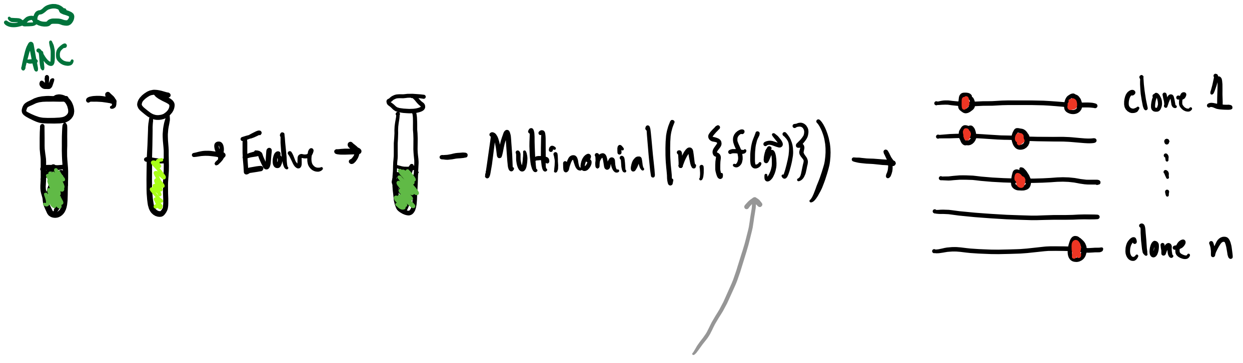
Chapter 9

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. $L=1$: $g=0,1$ $\Rightarrow$ $\underbrace{f(1) \equiv f}_{\text{mutant}}$, $\underbrace{f(0)=1-f}_{\text{WT}}$

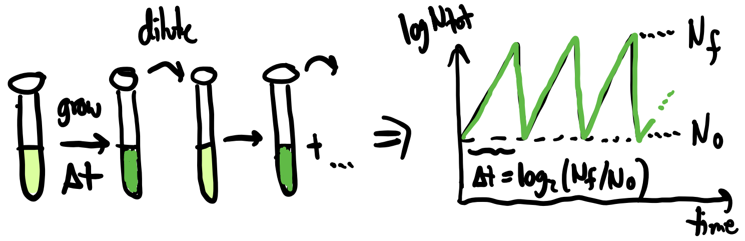
$L=2$: $\vec{g} = \underbrace{(0,0)}_{\text{WT}}, \underbrace{(1,0), (0,1)}_{\text{single mutants}}, \underbrace{(1,1)}_{\text{double mutant}}$

$L=3$: $\vec{g} = (0,0,0), \underbrace{(1,0,0)}_{\text{single}}, \ldots, \underbrace{(1,1,0)}_{\text{double}}, \ldots, \underbrace{(1,1,1)}_{\text{triple mutant}}$

$\vdots$

etc.

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



$$\boxed{\frac{\partial f(\vec{g})}{\partial t} = ???}$$

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

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

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

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

depends on $Z_{\vec{g}}$ @ <u>other</u> $\vec{g}'$!

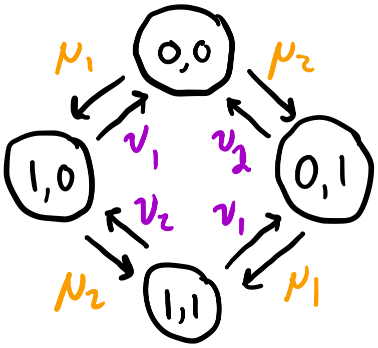
⇒ <u>correlations</u> between $\delta f(\vec{g})$ & $\delta f(\vec{g}')$ ⇒ keeps $f(\vec{g}, t)$ normalized!

$$\sum_{\vec{g}} f(\vec{g}, t+\delta t) = \cancelto{1}{\sum_{\vec{g}} f(\vec{g})} + \sum_{\vec{g}} \sqrt{\frac{f(\vec{g})\,\delta t}{N_e}}\, Z_{\vec{g}} - \cancelto{1}{\sum_{\vec{g}} f(\vec{g})} \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')\,\delta t}{N_e}}\, Z_{\vec{g}'}$$

$= 1$ ⇒ 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:

