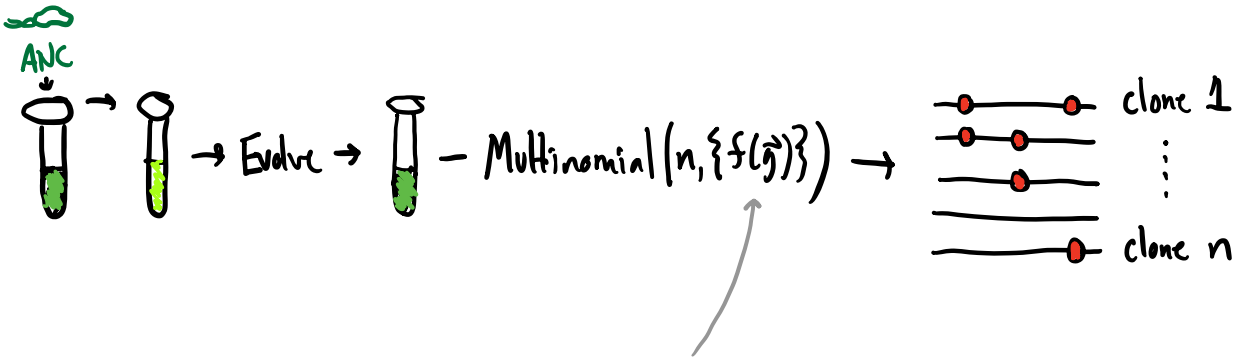
Chapter 9

Multi-locus models of evolution

Next Steps: now that we have methods for measuring genomes (or amplicons)



$\Rightarrow$ need models to predict $f(\vec{g})$'s that arise during evolution

For genome of length $L \Rightarrow 2^L$ possible genotypes

e.g. $L=1$: $g=0,1 \Rightarrow$ $f(1) \equiv f$, $f(0) = 1-f$ (mutant, WT)

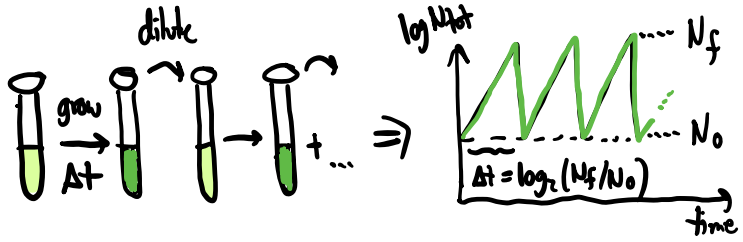
$L=2$: $\vec{g} =$ $(0,0)$, $(1,0)$, $(0,1)$, $(1,1)$ (WT, single mutants, double mutant)

$L=3$: $\vec{g} =$ $(0,0,0)$, $(1,0,0)$, ..., $(1,1,0)$, ..., $(1,1,1)$ (single, double, triple mutant)

⋮

etc.

Can we generalize our serial dilution (+ diffusion) models?



$$\frac{df(\vec{g})}{dt} = ???$$

(1) Genetic drift: first assume no growth rate differences... (+ no mutations)

⇒ After 1 day of growth (before dilution):

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

(i.e. no change in freqs ✓)

⇒ After dilution step:

(i) $n(\vec{g}, t+\Delta t) \sim \text{Poisson}(\bar{N}_0 \cdot f(\vec{g}))$ (sampling)

(ii) $f(\vec{g}, t+\Delta t) = \frac{n(\vec{g}, t+\Delta t)}{\sum_{\vec{g}'} n(\vec{g}', t+\Delta t)}$ (re-normalize)

$\Rightarrow$ if repeat our Taylor expansions from Ch 4 ($\bar{N}_0$ large):

$$n(\vec{g}, t+\Delta t) \sim \bar{N}_0 f(\vec{g}, t) + \sqrt{\bar{N}_0 f(\vec{g}, t)} \cdot Z_{\vec{g}}$$

where $Z_{\vec{g}} \overset{iid}{\sim} \text{Gaussian}(0,1)$

$\Rightarrow$ Taylor expand $f(\vec{g}) = \frac{n(\vec{g})}{\sum_{\vec{g}'} n(\vec{g}')}$:

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

depends on $Z_{\vec{g}}$ @ other $\vec{g}'$!

