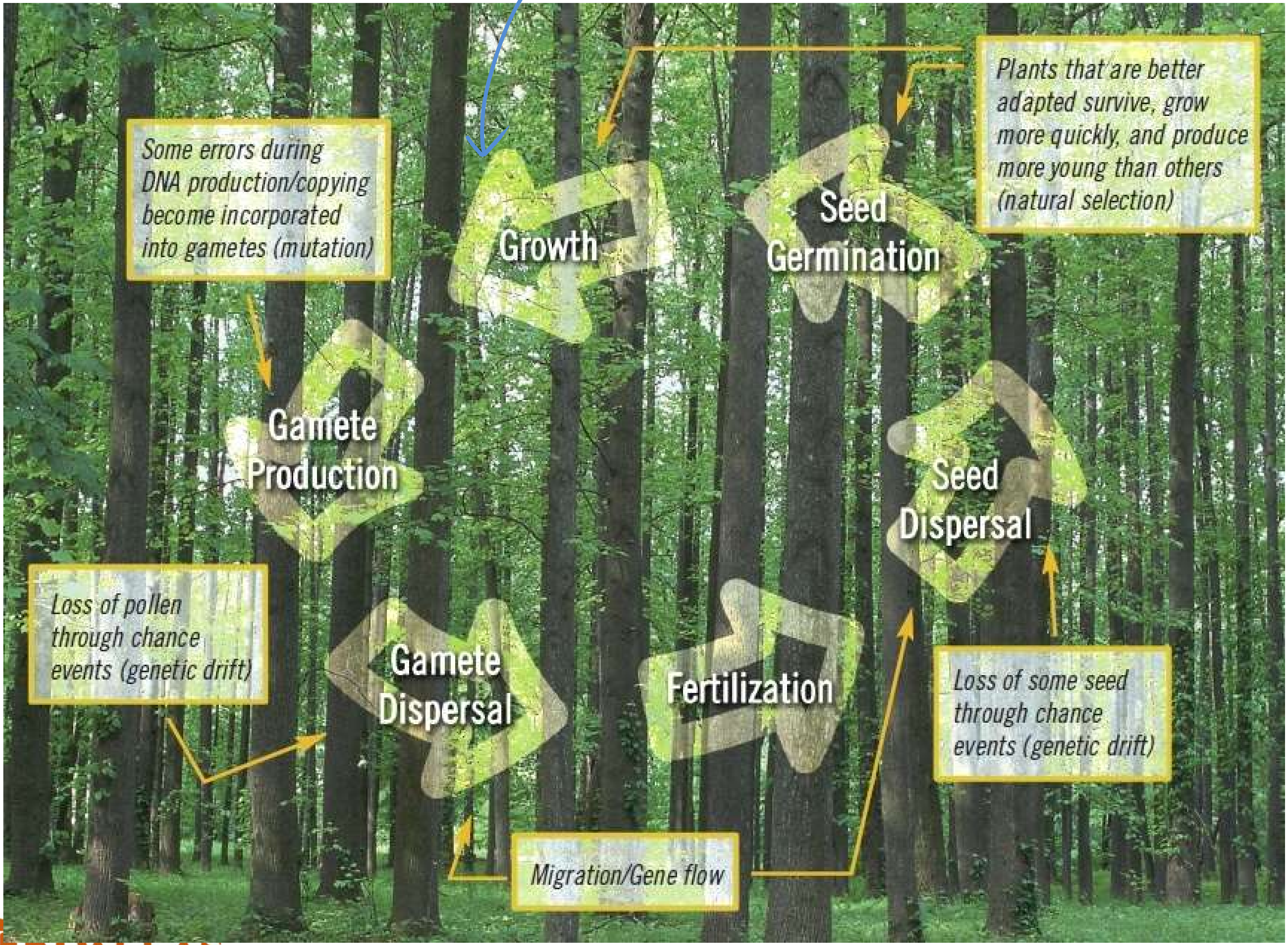
Evolutionary population genetics : a brief introduction

Pradipta Ray,

BIOL 6385 / BMEN 6389,

The University of Texas at Dallas

(some material based on content by PR in Eric Xing's 10-810 Carnegie Mellon class) accidental death (drift)



Some errors during DNA production/copying become incorporated into gametes (mutation)

Growth

Seed Germination

Plants that are better adapted survive, grow more quickly, and produce more young than others (natural selection)

Gamete Production

Seed Dispersal

Loss of pollen through chance events (genetic drift)

Gamete Dispersal

Fertilization

Loss of some seed through chance events (genetic drift)

Migration/Gene flow

usda.gov

Forces shaping observations

- Finite sample size : we cant sample the whole population
- Sample bias : is our sample representative ?

Fixing

- Fate of an allele in long run : evolutionary race among alleles
  - either dies out : “fixed” at 0 (known as loss)
  - intermediate situation : “fixed” at intermediate value, determined by equilibrium distribution of stochastic process (known as equilibrium)
  - wipes out all other alleles (becomes monoallelic) : “fixed” at 1 (known as fixing)

EVEN AT EQUILIBRIUM, ALLELES COULD STILL BE FIXED OR LOST UNDER DRIFT

Mutation : a cursory look

Mutation models : how many alleles

- Bi – allelic model : Two alleles exist – a mutation will change the allele from type A to type B (or a deleterious new allele which will vanish)
- Multi – allelic model : Many alleles exist – effect of mutation may be difficult to predict without explicit model
- Infinite – allelic model : Every mutation creates a new allele (convenient, not true)

Mutation models : how many sites

Infinite sites model :

- Every mutation happens at a new locus
- Expected no of substitutions / site << 1
- Plausible assumption for low mutation rate, short evolutionary history studies

Mutation models : parameters

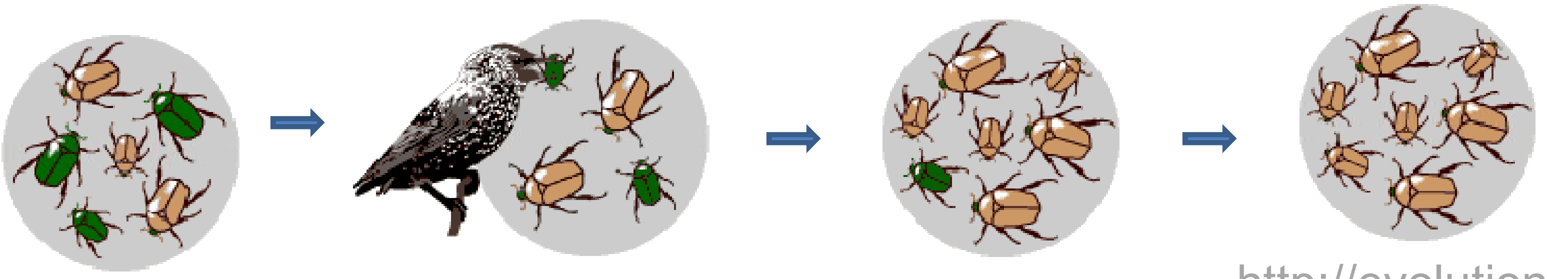
- Typical assumption : all mutations equally likely to occur (does not mean all mutations equally likely **to survive**)

N individuals → $2N$ alleles → $2\mu N$ mutations in every generation

Selection : a cursory look

Selection models : fitness

- Natural selection : Some changes are more important for survival or lineage propagation based on environmental and other factors : fitness fn selects some traits over others



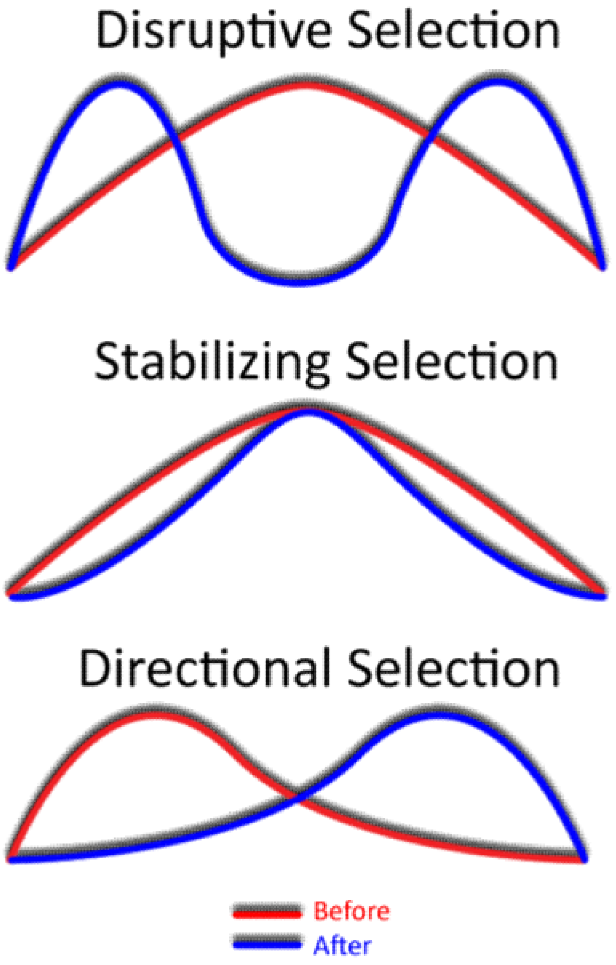
http://evolution.berkeley.edu/evosite/

Selection models : phenotypic selection

- Impossible to enumerate (aleph – 2 kinds !)

- Difficult to parameterize (curve fitting)

- Tricky to estimate (how many samples are enough depends on complexity of curve)



Wikipedia

Selection models: genomic selection

- At the phenotypic level : various kinds of selectional forces are at play
- At the genomic level : these translate to 3 basic kinds :
  – Positive selection : Advantageous changes are accelerated
  – Negative / purifying selection : Deleterious changes are discarded
  – Neutral selection : Selection plays no part

Facts about selection

- Any preference for one kind of change over another (simple eg: transition vs transversion)
- Operates at every granularity: nucleotide, codon, protein, etc
- Operates at both allelic and genotypic / haplotypic level
- Operates at every resolution: population, subspecies, species : one of the driving forces of speciation

A myth about selection

- Selection causes all but the fittest allele to vanish = myth (but it potentially could !)

— WHY? WE'LL DISCUSS THIS SHORTLY

- Think HKY 85 model:

$$Q = \begin{pmatrix} * & \kappa\pi_C & \pi_A & \pi_G \\ \kappa\pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \kappa\pi_G \\ \pi_T & \pi_C & \kappa\pi_A & * \end{pmatrix}$$

κ → some form of selection

Equilibrium distr. → $\pi_A \; \pi_C \; \pi_G \; \pi_T$

not $1 \quad 0 \quad 0 \quad 0$, etc.

Selectional models

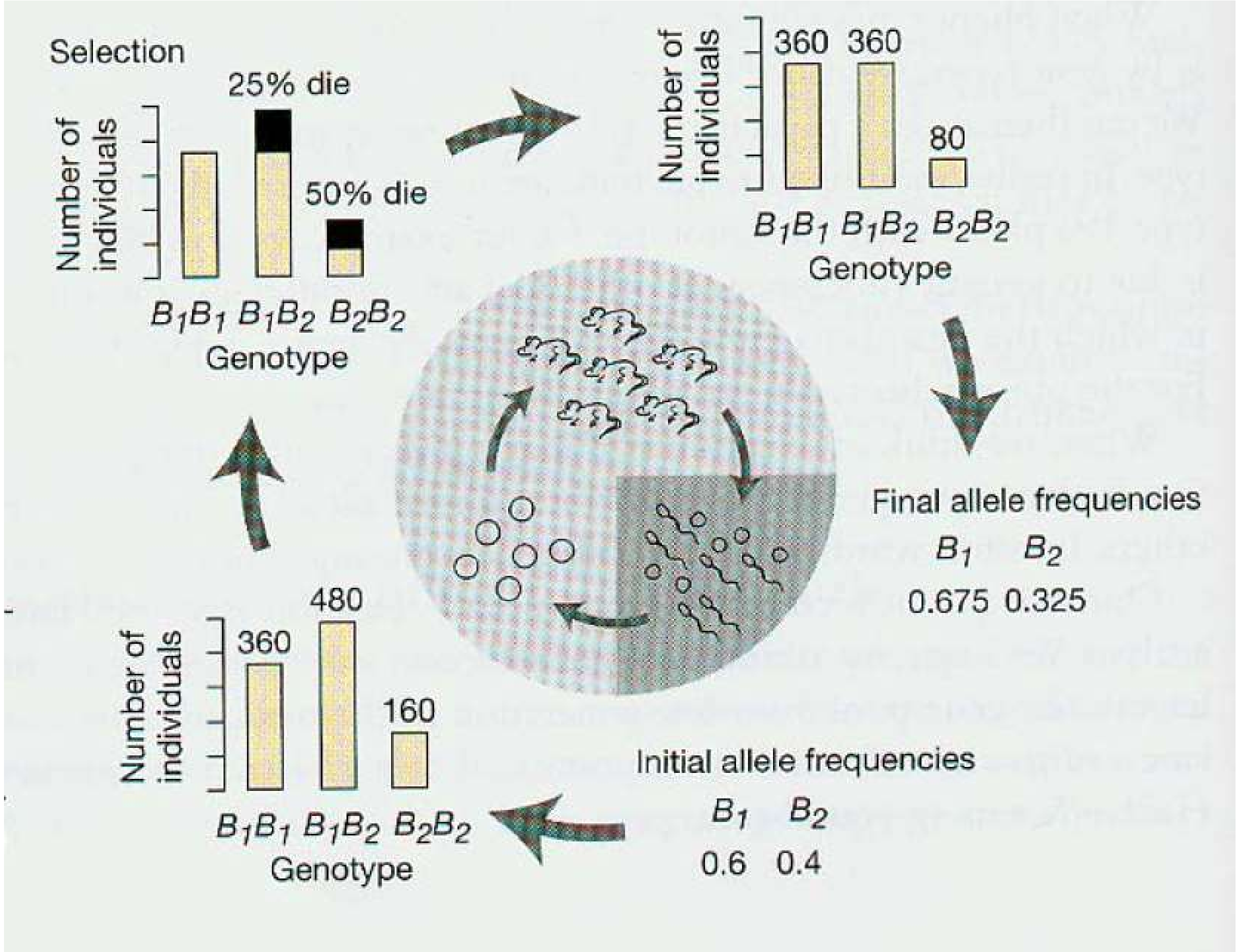
- Genotype based model
  - each genotype a_i has a different selectional coefficient s_i

↓

FACTOR THAT BIASES PROPNAL SAMPLING

- Can we incorporate our sampling bias into selectional models ?

How selection works



Pearson Prentice Hall

Modelling selection

- Fitness = expected no of offspring in the next generation
  - for the neutral model with fixed population, all genotypes have fitness = 1
- Various models of fitness
  - Dominant disease
  - Recessive disease
  - Heterozygous advantage
  - Directed selection

AA	Aa	aa
$1-s$	1	1
1	1	$1-s$
$1-t$	1	$1-s$
1	$1-hs$	$1-s$

LETHAL RECESSIVE
$\Rightarrow s = 1$

Drift : a cursory look

Drift

- What is drift
  - random, directionless fluctuations in allele frequency from generation to generation

- It is the act of randomly sampling a finite no of alleles from the previous generation
  - we don’t observe the whole population : can we incorporate fluctuations / errors due to our finite sample size into drift models ?

Parameterizing drift

- Drift is a result of finite sampling
  - sample size should be our parameter

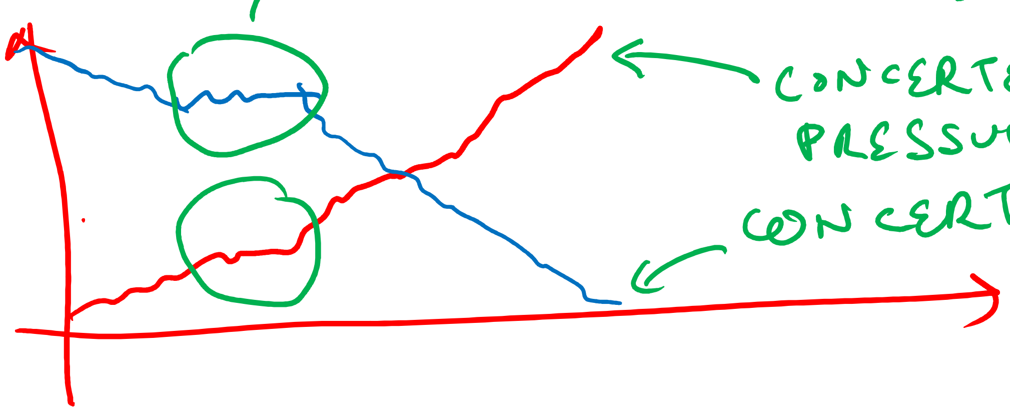
- What is our sample size ?
  - N individuals, 2N alleles
  - Population size is fixed : sample size = 2N
  - We will see later how to handle situations where population size is not fixed

Selection and drift

ROLE OF SELECTION: MORE PROMINENT IN LARGER POP.