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

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

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{mut} \equiv \sum_{\vec{g}'}\Big[\underbrace{M(\vec{g}'\to\vec{g})}\, f(\vec{g}') - \underbrace{M(\vec{g}\to\vec{g}')}\, f(\vec{g})\Big]$$

$2^L \times 2^L$ <u>matrix</u> of mut'n rates

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

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

③ Selection (growth rate differences)

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

⇒ then after 1 cycle of growth:

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

⇒ if $X(\vec{g})\Delta t \ll 1$ (continuum limit) ⇒ Talylor 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! $\langle \bar{X}(t) f(\vec{g},t)\rangle \neq \langle \bar{X}(t)\rangle \langle f(\vec{g},t)\rangle$

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

$\Rightarrow$ 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)

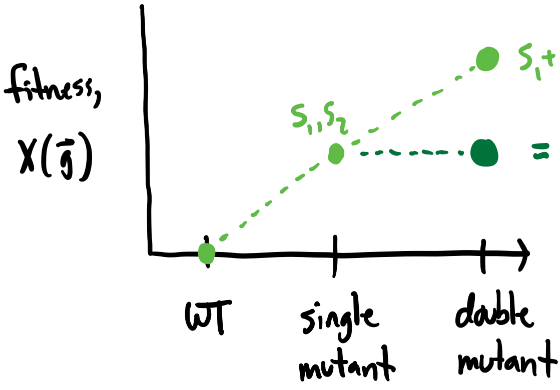
$X(1,0) \equiv s_1$
$X(0,1) \equiv s_2$
} could <u>measure</u>, e.g. gene deletion screen (HW2)

$$X(1,1) \equiv \,? \equiv \underbrace{s_1 + s_2}_{\text{"additive part"}} + \underbrace{\epsilon}_{\text{"epistasis"}}$$

(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:

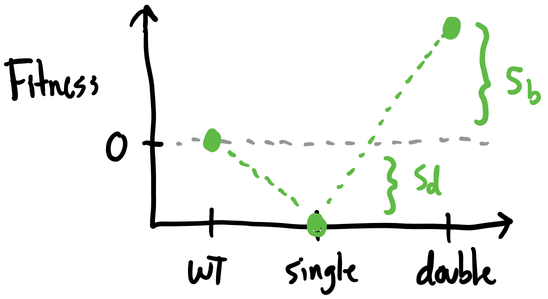


s_1+s_2 = "additive", e.g. orthogonal pathways in a cell

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

2 stop codons in same gene...

⇒ people often interested in scenarios like:



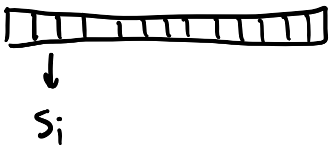
"fitness valley crossing"

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:

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

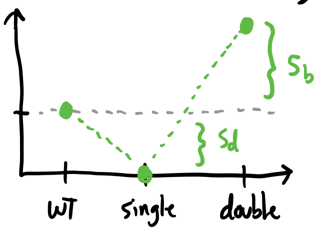
⇒ hard to parametrize in general (active area of research!)

⇒ in practice, people often use:

Additive model ($L\gg1$)

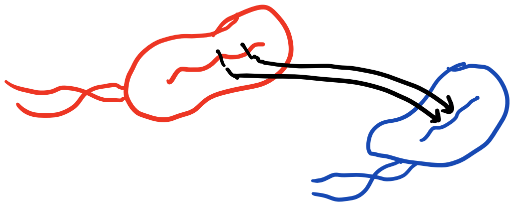
$$X(\vec{g}) \approx \sum_{\ell=1}^{L} s_\ell g_\ell$$

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



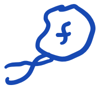
⇒ other new bit of biology for $L \geq 2$:

⑤ Recombination (exchange of genetic material between different individuals)



Many different mechanisms!

