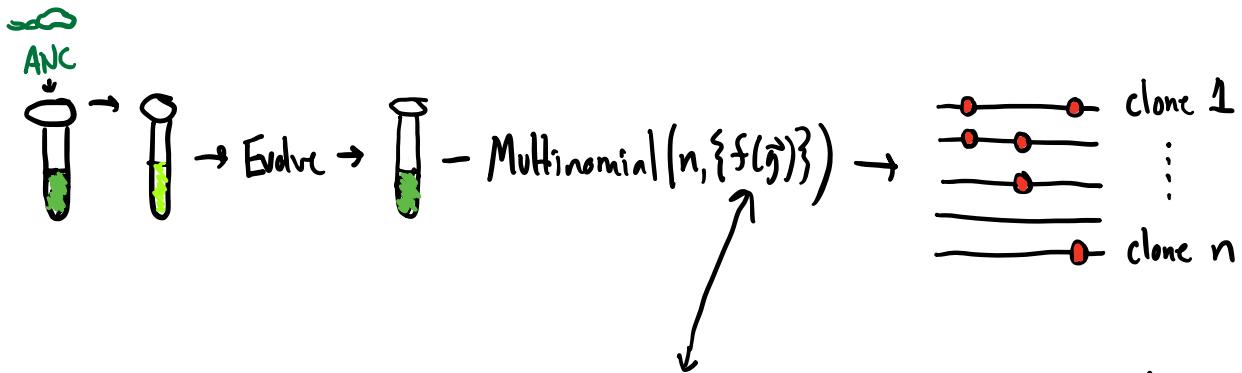


# **Chapter 9**

## **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

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

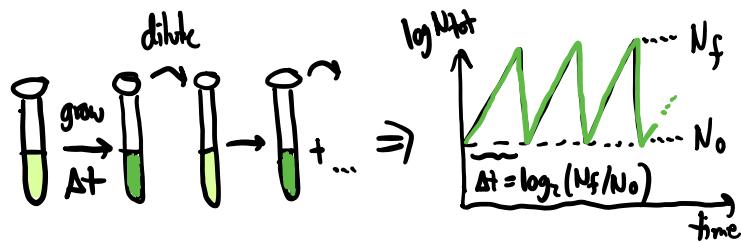
$$L=2: \vec{g} = \underbrace{(0,0)}_{\text{WT}}, \underbrace{(1,0), (0,1)}_{\text{single mutants}}, \underbrace{(1,1)}_{\text{double mutant}}$$

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

⋮

etc.

Can we generalize our serial dilution ( $\rightarrow$  diffusion) models?



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

① Genetic drift: first assume no growth rate differences...  
 $\rightarrow$  (no mutations)

$\Rightarrow$  After 1 day of growth:

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

$\Rightarrow$  After dilution step:

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

$\Rightarrow$  if repeat our Taylor expansions from before ( $N_0 \gg \text{large}$ )

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

where  $Z_{\vec{g}}$  are  $N(0,1)$  random variables w/

$$\langle Z_{\vec{g}} \rangle = 0, \quad \langle Z_{\vec{g}} Z_{\vec{g}'} \rangle = \begin{cases} 0 & \text{if } \vec{g} \neq \vec{g}' \\ 1 & \text{if } \vec{g}' = \vec{g} \end{cases}$$

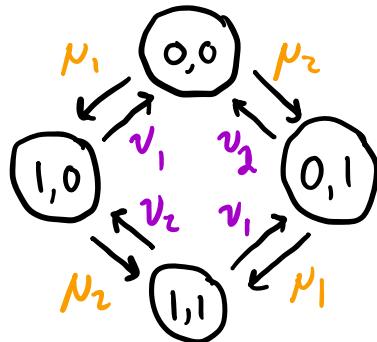
$\Rightarrow$  correlations ensure that  $f(\vec{g}, t)$  is normalized:

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

$$= 1$$

② Mutations:

$\Rightarrow$  easiest to start w/  $L=2$ :



$\Rightarrow$  key feature: can only move by ~1 step @ a time

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

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

$\Rightarrow$  continuum limit:

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

+ noise from drift.

linear in \*  
genotype freqs.

$\Rightarrow$  larger L's are similar, but more work to write out...