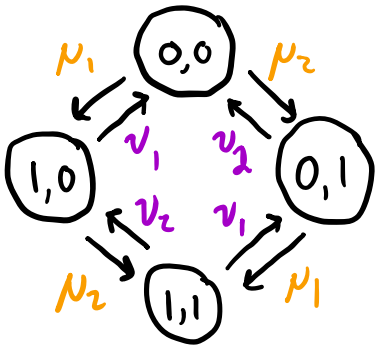
$\Rightarrow$ correlations between $\delta f(\vec{g})$ & $\delta f(\vec{g}')$ $\Rightarrow$ keeps $f(\vec{g}, t)$ normalized!

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

$= 1$ $\Rightarrow$ stays normalized @ later times

② Mutations:

⇒ easiest to start w/ $L=2$:



⇒ key feature: can only move by ~1 step @ a time

⇒ generalizing results from $L=1$ case, after 1 dilution:

$$n(1,0,t+\Delta t) \sim \text{Poisson}\Big(N_0 f(1,0,t) + N_0 \Delta t\big[\overbrace{\mu_1 f(0,0,t) + \nu_2 f(1,1,t)}^{\text{mutations into genotype}}\big] - N_0 \Delta t\big[\underbrace{\mu_2 f(1,0,t) + \nu_1 f(1,0,t)}_{\text{mutations out of genotype}}\big]\Big)$$

⇒ continuum limit (i.e. Taylor expansions):

$$\delta f(1,0)_{mut} = \Big[\mu_1 f(0,0) + \nu_2 f(1,1) - \mu_2 f(1,0) - \nu_1 f(1,0)\Big]\delta t$$

(+ noise from drift)

linear in * genotype freqs

⇒ larger L's are similar, but more work to write out...

⇒ one way is:

$$\left[\frac{\delta f(\vec{g})}{\delta t}\right]_{mut} \equiv \sum_{\substack{\text{nearest} \\ \text{neighbors} \\ \vec{g}'}} \sum_{\ell=1}^{L} \Big[\overbrace{\mu_\ell f(\vec{g}')\, g_\ell (1-g'_\ell) + \nu_\ell f(\vec{g}')(1-g_\ell)\, g'_\ell}^{\text{mutations } \underline{\text{into}} \text{ genotype}} \Big]$$

$$- \sum_{\ell=1}^{L} \Big[\overbrace{\mu_\ell f(\vec{g})(1-g_\ell) + \nu_\ell f(\vec{g})\, g_\ell}^{\text{mutations } \underline{\text{out of}} \text{ genotype}} \Big]$$

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{mut} \equiv \sum_{\vec{g}'} \Big[\underbrace{M(\vec{g}' \to \vec{g})}_{2^L \times 2^L \text{ matrix of mut'n rates}} f(\vec{g}') - \underbrace{M(\vec{g} \to \vec{g}')}_{2^L \times 2^L \text{ matrix of mut'n rates}} f(\vec{g}) \Big]$$

Note: mutation matrix normalized s.t. $\sum_{\vec{g}} \left(\frac{\delta f(\vec{g})}{\delta t}\right)_{mut} = 0$

⇒ ensures that $\sum_{\vec{g}} f(\vec{g}, t+\delta t) = \sum_{\vec{g}} f(\vec{g}, t) \,(\to 1) + \sum_{\vec{g}} \delta f_{mut}(\vec{g}) \,(\to 0) = 1$

③ Selection (growth rate differences)

If growth rate of genotype $\vec{g}$ is $\equiv r + X(\vec{g})$

$\Rightarrow$ then after 1 cycle of growth:

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

$\Rightarrow$ if $X(\vec{g})\Delta t \ll 1$ (continuum limit) $\Rightarrow$ Taylor expand:

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

where $\bar{X}(t) \equiv \sum_{\vec{g}} X(\vec{g}) f(\vec{g},t)$ (population mean fitness)

* note: not an ensemble avg! $\langle \bar{X}(t) f(\vec{g},t) \rangle \neq \langle \bar{X}(t) \rangle \langle f(\vec{g},t) \rangle$