- Directional pressure : $E(X_{t+1} - X_t) >, =, \text{or} < 0$
  - based on advantageous or deleterious allele
- Under only selection and drift :

IN SOME STRETCHES, DRIFT OVERSHADOWS SELECTION

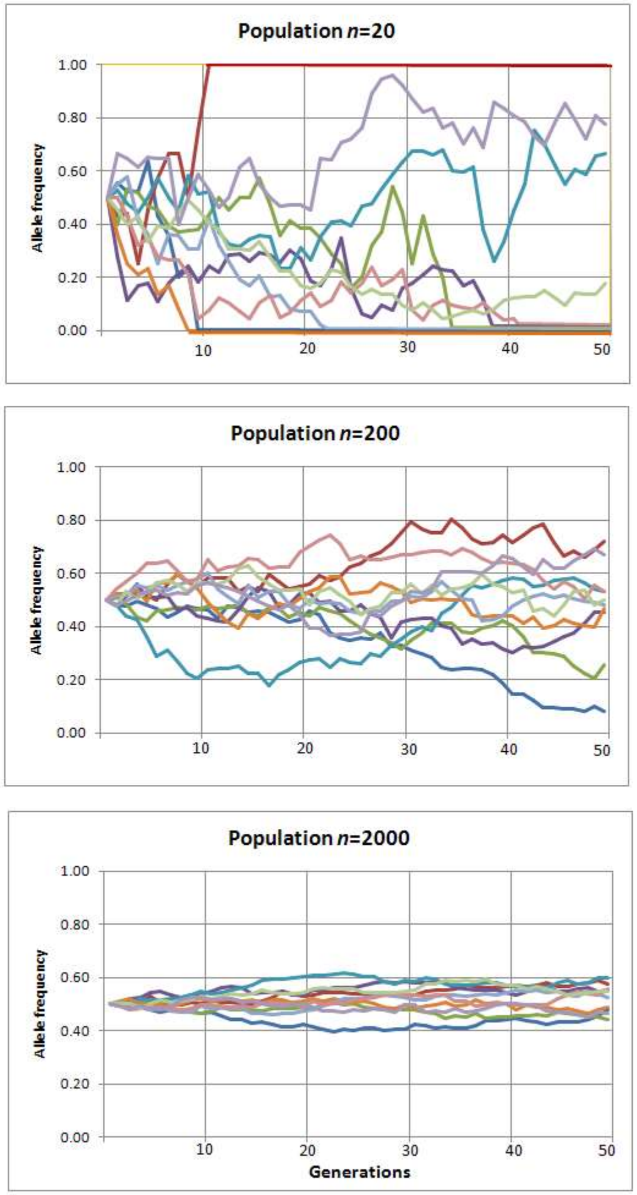


CONCERTED UPWARDS PRESSURE

CONCERTED DOWNWARDS PRESSURE

Population size

- Smaller population
  - greater chance of losing allele by drift alone : may undo evolutionary optimization by removing advantageous allele before selection can play a role
- Larger population
  - lesser role of drift, greater role of selection in variation reduction
  - greater no of mutations if rate is fixed

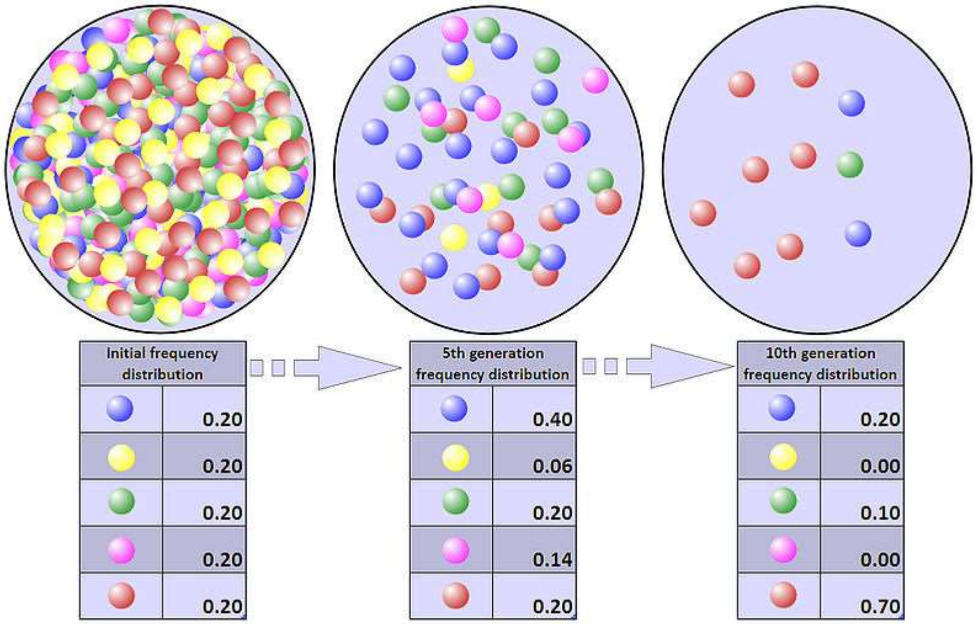


Wikipedia

Simulation under only drift

Population bottleneck

- Loss of alleles
- Alleles driven to irrecoverable frequencies : absorbed to 0
- Founder / bottleneck effects



	Initial frequency distribution	5th generation frequency distribution	10th generation frequency distribution
🔵	0.20	0.40	0.20
🟡	0.20	0.06	0.00
🟢	0.20	0.20	0.10
🟣	0.20	0.14	0.00
🔴	0.20	0.20	0.70

Wikipedia

In a world with only variation reduction

- Sooner or later, all but one allele will go extinct (which one may not be predictable : no stationary distribution)

- If the population size is infinite, then drift will have no role in the long term prospects of an allele [only the fittest allele will survive]

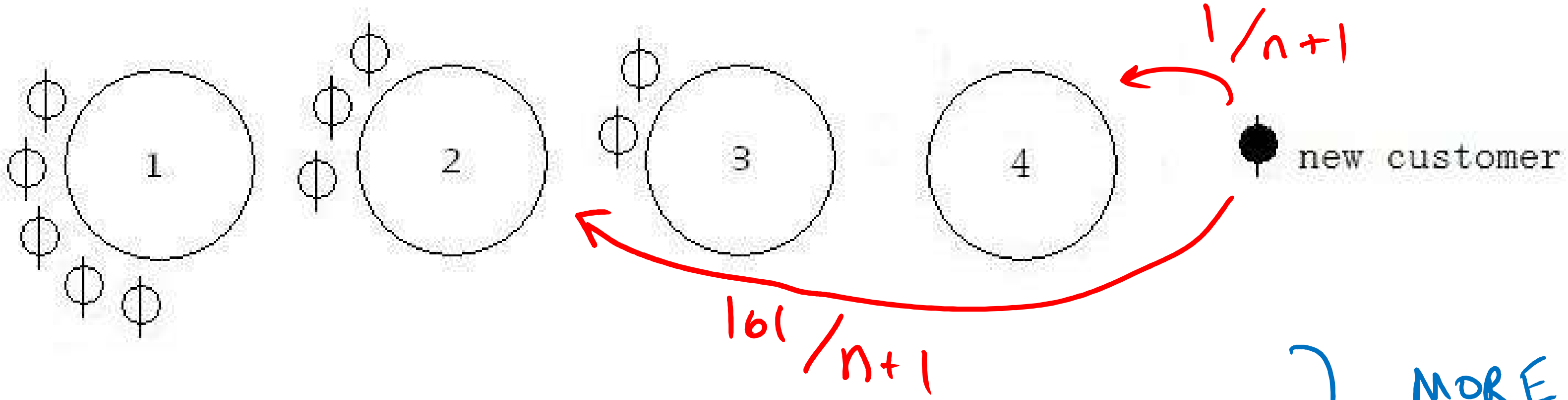
Static models

- Analyzing the allele frequencies at a particular time snapshot (one single generation's allele frequencies are modelled)

- Modelling allele diversity : how much diversity can we expect in a population ?
  - should it depend on the mutation rate ?
  - should it depend on the population size ?

Chinese restaurant process : modelling a diverse population

- Chinese restaurant process : infinite alleles : static or dynamic model ? $\Pr(B_n = B) = \dfrac{\prod_{b\in B}(|b|-1)!}{n!}$



byu.edu

OLD TBL

$$\frac{|b| - \alpha}{n + \theta}$$

NEW TBL

$$\frac{\theta + |B|\,\alpha}{n + \theta}$$

MORE GENERIC MODEL

Expected number of tables

- Relating allelic diversity with population size

$$\frac{\Gamma(\theta+n+\alpha)\Gamma(\theta+1)}{\alpha\Gamma(\theta+n)\Gamma(\theta+\alpha)}-\frac{\theta}{\alpha}.$$

- Looks at a particular snapshot in size, when population = n
    - Whether static or dynamic depends on your question : are you modelling evolution of no of alleles ? Evolution of population in terms of allele frequency ?

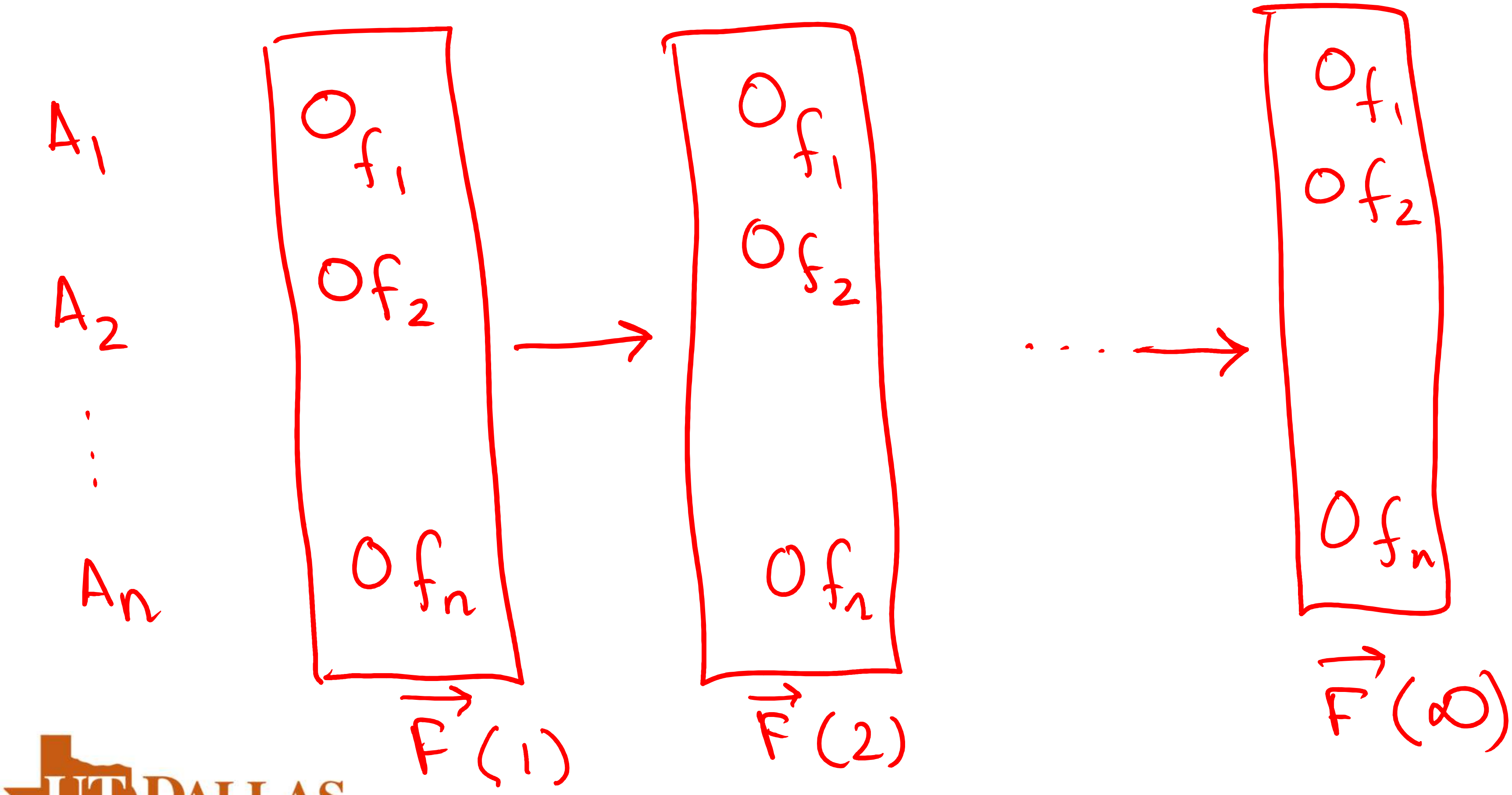
Dynamic model : multiple snapshots in time

- How do populations change at each table as more members join , and as a function of the number of tables ?

- Will the largest table will stay the largest after a doubling of the population ?

Standard dynamic pop genetic models

- Allele frequency models (absolute or relative)



Variations on the model

- 2 near-“absorbing states” : 0 and 1
  - not truly absorbing if mutation can occur
- Continuous valued
  - relative frequencies of alleles modelled
- Discrete valued
  - Granularity of relative frequencies determined by population
- Discrete time : generations
  - overlapping generations : eg human
  - non overlapping generations : eg annual plants

  [why not use generations in phylogenies ?]
- Continuous time : brownian motion, diffusion process

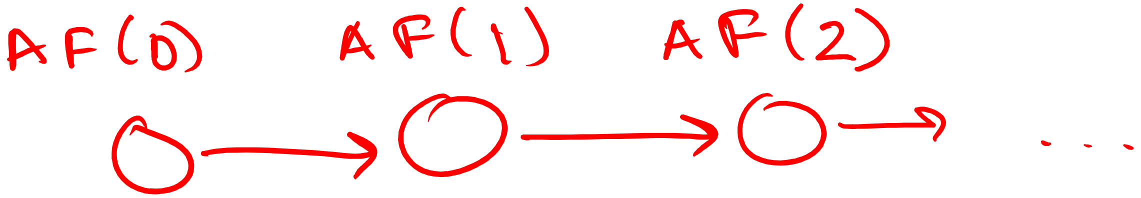
Standard pop genetic models

- Models may not necessarily be Markovian

- Imagine a situation where an individual's fecundity is bounded
  - the next generation's allele frequencies will also depend on the history of the population, not just the current allele frequencies

Standard pop genetic models

- What should Markovian models look like ?

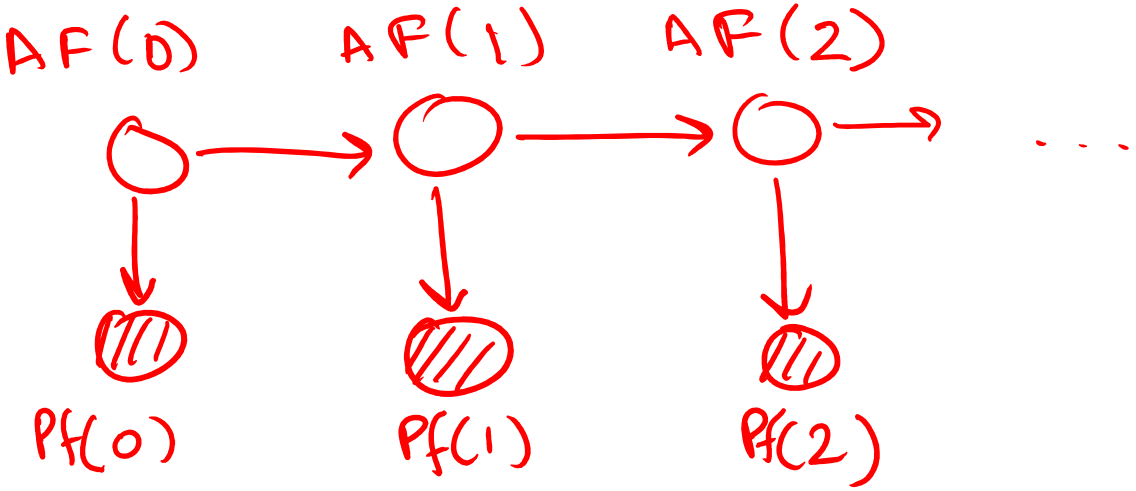


Discrete: time unit = generations

- Bayesian networks are a good way to visualize the dependencies

Standard pop genetic models

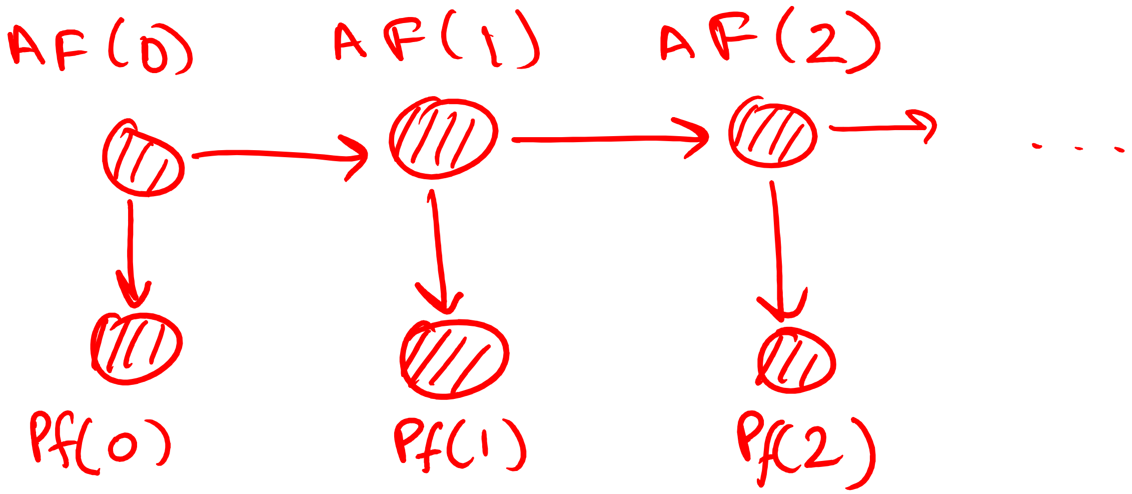
- What is observed ?