$\Rightarrow$  one way is:

$$\left[ \frac{\delta f(\vec{g})}{\delta t} \right]_{\text{mut}} = \sum_{\substack{\text{nearest} \\ \vec{g}' \\ \text{neighbors}}} \sum_{l=1}^L \left[ \underbrace{\mu_e f(\vec{g}') g_e (1-g'_e) + \nu_e f(\vec{g}') (1-g_e) g'_e}_{\text{mutations into genotype}} \right]$$

$$- \sum_{l=1}^L \left[ \underbrace{\mu_e f(\vec{g}) (1-g_e) + \nu_e f(\vec{g}) g_e}_{\text{mutations out of genotype}} \right]$$

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

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

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

### 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:

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

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

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



note: not an ensemble avg

population mean fitness

$$\langle \bar{X}(t) f(\vec{g}, t) \rangle \quad \leftarrow \quad \bar{X}(t) = \sum_{\vec{g}} X(\vec{g}) f(\vec{g}, t)$$

$$\neq \langle \bar{X}(t) \rangle \langle f(\vec{g}, t) \rangle$$

Normalized:  $\sum_{\vec{g}} f(\vec{g}, t + \delta t) = \sum_{\vec{g}} f(\vec{g}, t) + \sum_{\vec{g}} X(\vec{g}) f(\vec{g}, t) - \sum_{\vec{g}} f(\vec{g}) \sum_{\vec{g}'} X(\vec{g}') f(\vec{g}') = 1$

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

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

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

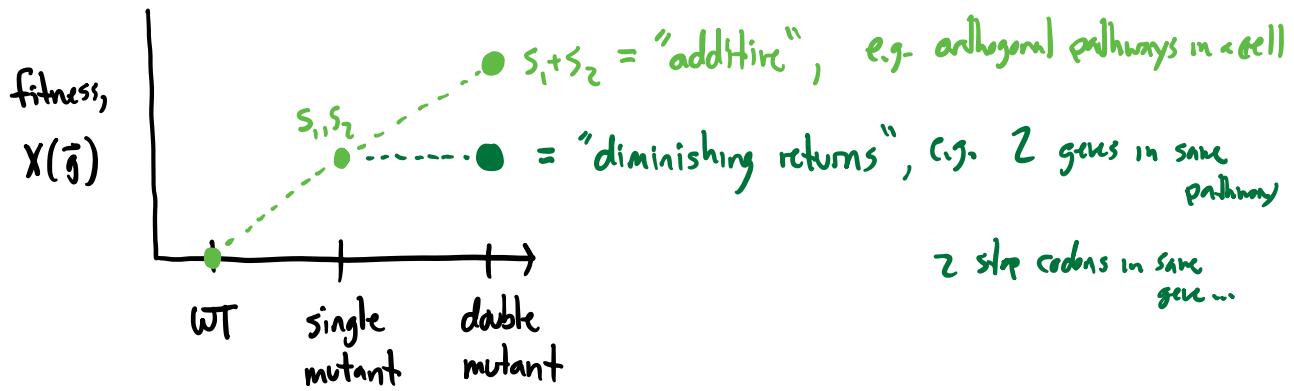
$$X(0,0) \equiv 0 \quad (\text{convention})$$

$$\begin{aligned} X(1,0) &\equiv S_1 \\ X(0,1) &\equiv S_2 \end{aligned} \quad \left. \begin{array}{l} \text{could measure, e.g. gene deletion screen} \\ (\text{HW2}) \end{array} \right\}$$

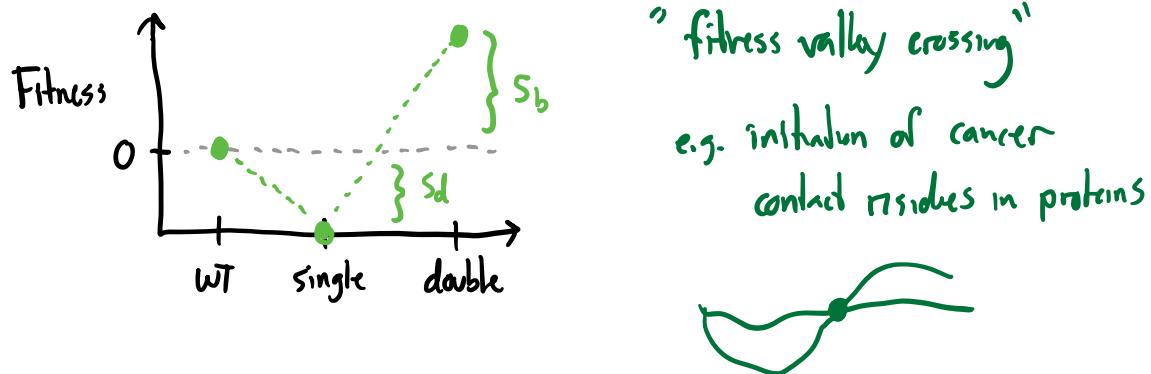
$$X(1,1) \equiv ? \equiv \underbrace{S_1 + S_2}_{\text{"additive part"}} + \underbrace{\epsilon}_{\substack{\text{"epistasis"} \\ (\text{how much deviation from additivity})}}$$

