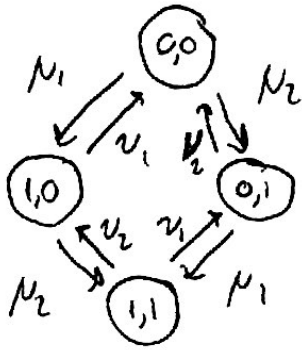
$$\langle z_{\vec{g}} \rangle \equiv 0 \text{ and } \langle z_{\vec{g}} z_{\vec{g}'} \rangle \equiv \delta_{\vec{g}, \vec{g}'}$$

why funny form w/ correlations? Ensures that $f(\vec{g}, t)$ stays normalized:

$$\sum_{\vec{g}} f(\vec{g}, t+\delta t) = \sum_{\vec{g}} f(\vec{g}, t) + \sum_{\vec{g}} \sqrt{\frac{f(\vec{g})\delta t}{N_e}} z_{\vec{g}} - \left[\sum_{\vec{g}} f(\vec{g}) \right] \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')\delta t}{N_e}} z_{\vec{g}'} = 1 \checkmark$$

3

2 Mutations: easiest to start w/ $L=2$ case:



\Rightarrow important point is that you can only move by ~ 1 step @ a time

e.g. for $\vec{g} = (1,0)$, after one dilution cycle:

$$n(1,0,t+\Delta t) = \text{Poisson} \left[N_0 \left(f(1,0,t) + \Delta t \left[\underbrace{\mu_1 f(0,0,t)}_{\text{muts from WT}} + \underbrace{\nu_1 f(1,1,t)}_{\text{back muts}} - \underbrace{\mu_2 f(1,0,t) - \nu_2 f(1,0,t)}_{\text{mutations out of genotype to other } \vec{g}'} \right] \right) \right]$$

$$\Rightarrow \text{continuum limit: } \delta f(1,0) = \left[\mu_1 f(0,0) + \nu_1 f(1,1) - \mu_2 f(1,0) - \nu_2 f(1,0) \right] \delta t$$

(again, linear in ~~mutator~~ genotype freqs)

+ noise.

\Rightarrow ~~extension~~ extension to other genotypes (a larger L) is similar, just slightly tricky to write in compact way.

one way to write it is:

4

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{\text{mut}} \equiv \sum_{\substack{\vec{g}' \text{ s.t.} \\ |\vec{g}-\vec{g}'|=1}} \sum_{e=1}^L \left(\mu_e f(\vec{g}') g_e (1-g_e) + \nu_e f(\vec{g}') (1-g_e) g_e \right)$$

↙ incoming mutations

$$= \sum_{e=1}^L \left(\mu_e f(\vec{g}) (1-g_e) + \nu_e f(\vec{g}) g_e \right) \equiv M(\{f(\vec{g})\})$$

↑ outgoing mutations

↑ linear operator.

Note: mutations are normalized so that $\sum_{\vec{g}} \left(\frac{\delta f(\vec{g})}{\delta t}\right)_{\text{mut}} = 0$
 (inflow cancels outflow in whole pop'n)

③ Selection (growth rate differences):

If growth rate of each genotype is $\equiv r + X(\vec{g})$,
 then after one cycle of growth:

$$f(\vec{g}) \rightarrow \frac{f(\vec{g}) e^{(r+X(\vec{g}))\Delta t}}{\sum_{\vec{g}'} f(\vec{g}') e^{(r+X(\vec{g}'))\Delta t}} = \frac{f(\vec{g}) e^{X(\vec{g})\Delta t}}{\sum_{\vec{g}'} f(\vec{g}') e^{X(\vec{g}')\Delta t}}$$

In limit that $X(\vec{g})\Delta t \ll 1$ (continuum limit), this becomes

(5)


$$f(\vec{g}, t + \delta t) = f(\vec{g}, t) + [X(\vec{g}) - \bar{X}(t)] f(\vec{g}, t) \delta t$$