  - in the old days, phenotype …

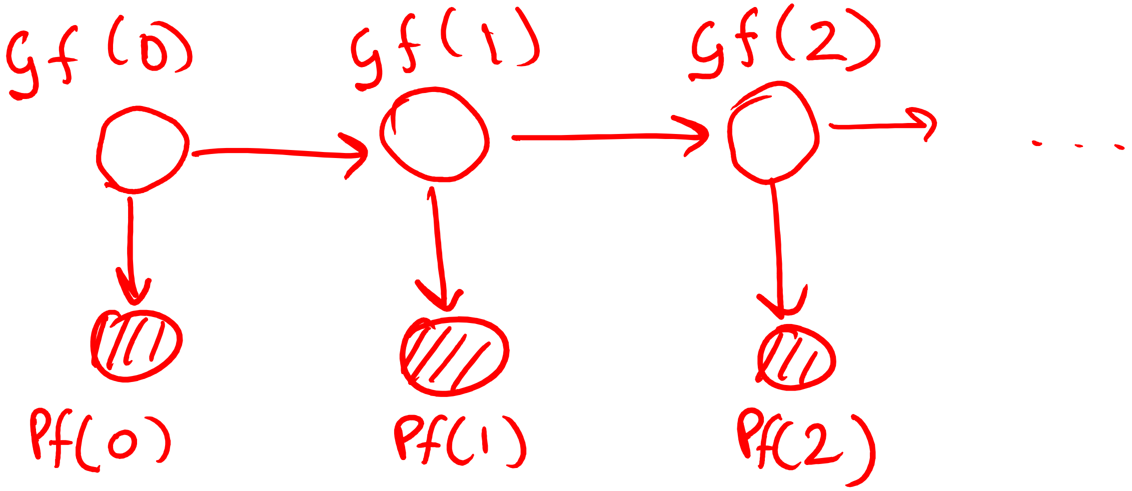
Standard pop genetic models

- What is observed ?



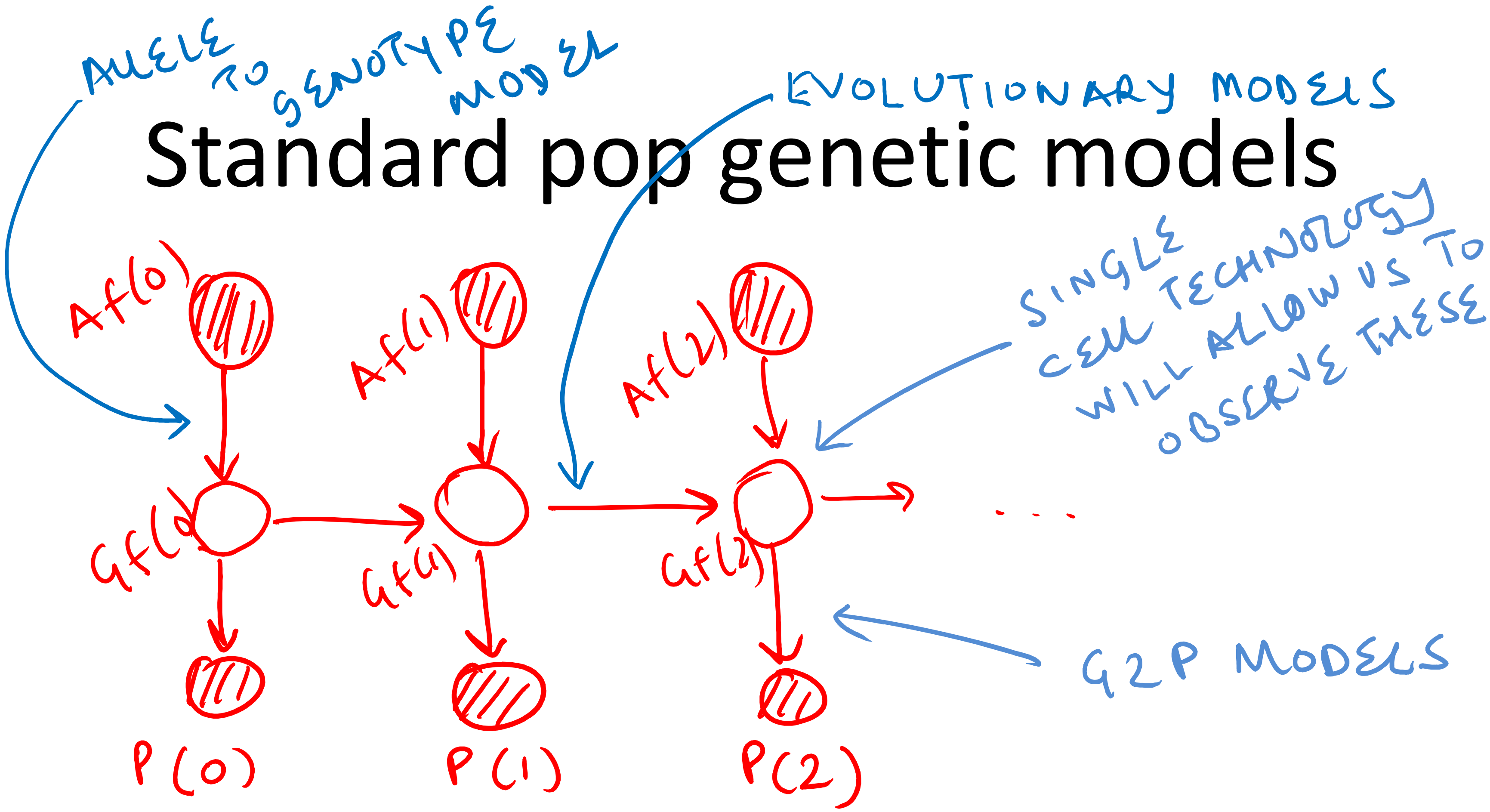
  - these days, we observe both …

Standard pop genetic models



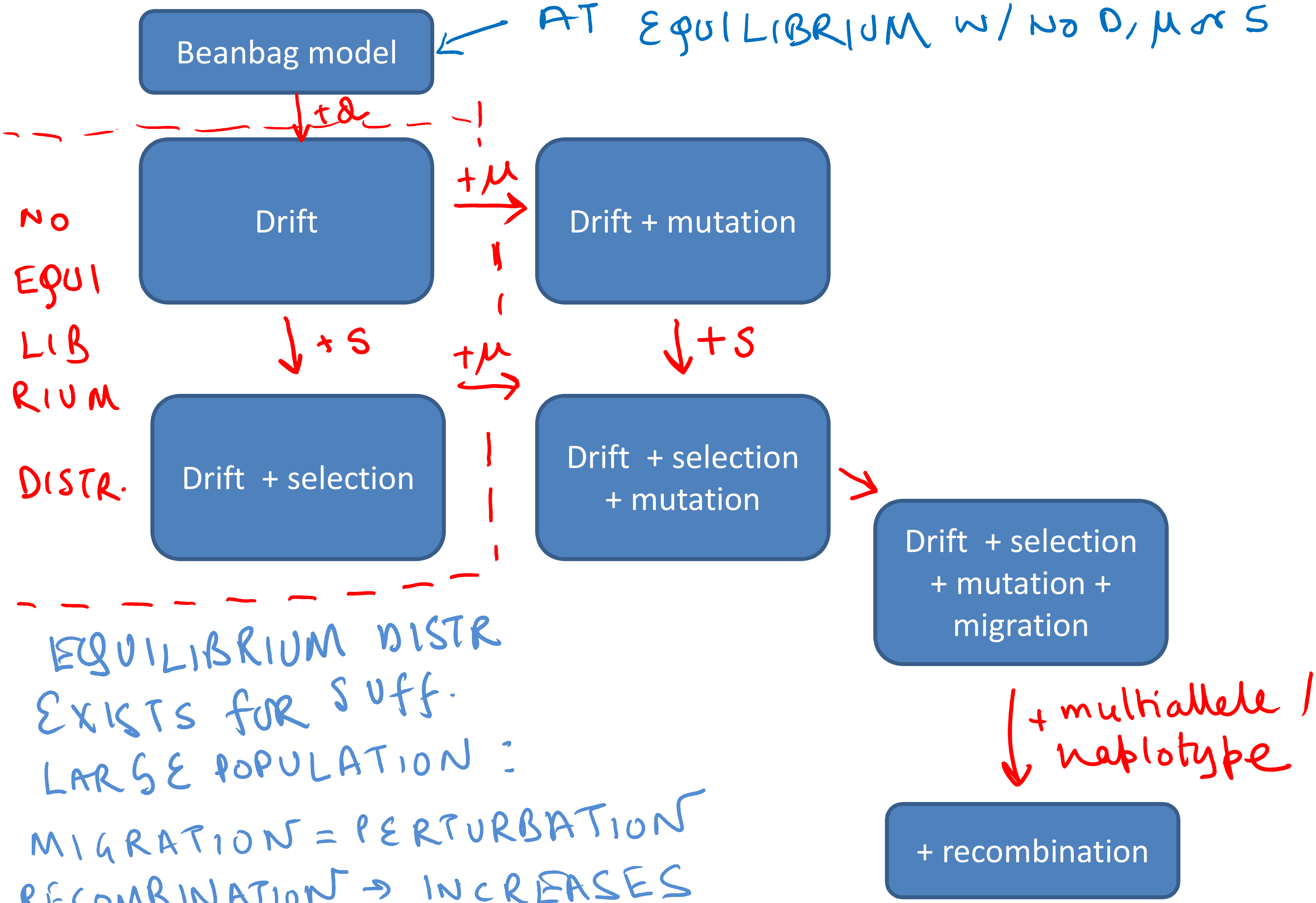
- Evolution is really operating at the level of genotype / haplotype

Standard pop genetic models



- Evolution is really operating at the level of genotype / haplotype
- A2G models : census of alleles, but we need genotype

AT EQUILIBRIUM w/ NO D, μ or S



Beanbag model

Drift

Drift + mutation

Drift + selection

Drift + selection + mutation

Drift + selection + mutation + migration

+ recombination

EQUILIBRIUM DISTR EXISTS FOR SUFF. LARGE POPULATION:

MIGRATION = PERTURBATION

RECOMBINATION → INCREASES VARIATION IN MULTIALLELIC MODEL

Classic idealized pop models

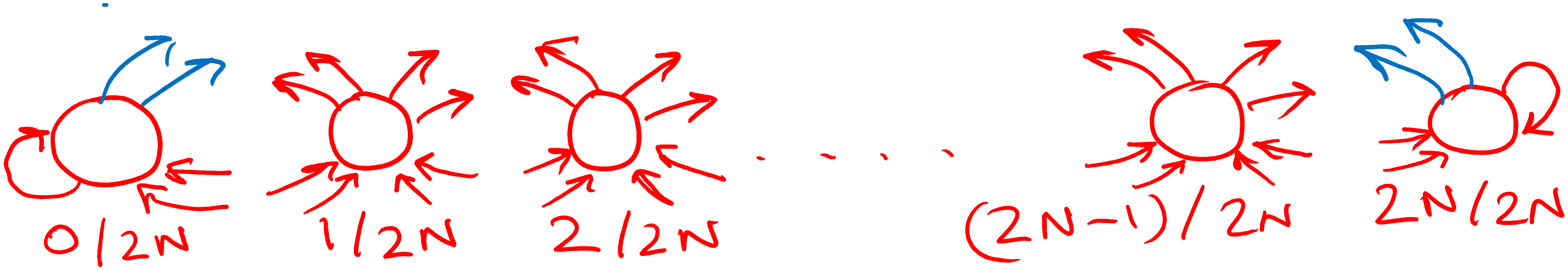
- Wright – Fisher model :
  - Mendelian inheritance
  - no sexual selection (allele frequencies don't differ in sexes)
  - no overlapping generations (in the most complicated deviation from model, we use continuous time)
  - sex ratio = 1
  - effective population = actual population
  - fixed population
  - no selection
  - finite population (discrete valued stoch process)
- Other popular models may trade off such oversimplification for less tractable inference

find APPROXIMATIONS UNDER VIOLATIONS TO THESE ASSUMPTIONS LATER

we'll violate these soon!

State space diagram under the Wright - Fisher model

- More connectivity than a random walk
- Bi allelic state space for finite population

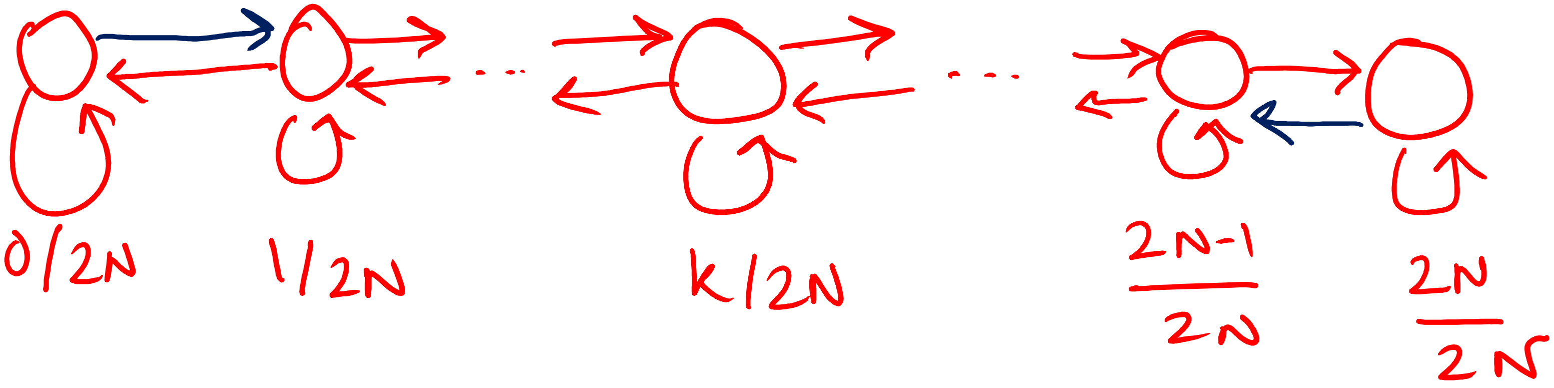


BLUE TRANSITIONS ALLOWED IF MUTATION IS MODELLED

Classic idealized pop models

- Other models : Moran model
  - overlapping generations
  - one or two individuals (depending on evolutionary model) selected to reproduce by sampling
  - new individual created and added to pool
  - one individual is killed to retain constant population

State space diagram for the Moran model

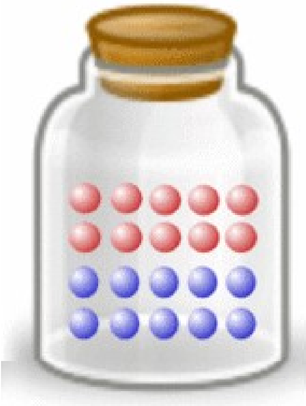


RANDOM WALK!

