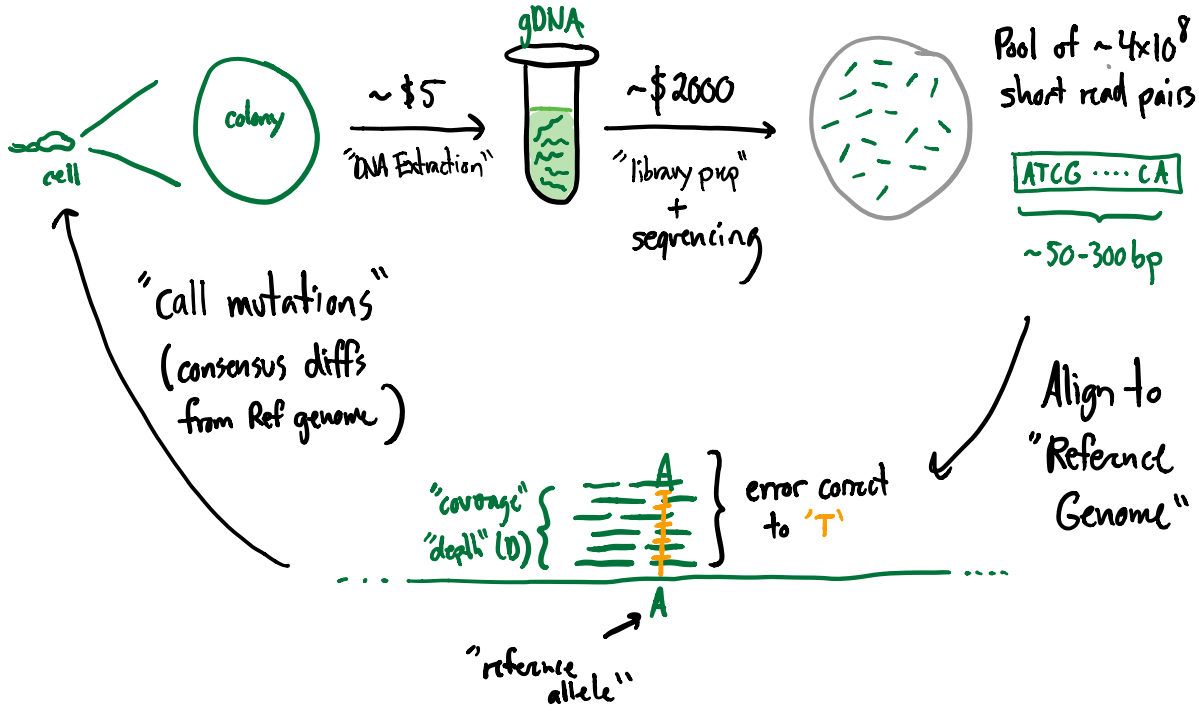
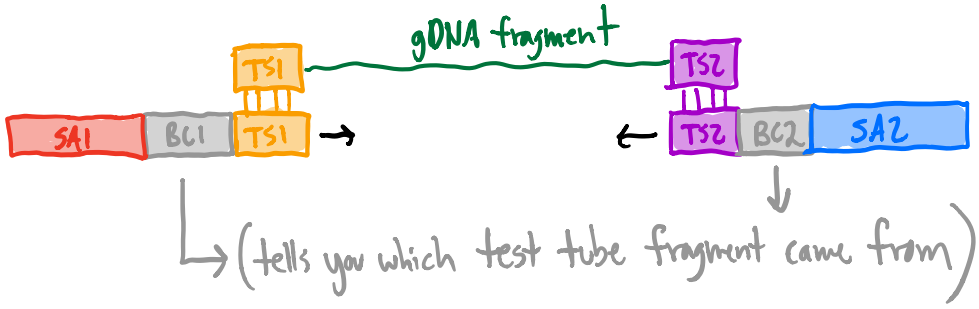
Announcements: PSET 3 DUE 2/23, Office Hours Today (12:30-2pm)
Anita Problem Session Fri (Slack)