↳ population mean fitness:

$$\bar{X}(t) \equiv \sum_{\vec{g}} X(\vec{g}) f(\vec{g}, t)$$

[note: not an ensemble average, i.e.
 $\langle \bar{X}(t) f(\vec{g}, t) \rangle \neq \langle \bar{X}(t) \rangle \langle f(\vec{g}, t) \rangle$]

⇒ intuitive interpretation: genotypes w/ above average fitness are amplified
those w/ below average fitness are eliminated

⇒ again, selection term is normalized so that 

$$\sum_{\vec{g}} f(\vec{g}, t + \delta t) = \sum_{\vec{g}} f(\vec{g}, t) + \sum_{\vec{g}} X(\vec{g}) f(\vec{g}, t) \delta t - \bar{X} \sum_{\vec{g}} f(\vec{g}, t) \delta t = 1$$

⇒ so far similar to single locus case, but w/ higher dimensions.

⇒ $L \geq 2$ introduces 2 new biological features.

4 "Epistasis": properties of $\vec{g} \rightarrow X(\vec{g})$ map ("fitness landscape")

Easiest to motivate w/ $L=2$ case (e.g. 2 gene deletions)

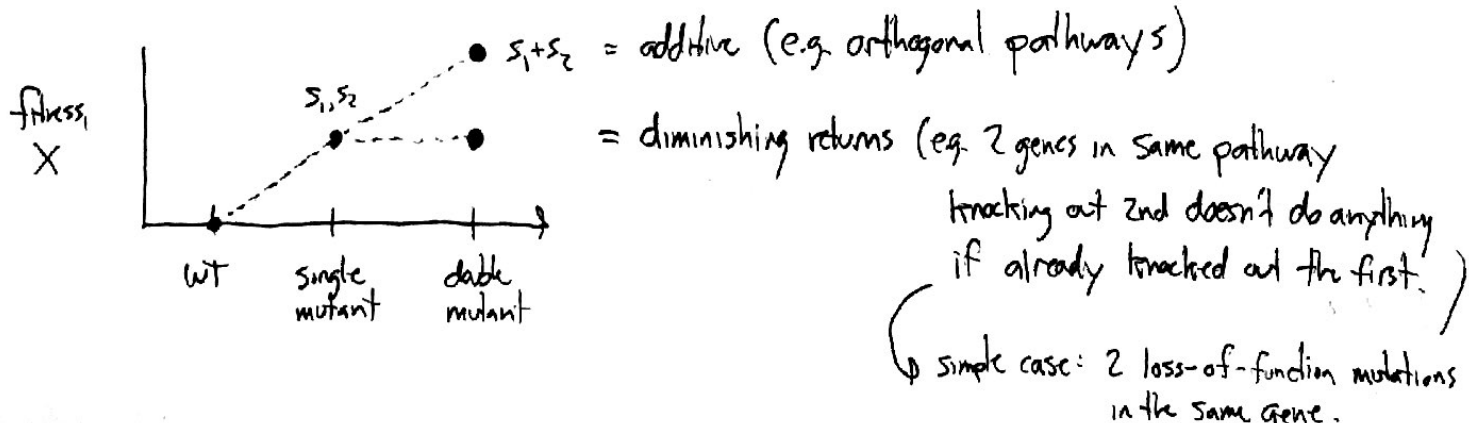
$X(0,0) \equiv 0$ (convention)

$X(1,0) \equiv S_1$
 $X(0,1) \equiv S_2$ } could measure, e.g. in gene deletion screen (homework)

$X(1,1) = ? \equiv \underbrace{S_1 + S_2}_{\text{"additive" part}} + \epsilon$ ← "epistasis" i.e. how much deviation from additivity

Lots of vocab to describe epistasis depending on relative values of ϵ, S_1, S_2
e.g. $\epsilon > 0$ = "positive epistasis", $\epsilon < 0$ = "negative epistasis"
but also "sign epistasis", "diminishing returns epistasis", "robustness", etc. etc.

often easiest to just draw picture:

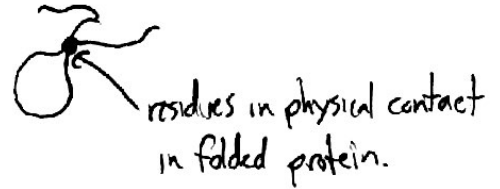


often people are interested in scenario like



"fitness valley crossing"

e.g. thought to be relevant for initiation of cancer or changing contact residues in proteins:



gets even more complicated for $L > 2$:

$$X(\vec{g}) \equiv \sum_{e=1}^L s_e g_e + \epsilon(\vec{g})$$

additive part
(coupon collecting)
model of evolution.

epistasis \approx $\underbrace{\sum_{e=1}^L \sum_{e'=1}^L \epsilon_{ee'} g_e g_{e'}}_{\text{"pairwise epistasis"}} + \underbrace{\sum \sum \sum \epsilon_{eee} g_e g_{e'} g_{e''}}_{\text{"higher order epistasis"}} + \dots$

can sometimes measure pairwise epistasis, e.g. in double deletion screens.

\Rightarrow geneticists use these to detect when 2 genes might be in same pathway (by searching for genes w/ $|\epsilon| \geq 0$).

\Rightarrow but in general, epistasis is hard to measure, & not clear that pairwise is really sufficient for biology. E.g. 3 loss-of-function mutations in same gene = pairwise description cannot handle.

this is an active area of research (both theory & exp.)

8

⇒ In practice, people typically work w/ additive model (for $L \gg 1$)
or draw pictures (for $L \sim \mathcal{O}(1)$)

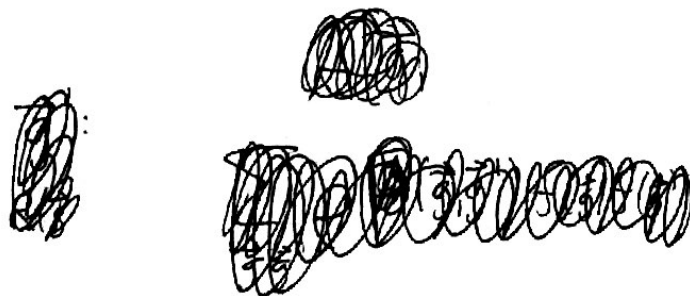
↓
(Sometimes coarse-graining sites into "modules" e.g. genes, pathways)

⇒ additive model may not be as bad as it looks @ first glance:
saw yesterday that genomes in same pop'n can only get so far apart, so mutations only need to be additive in some local neighborhood

⇒ epistasis then relevant only on longer evolutionary timescales.
(rather than pop-gen timescales)

⇒ in either case, need theory to tell us when evolutionary dynamics look different from what additive model can produce

⇒ we will see this is still pretty complicated on its own!



other new piece of biology that enters for $L \geq 2$ is ~~recomb~~

9

⑤ Recombination: (exchange of genetic material between diff individuals in the population)

Many different mechanisms (details often complicated & not fully understood in all cases)



many share same basic flavor:

① focal individual f is ~~is~~ chosen to undergo recombination

(probability ρ per individual per generation) → e.g. mating phage uptake of extra cellular DNA.