$\Rightarrow$ stays normalized: $\sum_{\vec{g}} f(\vec{g},t+\delta t) = \underbrace{\sum f(\vec{g},t)}_{1} + \sum_{\vec{g}} X(\vec{g}) f(\vec{g},t) - \underbrace{\sum_{\vec{g}} f(\vec{g}) \sum_{\vec{g}'} X(\vec{g}') f(\vec{g}')}_{1} = 1$

$\Rightarrow$ 2 new biological features that enter for $L \geq 2$:

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

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

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

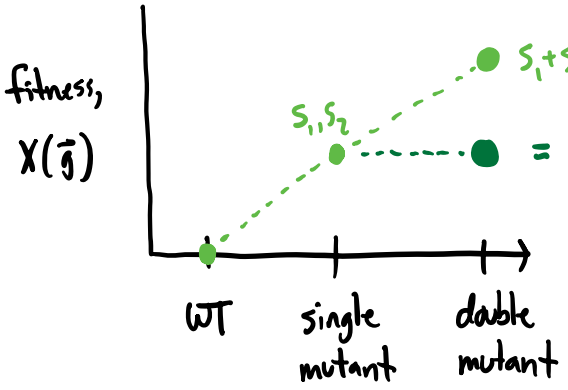
$X(1,0) \equiv s_1$
$X(0,1) \equiv s_2$
} could measure, e.g. gene deletion screen (HW2)

$X(1,1) \equiv ? \equiv s_1 + s_2 + \epsilon$

"additive part" "epistasis"
(how much deviation from additivity)

e.g. "$\epsilon > 0$" $\Rightarrow$ "positive epistasis"
"$\epsilon < 0$" $\Rightarrow$ "negative epistasis" $\Rightarrow$ "sign epistasis" etc. etc.

Often easiest to express w/ picture:

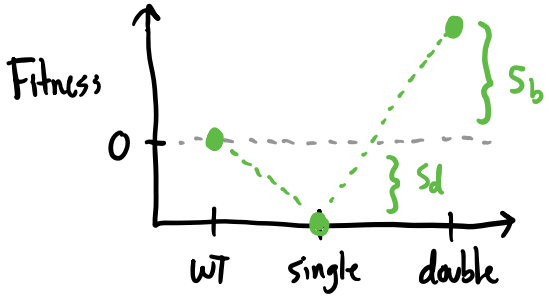


s_1+s_2 = "additive", e.g. orthogonal pathways in a cell

= "diminishing returns", e.g. 2 genes in same pathway

2 stop codons in same gene...

⇒ people often interested in scenarios like:



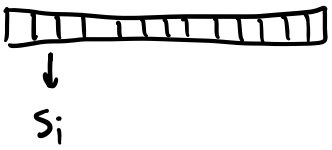
"fitness valley crossing"

e.g. initiation of cancer

contact btw AA's in proteins

↳ (Problem 6 of HW 3)

⇒ gets even more complicated for $L>2$:



$$X(\vec{g}) \equiv \underbrace{\sum_{\ell=1}^{L} s_\ell g_\ell}_{\text{additive part ("coupon collecting")}} + \underbrace{\epsilon(\vec{g})}_{\text{epistatic part.}}$$

⇒ can write as Taylor expansion around WT:

$$\epsilon(\vec{g}) = \underbrace{\sum_{\ell=1}^{L}\sum_{\ell'=1}^{L} \epsilon_{\ell\ell'} g_\ell g_{\ell'}}_{\text{"pairwise epistasis"}} + \underbrace{\sum_{\ell}\sum_{\ell'}\sum_{\ell''} \epsilon_{\ell\ell'\ell''} g_\ell g_{\ell'} g_{\ell''} + \dots}_{\text{"higher order epistasis"}}$$

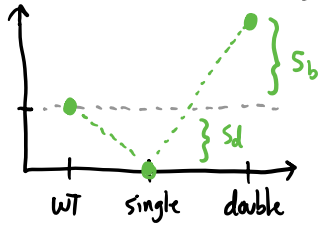
⇒ hard to parameterize in general (active area of research!)

