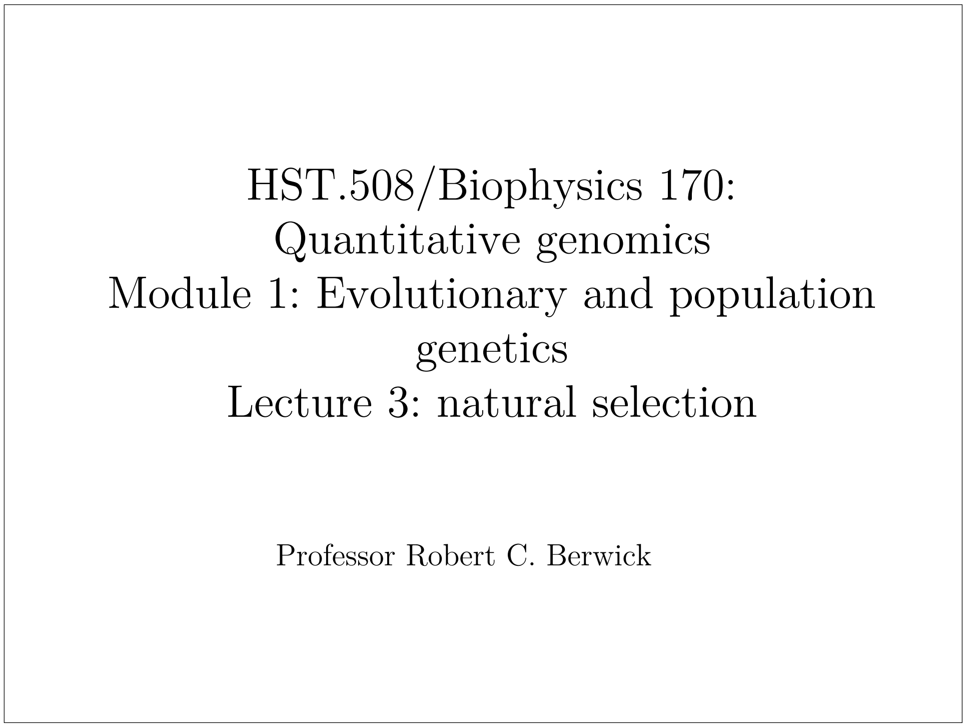
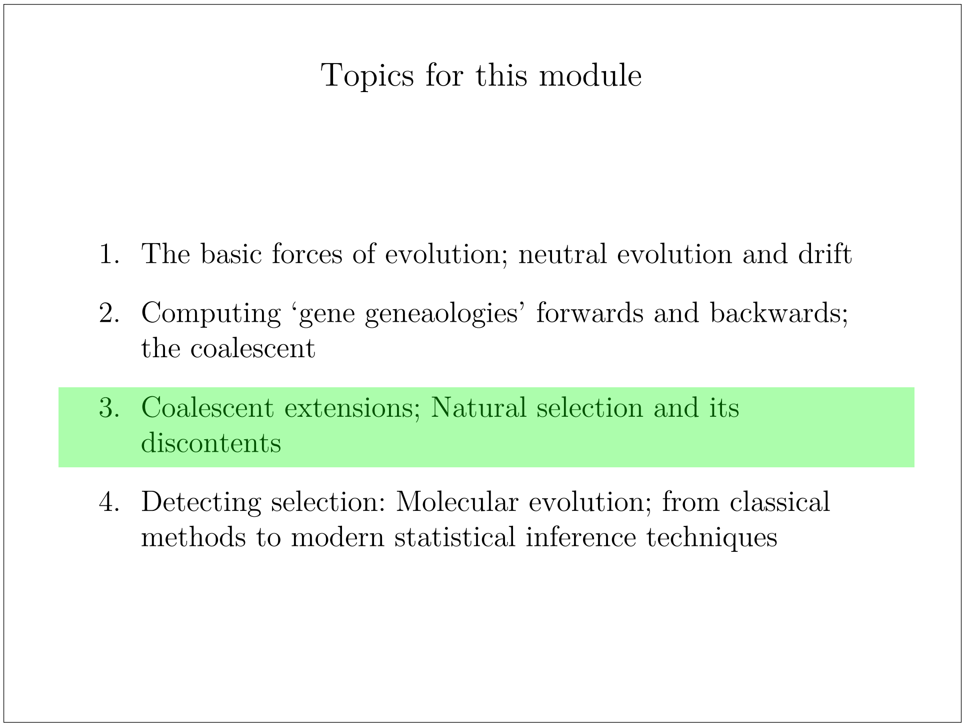
Harvard-MIT Division of Health Sciences and Technology
HST.508: Quantitative Genomics, Fall 2005
Instructors: Leonid Mirny, Robert Berwick, Alvin Kho, Isaac Kohane



# HST.508/Biophysics 170: Quantitative genomics Module 1: Evolutionary and population genetics Lecture 3: natural selection

Professor Robert C. Berwick



## Topics for this module

1. The basic forces of evolution; neutral evolution and drift
2. Computing ‘gene geneaologies’ forwards and backwards; the coalescent
3. Coalescent extensions; Natural selection and its discontents
4. Detecting selection: Molecular evolution; from classical methods to modern statistical inference techniques

# Agenda for today

1.Coalescing the coalescent: the Great Obsession; adding complications like demographics, recombination; how you can *use* the coalescent (simulation, estimation, testing)

2. Natural selection: from the basic dynamical system equation to the diffusion approximation: how can genes survive?

# Coalescent Summary

1. Coalescent theory describes the *genealogical* relationships among individuals in a Wright-Fisher population

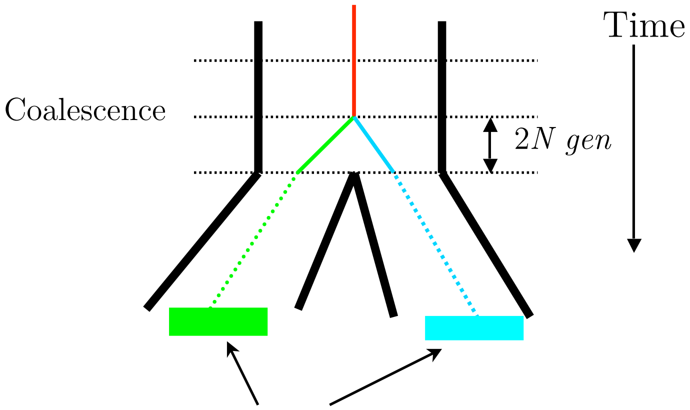
2. *Sample*, rather than *population.*

*3.* *Retrospective* (how did things get to be the way they are?) rather than *prospective* (what happens if?) – better for our situation of sampling from data.

3. That is: the coalescent model *differs* from the ‘classical’ random sampling gene pool model in that it gives us the opportunity to *start* with polymorphism data and work backwards – start with simplest model, if doesn’t work, change the model

4. Separate *demography* (coalescent) from *genetics* (mutation) - allows to *separate* the two & so gives us basic test statistics for diversity/variation ($\theta, \pi$ )

## The Great Obsession: variation (polymorphism) entangled with descent



Observed sample variation: is it from descent (tree), or from biology?

## Two ‘competing’ stochastic processes intertwined

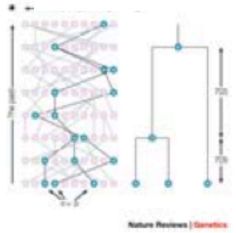
**1. Gene trees:** How long until sample sequences have common ancestor (*coalesce*)?

Answer: the coalescent models the geneology of a sample of $n$ individuals drawn from a (putative) population of size $N$ as a random bifurcating tree. The $n$-1 coalescent times $T(n)$, $T(n\text{-}1)$, ..., $T(1)$ are (to an approximation) mutually independent, exponentially distributed random variables

*Rate* of coalescence for two lineages is (scaled) at 1, where this is 2$N$ generations; *Total rate*, for $k$ lineages is ‘$k$ choose 2’

**2. Genetics:** sprinkle in Poisson mutation process with rate $\lambda = ut$, then what is expected distribution of variation?

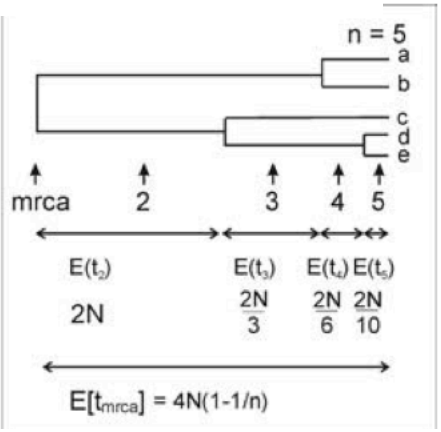
## Expected time to coalescence



Nature Reviews | Genetics

Rosenberg and Nordborg, 2002

As the sample size increases towards 2N, $E[t_{mrca}]$ approaches 4N, which equals the fixation time for a newly arisen mutation



$E[t_{mrca}] = 4N(1-1/n)$

## Time to coalescence for *n* sequences or genes

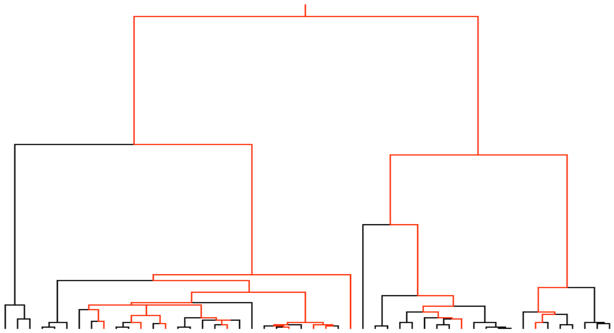
$$\Pr\{\text{coalescence given } n \text{ lineages}\} = \frac{n(n-1)}{2}\frac{1}{2N_e}$$