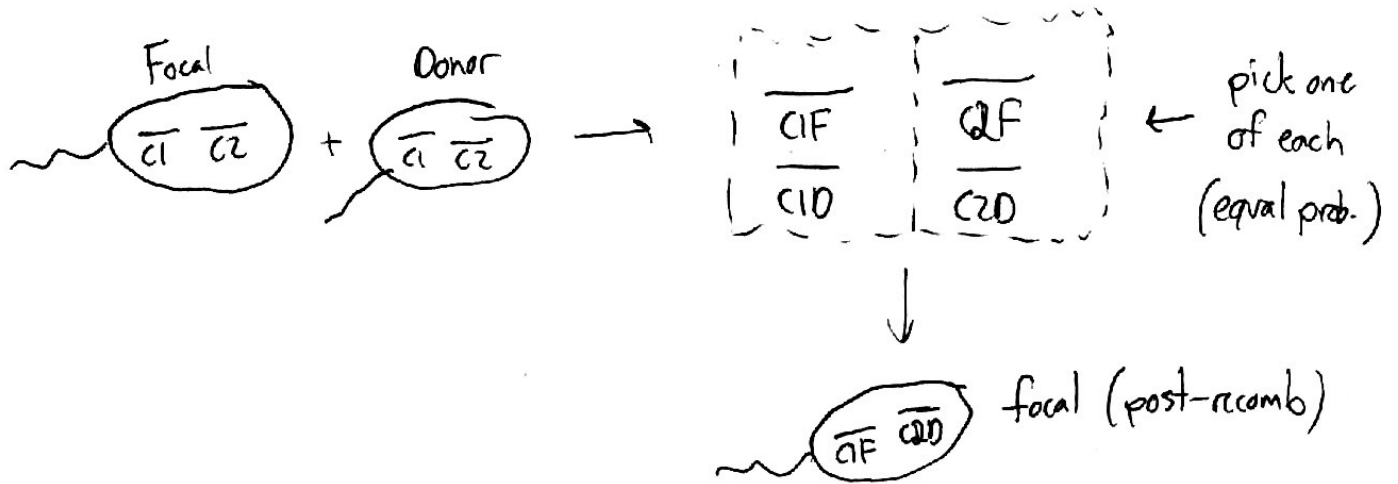
② donor individual d is chosen to donate some portion of genome

(probability $\sim 1/N \Rightarrow$ prob $f(\vec{g})$ for any individual of that genotype)

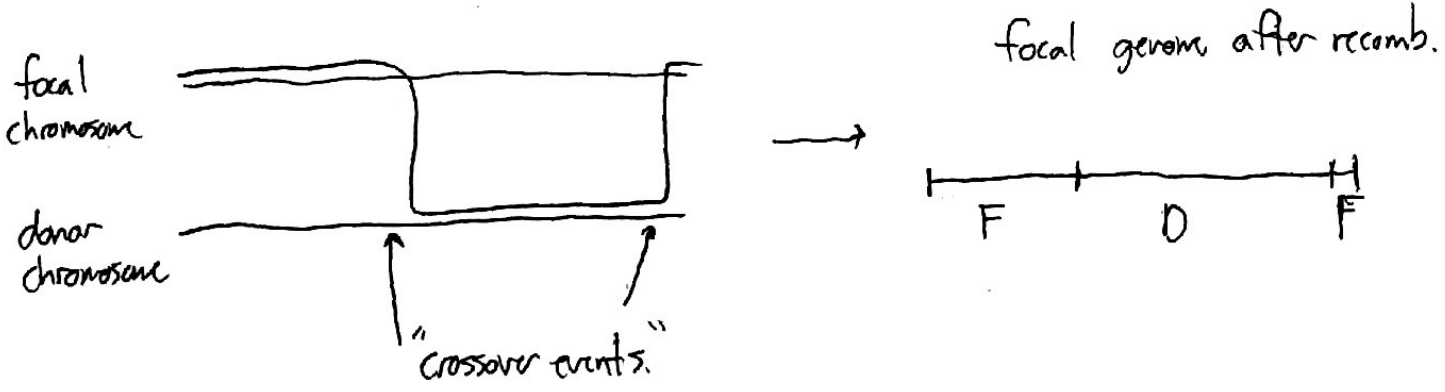
③ Some piece of donor's DNA is integrated into focal genome

\Rightarrow different mechanisms enter here.

(a) Reassortment (e.g. different chromosomes, e.g. humans, yeast, influenza)



(b) Crossover recombination (w/in chromosomes in humans)