⇒ in practice, people often use:

Additive model ($L \gg 1$)

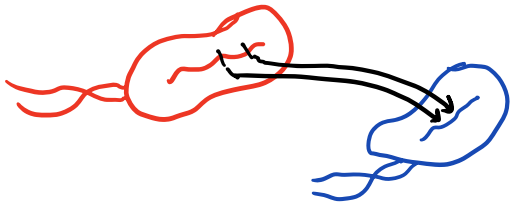
$$X(\vec{g}) \approx \sum_{\ell=1}^{L} s_\ell g_\ell$$

Pictures ($L \sim \mathcal{O}(1)$)



$\Rightarrow$ other new bit of biology for $L \geq 2$:

(5) Recombination (exchange of genetic material between different individuals)



Many different mechanisms!

$\Rightarrow$ but many share same basic behavior:

(1) Focal individual (f) is chosen to undergo recombination

$\Rightarrow$ w/ probability ρ per individual per-gen → e.g. mating
viruses/phage
uptake of DNA
cellular DNA, d

(2) Donor individual (d) is chosen to donate portion of genome

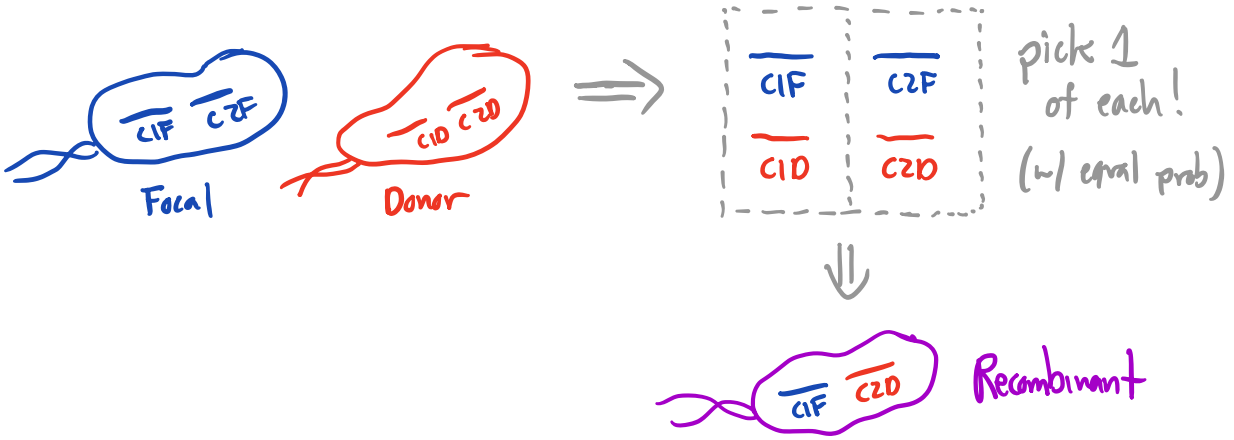
$\Rightarrow$ probability $\sim 1/N$ $\Rightarrow$ $f(\vec{g})$ for any individual of that genotype.

(3) some piece of donor's DNA is integrated into focal genome

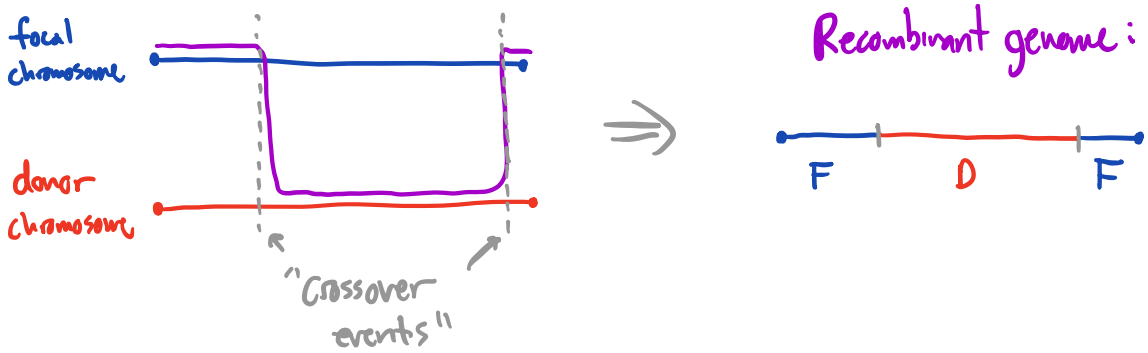
⇒ producing "recombinant"