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

Often easiest to express w/ picture:



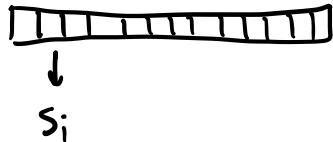
⇒ people often interested in scenarios like:



$\Rightarrow$  gets even more complicated for  $L > 2$ :

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

additive part  
("coupon collecting")
epistatic part.



$\Rightarrow$  can write as Taylor expansion around WT:

$$E(\vec{g}) = \underbrace{\sum_{e=1}^L \sum_{e'=1}^L e_{ee'} g_e g_{e'}}_{\text{"pairwise epistasis"}} + \underbrace{\sum_{e=e'}^L \sum_{e''=1}^L e_{eee''} g_e g_{e'} g_{e''}}_{\text{"higher-order epistasis"}} + \dots$$

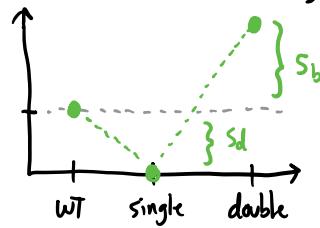
$\Rightarrow$  hard to parameterize in general (active area of research!)

$\Rightarrow$  in practice, people often use:

Additive model ( $L \gg 1$ )

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

Pictures ( $L \sim O(1)$ )

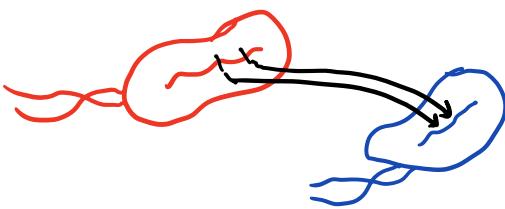


$\Rightarrow$  other new bit of biology for  $L \geq 2$ :

⑤

## Recombination

(exchange of genetic material  
between different individuals)



Many different mechanisms!

$\Rightarrow$  but many share same basic behavior:

① Focal individual  is chosen to undergo recombination

$\Rightarrow$  w/ probability  $\rho$  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