BLUE EDGES ALLOWED ONLY IF MUTATION IS MODELLED

Urn models of pop evolution

- George Polya
- Notion of sampling a non-homogeneous population
- Variations on a theme
  - Without replacement / with replacement / with replacement and duplication / new urn

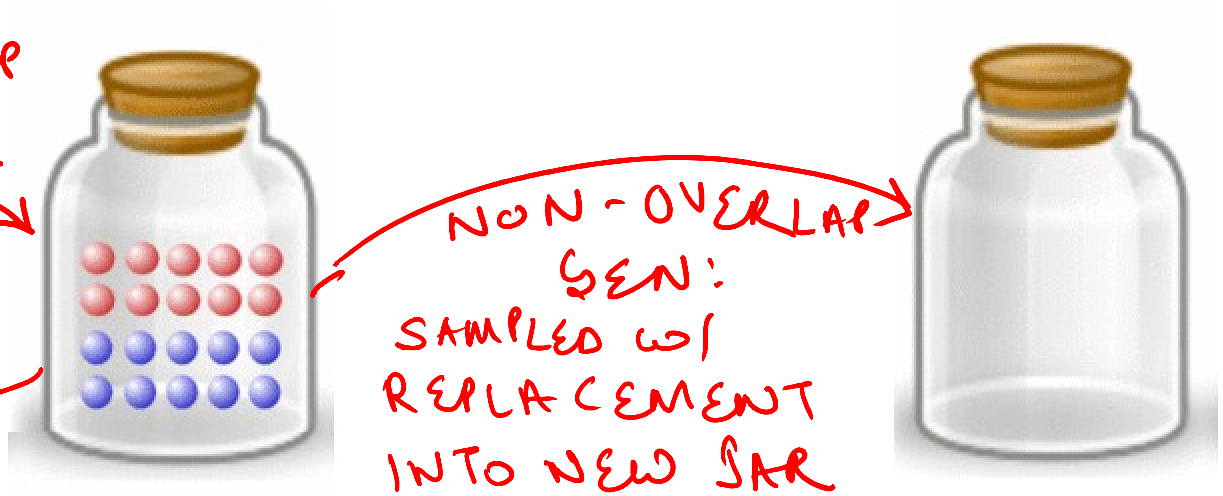


wikipedia

Carnegie Foundation

Analogous action on urns

- Overlapping / non overlapping generations



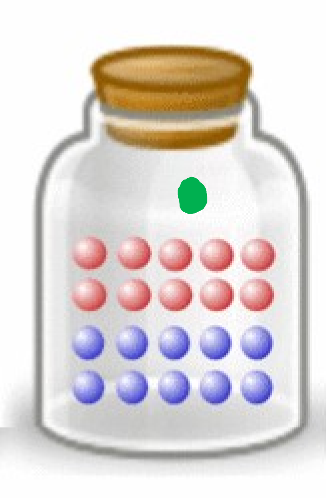
OVERLAP GEN: PUT INTO SAME JAR W/ DUPLICATION

NON-OVERLAP GEN: SAMPLED W/ REPLACEMENT INTO NEW JAR

← DEATH: SAMPLE WITHOUT REPLACEMENT

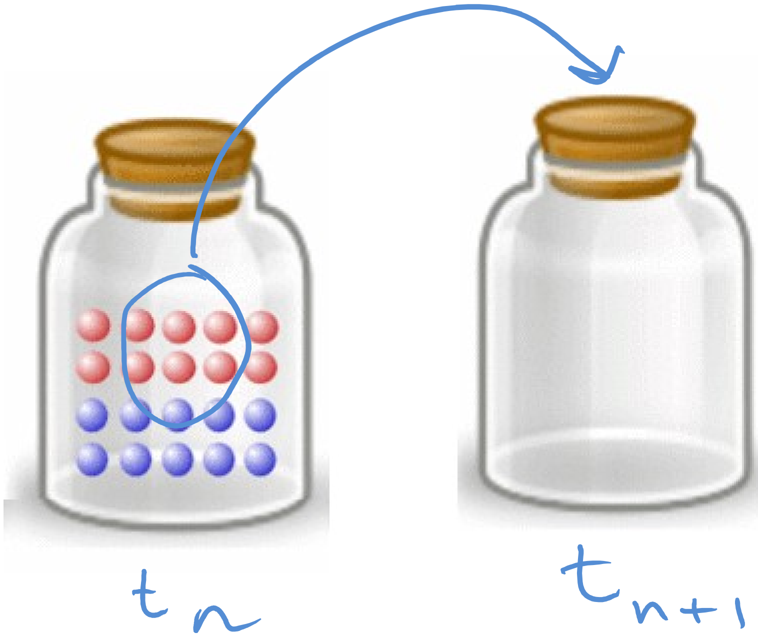
Analogous action on urns

- Mutation
  - new color : infinite alleles model
  - under no such assumption, the duplication merely causes the ball to change color



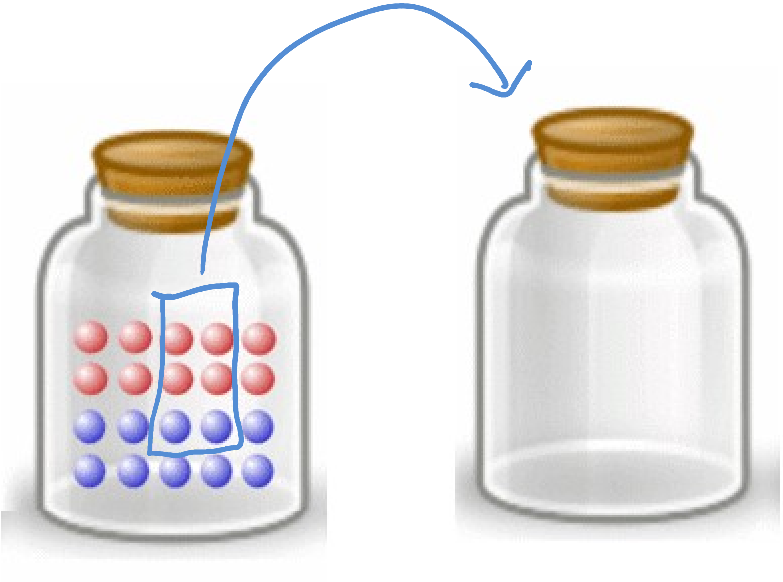
Analogous action on urns

- Drift: Random fluctuations in frequency from generation to generation : it is the act of the random sampling
  - how is death implicitly modelled in W F models ?



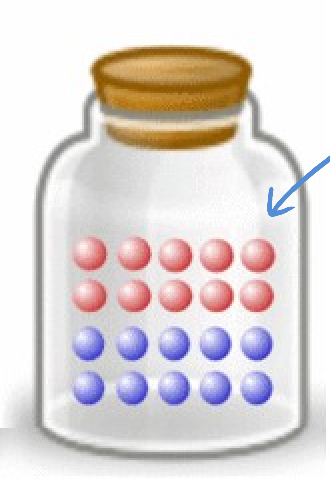
Analogous action on urns

- Selection : Sampling is disproportionate to no of balls of each color



Analogous action on urns

- Population effects : the effects of a finite sized jar



balls are indivisible during sampling

Which model should I choose ?

- Depends on your need
  - modelling Drosophila with overlapping generations (Moran model)
  - modelling populations where non overlapping generations are a good approximation (W – F model : we will be studying this from now on)
  - more complicated situations (diffusion model : we will study the basics of the diffusion process at the end of the lecture)

Bean bag models & deviations

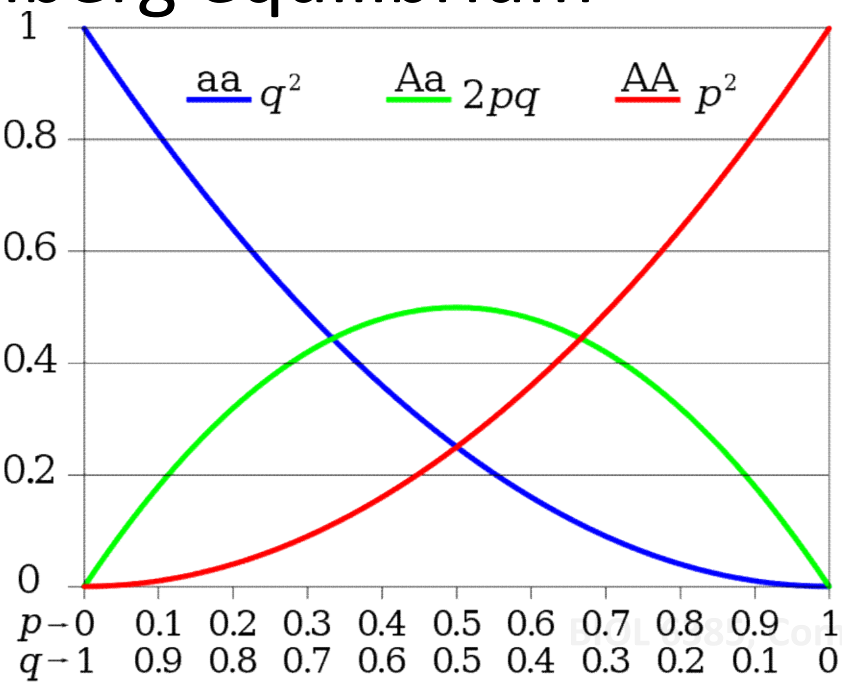
Null model of population genetics

- Simplest dynamic model of allele frequencies
- Assumption : allele (and thus genotype) frequencies are in equilibrium
  - no forces at work : drift, selection, mutation, etc

- Beanbag model : alleles only move around like beans in a bean bag from gen to gen
  - even without mutation or selection, can really happen only in an infinite (v large) population where drift doesn't affect allele frequency

Genotype frequencies

- Why model genotype frequencies ?
  - remember, traits determined by genotypes (selection acts on genotypes)
- Hardy Weinberg equilibrium



$p + q = 1$

Why ?

- Alleles segregate independently
- Even if you start with genotype frequencies not at HWE, in one generation genotype frequencies will be at HWE
- Allele frequencies are the same for both sexes (no sexual selection) : could allele frequencies be different in the two genders at HWE ?

Key properties

- Allele frequency is not affected by alleles segregating into different genotypes

- For pure dominance models, allele frequencies can be estimated from phenotype frequencies

Sex linked alleles

- p & q for heterogametic sex
- p^2, 2pq and q^2 for homogametic sex

CENSUS: $x_1 \rightarrow A$, $x_2 \longrightarrow a$

freq. of sex-linked allele.

At equilibrium, what are the diff. genotypes and their frequencies?

HWE : multiple alleles

CONSIDER: $(p_1V_1+p_2V_2+\cdots+p_nV_n)(p_1V_1+p_2V_2+\cdots+p_nV_n)$

→ what is the coefficient of V_iV_j ?

$P\{A_i$ observed x_i times, A_j observed x_j times$\}$

$$= p(A_iA_j) = C\,p_ip_j \quad \left[C = \frac{(x_i+x_j)!}{x_i!\,x_j!}\right]$$

$\therefore$ $i\neq j \Rightarrow p(A_iA_j)=2p_ip_j$, $i=j \Rightarrow p(A_iA_i)=p_i^2$

FOR HIGHER PLOIDY, THE COEFF. OF $\{A_i$ observed x_i times for $i=1,2,\ldots,m\}$

COUNTING $\rightarrow \left(\dfrac{x!}{x_1!\,x_2!\cdots x_m!}\right)$

Ewens sampling formula : distributions over partitions

- Finite sample size

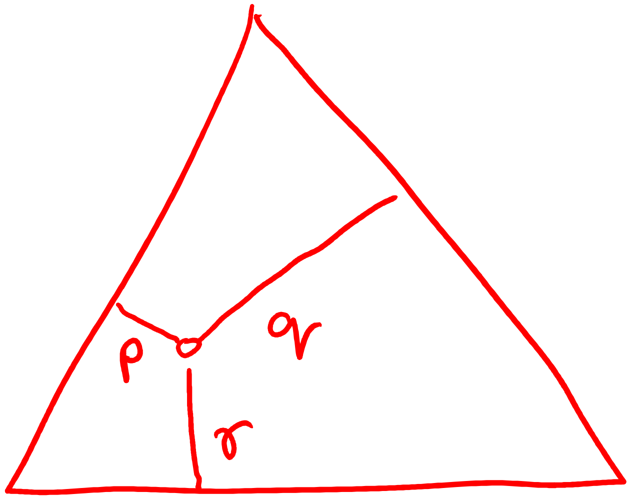
$$\Pr(a_1, \ldots, a_n; \theta) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{a_j}}{j^{a_j} a_j!},$$

No selection

- Role of theta

Space of allele frequencies

The simplex for 3 alleles

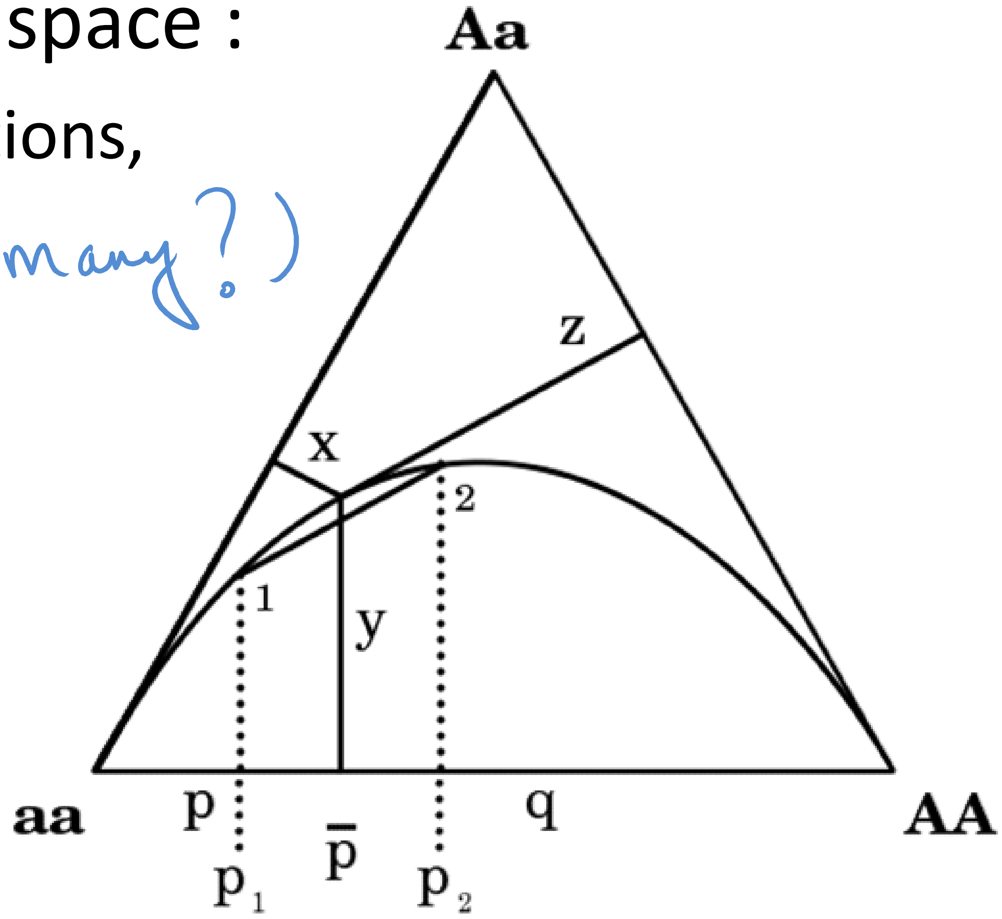


$p+q+r=1$

Corresponding genotype space

- Higher dimensional space :
  - choose 2 from k options, with repeats (How many?)