Number of pairs of lineages

Probability of a given pair coalescing

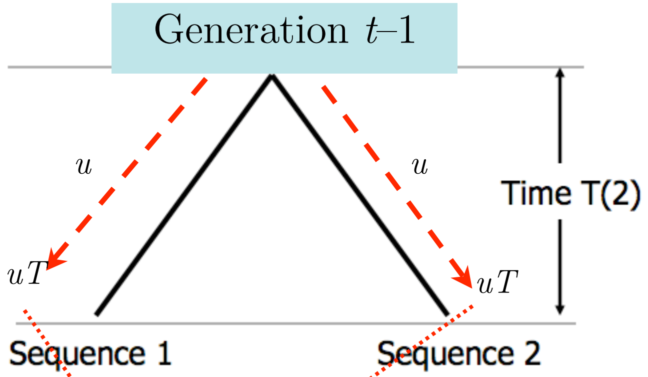
$$E[T_{co}] = \frac{4N_e}{n(n-1)}$$

## The structure of the basic coalescent



Expected time to coalescence for 2 genes is $2N$; variance $2N(2N–1)$
For $n$ sequences or genes…
If time is measured in units of $2N$ generations, by t'= $t/2N$
$E[T_{\mathrm{k,\ k-1}}]$= 1/($k$ choose 2); variance is square of this
Time to MRCA for all genes is sum of these times, or 2(1-1/$n$) [again in units of time measured in $2N$, i.e., $4N$(1-1/$n$) unscaled time

## Estimating nucleotide divergence as $\theta$



Notation: $T_i$= time to collapse of $i$ genes, sequences,…

$$2uT = 2u\ 2N = 4uN = \theta$$

Expected # mutations, $n$ allele or sequence case

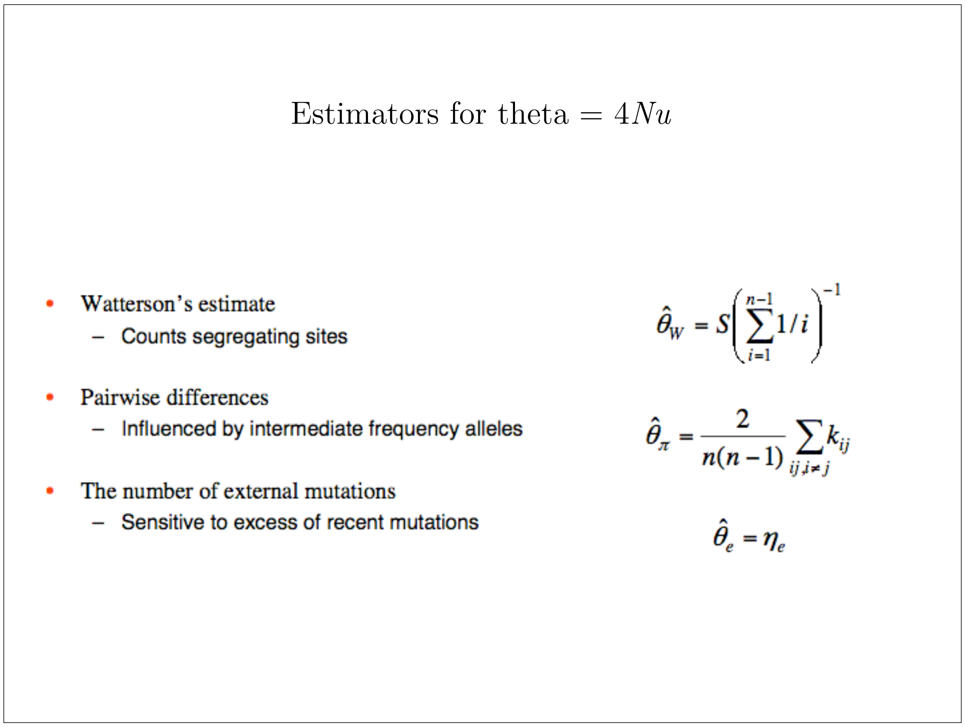
$$E[\text{no. mutations}] = \underset{\substack{\nearrow \\ \text{Time}}}{\frac{4N_e}{n(n-1)}} \times \underset{\substack{\nwarrow \\ \text{Total mutation rate}}}{n \times u}$$

$$= \frac{\theta}{n-1}$$

So this gives us the expected amount of sequence diversity

Summary results for basic coalescent

- Expected time to coalesce, for 2 alleles, $2N$
- Expected time to coalesce, all $k$ alleles (hence avg fixation time) $E[T_C] = \sum_{i=2}^{n} iE[T_i] = 4N\sum_{i=2}^{n}\frac{1}{i-1}$
- Expected # of segregating sites $E[S_N] = uE[T_C] = \theta\sum_{i=2}^{n}\frac{1}{i-1}$
- Expected amount of sequence diversity



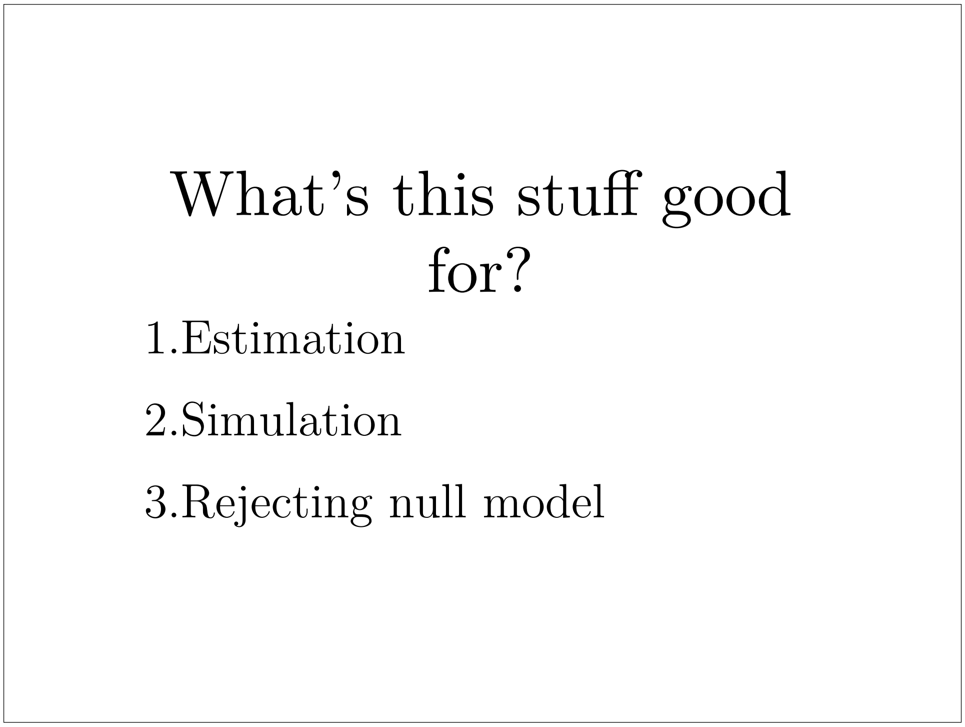
## Estimators for theta $= 4Nu$

- Watterson's estimate
  - Counts segregating sites
- Pairwise differences
  - Influenced by intermediate frequency alleles
- The number of external mutations
  - Sensitive to excess of recent mutations

$$\hat{\theta}_W = S\left(\sum_{i=1}^{n-1} 1/i\right)^{-1}$$

$$\hat{\theta}_\pi = \frac{2}{n(n-1)} \sum_{ij,i\neq j} k_{ij}$$

$$\hat{\theta}_e = \eta_e$$



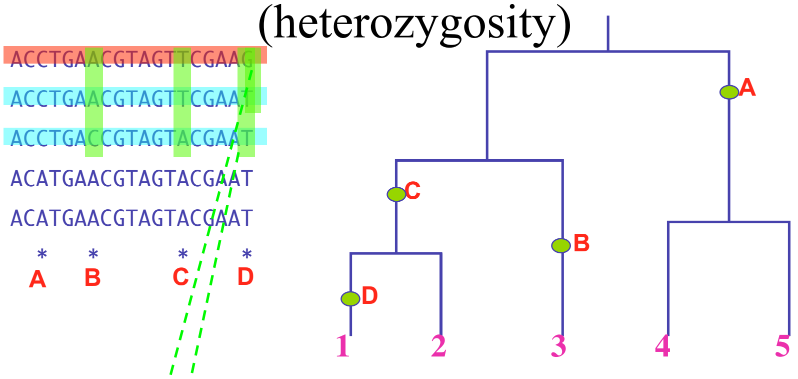
## What's this stuff good for?

1. Estimation
2. Simulation
3. Rejecting null model

Basic estimation idea: find coalescents that are 'improbable' to detect interesting (i.e., unlikely) patterns of mutations

This helps us untangle two sources of variation: gene/sequence *tree divergence* from *polymorphism*

## $\Theta_T$ estimated from pairwise differences (heterozygosity)



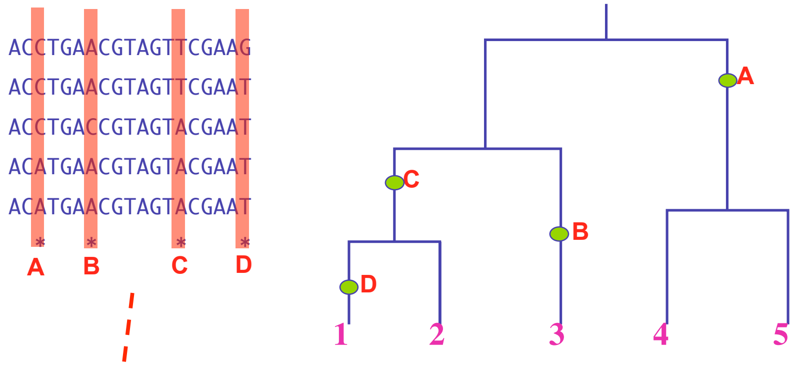
```
ACCTGAACGTAGTTCGAAG
ACCTGAACGTAGTTCGAAT
ACCTGACCGTAGTACGAAT
ACATGAACGTAGTACGAAT
ACATGAACGTAGTACGAAT
```

A mutation on an *interior* branch will have higher weight

$\Theta_T$ = Average Pairwise Distance (just the average heterozygosity)

$= (1+3+3+3+2+2+2+2+2)/10=2$

## $\Theta_W$=4Nμ estimated from # segregating sites



```
ACCTGAACGTAGTTCGAAG
ACCTGAACGTAGTTCGAAT
ACCTGACCGTAGTACGAAT
ACATGAACGTAGTACGAAT
ACATGAACGTAGTACGAAT
```