⇒ but many share same basic behavior:

① Focal individual  is chosen to undergo recombination

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

② Donor individual is chosen to donate portion of genome

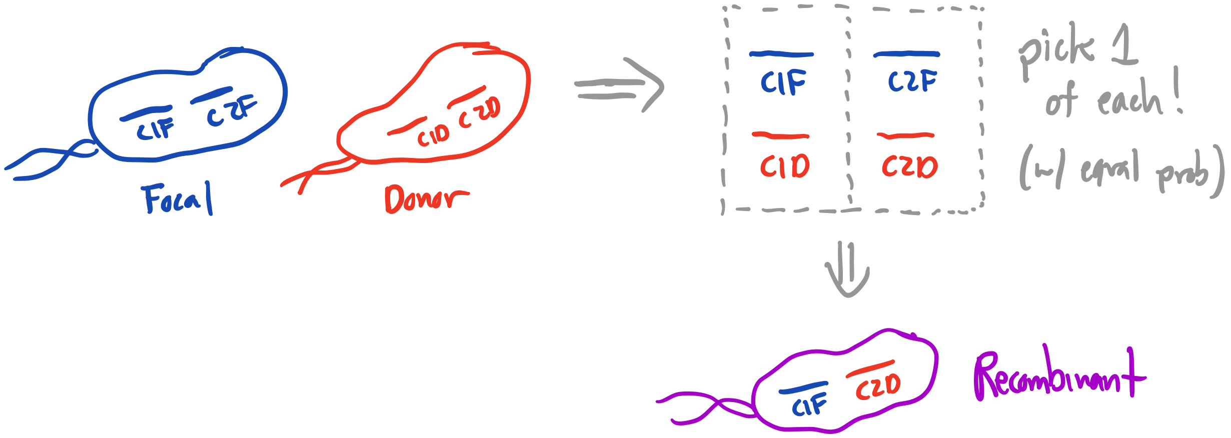
⇒ probability $\sim 1/N$ ⇒ $f(\vec{g})$ for any individual of that genotype.

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

⇒ producing "recombinant"

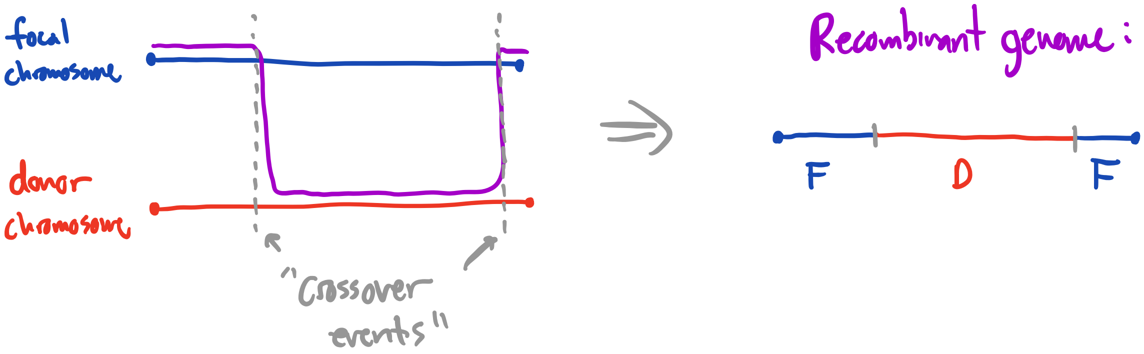
⇒ different mechanisms enter @ this step:

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

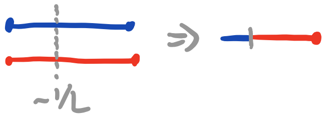


pick 1 of each! (w/ equal prob)

ⓑ **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 pairs of sites in same genome!

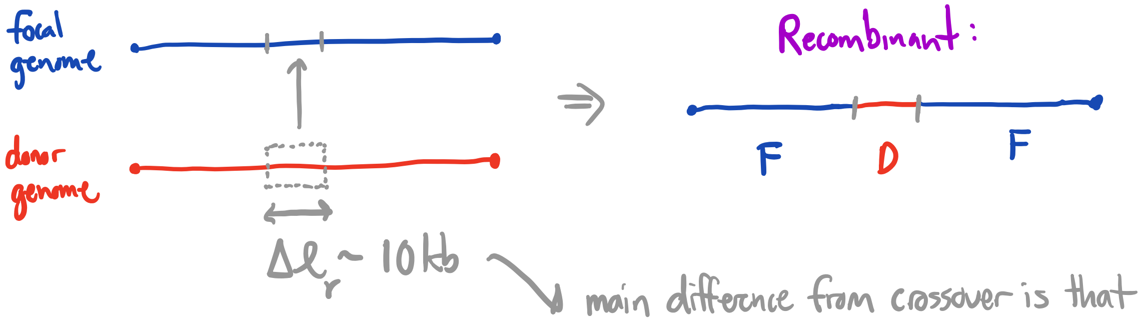
⇒ e.g. in humans ⇒ $L_{chrom} \sim 10^8$ bp (x 23 chromosomes)