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

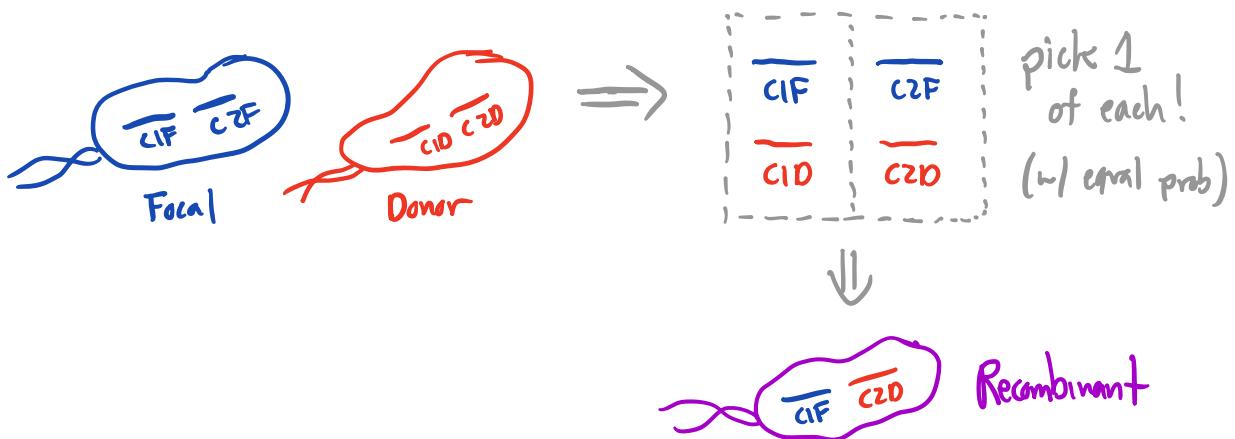


③ Some piece of donor's DNA is integrated into focal genome

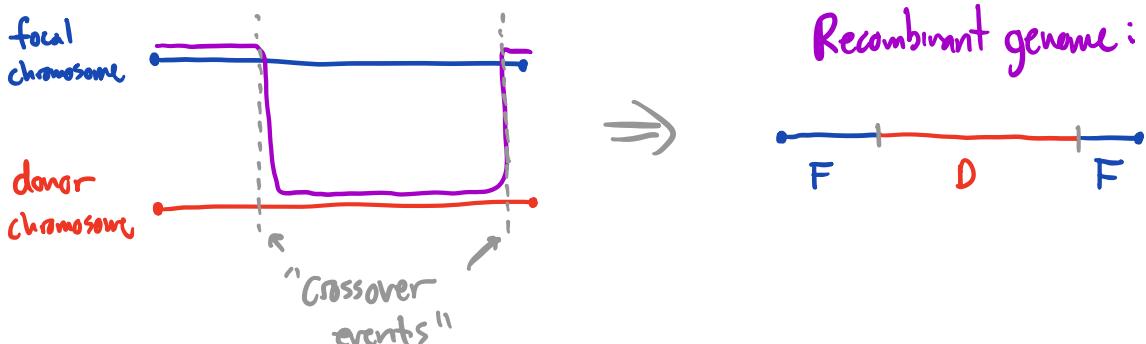
⇒ producing "recombinant"

⇒ different mechanisms enter @ this step:

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

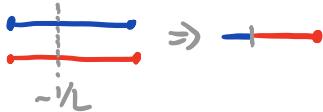


b) Crossover Recombination (e.g. w/in chromosomes in humans)



$\Rightarrow$  often modeled w/  $\sim 1$  crossover per recombination event

w/ location chosen uniformly across chromosome



$\Rightarrow$  in practice, "hot spots" + "cold spots"  $\Rightarrow$  "recombination map"

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

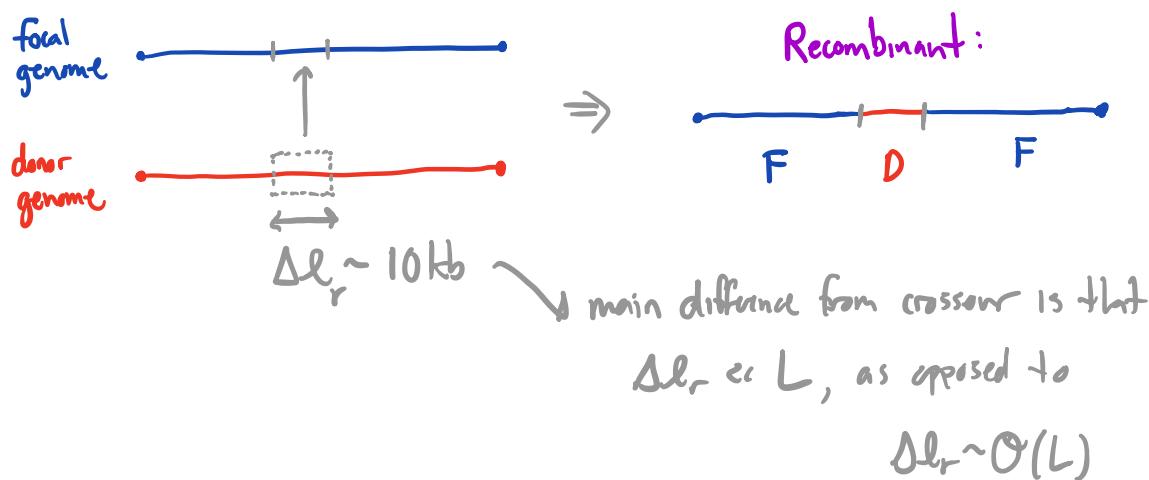
$\Rightarrow$  e.g. in humans  $\Rightarrow L_{\text{chrom}} \sim 10^8$  bp

$\Rightarrow p(\text{recomb}) \sim 100\%$ , if 2 ends of same chrom

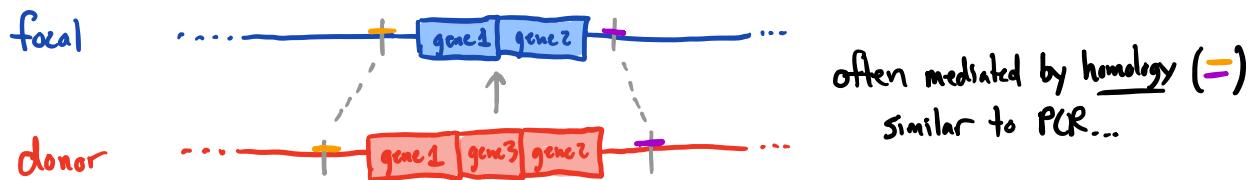
$\Rightarrow p(\text{recomb}) = 10^{-8}$  if neighboring base pairs