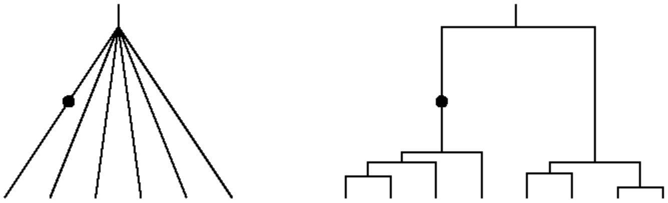
Expected number of segregating sites:

$$S_n = \Theta_W \sum_{i=1}^{k-1} \frac{1}{i}$$

$\Theta_W$= 4/(1+1/2+1/3+1/4)=24/11=2.1818

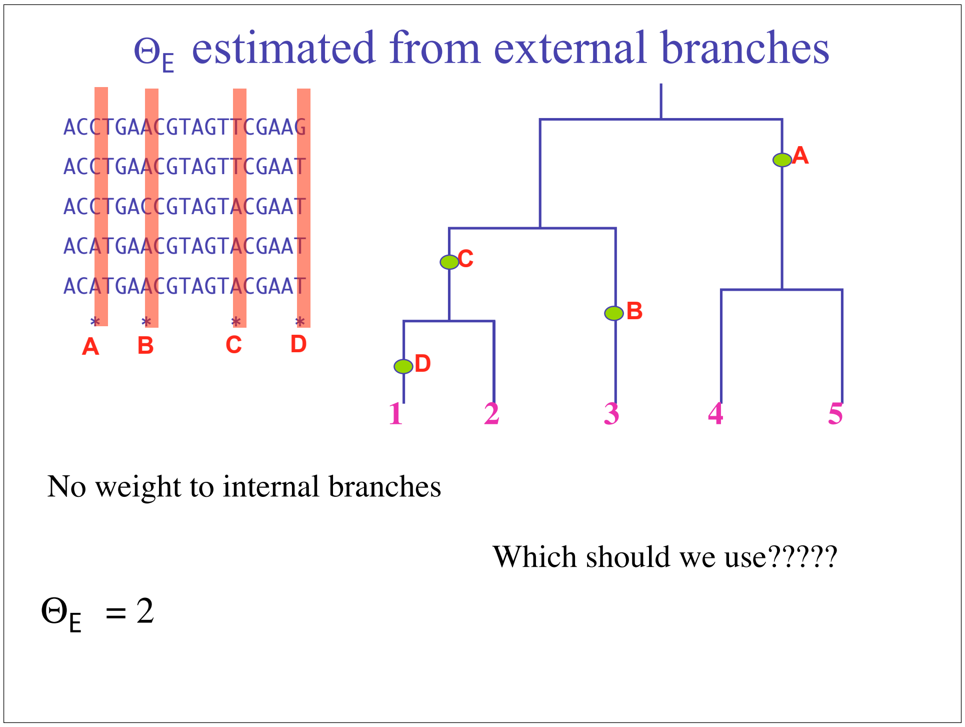
Watterson, 1975

## Different coalescent patterns (relative branch lengths) yield different estimates for theta even though total branch length is the same and # segregating sites remains the same



Second type of mutation counted more times when calculating the average pairwise distance – typical when there’s a ‘burst’ after a population bottleneck

Use the *difference* between the two estimates to figure out a statistical measure that can pick out these two patterns



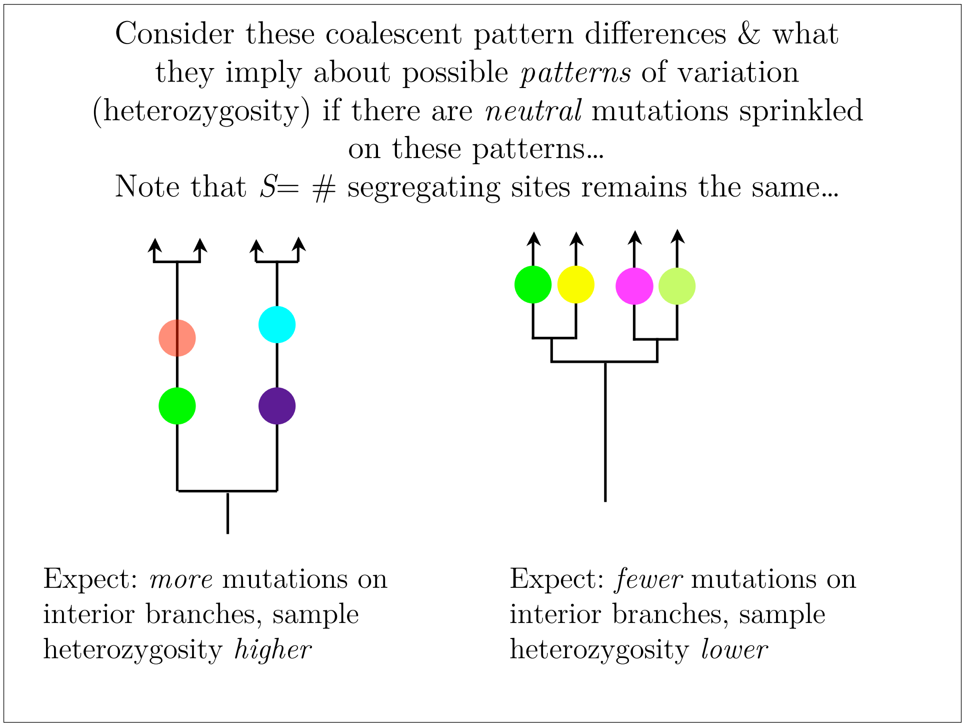
## $\Theta_E$ estimated from external branches

```
ACCTGAACGTAGTTCGAAG
ACCTGAACGTAGTTCGAAT
ACCTGACCGTAGTACGAAT
ACATGAACGTAGTACGAAT
ACATGAACGTAGTACGAAT
  *   *      *    *
  A   B      C    D
```

No weight to internal branches

Which should we use?????

$\Theta_E = 2$

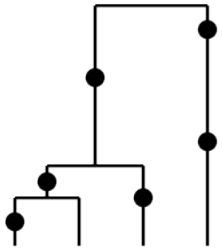


Consider these coalescent pattern differences & what they imply about possible *patterns* of variation (heterozygosity) if there are *neutral* mutations sprinkled on these patterns…
Note that $S$= # segregating sites remains the same…

Expect: *more* mutations on interior branches, sample heterozygosity *higher*

Expect: *fewer* mutations on interior branches, sample heterozygosity *lower*

## Two estimates of theta



$$E[\pi] = \theta$$

$$E[S] = \theta \sum_{i=1}^{n-1} \frac{1}{i}$$

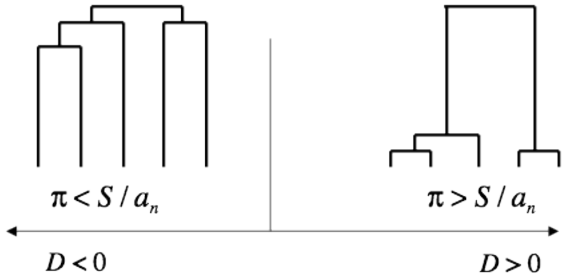
$$D = \frac{\pi - S/a_n}{\sqrt{\mathrm{Var}(\pi - S/a_n)}} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima (1989)

## Use of Tajima's *D*

$$D = \frac{\pi - S/a_n}{\sqrt{\mathrm{Var}(\pi - S/a_n)}} \qquad a_n = \sum_{i=1}^{n-1} \frac{1}{i}$$

Tajima (1989)



## Human mitachondrial DNA

Ingman *et al.* (2000)

52 complete molecules from a worldwide sample
(linguistic groups)
521 segregating sites excluding D-loop

$$\pi = 44.2$$

$$a_{52} = 4.52$$

$$S / a_{52} = 115.3$$

$$\sqrt{\hat{\mathrm{V}}(d)} = 31.8$$

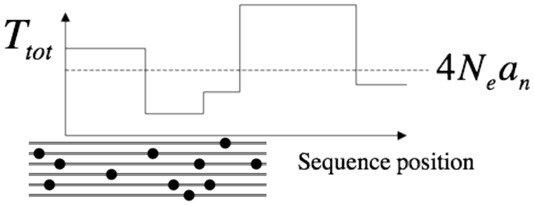
$$D = \frac{44.2 - 115.3}{31.8} = -2.23$$

Probability of observing such an extreme value under neutrality = 0.01

Human mtDNA have an excess of low-frequency variants

## Factors affecting test power

- The number of mutations in the sample is of critical importance
  - In general, sequencing a large region is more important than sequencing many individuals
- Recombination reduces the possibility of drawing trees from sequences, but evens out evolutionary stochasticity



## Example 1 – human mitochondrial DNA

52 complete molecules

521 segregating sites

$$D = \frac{\Theta_T - \Theta_W}{\mathrm{Std}(\Theta_T - \Theta_W)}$$

$\Theta_T = 44.2$ $\Theta_W = 115.3$

$\mathrm{Std}(\Theta_T - \Theta_W) = 31.8$

$D = -2.23$ ($P<0.01$)

Ingman et al. 2000

## Example 2 – human Y-chromosome

3 Y-chromosome genes, 40 kb of sequence in 53 males

47 polymorphic sites

Tajima’s *D*: -2.3, -2.0 and –1.8 highly significant

TMRCA: No growth: 84,000 (55,000-149,000)
Exponential: 59,000 (40,000-140,000)

With exponential growth more mutations are recent and therefore estimated TMRCA is smaller

Thomson et al. 2000, Shen et al. 2000

Applications– Simulation for model testing

- Ex: ~1400bp at Sod locus in Drosophila

  10 taxa

  5 were identical. The other 5 had 55 mutations

Q: Is this a chance event, or is there selection for this haplotype?