- Under HWE :
  - constrained space



Test for HWE : categorical tests

- Measuring deviations in the simplex

	AA	aa	Aa
OBSERVED	n_{AA}	n_{aa}	n_{Aa}
ESTIMATED	$p^2 \cdot n$	$(1-p)^2 n$	$2p(1-p) n$

$$n = n_{AA} + n_{Aa} + n_{aa}$$

- Perform a categorical test : are the 2 rows drawn from the same distribution ?
  - eg chi square (degrees of freedom = no of genotypes – no of alleles)

So, we arent in HWE, …

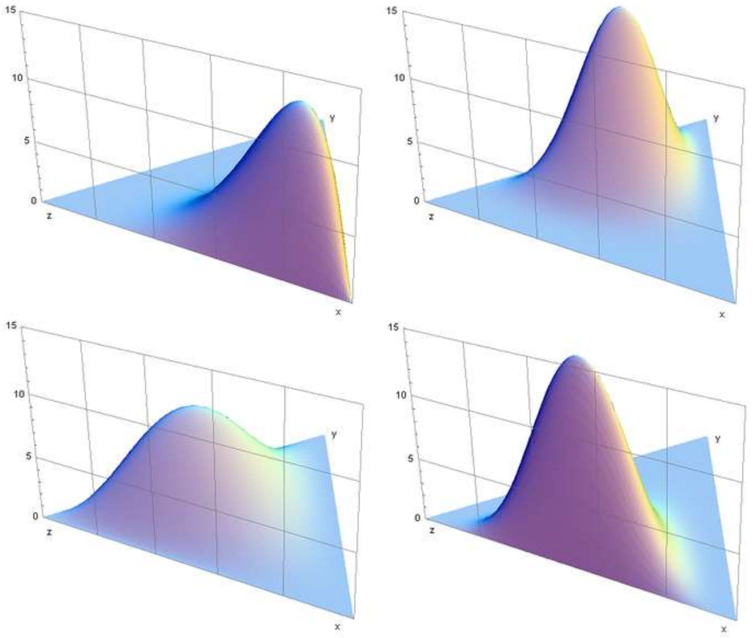
- what next?

- A categorical test only tells us whether the population is under HWE, doesn’t tell us the likelihood of observing a non-HWE equilibrium

- Without explicitly modelling the different kinds of forces, we may put a prior over the space of genotype frequencies, based on sampling or prior knowledge

Dirichlet distributions : non-HWE equilibrium models

- For HWE violations, we want to move away from the (p^2, 2pq, q^2) parameterization
- “Pushing” the point on the simplex to a region of the simplex : Dirichlet distrn



wikipedia

$$P(\vec{x}|\alpha) = \int P(\vec{x}|\theta)\, P(\theta|\alpha)\, d\theta$$

$$= \int \prod_{j=1}^{m} \theta_j^{N_j(\vec{x})} \left(\frac{1}{c(\alpha)} \prod_{j=1}^{m} \theta_j^{\alpha_j - 1} \right) d\theta$$

$$= \frac{1}{c(\alpha)} \int \prod_{j=1}^{m} \theta_j^{N_j(\vec{x}) + \alpha_j - 1}\, d\theta$$

$$= \frac{c\left(N(\vec{x}) + \alpha\right)}{\alpha}, \text{ s.t.}$$

$$c(\beta) = \int \prod_{j=1}^{m} \theta_j^{\beta_j - 1}\, d\theta = \frac{\prod_{j=1}^{m} \Gamma(\beta_j)}{\Gamma(\beta_s)}, \quad \beta_s = \sum_{j=1}^{m} \beta_j$$

$$P(X_{n+1} = k \mid X_{1:n}, \alpha) = \frac{N_k(X_{1:n}) + \alpha_k}{n + \alpha_s}$$

Moving on from the bean-bag model

- We want to model the evolutionary dynamics of the allele frequencies
  - population may not be in equilibrium : we may want to characterize the trajectory such populations take towards their long run configurations
  - even if it is in equilibrium, we may want to find out the nature of forces (mutation, drift, selection) acting on it

Simplest violation of the bean bag model

- Allele populations remain in equilibrium if sampling can be done faithfully at every generation

- Finite populations
  - finite sampling comes up with distribution with errors wrt original population distribution

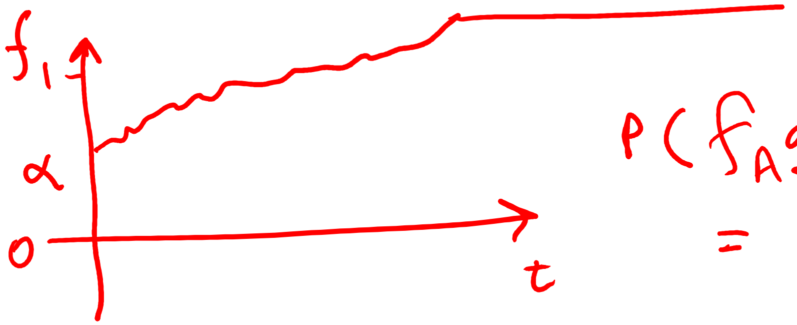
Drift only models

Drift only model : long run

- No stationary distribution
  - What is the probability that drift will fix one allele and not the other : probability of fixation ?
  - Proportional to the relative frequencies
  - Remember gambler's ruin and the random walk

PROB OVER MANY RUNS

WE WILL DISCUSS THIS AGAIN WHEN STUDYING COALESCENTS



$P(f_A \text{ goes to } 1)$
$= \alpha$

Drift only model: heterozygosity

- With no mutation, identity by descent
- H_t = Pr of picking two different alleles in the population at time t
- For bi allelic model, $H_0 = 2\, x_0\, (\, 1 – x_0\,)$

$$E_{x_1 | x_0}(H_1) = \sum_{x_1} 2 x_1 (1-x_1) P(X_1 = x_1 | X_0 = x_0)$$

$$= H_0 \left(1 - \frac{1}{2N}\right)$$

expectation over what?

$$\longrightarrow E(H_t) = H_0 \left(1 - \frac{1}{2N}\right)^t$$

Allele diversity retained only for N → ∞

Variance in the sampling process

$$\Delta x = x_t - x_{t-1}$$

$E(\Delta x) = 0 \Leftarrow$ TRULY RANDOM, NOT DIRECTIONAL

$V(\Delta x) \uparrow$ as $N \downarrow$ $\Leftarrow$ Effect of DRIFT INCREASES WITH SMALLER POPULATION

$$V(\Delta x) = \frac{x(1-x)}{2N}$$

$2N$ pop. → allele pop: K_t , $2N - K_t$

rel. freq: x , $1 - x$

$$K_{t+1} \sim \text{Bin}(2N, x)$$

$$V(\Delta x) = V(x_{t+1} - x_t)$$

observed

under $p(X_{t+1} | X_t)$

$$= V(x_{t+1} - \text{const.}) = V(x_{t+1})$$

$$= V\left(\frac{K_{t+1}}{2N}\right)$$

$$= \frac{1}{(2N)^2} V(K_{t+1})$$

$$= \frac{1}{(2N)^2} \cdot \underbrace{2N \cdot x \cdot (1-x)}_{\uparrow \text{Var. of binomial}}$$

$$= \frac{x(1-x)}{2N}$$

Modelling more complex directional change

- Under drift, expected value of change in allele frequency in one generation = 0

- However, empirically, we know that allele frequencies show directed change : selection

Selection only models

Selection only models

- N is assumed to be large : drift has little effect
- Usually variation decreasing force

- But, what happens if the heterozygous allele has maximum fitness ?
  - variation increasing force : if variation is thought of as degree of heterozygosity of the population

- How to model changing allele frequencies ?

Selection only models

CHANGES EVERY GEN IF ALLELE FREQ IS NOT AT EQUIL.

- Important notion : mean fitness of population

	AA	Aa	aa	$\bar{\omega}$
fitness	ω_{AA}	ω_{Aa}	ω_{aa}	$= x^2\omega_{AA} + 2x(1-x)\omega_{Aa} + (1-x)^2\omega_{aa}$
Orig. freq.	x^2	$2x(1-x)$	$(1-x)^2$	
Expected no. of offspring	$\omega_{AA}x^2$	$\omega_{Aa}2x(1-x)$	$\omega_{aa}(1-x)^2$	
freq of genotype in next gen.	$\frac{x^2\omega_{AA}}{\bar{\omega}}$	$\frac{2x(1-x)\omega_{Aa}}{\bar{\omega}}$	$\frac{\omega_{aa}(1-x)^2}{\bar{\omega}}$	

Change in allele frequencies

$x' \leftarrow$ allele freq. in next gen.

$$x' = \left[2x^2 \cdot \frac{\omega_{AA}}{\bar{\omega}} + 2x \frac{(1-x)}{\bar{\omega}} \omega_{AB} \right] \times \frac{1}{2}$$

$\rightarrow$ No. of alleles = 2 No. of genotypes

$$= \frac{x}{\bar{\omega}} \left(x\,\omega_{AA} + (1-x)\,\omega_{AB} \right)$$

$$= x \frac{\bar{\omega}_A}{\bar{\omega}}$$

$\leftarrow$ Mean fitness of allele A

$$\Delta x = x' - x = \frac{x}{\bar{\omega}} \left(\bar{\omega}_A - \bar{\omega} \right) = \frac{x(1-x)(\bar{\omega}_A - \bar{\omega}_B)}{\bar{\omega}}$$

Adding to the mix

- So far, we have modelled how existing alleles compete with each other over generations
- But how do these different alleles get created ?
- Modelling the primary driving force of polymorphism : mutations

Mutation only models

Frequency of mutation

- Mutation rate : no of de novo mutations as a fraction of the total population

Mutation rate = μ

fraction of population with mutation after 1 gen = μ

after n gen = $1 - (1-\mu)^2$

- Does this really happen ?

No, selection & drift drive most mutations to 0 frequency (Loss)

Mutation only models

- May not be very interesting to look at long run probabilities : no drift or selection to balance allele frequencies at equilibrium : no drift or selection to drive them to loss or fixing
- Instead, more realistic models will try to model mutation in a setting where alleles are lost due to drift or selection

Drift – mutation models

Fixation prob of neutral mutation

- Initial prob of novel mutation = 1 / (2N) (no two mutations are same under infinite alleles model)
- Remember drift only models : the probability of fixing this mutation would be its starting relative frequency = 1 / (2N) ← INFINITE ALLELE MODEL
- Under infinite sites model, fixation rate of *any* new mutation (from a generation) :

$$2N\mu \times \frac{1}{2N} = \mu$$

(INDEP. OF POPULATION !)

Mutation with drift : neutral model

for driftonly model, $G_t = 1 - H_t$ [ASSUME INFINITE ALLELES]

$$G_{t+1} = \frac{1}{2N} + \left(1 - \frac{1}{2N}\right)G_t$$

With a mutation rate μ [FRACTION of ALLELES UNDERGOING MUTATION IN 1 GEN]

$$G_{t+1} = (1-\mu)^2 \left[\frac{1}{2N} + \left(1 - \frac{1}{2N}\right)G_t\right]$$

At equilibrium, $G_{t+1} = G_t$ & $\mu \ll N$

$$G_{eq} = \frac{1}{1+4N\mu}, \quad H_{eq} = \frac{4N\mu}{1+4N\mu}$$

Problems !

- Can we estimate N (population) if we know mutation rate and heterozygosity
  - heterozygosity : sample population
  - mutation rate = substitution rate (why ? later …)
- We get wrong answers for N using well established data sets for humans (we get N = 6000) and Drosophila (we get N = 200,000)
  - why ? Real populations may not be following W – F model

Effective population size

- Ne may be much less than N
- How to estimate Ne

- Variance in no of offspring

$$N_e = N/\sigma^2$$

- Fluctuating population

$$N_e = \frac{1}{\frac{1}{t}\sum_i \frac{1}{N_i}}$$

← harmonic mean (affected by min., models pop. bottleneck)

- Gender skew

$$N_e = \frac{4N_mN_f}{N_m + N_f}$$

SIMPLE WAY TO CHECK: SIMULATE UNDER WRIGHT FISHER

What is N_e anyway ?

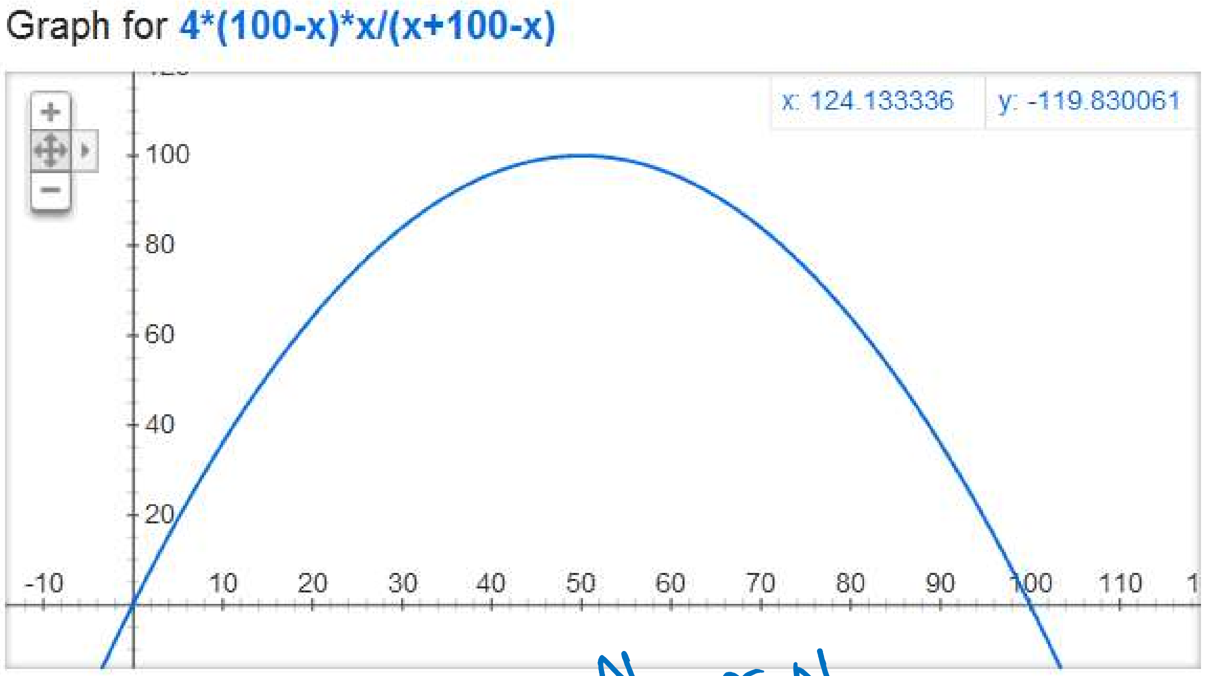
- It is the population of an ideal W-F model that would approximate the population dynamics of the current population under study.

Gender imbalance and effective population

- For organisms with matriarchal or patriarchal clans, the approximation should be different

N_e

$= \frac{4N_mN_f}{N_m+N_f}$



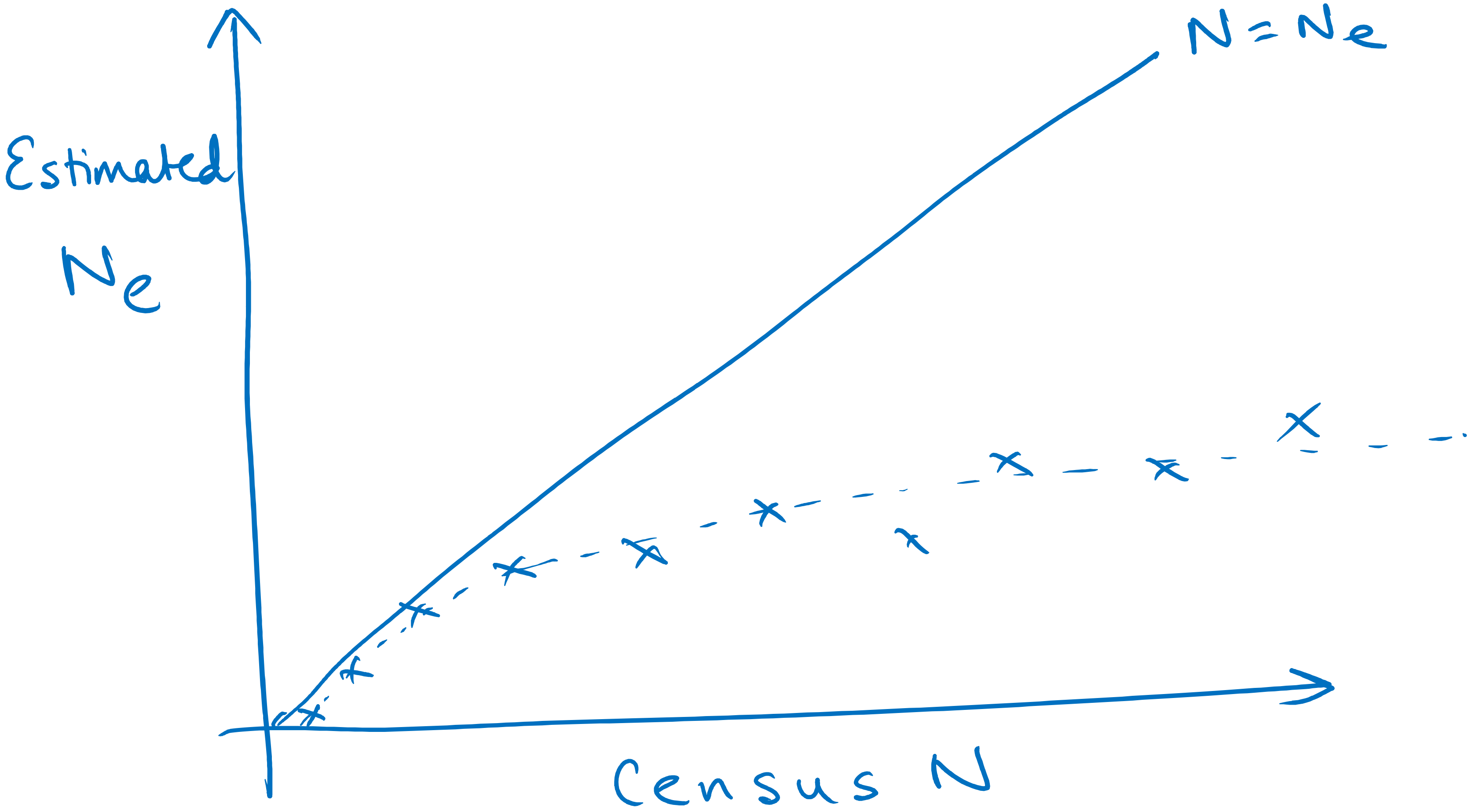
Graph for 4*(100-x)*x/(x+100-x)

x: 124.133336 y: -119.830061

N_f or N_m

Remember, $N_m+N_f = N$

N versus Ne



Mutation – selection models

Continuous allele frequencies

- From now on, we will consider that allele frequencies can be modelled as continuous
- We can now take derivatives wrt the allele frequency !