⇒ different mechanisms enter @ this step:

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

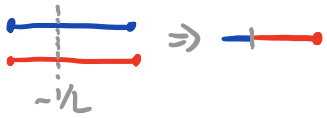


(b) **Crossover Recombination** (e.g. w/in chromosomes in humans)



⇒ often modeled w/ ~1 crossover per recombination event

w/ location chosen uniformly across chromosome



⇒ in practice, "hot spots" & "cold spots" ⇒ "recombination map"

⇒ effective recombination rates vary over many orders of magnitude for different sites in same genome!

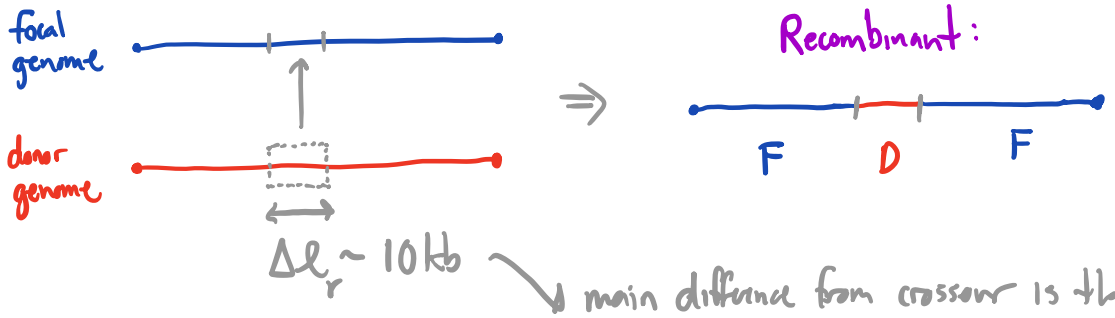
⇒ e.g. in humans ⇒ $L_{chrom} \sim 10^8$ bp

⇒ P(recomb) ~ 100% if 2 ends of same chrom

⇒ P(recomb) ~ 10^{-8} if neighboring base pairs

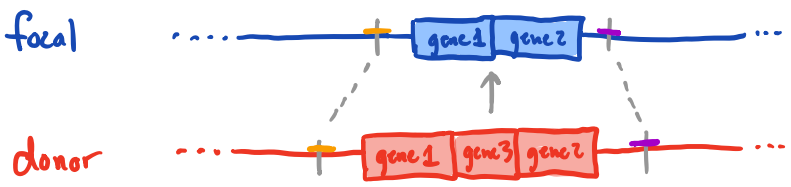
(c) "Horizontal gene transfer" / "gene conversion"

⇒ lingo is a little controversial, but basic idea pretty simple:



main difference from crossover is that $\Delta \ell_r \not\propto L$, as opposed to $\Delta \ell_r \sim \mathcal{O}(L)$

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

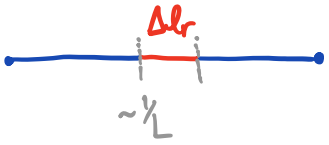


often mediated by homology (=)
similar to PCR...

⇒ active area of research!

⇒ but in this class, will mostly focus on "core genome"

⇒ simplest HGT model:



Δl_r = const, location ~ uniform

So far: individual-based picture...

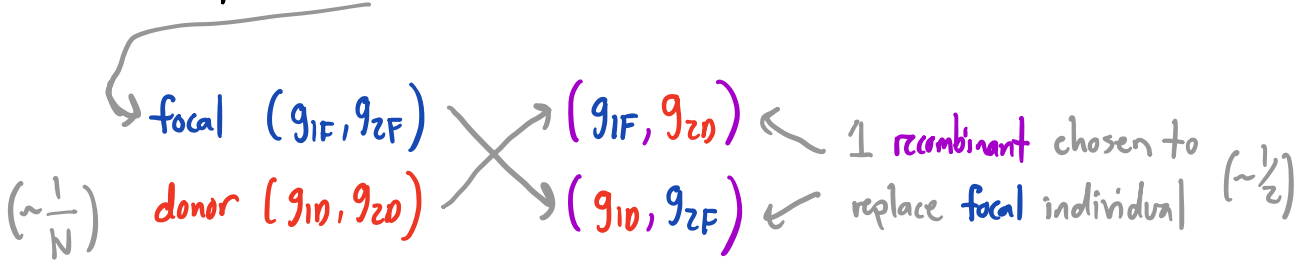
⇒ can we translate to continuum limit?