## Simulation results

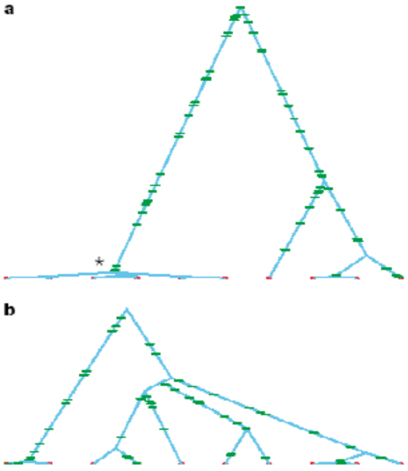
1. 10000 coalescent simulations were performed on 10 taxa

2. 55 mutations placed on the coalescent branches

3. Count the number of times 5 lineages are identical

4.This event happened in only 1.1% of the cases

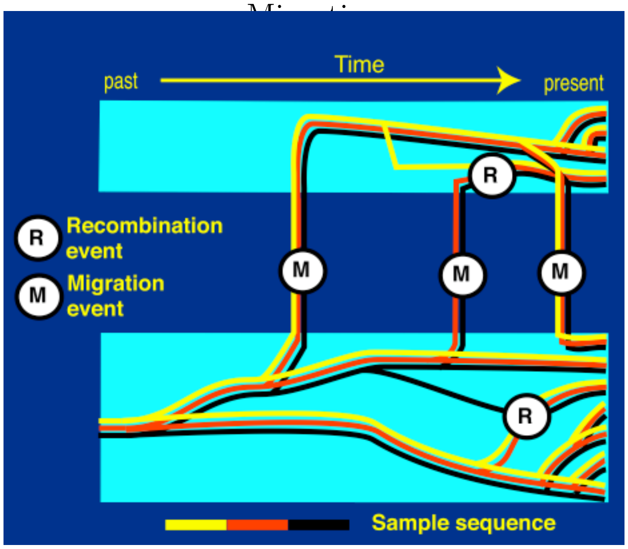
Conclusion: selection, or some other mechanism explains this data – not the neutral mutations

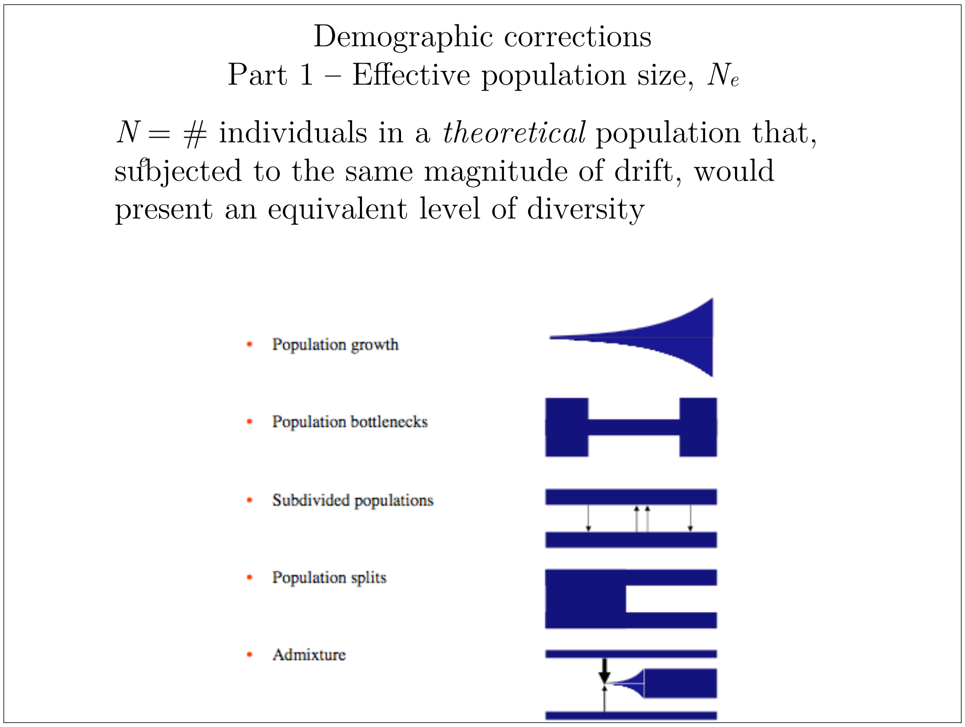


## Extensions to the mathematical/computational model

1. *Effective* population size, not census population size
2. Demographic changes generally: population flux, migration, gene flow
3. Recombination – turns the trees into general networks.
4. Selection–gene copies no longer act ‘independently’
5. Statistical– to get confidence limits, etc., must simulate over many generated ‘trees’ – use likelihood methods (Computer packages: Lamarc; Simcoal2; …)

## Complications make the simple coalescent look more complicated! Population flow, Recombination, Migration

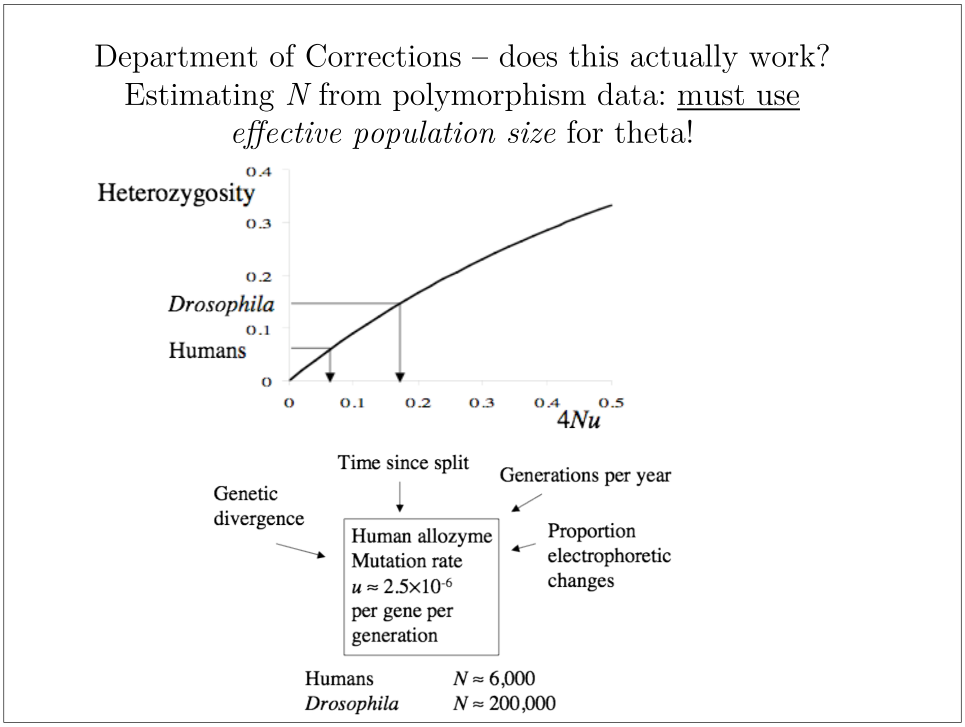




## Demographic corrections
## Part 1 – Effective population size, $N_e$

$N$ = # individuals in a *theoretical* population that, subjected to the same magnitude of drift, would present an equivalent level of diversity

- Population growth
- Population bottlenecks
- Subdivided populations
- Population splits
- Admixture



## Department of Corrections – does this actually work?
## Estimating $N$ from polymorphism data: must use *effective population size* for theta!

Time since split

Genetic divergence

Generations per year

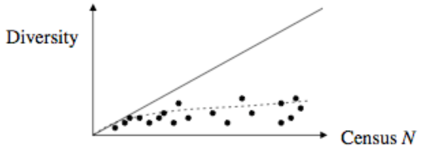
Human allozyme Mutation rate $u \approx 2.5\times10^{-6}$ per gene per generation

Proportion electrophoretic changes

| | |
|---|---|
| Humans | $N \approx 6{,}000$ |
| *Drosophila* | $N \approx 200{,}000$ |

## Patching the model; demographics *matters*

- Levels of polymorphism vary less between species than the census population size



- The rate of genetic drift varies due to
  - Inbreeding, skewed sex ratios, fluctuating population size, variation in family size
- Many biologically realistic complications can be modelled by a coalescent process with a smaller EFFECTIVE population size

$$N \rightarrow N_e \qquad E[\pi] = 4N_e u \qquad \theta = 4N_e u$$

## Effective population size, $N_e$

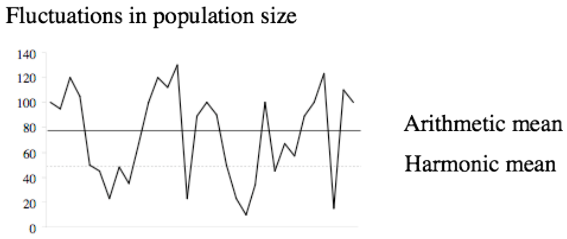
Working definition: the size of an ideal population that has the *same properties with respect to genetic drift* as the actual population does

Lots of ways to define what's in italics…

1. Variance adjustment
2. Inbreeding adjustment

(first related to # of individuals in offspring generation; second related to # of individuals in parental generation)

## Fluctuating population size



$$N_e = \frac{1}{\frac{1}{t}\sum \frac{1}{2N_i}}$$

Example: if population size is 1000 w/ pr 0.9 and 100 w/ pr 0.1, arithmetic mean is 901, but the harmonic mean is (0.9 x 1/1000 + 0.1 x 1/10)$^{-1}$ = 91.4, an order of magnitude less!

Thus, if we have a population (like humans, cheetahs) going through a 'squeeze', this *changes* the population sizes, hence θ

But *Why* do we use the harmonic mean???