Rate of change of allele frequencies

- As N is assumed to be large, allele frequencies can be modelled as continuous. Derivatives wrt x can be taken : CONTINUOUS VALUED, DISCRETE TIME MODEL

$$E_s[\Delta x] = \frac{x(1-x)}{2\bar{w}} \frac{d\bar{w}}{dx}$$

Max selectional force for intermediate allele freq.

$\bar{w}$ → mean fitness

allele frequency increases if allele increases pop. fitness

Rate of change of allele frequencies

$$\frac{d\bar{\omega}}{dx} = \frac{d}{dx}\left[x^2 \omega_{AA} + 2x(1-x)\,\omega_{AB} + (1-x)^2\,\omega_{BB}\right]$$

$$= 2x\,\omega_{AA} + 2(1-x)\,\omega_{AB} - 2x\,\omega_{AB}$$
$$- 2(1-x)\,\omega_{BB}$$

$$2\left(\bar{\omega}_A - \bar{\omega}_B\right) = 2\left(x\,\omega_{AA} + (1-x)\,\omega_{AB}\right.$$
$$\left. - x\,\omega_{AB} - (1-x)\,\omega_{BB}\right)$$

$$= \frac{d\bar{\omega}}{dx}$$

Selection – mutation balanced model

	GENOTYPE		
	AA	Aa	aa
fitness	1	1	1-s
Before selection	p^2	$2pq$	q^2
After selection (UN NORMALIZED)	p^2	$2pq$	$q^2(1-s)$

RECESSIVE DISEASE MODEL (MULT SCLEROSIS)

How "s" can be embirically calculated in F_2

Selection – mutation balanced model

- Balancing of allele frequency by mutation and selection
  - Why is drift not considered here ?

<u>Small q</u>

$$E_s(\Delta q) = \frac{q(1-q)}{2\bar{w}} \frac{d\bar{w}}{dq} \approx -sq^2(1-q) \approx sq^2$$

$$E_m(\Delta q) = \mu(1-q) \approx \mu$$

EQUIL: $E_s(\Delta q) = -E_m(\Delta q)$

$$q_{EQUIL} = \sqrt{\frac{\mu}{s}}$$

Working this out for other models

- Can be worked out for other sets of selection coefficients
  - eg, another model

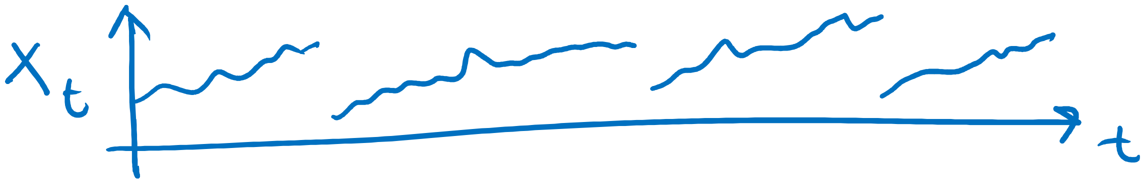
AA	Aa	aa
1-s	1	1

$$q_{EQUIL} = \frac{\mu}{s}$$

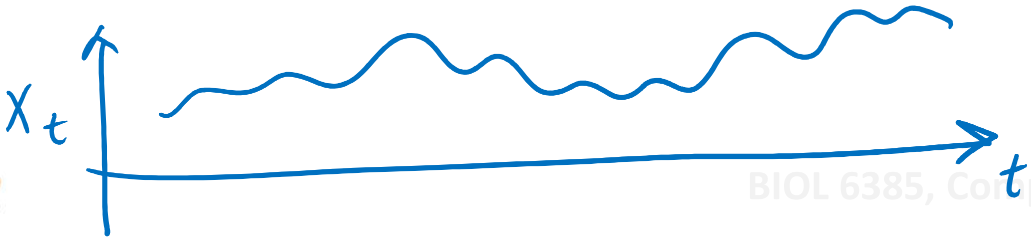
Drift – mutation – selection models

Continuous-valued, continuous-time stochastic processes

- Continuous valued, continuous time processes :
  - discontinuous in time/jump : sample paths discont



  - many notions of continuous in time : Sample – continuous : all sample paths *almost surely* continuous (eg diffusion process)



Diffusion process

CHAPMAN KOLMOGOROV

$$Pr(X_t = a) = \sum_b Pr(X_{t-k} = b)\, Pr(X_t = a \mid X_{t-k} = b)$$

FOR US:

$$Pr(X_t = i) = \sum_{\Delta i} Pr(X_{t-1} = i + \Delta i)\, Pr(X_t = i \mid X_{t-1} = i + \Delta i)$$

(FORWARD KOLMOGOROV EQN)

for CONTINUOUS-VALUED, CONTINUOUS-TIME PROCESS:

$\phi(x;t)$ → density fn for observing allele frequency x at time t

$\boxed{g(x-\epsilon, \epsilon, \delta t)}$ → Prob. of going from $x-\epsilon$ to $(x-\epsilon)+\epsilon$ in time δt

EMBODIES THE PROCESS

$$\phi(x; t+\delta t) = \int \phi(x-\epsilon; t)\, g(x-\epsilon, \epsilon, \delta t)\, d\epsilon$$

Diffusion process

TAYLOR APPROX:

$$\phi(x;t+\delta t) = \int \left[\underset{\substack{\uparrow \\ \phi(x-\epsilon;t)}}{\phi()}\; \underset{\substack{\nwarrow \\ g(x-\epsilon,\epsilon,\delta t)}}{g()} - \epsilon \frac{\partial\{\phi()g()\}}{\partial x} + \frac{\epsilon^2}{2}\frac{\partial^2\{\phi()g()\}}{\partial x^2} \right] d\epsilon$$

$$= \phi() \underbrace{\int g()\, d\epsilon}_{\text{sum to 1}} - \frac{\partial}{\partial x}\left[\phi() \int \epsilon g()\, d\epsilon\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\phi() \int \epsilon^2 g()\, d\epsilon\right]$$

(switching order of integration, diffn & summing)

Expected change to x in time δt

$$\frac{\phi(x;t+\delta t) - \phi(x;t)}{\delta t} = -\frac{\partial}{\partial x}\left[\phi(x,t)\frac{1}{\delta t}\boxed{\int \epsilon g(x-\epsilon,\epsilon,\delta t)\, d\epsilon}\right] + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left[\phi(x,t)\frac{1}{\delta t}\boxed{\int \epsilon^2 g(x-\epsilon,\epsilon,\delta t)\, d\epsilon}\right]$$

Expected change to x^2 in time δt

Diffusion process

$$M(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \int \epsilon \, g(x-\epsilon, \epsilon, \delta t) \, d\epsilon$$

$$V(x,t) = \lim_{\delta t \to 0} \frac{1}{\delta t} \int \epsilon^2 \, g(x-\epsilon, \epsilon, \delta t) \, d\epsilon$$

[∵ for small $E(Y)$, $V(Y) \approx E(Y^2)$]

$$\frac{\partial \phi(x;t)}{\partial t} = -\frac{\partial}{\partial x}\left\{\phi(x;t)\, M(x,t)\right\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left\{\phi(x;t)\, V(x,t)\right\}$$

$$\phi(x;t) = \frac{\text{Const}}{V(x,t)} \exp\left(2 \int \frac{M(x,t)}{V(x,t)} \, dt\right)$$

Modelling the diffusion process

- We are now dealing with densities, not probabilities

- So far, preference for one kind of change over another was exclusively modelled through selection
  - now for each kind of mutation (A ➔ B) , we have a mutation rate (may not be agnostic to nature of change)

Modelling the diffusion process

- Model mutation rates

- Model selection co efficients

- Model the functions of mean and variance of the rate of change of the alleles

  $M(x,t)$ & $V(x,t)$

  – additional parameters may be needed (eg. variance contributed by selection ?)

Equilibrium frequencies : adaptive mutation

AA	Aa	aa
1	1+s	1+2s

$$\begin{tikzcd} A \arrow[r, bend left, "\mu"] & a \arrow[l, bend left, "\nu"] \end{tikzcd}$$

selection

$$M_{\delta x} = s x(1-x)$$
$$+\left[(1-x)\mu - x\nu\right]$$

mutation

$$V_{\delta x} = \frac{x(1-x)}{2N_e}$$

drift

$$\phi(x) = \frac{C}{V_{\delta x}} e^{2\int \frac{M_{\delta x}}{V_{\delta x}} dx}$$

$$= C e^{4N_e s x} \boxed{x}^{4N_e\mu - 1} (1-x)^{4N_e\nu - 1}$$

At equilibrium freq is $\phi(x)$

PRESENT FREQUENCY

Fixation prob : adaptive mutation

- Diffusion process

AA	Aa	aa
1	1+s	1+2s

Prob of fixation = $u(x)$

$u'(x) = du(x)/dx$

BACKWARD KOLMOGOROV EQN:

$$u(x) = \int_{\delta x} u(x+\delta x)\, g(\delta x, x, \delta t)$$

$$\begin{tikzcd} A \arrow[r, bend left] & a \arrow[l, bend left] \end{tikzcd}$$

UNDER INFINITE ALLELE MODEL, WHAT ARE THE MUTATION RATES?

$$u(x+\delta x) \approx u(x) + \delta(x)\, u'(x) + \frac{\delta(x)^2}{2} u''(x)$$

[TAYLOR SERIES]

SUBSTITUTING:

$$u(x) \approx \int_{\delta x} u(x)\, g(\delta x, x, \delta t) + \int_{\delta x} \delta x\, u'(x)\, g(\delta x, x, \delta t) + \int_{\delta x} \frac{\delta(x)^2}{2} u'' x\, g(\delta x, x, \delta t)$$

Fixation prob of adaptive mutation

$$M_{\delta x} = \int \delta x \; g(\delta x, x, \delta t) \leftarrow \text{MEAN}$$

$$V_{\delta x} = \int (\delta x)^2 \; g(\delta x, x, \delta t) \leftarrow \text{VARIANCE}$$

$$u\cancel{(x)} = u\cancel{(x)} + u'(x) M_{\delta x} + u''(x) V_{\delta x}$$

$$u'(x) M_{\delta x} + u''(x) V_{\delta x} = 0$$

$$\boxed{sx(1-x)}\, u'(x) = \boxed{\frac{-x(1-x)}{4N_e}}\, u''(x)$$

$$4N_e s \, u'(x) = -u''(x)$$

REMEMBER OUR INTEGRATING FACTOR?

$$u(x) = C \cdot e^{-4N_e s x} + D$$

BOUNDARY $u(0) = 0, \quad u(1) = 1$

$$u(x) = \frac{e^{-4N_e s x} - 1}{e^{-4N_e s} - 1}$$

Rate of evolution

$$u\left(\frac{1}{2N}\right) \approx \frac{2sN_e/N}{1-e^{-4N_e s}} \approx 2sN_e/N$$

= Chance of fixation of new mutation

- Rate of evolution = Rate of observed mutations

  = Rate of mutation X rate of fixation

IS THIS THE SAME AS RATE OF SUBSTITUTION?

$$\approx (2\mu \cdot N) \times 2s\,N_e/N$$

$$= \underbrace{4\mu \cdot s \cdot N_e}$$

← RESULT RELATING DRIFT, SELECTION & MUTATION!

Paradigms of selection

- 3 regimes : based on value of Ne X s

$\lessapprox 1$ → NEUTRAL SELECTION

$\gg 1$ → POSITIVE SELECTION

$\ll 1$ → NEGATIVE SELECTION

N_e → effective population

s → selectional coefficient, expected no. of offspring

Neutral Theory

- Motoo Kimura (1968)
- Very large fraction of fixed mutations (both within and between species) are the result of truly random processes (drift) and not of directed selection
  - previously, it was thought natural selection main driver of fixed mutations
  - do not confuse neutral theory with neutral model (which is any model of evolution under no / neutral selection)

Neutral Theory

- Functional sites : Most mutations deleterious and immediately removed by negative selection SUBST. RATE = $\mu_{NEUTRAL}$
- -ve correlation betw functional significance and substitution rate : more functionally significant ➔ more types of mutations likely to be deleterious / more types of mutations less likely to be fixed ➔ lower neutral mutation rate & lower substitution rate

Allele freq = $f(\mu_{NEUTRAL}, N_e)$

Explanation of molecular clock

- Neutral mutation rate is expected to be constant across species and lineage
- Completely random mutations would accrue linearly over time
  - Branch length = Expected no of substitutions = Substitution rate X time = Neutral mutation rate X time = constant X time

Evidence

- For neutral theory : Functionally important sites show lower substitution rates wrt functionally unimportant sites
- For neutral theory : molecular clock
- Against neutral theory : Only accounts for strongly deleterious and neutral mutations. Evidence exists of weakly deleterious mutations.

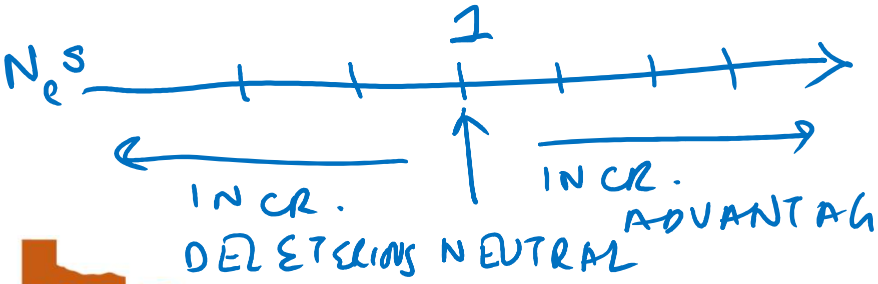
Selectionist – neutralist debate

- Ohta : Nearly – neutral theory
  - strongly deleterious alleles get wiped out
  - weakly deleterious alleles get fixed under mutation – selection balance

$N_e s \approx 1$ or $N_e s < 1$

Allele freq = $f(\mu, s, N_e)$

slightly higher than $\mu_{NEUTRAL}$



Identifying neutrality

- Biggest challenge in using neutral theory : which changes are neutral ?
- Question to address : which phenotypes are affected on which natural selection can act ?

A complicated situation

- What about mutations in transcription factor binding site ?
  - If the change increases binding affinity ?
  - If the change decreases binding affinity ?
  - If it causes no (negligible) change ?