Last time: Next-gen (Illumina) sequencing

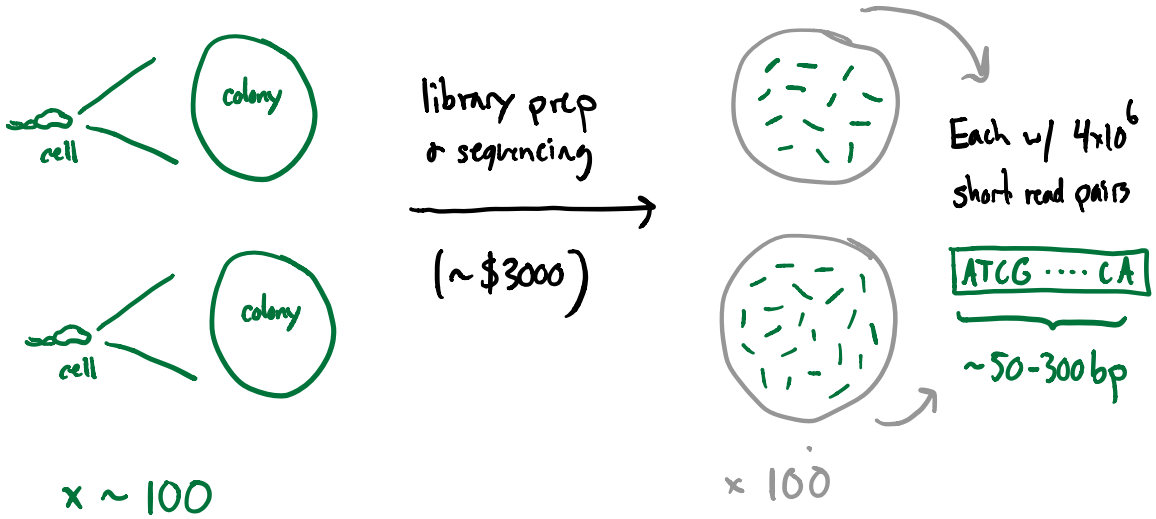


Overkill? ~4×10^8 reads $\Rightarrow 10^{11}$ bp $\Rightarrow 10^5$-fold coverage of E. coli genome

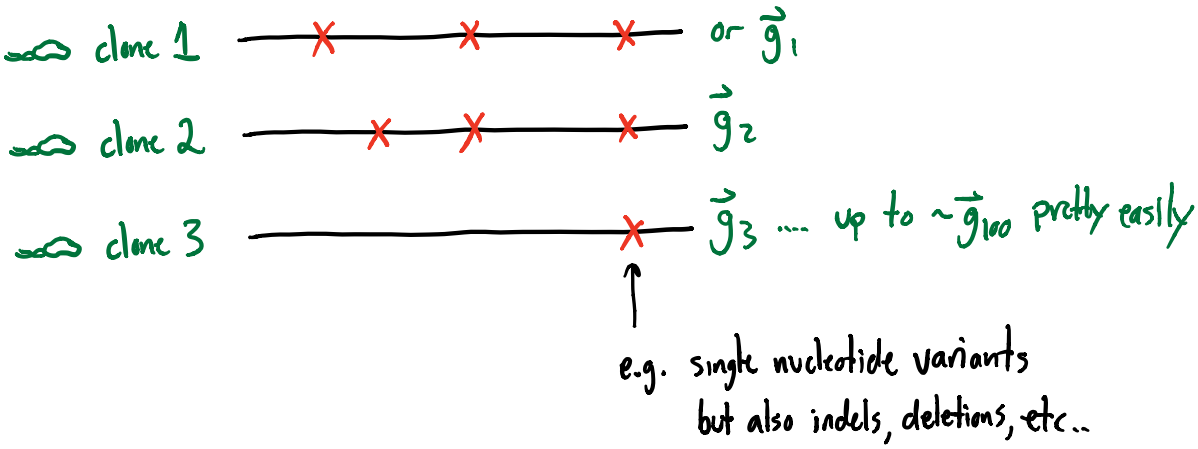
Solution: "multiplex": Add sample specific "barcode" sequence during library prep step



Upshot: can sequence ~100 E. coli libraries on one flow cell & get 300-fold coverage of E. coli genome



⇒ After aligning reads & detecting "true" mutations, get sequences of genomes:



How are sampled genomes related to dist'n of genomes in pop'n?