## In general:Variance effective population size

Let $Var(p)$ be the variance calculated for our actual population $N_e^{(v)}$ = effective population size adjusted for this variance

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

$$N_e^{(v)} = \frac{p(1-p)}{2\widehat{Var}(p)}$$

## Effective population size must be used to 'patch' the Wright-Fisher model to keep the *variance* the same

*Standard* variance is $pq/N$

Variance for $N_1$ is $p(1–p)/2N_1$ with probability $r$
Variance for $N_2$ is $p(1–p)/2N_2$ with probability $1–r$
Average these 2 populations together, to get mean variance, 'solve' for $N_e$

$$Var[p'] = p(1-p)\left(\frac{r}{2N_1} + \frac{1-r}{2N_2}\right) \text{ or}$$

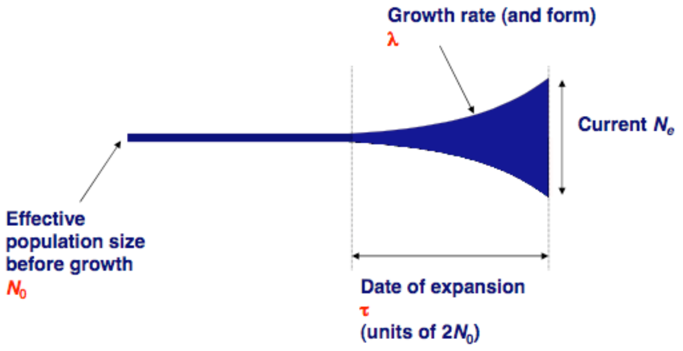
$$N_e = \frac{1}{r\frac{1}{N_1} + (1-r)\frac{1}{N_2}}$$

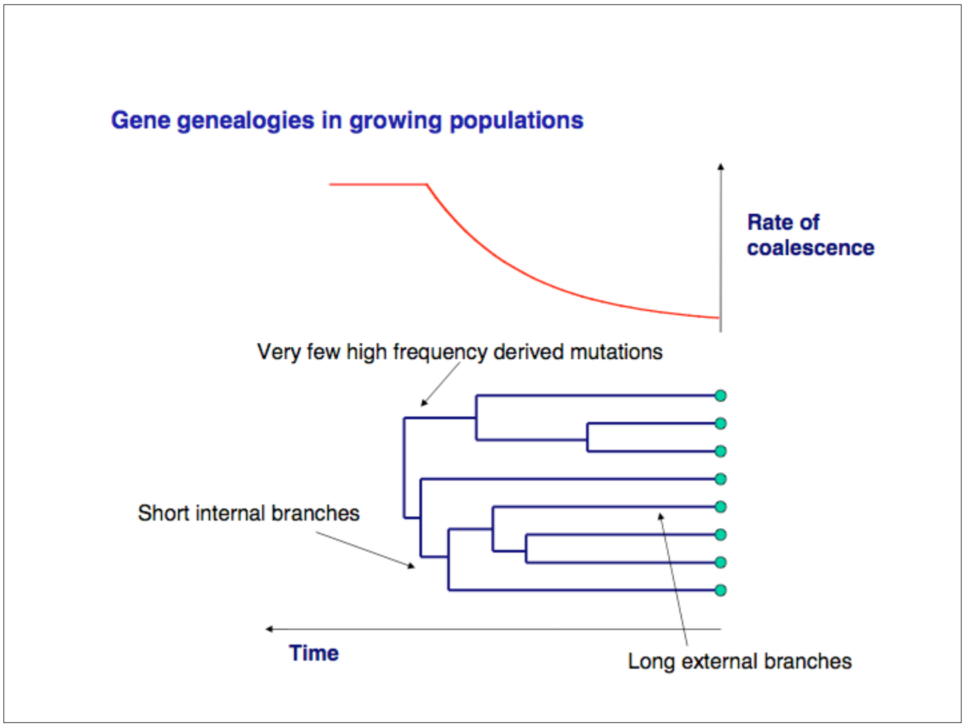
*i.e.*, the harmonic mean of the population sizes (the reciprocal of the average of the reciprocals) is used because it averages the variation properly!

*Always* smaller than the mean; *Much more sensitive to* small numbers

## Demographic Corrections, part 2: effects on coalescent

- Exponentially growing populations
  - Humans, *HIV*-1 (within patients), *HIV*-1 (worldwide)





## Gene genealogies in growing populations

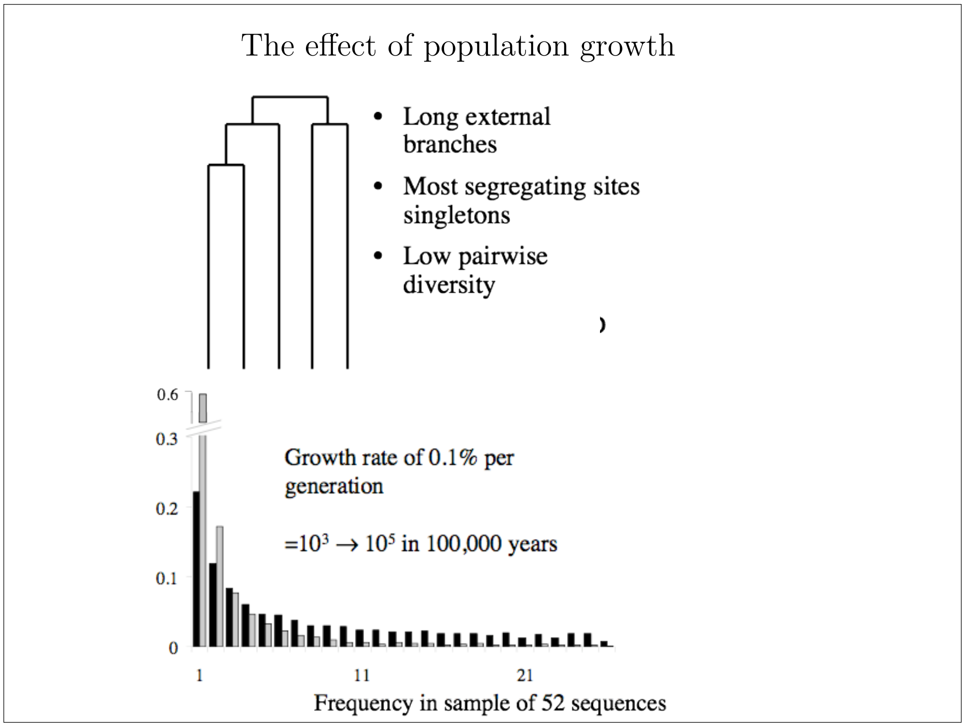
Rate of coalescence

Very few high frequency derived mutations

Short internal branches

Time

Long external branches

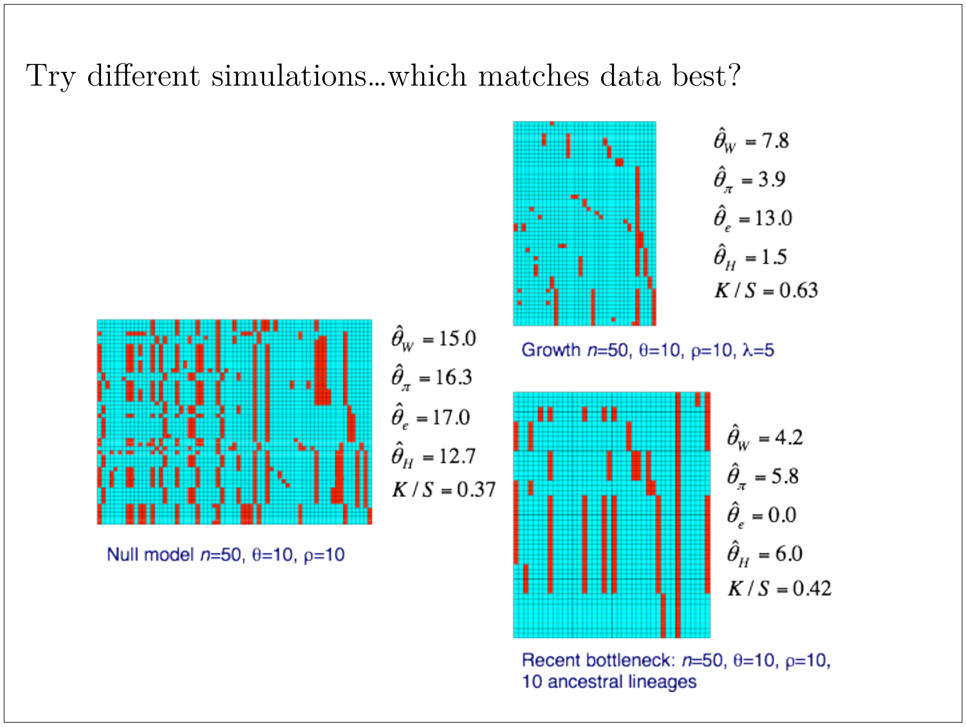


## The effect of population growth

- Long external branches
- Most segregating sites singletons
- Low pairwise diversity

Growth rate of 0.1% per generation

=$10^3 \rightarrow 10^5$ in 100,000 years



Try different simulations...which matches data best?

$\hat{\theta}_W = 15.0$

$\hat{\theta}_\pi = 16.3$

$\hat{\theta}_e = 17.0$

$\hat{\theta}_H = 12.7$

$K / S = 0.37$

Null model $n$=50, θ=10, ρ=10

$\hat{\theta}_W = 7.8$

$\hat{\theta}_\pi = 3.9$

$\hat{\theta}_e = 13.0$

$\hat{\theta}_H = 1.5$

$K / S = 0.63$

Growth $n$=50, θ=10, ρ=10, λ=5

$\hat{\theta}_W = 4.2$

$\hat{\theta}_\pi = 5.8$

$\hat{\theta}_e = 0.0$

$\hat{\theta}_H = 6.0$

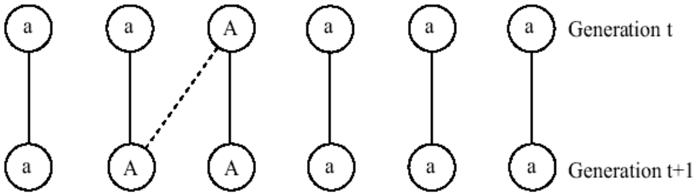
$K / S = 0.42$

Recent bottleneck: $n$=50, θ=10, ρ=10,
10 ancestral lineages

# Why is modeling selection hard with the coalescent?

Problem: Genealogical and mutation processes no longer independent!