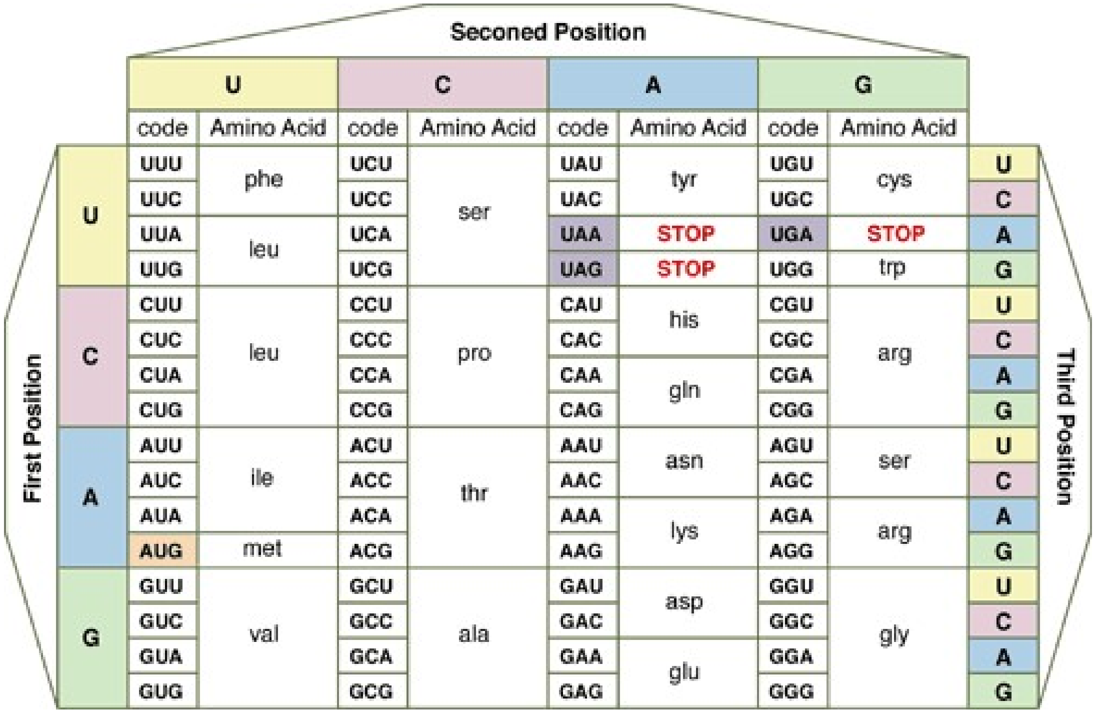
- Difficult to say due to compensatory binding sites nearby : difficult to quantify from binding alone : expression levels of genes need to be observed : still may not be enough

A less complicated situation : codons

- Simple situation : coding region : silent mutations (which do not change the coded amino acid) are termed neutral
- Other changes are deemed non – neutral
- For a MSA, no. and nature of mutations need to be figured out on the tree relating the sequences (or averaged over trees)

Codon table

- Synonymous & non synonymous mutations



First Position	Seconed Position								Third Position
	U		C		A		G		
	code	Amino Acid	code	Amino Acid	code	Amino Acid	code	Amino Acid	
U	UUU	phe	UCU	ser	UAU	tyr	UGU	cys	U
	UUC		UCC		UAC		UGC		C
	UUA	leu	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG	STOP	UGG	trp	G
C	CUU	leu	CCU	pro	CAU	his	CGU	arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	ile	ACU	thr	AAU	asn	AGU	ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lys	AGA	arg	A
	AUG	met	ACG		AAG		AGG		G
G	GUU	val	GCU	ala	GAU	asp	GGU	gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	glu	GGA		A
	GUG		GCG		GAG		GGG		G

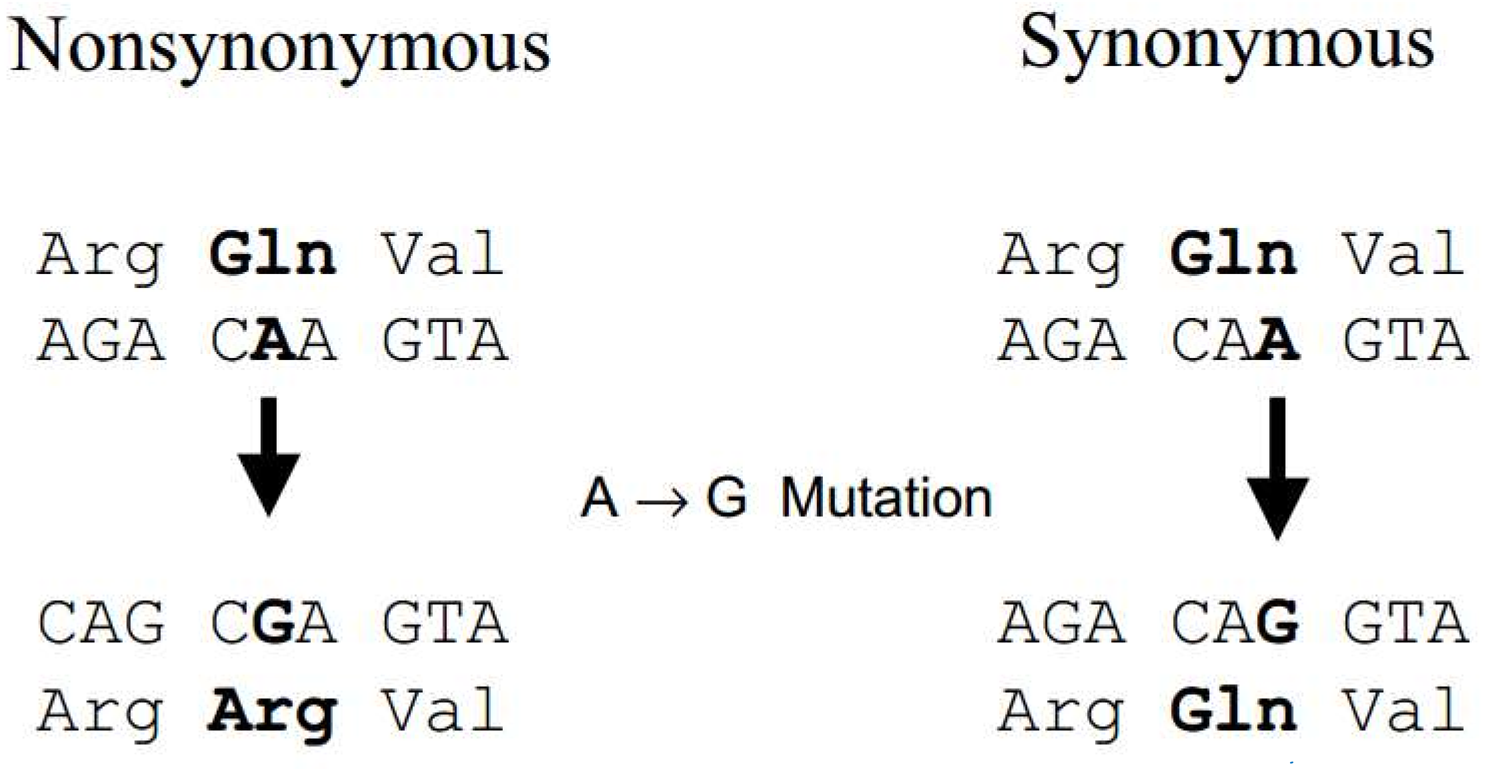
Detecting selection in codons

- Goal is to identify regions in genes where rate of amino acid change (rate of non synonymous mutation) is greater or lesser than the rate of neutral (synonymous) mutation.

dN / dS

- dN = no of non synonymous changes, dS = no of synonymous changes
- Ratio : > , = , < 1 : positive, neutral or negative selection
- How to put probabilities on such hard constraints ?
  - distribution of (dN – dS) for gold standard sets of neutral and non neutral sites

An example



Nonsynonymous

Arg **Gln** Val
AGA C**A**A GTA

CAG C**G**A GTA
Arg **Arg** Val

A → G Mutation

Synonymous

Arg **Gln** Val
AGA CA**A** GTA

AGA CA**G** GTA
Arg **Gln** Val

G McVean, Oxford Uni

Counting dN & dS

- Another example :

CAA	CAA	GTA	CAA	AGA
↓	↓		↓	
CAG	CGA	GTA	CAG	AGA
S	N		S	

$$\frac{dN}{dS} = \frac{1}{2} < 1$$

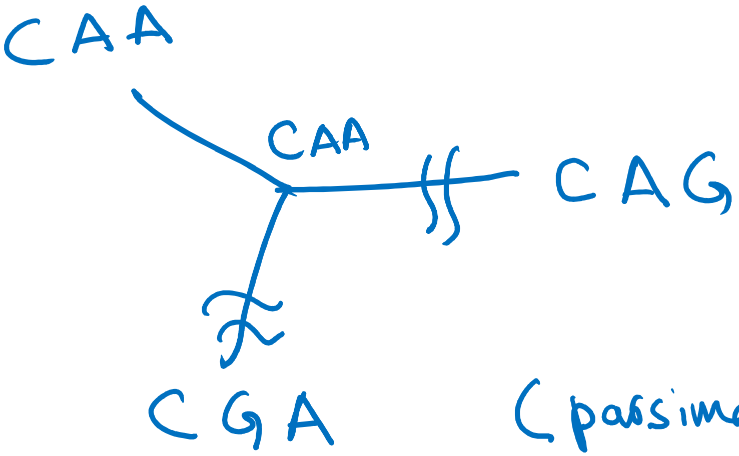
Counting dN & dS over a tree

- MSA

CAA
CAG
CGA

⇒

USE MODEL BASED TREE, find ANCESTRAL VALUES



(parsimony eg)

⇓ COUNT MUTATIONS

dN = 1

dS = 1

⇐ TAKE DECISION

$\frac{dN}{dS} = 1$, so neutral

Counting dN & dS

CAA   CAA   CAA

CGA   CAG   CGA

for pairwise, it is simple:

CAA   CAA   CAA

↕   ↕   ↕

CGA   CAG   CGA

N   S   N

dN = 2, dS = 1. We ARE DONE

FOR A MULTIPLE SEQ ALIGNMENT,
IT IS NOT SO SIMPLE.

$$\left.\begin{array}{ll} x_1 & CAA \\ x_2 & CGA \\ x_3 & CAG \\ x_4 & CAA \end{array}\right\}$$ what is dN,
what is dS?

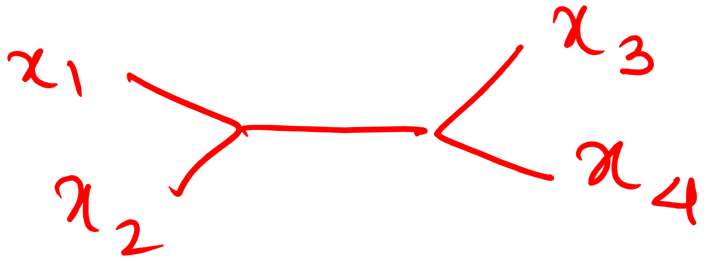
To count mutations, we want to first fix the level at which evolution is operating. [eg nucleotide, codon]

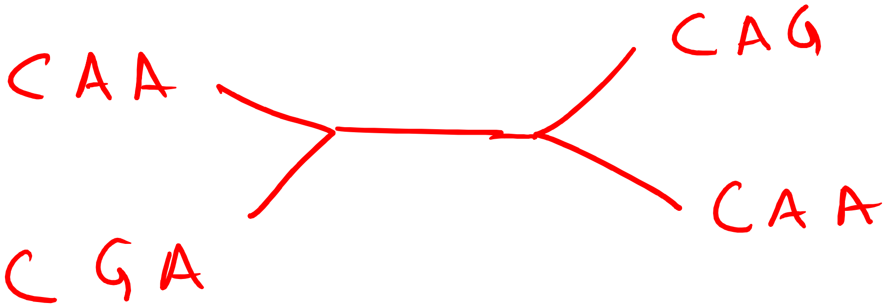
WE CAN USE NUCLEOTIDE LEVEL

Next we need a model of evolution [ML or Parsimony, etc.]

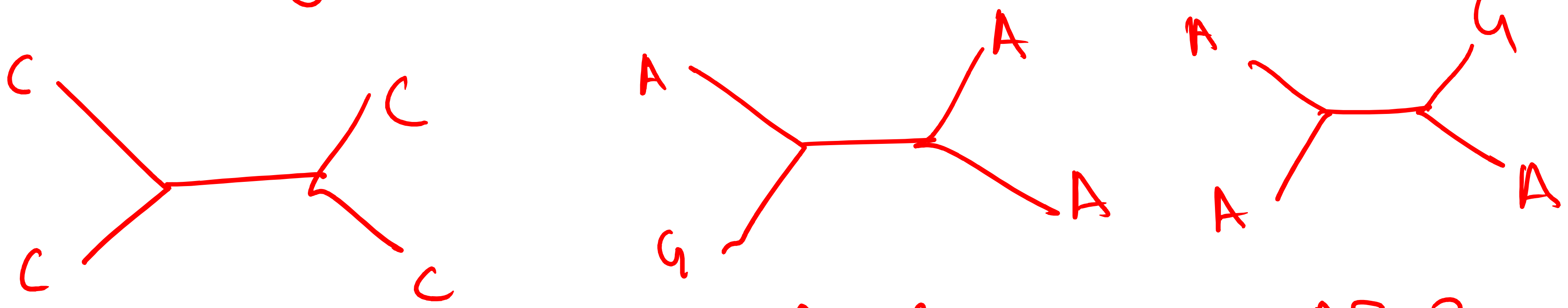
lets pick parsimony.

for simplicity, we assume topology is known.





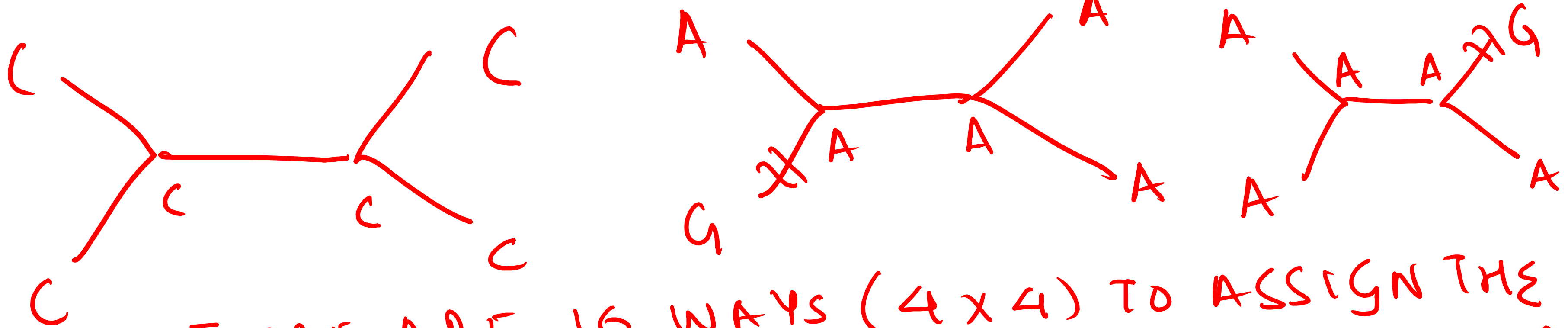
Now, we need to calculate the ancestral states for each position of alignment



Pos 1

Pos 2

Pos 3

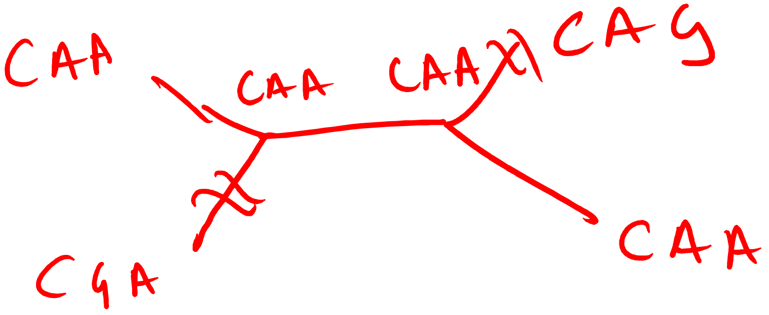


THERE ARE 16 WAYS (4 x 4) TO ASSIGN THE 2 ANCESTORS. FOR PARSIMONY, WE CHOOSE THE ASSIGNMENT(S) WITH MIN. # OF MUTATIONS

THESE ASSIGNMENTS FOR LEAST NO. OF MUTATIONS MAY NOT BE UNIQUE. FOR EG, THERE MAY BE 3 WAYS TO ASSIGN ANCESTORS FOR POSITION 2, AND 2 WAYS TO ASSIGN ANCESTORS FOR POSITION 3.

NOW, TO PUT BACK THE NUCLEOTIDES:

UNDER INDEPENDENCE MODEL OF EACH SITE:



WE WERE LUCKY, THE ASSIGNMENTS WERE UNAMBIGUOUS. NOW ITS EASY TO COUNT.