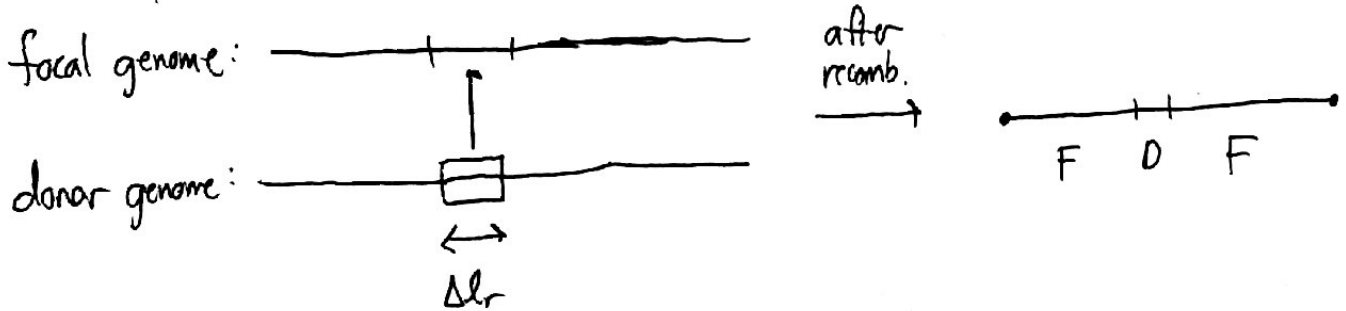


⇒ often modeled w/ ~ 1 crossover per recombination event, ~~prob~~ w/ location chosen uniformly across chromosome.
 (in practice, "hot spots" or "cold spots" → active area of research!)

⇒ note: this means that crossover rates can vary over many orders of magnitude in same genome (e.g. 2 sites @ opposite ends of chromosome (prob=1) vs 2 sites next to each other (prob= $\frac{1}{L} \sim 10^{-8}$!))

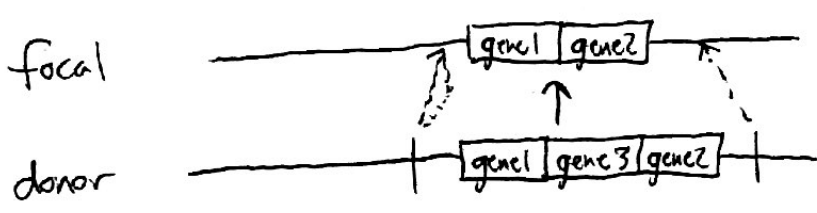
③ "Horizontal gene transfer" / "gene conversion"

↳ note: lingo is a little controversial, basic idea is simple though:



⇒ main difference from crossover recomb. @ this level is that $\Delta L_r \ll L$ in this case, while for crossover, $\Delta L_r \sim \mathcal{O}(L)$ (\sim crossover / recomb.)

⇒ also a mechanism for gaining & losing new genes ("accessory genome")



(recombination often mediated by homology @ ends of fragment - similar to PCR primers) cares less about what is in middle.

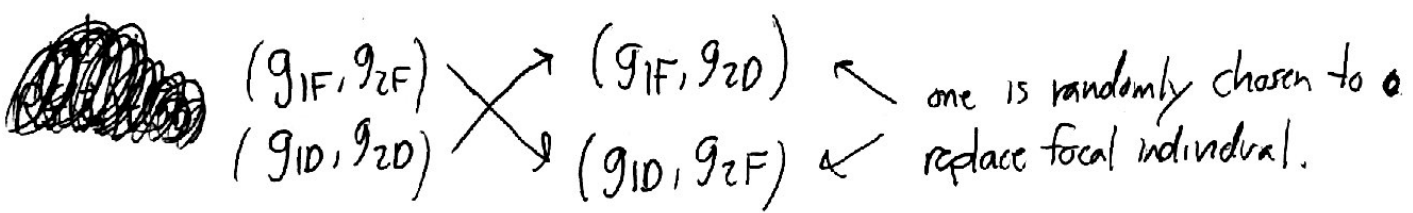
this is also active area of research, but won't consider it too much in this class (will focus on simple picture of "core" genome)

⇒ simplest model is $\Delta L_r = \text{const}$, location = uniform across genome

So far, have described these mechanisms @ level required to simulate them in individual based simulations.