⇒ P(recomb) ~ 100% if opp. ends of same chrom (or diff chroms)

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

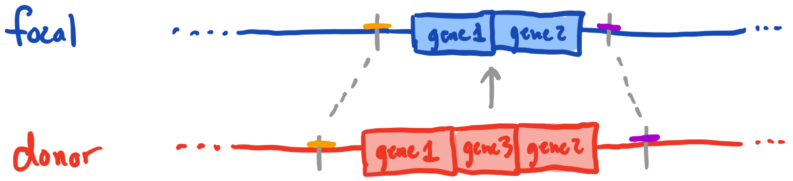
Ⓒ "Horizontal gene transfer (HGT)" / "gene conversion"

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



main difference from crossover is that $\Delta\ell_r \ll L$, as opposed to $\Delta\ell_r \sim O(L)$

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

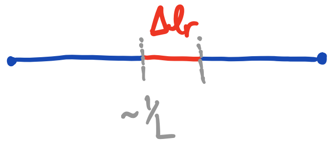


often mediated by homology (=)
similar to PCR...

⇒ active area of research!

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

⇒ simplest HGT model:



$\Delta\ell_r$ = const, location ~ uniform

So far: individual-based picture...

⇒ can we translate to continuum limit?

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

⇒ 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 $\rho, L, \Delta\ell r, \ldots$ etc.]

$$\begin{tikzcd}
 & \text{focal } (g_{1F}, g_{2F}) \arrow[dr] & (g_{1F}, g_{2D}) & \text{1 recombinant chosen to} \arrow[l] & \\
(\sim \frac{1}{N}) & \text{donor } (g_{1D}, g_{2D}) \arrow[ur] & (g_{1D}, g_{2F}) & \text{replace focal individual} \arrow[l] & (\sim \frac{1}{2})
\end{tikzcd}$$

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

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

case 1 (of 16):

$$\begin{tikzcd}
F \ (1,1) \arrow[dr] & (1,0) & \Rightarrow \text{rate} \quad Rf(1,1)f(0,0) \cdot \frac{1}{2} \\
D \ (0,0) \arrow[ur] & (0,1) & \Rightarrow \quad Rf(1,1)f(0,0) \cdot \frac{1}{2}
\end{tikzcd}$$

case 2 (of 16):

$$\begin{tikzcd}
F \ (0,0) \arrow[dr] & (0,1) \\
D \ (1,1) \arrow[ur] & (1,0)
\end{tikzcd}$$

same!

Case 3 (of 16):

$$\begin{tikzcd}
(1,1) \arrow[dr] & (1,0) \arrow[r, Rightarrow] & Rf(1,1)f(1,0)\,\frac{1}{2} \\
(1,0) \arrow[ur] & (1,1) \arrow[r, Rightarrow] & Rf(1,1)f(1,0)\,\frac{1}{2}
\end{tikzcd}$$

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

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

same!

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

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

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

⇒ 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:

$$\left(\frac{\delta f(\vec{g})}{\delta t}\right)_{rec} = \rho \overbrace{\sum_{\vec{g}_F,\vec{g}_O} \underbrace{T(\vec{g}_F,\vec{g}_O\to\vec{g})}_{\text{"recombination kernel"} \to \text{"tensor"}} \underbrace{f(\vec{g}_F)\, f(\vec{g}_O)}_{\text{nonlinear!}}}^{\text{incoming recombinants}} - \overbrace{\rho f(\vec{g})}^{\text{outgoing recombinants.}}$$

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

Putting everything together, general multilocus model looks like:

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

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

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

Problem: No exact solution for stationary dist'n, P_{fix}, etc.

— even for $L = 2$!