⇒ let $n_{\vec{g}}$ = # sampled clones w/ genome $\vec{g}$ (random from sampling)

$f(\vec{g})$ = frequency of genome $\vec{g}$ in population (random from evolution)

⇒ then $\Pr\left[\{n_{\vec{g}}\} \middle| n, \{f(\vec{g})\}\right]$ = multinomial distribution

(n ↑ total # clones sampled)

$$\propto \prod_{\vec{g}} \frac{f(\vec{g})^{n_{\vec{g}}}}{n_{\vec{g}}!}$$

⇒ Genotype space is huge!

⇒ often coarse-grain to summary statistics.

e.g. n_ℓ = # individuals w/ mutation @ site ℓ $= \sum_{\vec{g}} g_\ell n_{\vec{g}}$

(g_ℓ: 1 if mutant site, 0 else)

⇒ can show $\Pr[n_\ell | n, \{f(\vec{g})\}] = \text{Binomial}(n, f_\ell)$

$f_\ell \equiv \sum_{\vec{g}} g_\ell f(\vec{g})$

e.g. total # of mutations separating 2 genomes

⇒ since depends on genome length, often normalized by L:

$$\frac{\#\text{ mutations between 2 random clones}}{L} = \begin{cases} \text{"heterozygosity" } (\pi) & \text{if from same pop'n} \\ \text{"divergence" } (d) & \text{if from diff "species" (or isolated sub-pops)} \end{cases}$$

e.g. heterozygosity (π) in humans is $\sim 10^{-3}$

divergence (d) between humans & chimps is $\sim 10^{-2}$

heterozygosity (π) between E.coli in different humans is $\sim 10^{-2}$

Can we relate π to genotype distribution, $f(\vec{g})$?

Note that: ($g_{1\ell}, g_{2\ell} \in \{0,1\}$)

$$\pi = \frac{1}{L}\sum_{\ell=1}^{L}\left[\, g_{1\ell}(1-g_{2\ell}) + (1-g_{1\ell})g_{2\ell} \,\right]$$

$$\left\langle \pi \,\middle|\, \{f(\vec{g})\} \right\rangle_{\text{sampling}} = \frac{1}{L}\sum_{\ell=1}^{L}\left[\underbrace{\langle g_{1\ell}(1-g_{2\ell})\rangle}_{f_\ell(1-f_\ell)} + \underbrace{\langle (1-g_{1\ell})g_{2\ell}\rangle}_{(1-f_\ell)f_\ell} \right]$$

$$= \frac{1}{L}\sum_{\ell=1}^{L} 2f_\ell(1-f_\ell)$$

Remember: $f_\ell \equiv \sum_{\vec{g}} g_\ell f(\vec{g})$

$\Rightarrow$ averaging over f_ℓ:

$$\langle \pi \rangle = \frac{1}{L}\sum_{\ell=1}^{L}\langle 2f_\ell(1-f_\ell)\rangle = \frac{1}{L}\sum_{\ell=1}^{L}\int 2f_\ell(1-f_\ell)\,P(f_\ell)\,df_\ell$$

e.g. if genome is collection of neutral sites

$$\Rightarrow p(f_\ell) \approx \frac{2N_e\nu}{f_\ell} \quad \text{(quasi-stationary dist'n)}$$

$$\langle \pi \rangle = \int 2f(1-f)\,\frac{2N_e\nu}{f}\,df = \boxed{2N_e\nu}\ *$$

$\hookrightarrow$ fit $N_e \equiv \dfrac{\langle \pi \rangle}{2\nu} \approx \dfrac{10^{-3}}{2\times 10^{-8}}$

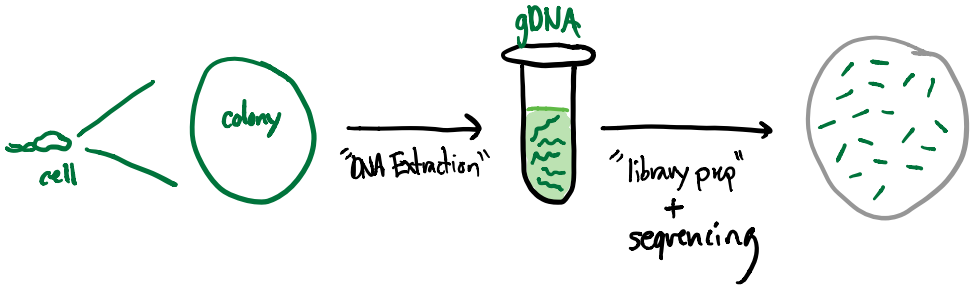
→ population can only spread out so far in sequence space!