Two alleles, A and a, A has an advantage of $s$
Mutation rate between types = $u$



Krone and Neuhauser 1997

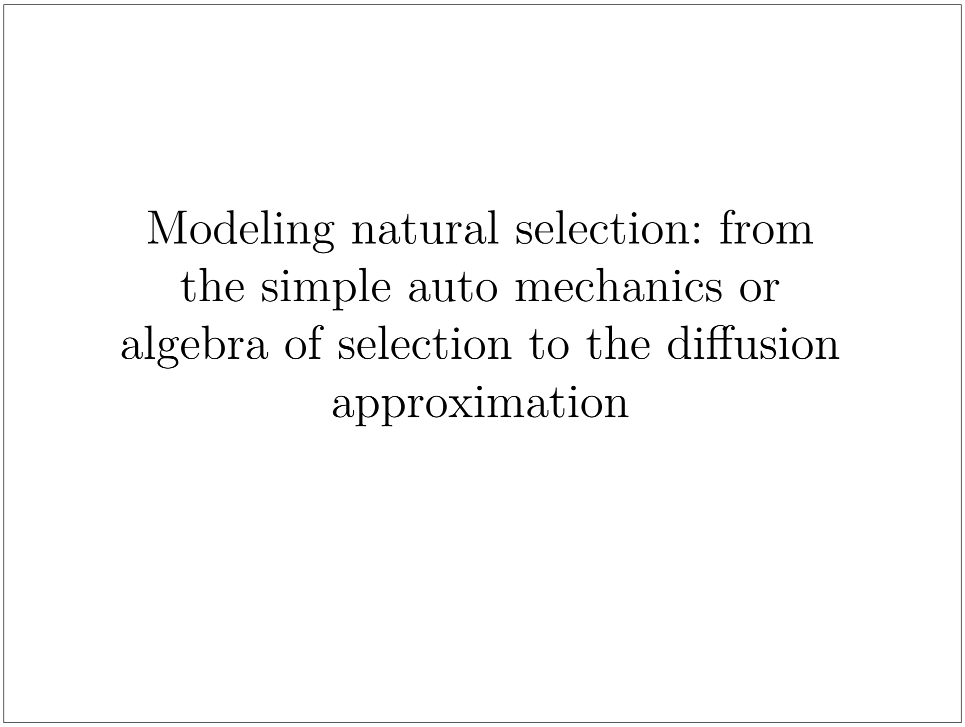
# Summary so far…

| | Whole genome effect | Local effect |
|---|---|---|
| Long external branches<br>(Tajima's $D < 0$) | Population growth<br>Very severe bottleneck | Directional selection |
| Long internal branches<br>(Tajima's $D > 0$) | Population subdivision<br>Less severe bottleneck | Balancing selection<br>Recent population mixing |

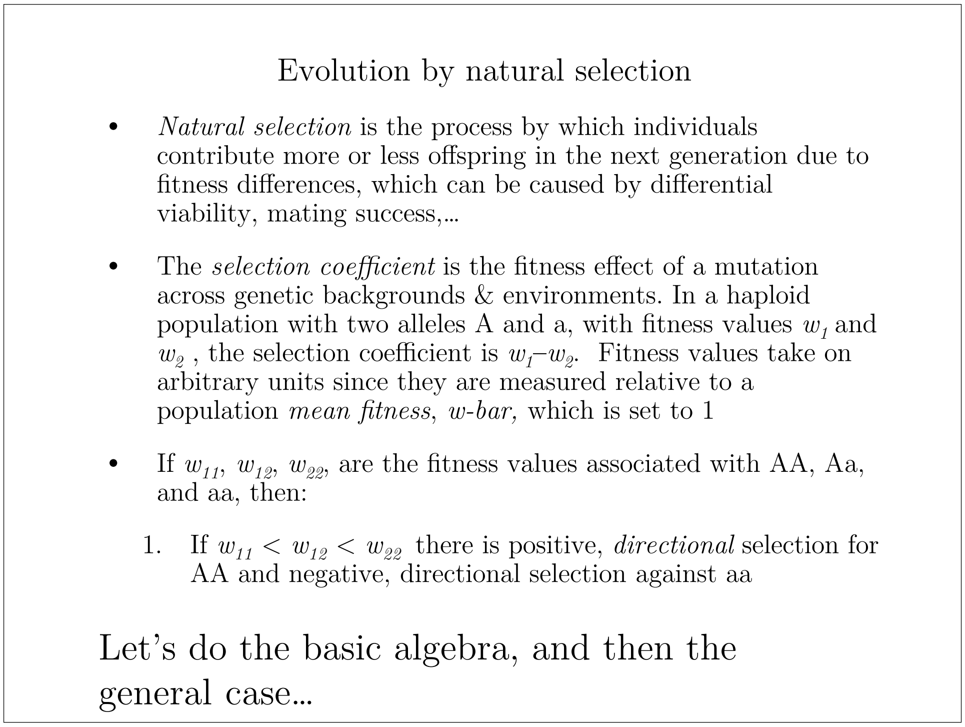
A strong bottleneck resembles population growth
A weaker bottleneck resembles directional selection for some loci
And balancing selection for other loci

This is where current computer packages take us!

Screenshots removed due to copyright reasons.
Please see:
University of Oxford, Department of Zoology,
Evolutionary Biology Group: http://evolve.zoo.ox.ac.uk/software.html



# Modeling natural selection: from the simple auto mechanics or algebra of selection to the diffusion approximation



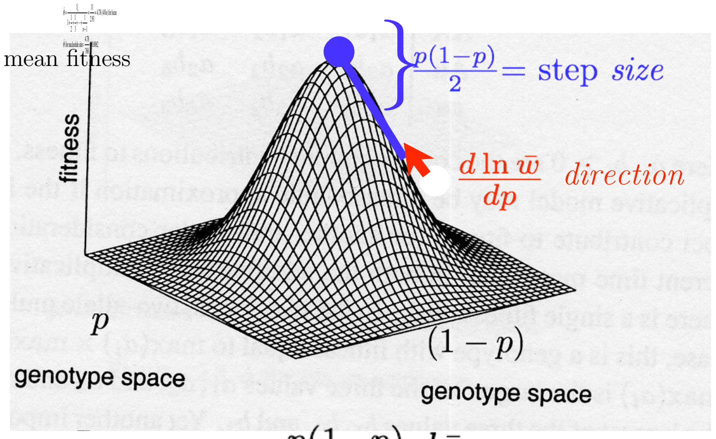
## Evolution by natural selection

- *Natural selection* is the process by which individuals contribute more or less offspring in the next generation due to fitness differences, which can be caused by differential viability, mating success,…
- The *selection coefficient* is the fitness effect of a mutation across genetic backgrounds & environments. In a haploid population with two alleles A and a, with fitness values $w_1$ and $w_2$ , the selection coefficient is $w_1$–$w_2$. Fitness values take on arbitrary units since they are measured relative to a population *mean fitness*, *w-bar,* which is set to 1
- If $w_{11}$, $w_{12}$, $w_{22}$, are the fitness values associated with AA, Aa, and aa, then:

  1. If $w_{11} < w_{12} < w_{22}$ there is positive, *directional* selection for AA and negative, directional selection against aa

Let's do the basic algebra, and then the general case…

## Sewall Wright's adaptive landscape: Understanding the formula


mean fitness
fitness
$\frac{p(1-p)}{2}$ = step *size*
$\frac{d\ln\bar{w}}{dp}$ *direction*
$p$
genotype space
$(1-p)$
genotype space

$$\Delta p = \frac{p(1-p)}{2\bar{w}}\frac{d\bar{w}}{dp}$$

$$\Delta p = \frac{p(1-p)}{2}\frac{d\ln(\bar{w})}{dp}$$

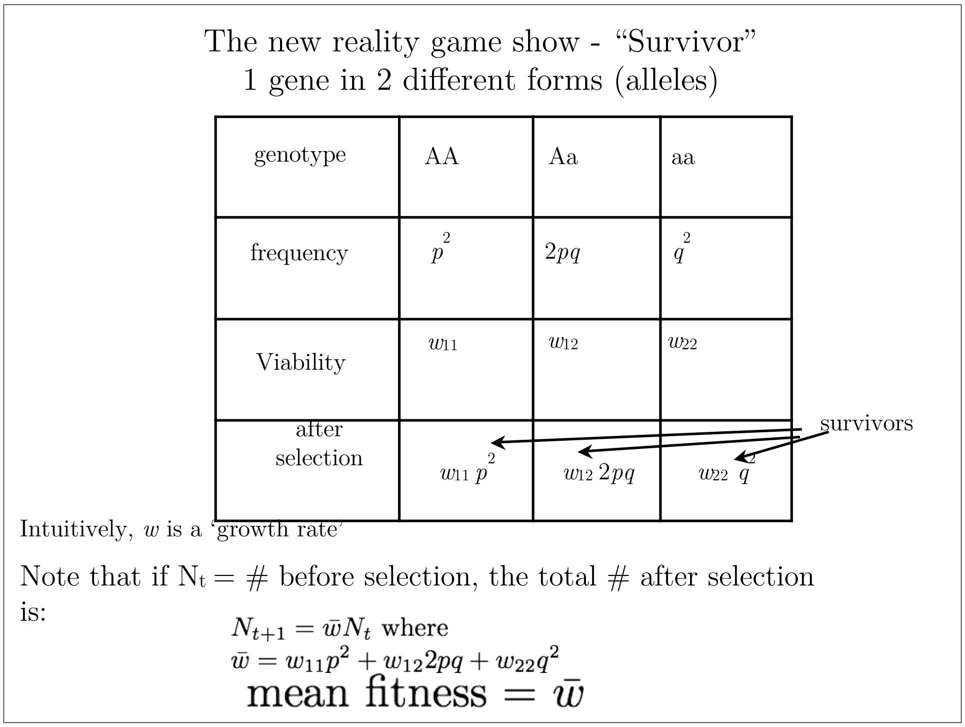
## Some dissection…

$$\Delta p = \frac{p(1-p)}{2}\,\frac{d(\bar{w})}{\bar{w}dp}$$

*Variance* component of allele A within genotype

Slope of fitness function divided by mean population fitness – a *potential function*?