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{rec} = ???$$

⇒ easiest to start w/ L = 2 case ⇒ $\vec{g} = (g_1, g_2)$

⇒ all mechanisms have same net effect:

$\Rightarrow$ w/ rate R [function of $\rho, L, \Delta h_r, \ldots$ etc.]



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

$\Rightarrow$ total inflow? $2^2 \times 2^2 = 16$ possible focal/donor combos

case 1 (of 16):

F $(1,1)$ $\rightarrow$ $(1,0)$ $\Rightarrow$ rate $Rf(1,1)f(0,0)\cdot\frac{1}{2}$

D $(0,0)$ $\rightarrow$ $(0,1)$ $\Rightarrow$ $Rf(1,1)f(0,0)\cdot\frac{1}{2}$

case 2 (of 16): F $(0,0)$ $\rightarrow$ $(0,1)$; D $(1,1)$ $\rightarrow$ $(1,0)$ — same!

Case 3 (of 16): $(1,1) \to (1,0) \Rightarrow Rf(1,1)f(1,0)\frac{1}{2}$

$(1,0) \to (1,1) \Rightarrow Rf(1,1)f(1,0)\frac{1}{2}$

$\Rightarrow$ after tabulating all 16 combinations (all 32 recombinants) can add them up to obtain:

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

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

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

$$\left(\frac{\delta f(0,1)}{\delta t}\right)_{rec} = \text{same.}$$

$\Rightarrow$ normalized so that $\sum_{\vec{g}} \delta f(\vec{g})_{rec} = 0$ ✓

⇒ harder to write down explicitly for $L > 2$ ....

but will have general form:

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{rec} = \rho \sum_{\vec{g}_F, \vec{g}_O} T(\vec{g}_F, \vec{g}_O \to \vec{g}) f(\vec{g}_F) f(\vec{g}_O) - \rho f(\vec{g})$$

incoming recombinants; outgoing recombinants.

"recombination kernel" → "tensor"

nonlinear!

⇒ unlike mutation, can create genotypes far from $\vec{g}$!

Putting everything together, general multilocus model looks like:

$$\frac{df(\vec{g})}{dt} = \underbrace{\left[X(\vec{g}) - \bar{X}(t)\right] f(\vec{g})}_{\text{selection (nonlinear)}} + \underbrace{\sum_{\vec{g}'} M(\vec{g}' \to \vec{g}) f(\vec{g}') - M(\vec{g} \to \vec{g}') f(\vec{g})}_{\text{mutation (linear, "local")}}$$

$$+ \underbrace{\rho \sum_{\vec{g}_F, \vec{g}_D} T(\vec{g}_F, \vec{g}_D \to \vec{g}) f(\vec{g}) - \rho f(\vec{g})}_{\text{recombination (nonlinear, non-local)}}$$

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

Problem: No exact solution for stationary dist'n, p_{fix}, etc.
— even for $L = 2$!

⇒ What do we do instead?!? ⇒ asymptotic approx's

Question: Given parameters ("knobs") $L, N, X(\vec{g}), M, \rho, T$

⇒ what are some limits where we might understand understand this SDE?