$\Rightarrow$ Variance of π $\Rightarrow$ much more complicated!

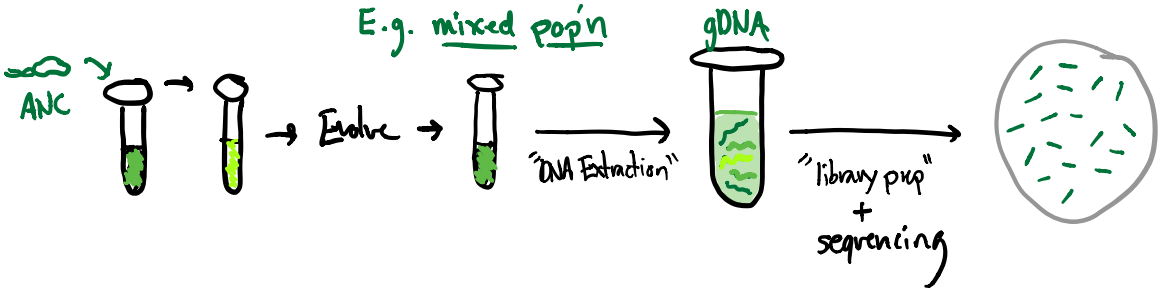
$\Rightarrow$ correlations between g_ℓ & $g_{\ell'}$

$\Rightarrow$ will see more later!

So far, have focused on clones



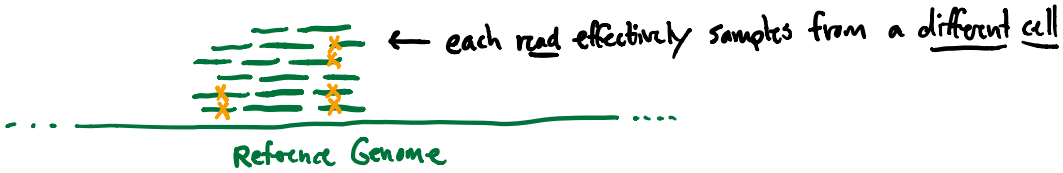
⇒ lots of other things we could put in our library prep...



⇒ known as "pooled sequencing" / "metagenomic sequencing"

⇒ in this case, assembly very hard!

⇒ Payoff comes from reference mapping:



e.g. if $A_\ell \equiv$ # reads w/ mutations @ site ℓ

$$\Rightarrow \Pr\left[A_\ell \,\middle|\, D_\ell, \{f(\vec{g})\}\right] = \text{Binomial}\left(D_\ell, f_\ell\right)$$

D_ℓ: total coverage @ site ℓ

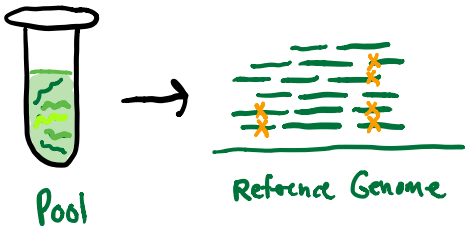
+ sequencing/PCR errors

$\Rightarrow$ since we can sequence ~100 E.coli genomes @ >100x coverage in 1 run of Illumina sequencing

$\Rightarrow$ can effectively sample ~100 clones ~100x more cheaply by sequencing pools!

(much cheaper way to track freqs of individual mut'ns)

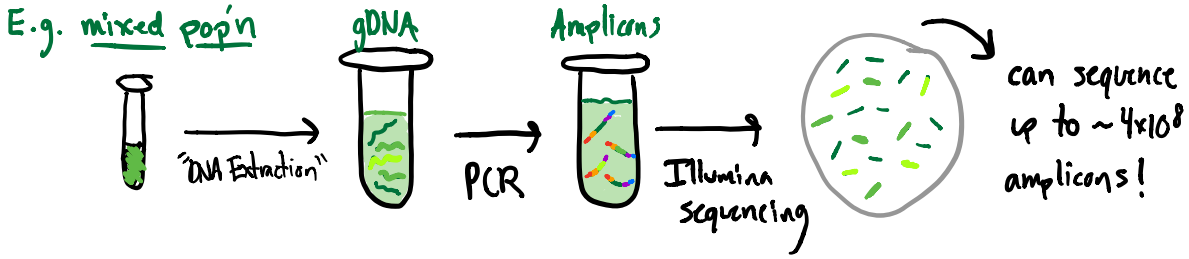
Downsides:



1. sequencing errors!
2. which mutations are in same cells? ("linkage information")

e.g.  vs

Can also sequence pools of amplicons:



Two common targets:

all bacteria have it

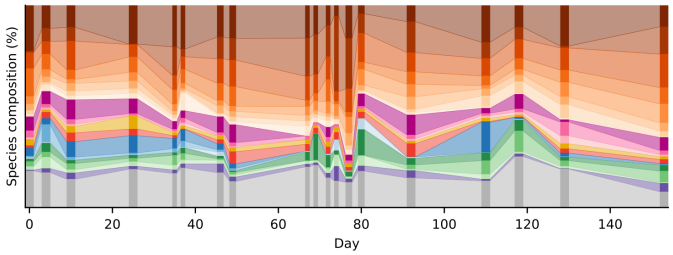
↳ track abundances of different species

can be added to lab strains

↳ track closely related strains

⇓

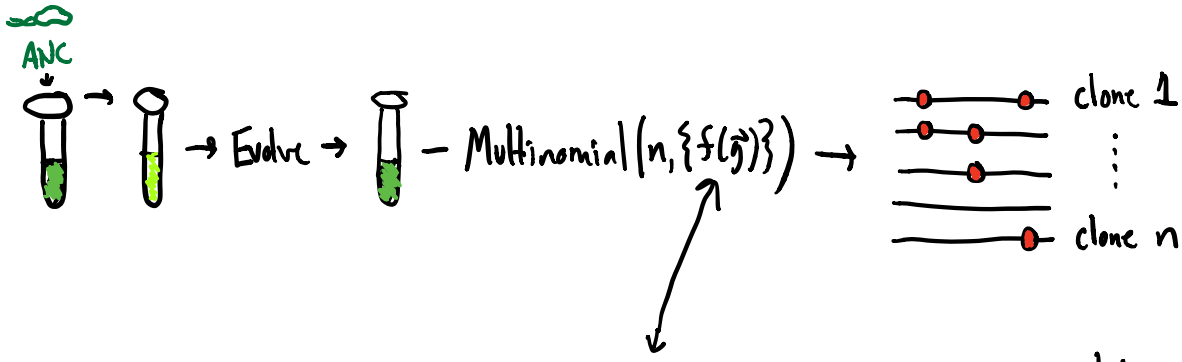
E.g. 1 person's microbiome over time



⇓

E.g. Homework Problem

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

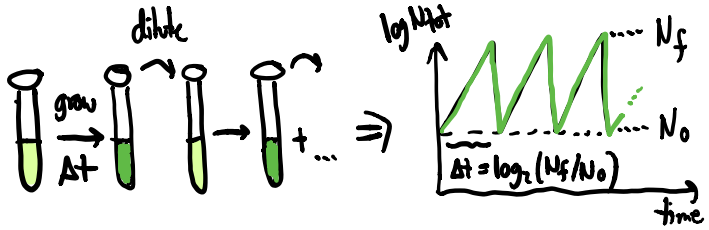
$L=2$: $\vec{g} =$ $(0,0)$ (WT), $(1,0), (0,1)$ (single mutants), $(1,1)$ (double mutant)

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

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

⇒ 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 ($N_e \gg$ 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/

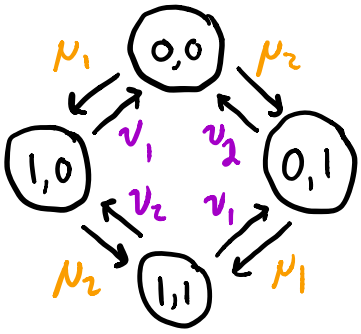
$$\langle Z_{\vec{g}} \rangle = 0, \quad \langle Z_{\vec{g}} Z_{\vec{g}'} \rangle = \begin{cases} 0 \text{ if } \vec{g} \neq \vec{g}' \\ 1 \text{ if } \vec{g}' = \vec{g} \end{cases}$$

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

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

$$= 1$$

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:

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

⇒ continuum limit:

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

+ noise from diff.

↖ linear in * genotype freqs.

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

⇒ one way is:

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

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

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

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

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

(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:

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

$\bar{X}(t)$: population mean fitness

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

note: not an ensemble avg

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

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