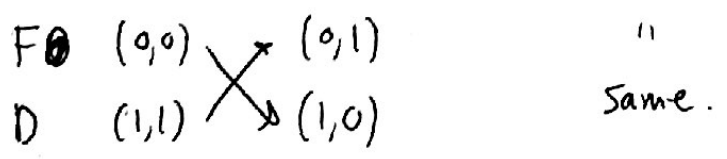
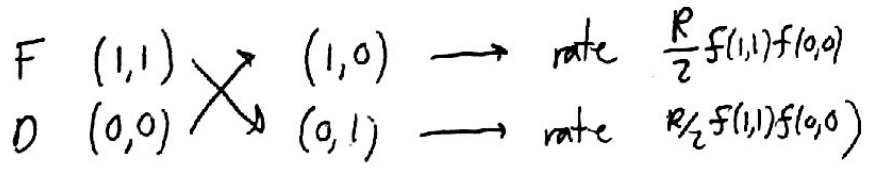
⇒ to pass to continuum ~~limit~~ ^{limit}, helpful to think about $L=2$ case

⇒ In this case, all mechanisms have same net effect w/ rate R , individual w/ genome \vec{g} (focal) undergoes recomb w/ donor individual (\vec{g}') & swaps sites:

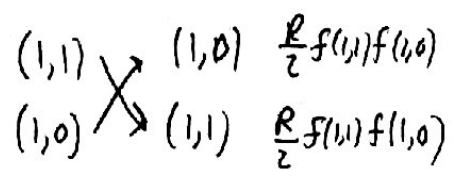


⇒ total outflow from recombination - $Rf(\vec{g})$

⇒ how many ways to create $f(\vec{g})$ from recomb? $4 \times 4 = 16$ possible genotype pairs:



same.



can go through & write out all 16 combinations,
and all 32 possible outputs, add them up, and find:

$$\left(\frac{\delta f(1,1)}{\delta t} \right)_{rec} = R f(0,1) f(1,0) - R f(1,1) f(0,0)$$

$$\left(\frac{\delta f(0,0)}{\delta t} \right)_{rec} = R f(0,1) f(1,0) - R f(0,0) f(1,1)$$

$$\left(\frac{\delta f(1,0)}{\delta t} \right)_{rec} = R f(0,0) f(1,1) - R f(1,0) f(0,1)$$

$$\left(\frac{\delta f(1,0)}{\delta t} \right)_{rec} = R f(0,0) f(1,1) - R f(0,1) f(1,0)$$

=> again, normalized so that $\sum_{\vec{g}} \delta f(\vec{g} + \delta t) = 0$

=> even harder to write down explicitly for $L > 2$, but will have general form:

$$\left(\frac{\delta f(\vec{g})}{\delta t} \right)_{rec} = \rho \sum_{\vec{g}_F, \vec{g}_D} P_r(\vec{g}_F, \vec{g}_D) \underbrace{f(\vec{g}_F) f(\vec{g}_D)}_{\text{nonlinear (hard!)}} - \rho f(\vec{g})$$

& can create genotypes very far from $f(\vec{g})$!!

Putting everything together, general multilocus model looks like:

$$\frac{df(\vec{g})}{dt} = \underbrace{\left[X(\vec{g}) - \bar{X}(t) \right] f(\vec{g})}_{\text{selection}} + \underbrace{\sum_{\vec{g}'} \left[M_{\vec{g} \rightarrow \vec{g}'} f(\vec{g}') - M_{\vec{g}' \rightarrow \vec{g}} f(\vec{g}) \right]}_{\text{mutation}}$$

$$+ e \left[\underbrace{\sum_{\vec{g}_F, \vec{g}_D} R_{\vec{g}_F \vec{g}_D} f(\vec{g}_F) f(\vec{g}_D) - f(\vec{g})}_{\text{recombination}} \right]$$

$$+ \left[\underbrace{\sqrt{\frac{f(\vec{g})}{N_e}} \eta(\vec{g}, t) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N_e}} \eta(\vec{g}', t)}_{\text{genetic drift}} \right]$$

⇒ ~~no exact solution for stationary dist'n, ~~equation~~ fixation prob. in general case. (even for L=2!) (analogous to multi-particle schrodinger eq. in physics)~~

what do we do instead?!?