### C) "Horizontal gene transfer" / "gene conversion"

$\Rightarrow$  lingo is a little controversial, but basic idea pretty simple:



$\Rightarrow$  also a mechanism for gaining + losing genes ("accessory genome")

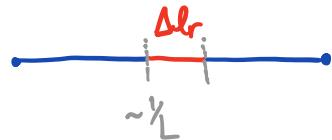


$\Rightarrow$  active area of research!

$\Rightarrow$  but in this class, will mostly focus on "core genome"

$\Rightarrow$  simplest HGT model:

$$\Delta l_r = \text{const}, \text{location} \sim \text{uniform}$$



So far: individual-based picture...

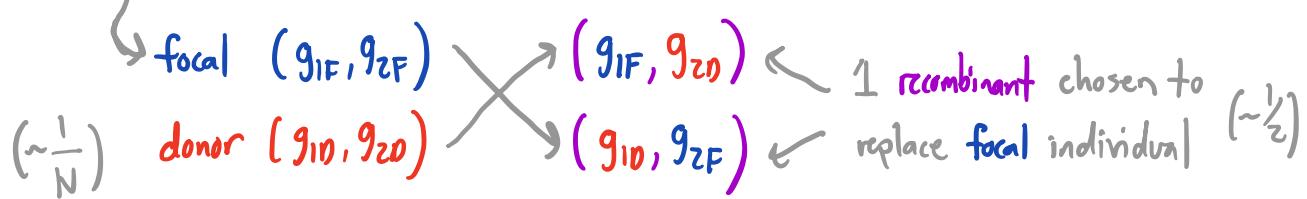
$\Rightarrow$  can we translate to continuum limit?

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

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

$\Rightarrow$  all mechanisms have same net effect:

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



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

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

case 1 (of 16) :

$$F(1,1) \xrightarrow{\quad} (1,0) \Rightarrow \text{rate } R f(1,1) f(0,0) \cdot \frac{1}{2}$$

$$D(0,0) \xrightarrow{\quad} (0,1) \Rightarrow R f(1,1) f(0,0) \cdot \frac{1}{2}$$

case 2 (of 16) :

$$F(0,0) \xrightarrow{\quad} (0,1)$$

$$D(1,1) \xrightarrow{\quad} (1,0)$$

same!

case 3 (of 16):  $(1,1) \xrightarrow{\cancel{\longrightarrow}} (1,0) \Rightarrow Rf(1,1)f(1,0) \frac{1}{2}$

$(1,0) \xrightarrow{\cancel{\longrightarrow}} (1,1) \Rightarrow Rf(1,1)f(1,0) \frac{1}{2}$

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

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

Same!

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

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

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

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

$\Rightarrow$  harder to write down explicitly for  $L > 2 \dots$

but will have general form:

$$\left( \frac{\delta f(\vec{g})}{\delta t} \right)_{rec} = e \sum_{\vec{g}_F, \vec{g}_D} \underbrace{T(\vec{g}_F, \vec{g}_D \rightarrow \vec{g})}_{\text{"recombination kernel"}} \underbrace{f(\vec{g}_F) f(\vec{g}_D)}_{\text{"tensor"}} - e f(\vec{g})$$

incoming recombinants      outgoing recombinants.

nonlinear!

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

Putting everything together, general multilocus model looks like:

$$\frac{df(\vec{g})}{dt} = \left[ X(\vec{g}) - \bar{X}(+) \right] f(\vec{g}) + \sum_{\vec{g}'} M(\vec{g} \rightarrow \vec{g}') f(\vec{g}') - M(\vec{g}' \rightarrow \vec{g}) f(\vec{g})$$

Selection (nonlinear)

mutation (linear, "local")

$$+ \rho \sum_{\vec{g}_F, \vec{g}_D} T(\vec{g}_F, \vec{g}_D \rightarrow \vec{g}) f(\vec{g}) - \rho f(\vec{g})$$

recombination  
(nonlinear, non-local)

$$+ \sqrt{\frac{f(\vec{g})}{N}} \eta(\vec{g}) - f(\vec{g}) \sum_{\vec{g}'} \sqrt{\frac{f(\vec{g}')}{N}} \eta(\vec{g}')$$