$\Rightarrow$ What do we do instead?!? $\Rightarrow$ asymptotic approx's

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

$\Rightarrow$ what are some limits where we might understand understand this SDE?

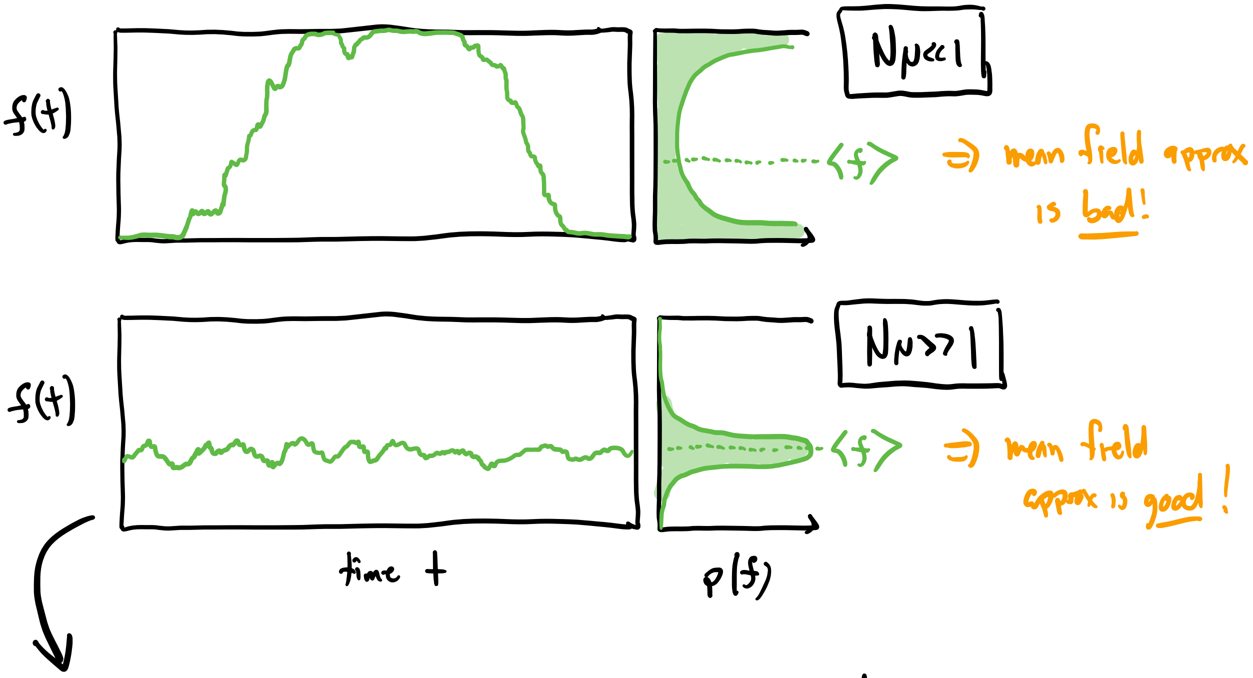
$$\boxed{\frac{\partial f(\vec{g})}{\partial t} = \sim (x - \bar{x}) + \sim L \times \mu + \sim \rho + \sim \frac{Z}{\sqrt{N}}}$$

① Obvious answer: $L=1$ ⇒ cheating! *

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

⇒ is this a good approx here?

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



⇒ mean field approx is bad!

⇒ mean field approx is good!

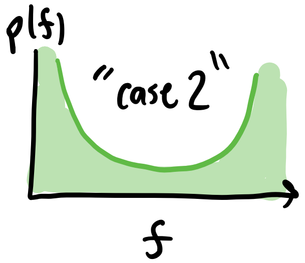
key feature: large # of individuals in both genotypes @ same time

⇒ so fluctuations are small.

⇒ e.g. for $L=2$, might be ok ⇒ but for $L \gg 1$ ⇒ $2^L \gg N$!

e.g. $L \sim 1000\text{bp}$ ⇒ $2^L \sim 10^{300}$!

⇒ large L will <u>always</u> look like (@ least in some dimensions)



⇒ noise always relevant!

Need to look for other approximations of SDE...

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

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 <u>Successive mutations regime</u> (i.e. treat mutation as small correction)

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