Why variance? Draw from pool of $A$, $a$ gametes many many times: binomial sampling – frequency of $A$ within a genotype is either 1, 1/2, or 0; variance is $p$(1-$p$)/2 ("heterozygosity")



The new reality game show - “Survivor”
1 gene in 2 different forms (alleles)

| genotype | AA | Aa | aa |
|---|---|---|---|
| frequency | $p^2$ | $2pq$ | $q^2$ |
| Viability | $w_{11}$ | $w_{12}$ | $w_{22}$ |
| after selection | $w_{11}\,p^2$ | $w_{12}\,2pq$ | $w_{22}\;q^2$ |

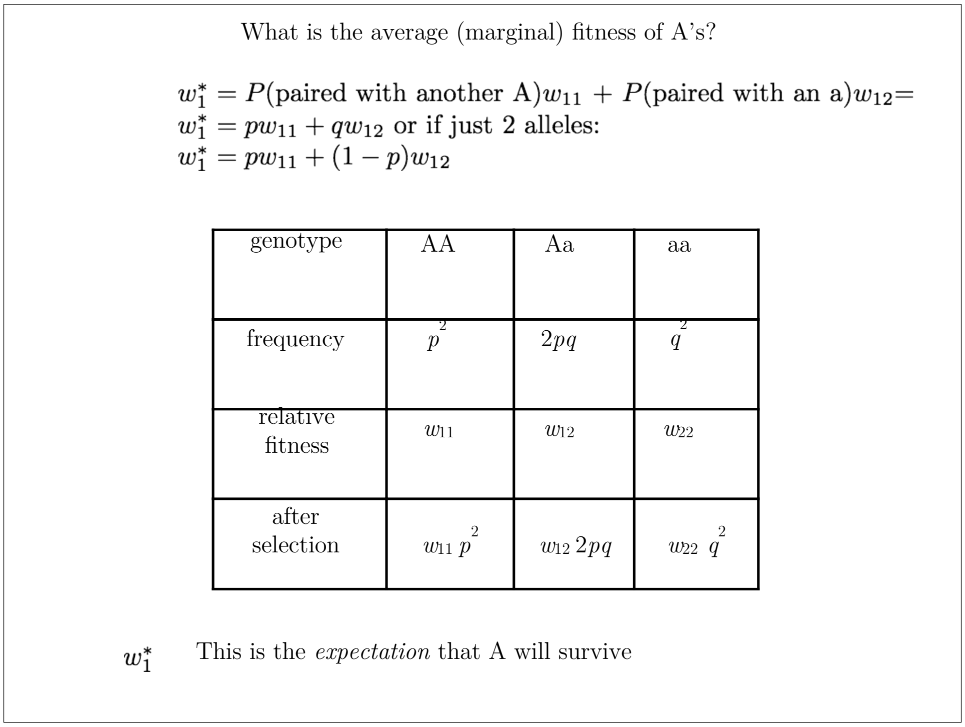
survivors

Intuitively, $w$ is a ‘growth rate’

Note that if $N_t$ = # before selection, the total # after selection is:

$$N_{t+1} = \bar{w}N_t \text{ where}$$
$$\bar{w} = w_{11}p^2 + w_{12}2pq + w_{22}q^2$$

$$\text{mean fitness} = \bar{w}$$

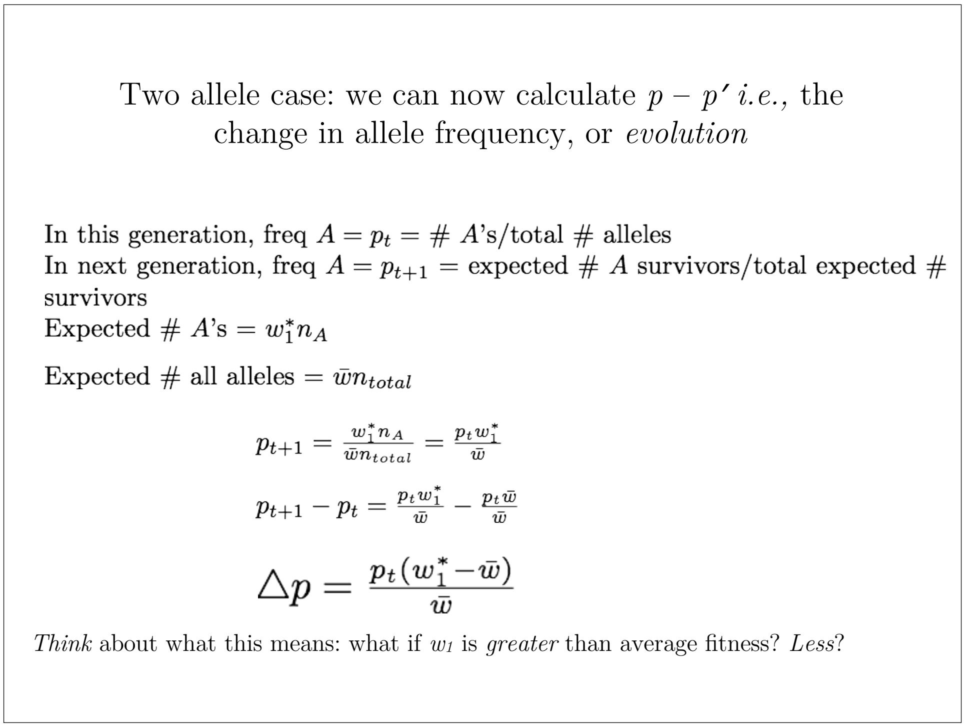


What is the average (marginal) fitness of A’s?

$$w_1^* = P(\text{paired with another A})w_{11} + P(\text{paired with an a})w_{12} =$$
$$w_1^* = pw_{11} + qw_{12} \text{ or if just 2 alleles:}$$
$$w_1^* = pw_{11} + (1-p)w_{12}$$

| genotype | AA | Aa | aa |
|---|---|---|---|
| frequency | $p^2$ | $2pq$ | $q^2$ |
| relative fitness | $w_{11}$ | $w_{12}$ | $w_{22}$ |
| after selection | $w_{11}\,p^2$ | $w_{12}\,2pq$ | $w_{22}\;q^2$ |

$w_1^*$ This is the *expectation* that A will survive



## Two allele case: we can now calculate $p - p'$ *i.e.,* the change in allele frequency, or *evolution*

In this generation, freq $A = p_t = \#$ $A$'s/total # alleles
In next generation, freq $A = p_{t+1} =$ expected # $A$ survivors/total expected # survivors
Expected # $A$'s $= w_1^* n_A$

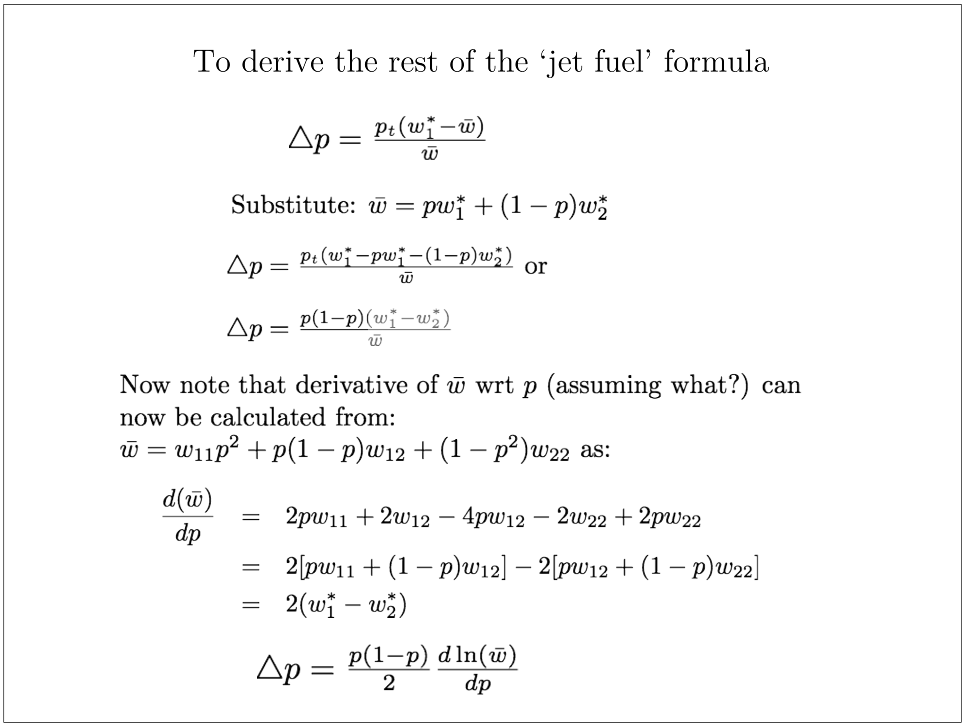
Expected # all alleles $= \bar{w} n_{total}$

$$p_{t+1} = \frac{w_1^* n_A}{\bar{w} n_{total}} = \frac{p_t w_1^*}{\bar{w}}$$

$$p_{t+1} - p_t = \frac{p_t w_1^*}{\bar{w}} - \frac{p_t \bar{w}}{\bar{w}}$$

$$\triangle p = \frac{p_t(w_1^* - \bar{w})}{\bar{w}}$$

*Think* about what this means: what if $w_1$ is *greater* than average fitness? *Less*?



## To derive the rest of the 'jet fuel' formula

$$\triangle p = \frac{p_t(w_1^* - \bar{w})}{\bar{w}}$$

Substitute: $\bar{w} = p w_1^* + (1-p) w_2^*$

$$\triangle p = \frac{p_t(w_1^* - p w_1^* - (1-p) w_2^*)}{\bar{w}} \text{ or}$$

$$\triangle p = \frac{p(1-p)(w_1^* - w_2^*)}{\bar{w}}$$

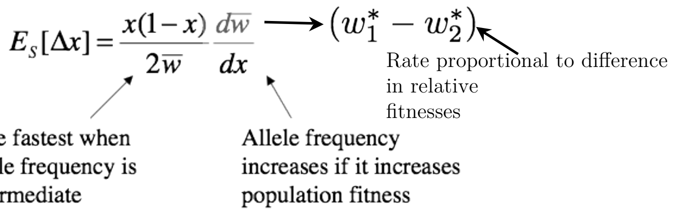
Now note that derivative of $\bar{w}$ wrt $p$ (assuming what?) can now be calculated from:
$\bar{w} = w_{11} p^2 + p(1-p) w_{12} + (1-p^2) w_{22}$ as:

$$\begin{aligned} \frac{d(\bar{w})}{dp} &= 2pw_{11} + 2w_{12} - 4pw_{12} - 2w_{22} + 2pw_{22} \\ &= 2[pw_{11} + (1-p)w_{12}] - 2[pw_{12} + (1-p)w_{22}] \\ &= 2(w_1^* - w_2^*) \end{aligned}$$

$$\triangle p = \frac{p(1-p)}{2} \frac{d \ln(\bar{w})}{dp}$$

## The ‘jet fuel’ formula

$$\triangle p = \frac{p_t(w_1^* - \bar{w})}{\bar{w}}$$



$$E_s[\Delta x] = \frac{x(1-x)}{2\overline{w}}\frac{d\overline{w}}{dx} \longrightarrow (w_1^* - w_2^*)$$

Rate proportional to difference in relative fitnesses

Rate fastest when allele frequency is intermediate

Allele frequency increases if it increases population fitness

Adaptation is not instantaneous:
The ratio of $p$ to (1-$p$) changes by $w_1/w_2$ every generation