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")  $L, N, X(\vec{g}), M, \rho, T$

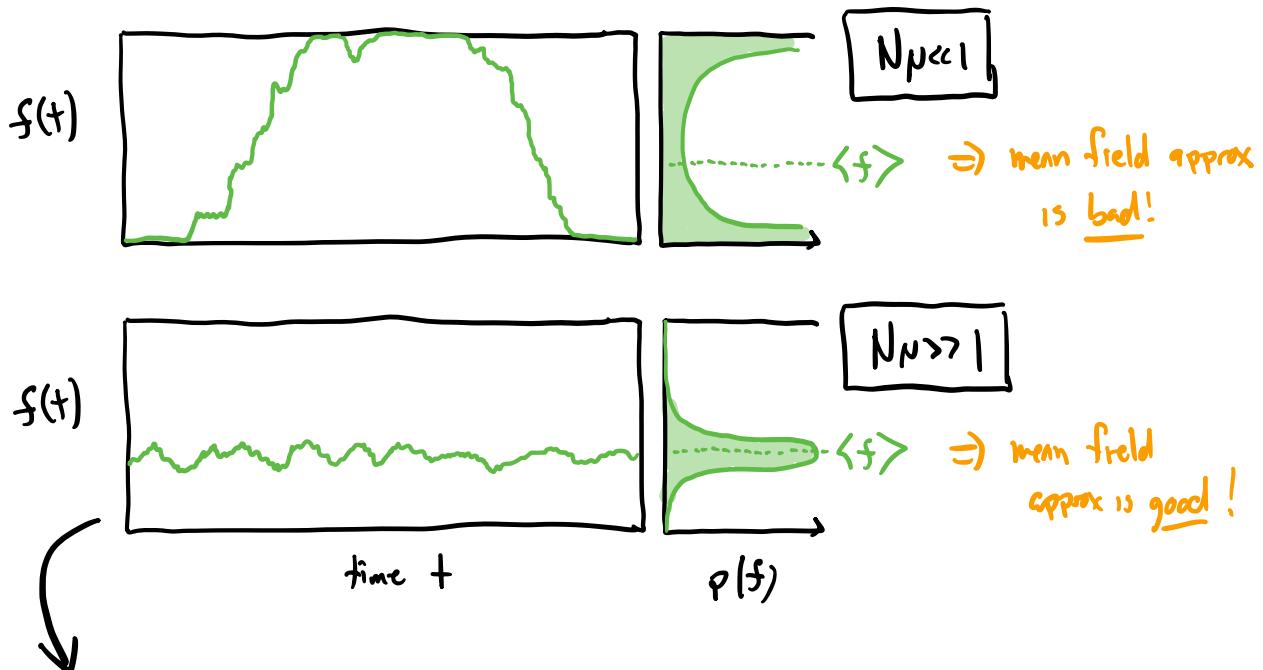
$\Rightarrow$  what are some limits where we might understand understand this SDE?

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

$$+ \sim \rho + \sim \frac{Z}{\sqrt{N}}$$

- ① Obvious answer:  $L=1 \Rightarrow$  cheating! \*
- ② in physics, might be primed to take  $N \rightarrow \infty$  limit ...  
("mean field approx") since @ least noise goes away ...  
 $\Rightarrow$  is this a good approx here?

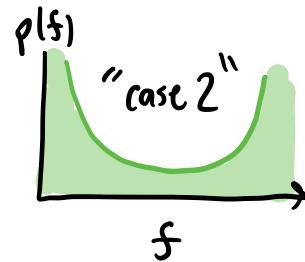
$\Rightarrow$  Recall for  $L=1$  case, 2 different regimes when  $t \rightarrow \infty$ :



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.  $L \sim 1000 \text{ bp} \Rightarrow 2^L \sim 10^{300}$ !

$\Rightarrow$  large  $L$  will always look like  
(@ least in some dimensions)



$\Rightarrow$  noise always relevant!

Need to look for other approximations of SDE ...

$$\frac{d\vec{x}(t)}{dt} = \sim (x - \bar{x}) + \sim L \cdot \mu + \sim \epsilon + \sim \frac{\vec{z}}{JN}$$

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

$\Rightarrow$  even if  $L \gg 1$ , if behavior "looks like"  $L=1$  case,  
 $\Rightarrow$  can use what we already know...

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

$\Rightarrow$  what if mutation rates are low enough that  
Only 1 or 2 genotypes are present @ a time?