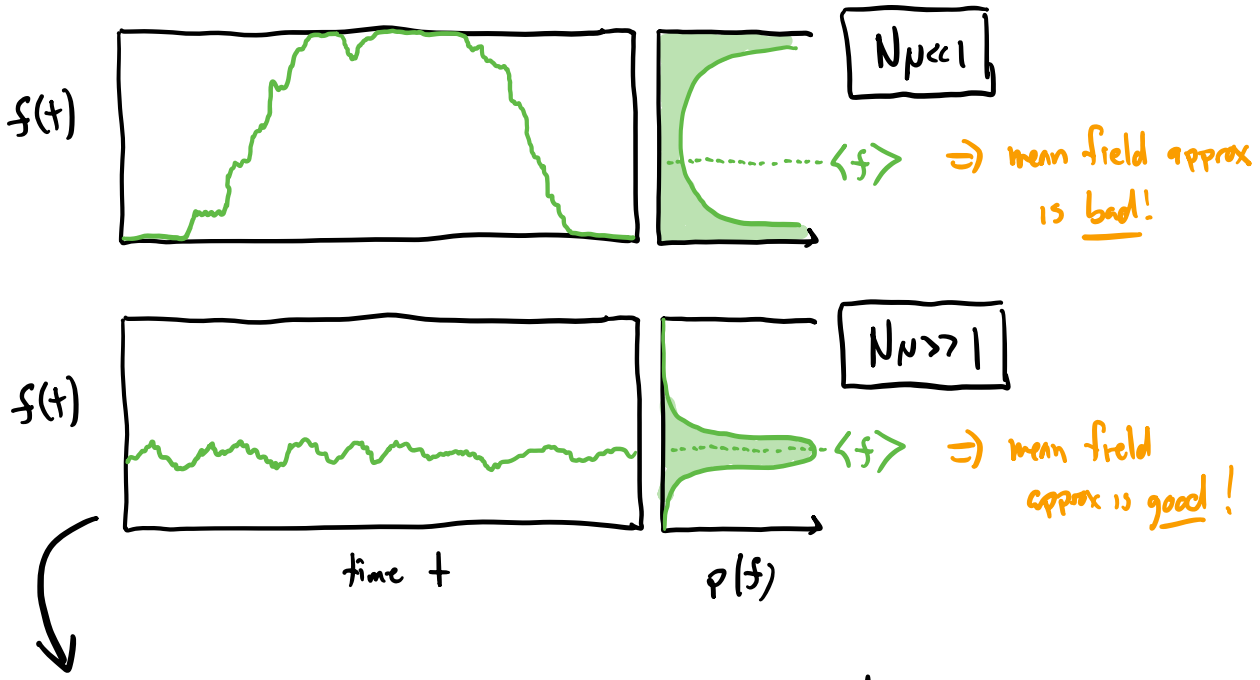
$$\frac{df(\vec{g})}{dt} = \sim(x - \bar{x}) + \sim L\times\mu + \sim\rho + \sim\frac{z}{\sqrt{N}}$$

① Obvious answer: $L=1 \Rightarrow$ cheating! *

② in physics, might be primed to take $N \to \infty$ limit ...
("mean field approx") since @ least noise goes away ...

$\Rightarrow$ is this a good approx here?

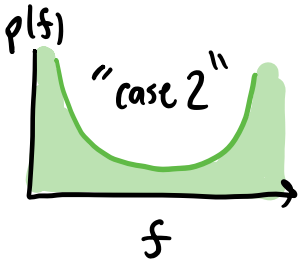
$\Rightarrow$ Recall for $L=1$ case, 2 different regimes when $t \to \infty$:



key feature: large # of individuals in both genotypes @ same time
$\Rightarrow$ so fluctuations are small.

$\Rightarrow$ e.g. for $L=2$, might be ok $\Rightarrow$ but for $L \gg 1 \Rightarrow 2^L \gg N$!
e.g. $L \sim 1000$ bp $\Rightarrow 2^L \sim 10^{300}$!

⇒ large L will always look like
(@ least in some dimensions)



⇒ noise always relevant!

Need to look for other approximations of SDE...

$$\frac{ds(\vec{g})}{dt} = \sim(x-\bar{x}) + \sim L\times\mu + \sim\varrho + \sim\frac{\mathbb{Z}}{\sqrt{N}}$$

Let's revisit our first idea (L=1)

⇒ even if L≫1, if behavior "looks like" L=1 case,

⇒ can use what we already know...

(3) Successive mutations regime (i.e. treat mutation as small correction)

⇒ what if mutation rates are low enough that only 1 or 2 genotypes are present @ a time?