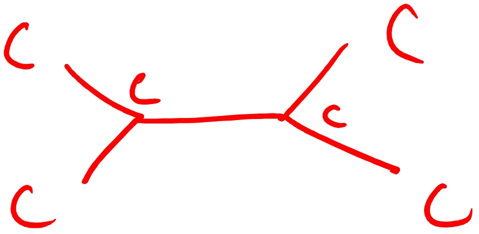
$$\text{CAA} \overset{A \leftrightarrow G}{\longleftrightarrow} \text{CGA} = 1\ N$$

$$\text{CAA} \underset{A \leftrightarrow G}{\longleftrightarrow} \text{CAG} = 1\ S$$

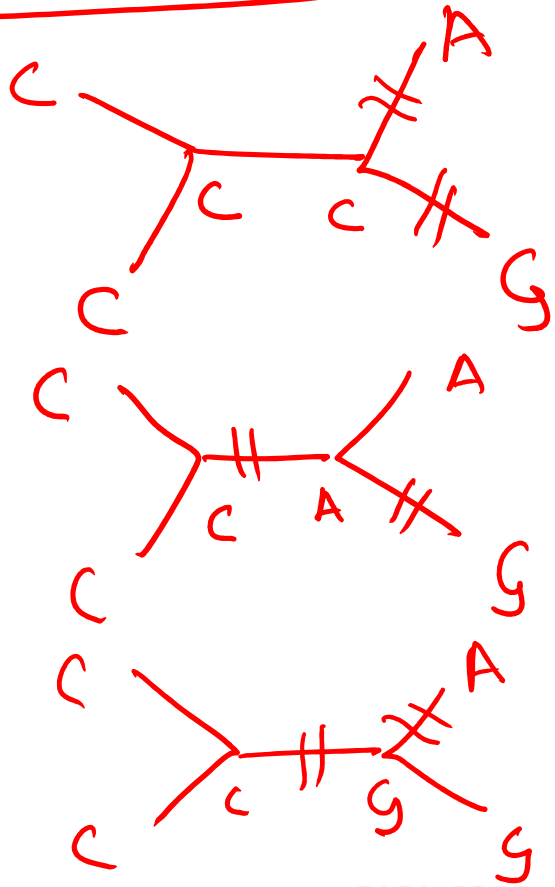
THERE ARE 2 WAYS IN WHICH THIS SITUATION GETS MORE COMPLICATED

(A) AMBIGUOUS ASSIGNMENTS

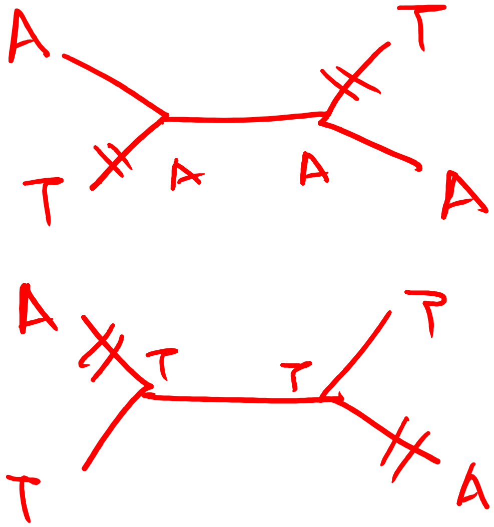
POS 1



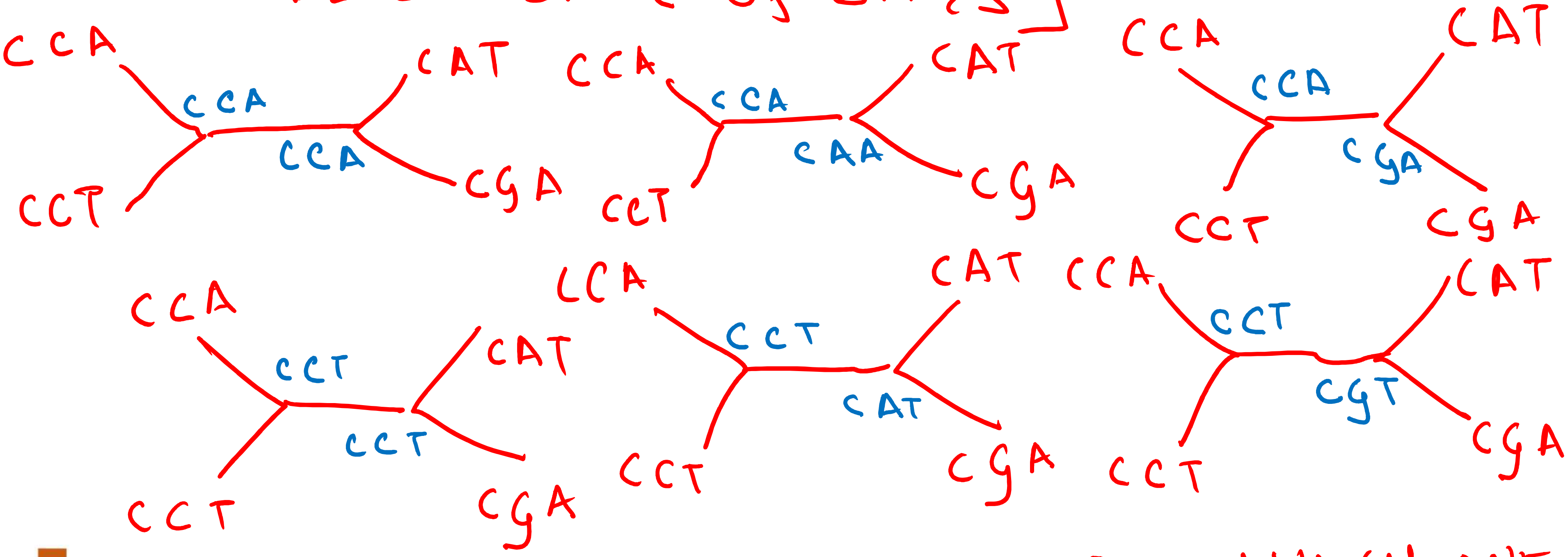
POS 2



POS 3



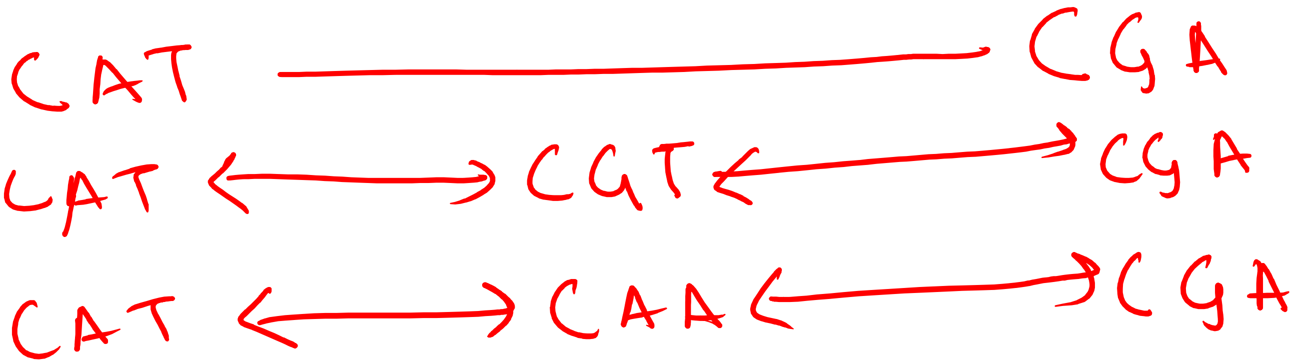
NOW, INSTEAD OF A SINGLE (1×1×1=1) POTENTIAL ANCESTRAL ASSIGNMENT YOU HAVE (1×3×2=6) SIX POTENTIAL ANCESTRAL ASSIGNMENTS. [∵ WE ASSUMED INDEPENDENCE OF SITES]



WILL GET SIX dN/dS RATIOS - WHICH ONE IS RIGHT?

(B) THERE IS ANOTHER PROBLEM

HOW ABOUT BRANCHES WITH MULT. SITES CHANGED?



ORDERING OF CHANGES MAY CHANGE dN/dS RATIO.

SO, EVEN FOR EACH TREE, WE HAVE MULTIPLE dN/dS RATIOS.

THIS IS WHEN YOU WISH YOU HAD USED MAXIMUM LIKELIHOOD MODELS.

IN GENERAL,

PROBLEM Ⓐ : SOLUTION: USE LIKELIHD FRAMEWORK TO PICK THE MOST LIKELY ANCESTORS (JOINT OR MARGINAL ?), OR SUM OVER LIKELIHD OF TREES

PROBLEM Ⓑ : ML TRAJECTORIES ARE DIFFICULT IN CTMP TO CALCULATE. EASIER TO DO VITERBI ANALYSIS OF JUMP CHAIN OF CTMP WHICH MODELS CODON EVOLUTION

Word of warning

- Remember, a mode of selection over a set of sites does not guarantee that the same mode of selection will operate on a subset of the sites !

McDonald – Kreitman test

- Synthesis : species and population genetic models : test for ancient selectional forces
- Between species and within species dN & dS compared by categorical tests

BETW SPECIES

WITHIN SPECIES

	Fixed	Polymorphic
Synonymous	D_s	P_s
Nonsynonymous	D_n	P_n

$D_n/P_n >> D_s/P_s$

implies +ve selection

$\approx D_s/P_s$ → neutral selection

$<< D_s/P_s$ → -ve selection

Notion behind the MK test

- Deleterious mutations may persist in populations for a few generations due to drift, very unlikely to become fixed.
  - contribute to polymorphism, but not divergence.
- Advantageous / adaptive mutations become fixed in populations pretty fast : contribute little to polymorphism, appear as fixed differences between species.
- Compare no of fixed to polymorphic differences for synonymous and nonsynonymous mutations deviations from the neutral theory can be detected

So, what can we do with these tools ?

- Given initial allele frequency, and selectional coefficients and mutation rates
  - predict probability of fixation and / or equilibrium frequencies
- Given allele frequencies in equilibrium
  - estimate heterozygosity and other notions of genetic variability and estimate effective population size, mutation rates, selection coefficients
- Given alleles and model of change for a set of loci, predict the nature and degree of selection

More complications

- Genetic hitchhiking
- Modelling multiple loci
- Models of recombination
  - linkage between loci

- Polymorphism as a function of recombination rates

What if …

- We are less interested in the evolutionary parameters
- More interested in the genealogy ?

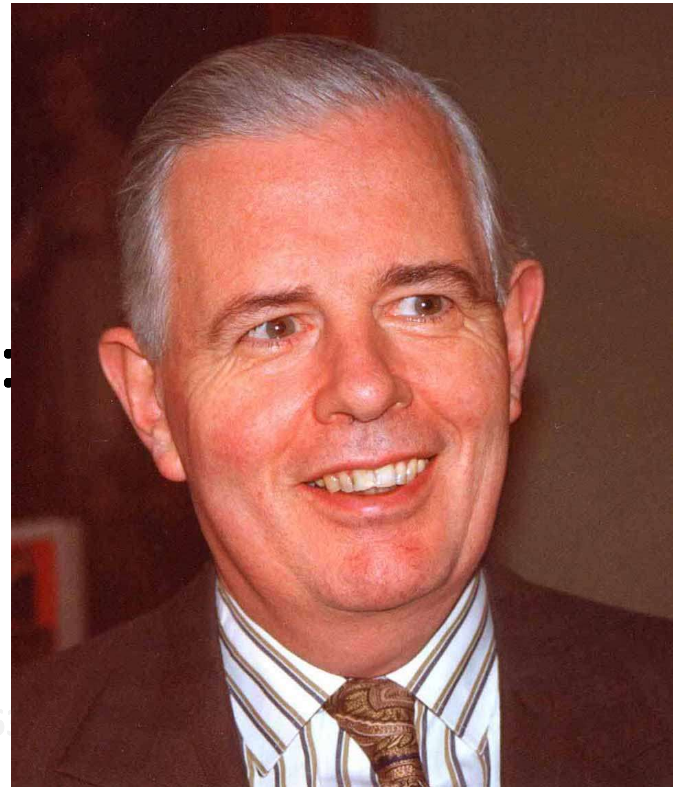
Coalescent theory

- Purely historical, not predictive
- Retrospective, may be generative

- Genealogical tree to MRCA

- (Bad) analogy in phylogenetics: tree reconstruction

John Kingman



Isaac Newton Institute

Coalescent theory

- Visualization of the coalescent : http://www.ucl.ac.uk/tcga/presentations/TCGAugss/TCGA_MW_Seminar4.ppt

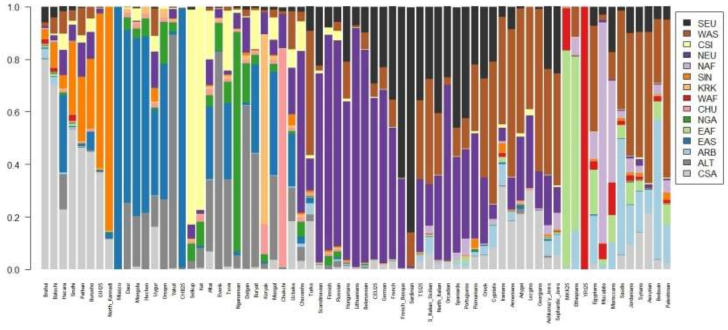
- Deriving the coalescent : http://bio.classes.ucsc.edu/bio107/Class%20pdfs/W05_lecture14.pdf

A few uses: genetic fingerprinting

- Pick a set of loci s.t. no of allelic configurations (genotypic or haplotypic) approaches the no of individuals in the population
  - Not enough selection, and sufficiently high rate of mutation that it is conserved across individuals (effective population is same for all alleles)

A few uses: reconstructing ancestry

- Paternal and maternal lineages : avoid confounding recombination
  - paternal : Y chromosome
  - maternal : mtDNA (mitochondrial eve)
- Distinguishing divergence from gene flow
- Admixture components : relative contributions of founder populations



Dodecad

That’s all, folks !

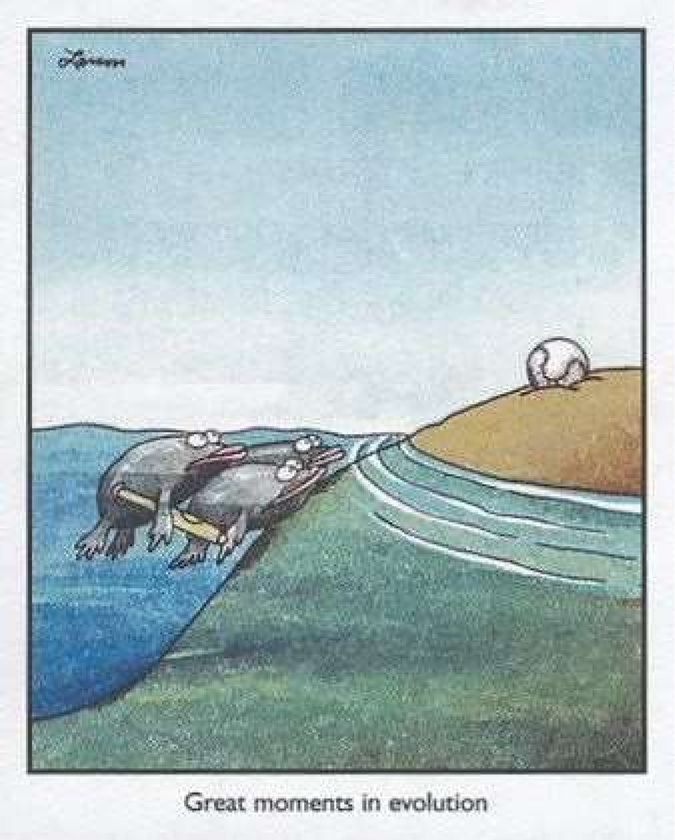
More reading (on the website)

Comparing different methods :

- Phylogenetic vs pop genetic
- Historic vs predictive
- ML vs Bayesian

etc …

Which one to use ?



Larson, The Far Side

Summary

- Population genetics: Toolkit for understanding a more fine-grained evolutionary picture, merges evolutionary theory with quantitative genetics (population genomics : whole genome view)
- Evolutionary process : cooking pot, alleles : ingredients, drift, mutation, selection, recombination, population structure and migration, **stochasticity** : recipe
- Changes in allele frequencies : outcome of the process !
- Often, the goal is to observe the outcome and make evidence-driven guesses about **missing pieces** of the recipe
  - **GENEALOGY ESTIMATION AND INFERENCE: identifying evolutionary relationships between individuals and using such relationships for inference**: estimating allele genealogy, coalescents, pedigree based inference
  - **POPULATION GENETICS: evolutionary forces**: mutation rates, selectional model, recombination rate, **demography:** migratory model, population size
  - **ASSOCIATION STUDIES (CLASSICAL GENETICS) : genotype – phenotype relationships**: phenotype-associated loci, epistasis model, quantitative trait models

(Some) things we didn’t cover

- Gene tree – species tree reconciliations
- Violating W-F models in additional ways : Inbreeding, migration, ancestry & demographic models
- Modelling multi locus dynamics : recombination
- Quantitative genetics

Acknowledgements

- Eric Xing
- Dannie Durand
- Gil McVean