After $t$ generations,

$$\frac{p_t}{(1-p_t)} = \frac{p_0}{(1-p_0)}\left(\frac{w_1}{w_2}\right)^t$$

## Getting a feel for the dynamics

Genotype:
AA Aa aa
Relative fitness: 1
1-$hs$ 1-$s$
1–$hs$= $w_{12}/w_{11}$
1–$s$ = $w_{22}$ /$w_{11}$

$s$= selection coefficient. Measure of fitness of AA relative to aa.
If positive, aa is *less* fit than AA; if negative, aa is *more* fit
$h$= heterozygous effect. Measure of fitness of heterozygote relative to selective difference between the two homozygotes –a measure of dominance:

$h$=0, A dominant, a recessive
$h$=1, a dominant, A recessive
0< $h$ <1 incomplete dominance
$h$<0 overdominance
h >1 underdominance

## Dynamical system analysis of ‘adaptive topography’ or mean fitness vs. $p$ - nondegenerate case

$$\frac{d(\bar{w})}{dp} = 2(w_1^* - w_2^*) = 0 \text{ or}$$

$$w_1^* = w_2^*, \text{ so}$$

$$w_{11}p + w_{12}(1-p) = w_{12}p + w_{22}(1-p) =$$

$$w_{11}p + w_{12} - w_{12}p = w_{12}p + w_{22} - w_{22}p$$

$$p[(w_{11} - w_{12}) + (w_{22} - w_{12})] = w_{22} - w_{12}$$

Equillibrium value of $p$

$$\hat{p} = \frac{w_{22} - w_{12}}{[(w_{11} - w_{12}) + (w_{22} - w_{12})]}$$

## The delta $p$ equation in these terms (relative fitnesses)

$$p' = \frac{p^2 w_{11} + pqw_{12}}{\bar{w}}$$

$$p' - p = \frac{p^2 w_{11} + pqw_{12} - p\bar{w}}{\bar{w}}$$

$$\Delta p = \frac{pqs[ph + q(1-h)]}{1 - 2pqhs - q^2 s}$$

where $\bar{w} = 1 - 2pqhs - q^2 s$

*h* determines where allele frequency ends up;
*s* determines how quickly it gets there

There turn out to be <u>three</u> kinds of selection:

dominant (AA> Aa > aa);
overdominant (Aa> AA, aa);
underdominant (AA, aa > Aa)

## Some dissection…

$$\triangle p = \frac{p(1-p)}{2} \frac{d(\bar{w})}{\bar{w} dp}$$

*Variance* component of allele A within genotype

Slope of fitness function divided by mean population fitness – a *potential function*?

Why variance? Draw from pool of $A$, $a$ gametes many many times: binomial sampling – frequency of $A$ within a genotype is either 1, 1/2, or 0; variance is $p$(1-$p$)/2 (“heterozygosity”)

## Some exploration, fitness AA is 1.0; Aa = 0.95, aa= 0.90

Screenshots removed due to copyright reasons.

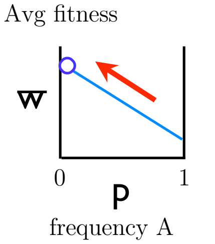
## Plot avg fitness vs $p$ to get feel for the dynamics…

Note that avg fitness is a quadratic function so it can have at most 1 minimum or maximum…

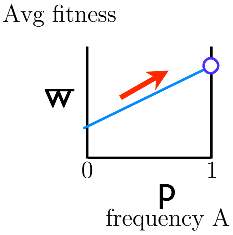
$$\begin{aligned}\bar{w} &= w_{11}p^2 + w_{12}2p(1-p) + w_{22}(1-p)^2\\ &= w_{11}p^2 + w_{12}2p - w_{12}2p^2 + w_{22} - w_{22}2p + w_{22}p^2\\ &= p^2[(w_{11}-w_{12}) + (w_{22}-w_{12})] - 2p[w_{22}-w_{12}] + w_{22}\end{aligned}$$

'Degenerate' case: quadratic mean fitness, with $w_{12} = (w_{11} + w_{22})/2$

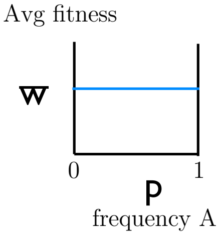
## One locus, 2 allele case: graphs, $p$ vs. $\bar{w}$



Directional selection

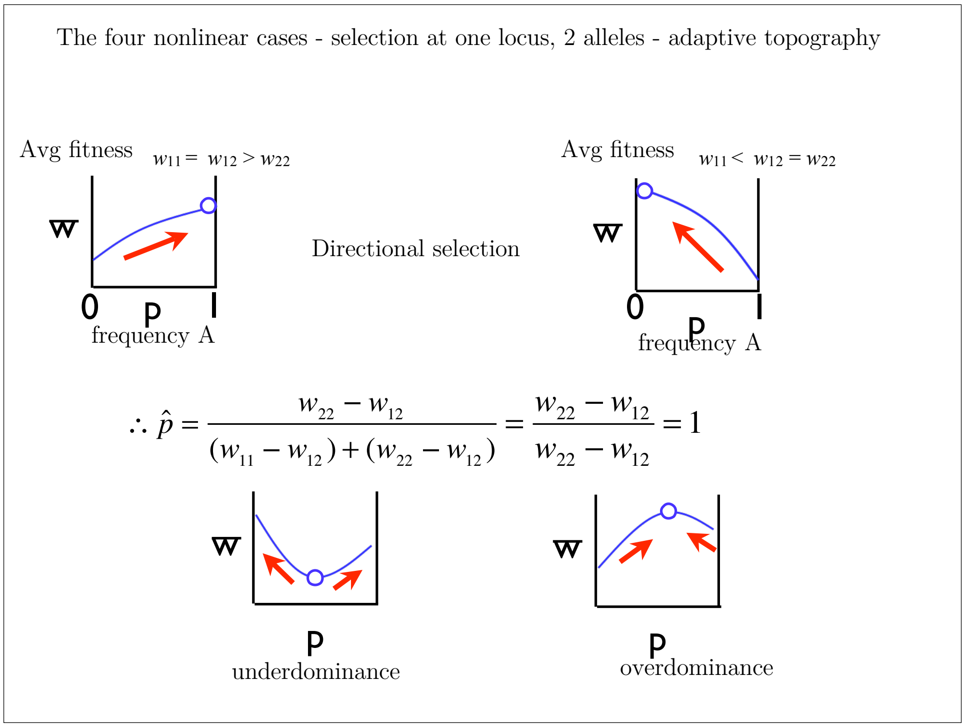


Directional selection



Zip selection

'Degenerate' case: quadratic mean fitness, with $w_{12} = (w_{11} + w_{22})/2$



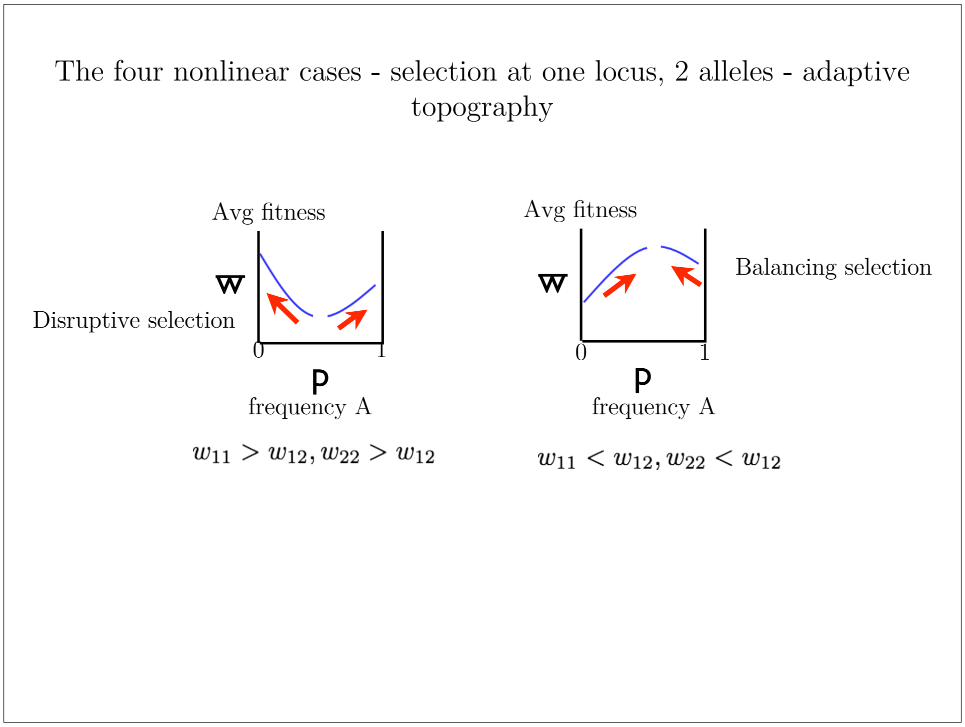
The four nonlinear cases - selection at one locus, 2 alleles - adaptive topography

Directional selection

$$\therefore \hat{p} = \frac{w_{22} - w_{12}}{(w_{11} - w_{12}) + (w_{22} - w_{12})} = \frac{w_{22} - w_{12}}{w_{22} - w_{12}} = 1$$

underdominance

overdominance



The four nonlinear cases - selection at one locus, 2 alleles - adaptive topography

Disruptive selection

Balancing selection

$w_{11} > w_{12}, w_{22} > w_{12}$

$w_{11} < w_{12}, w_{22} < w_{12}$

## The multiple allele jet-fuel formula

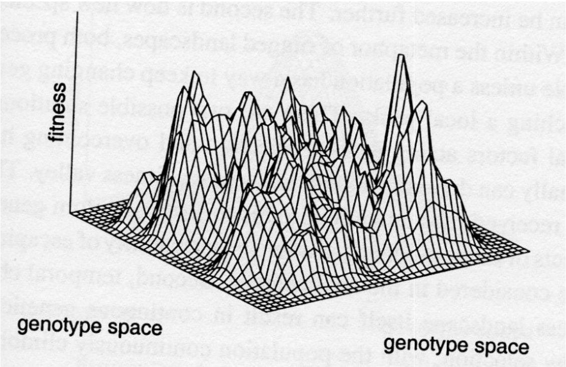
$$\Delta\vec{p}_i = \Delta\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \frac{1}{2\overline{w}}\begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_2p_1 & p_2(1-p_2) & -p_2p3 \\ -p_3p_1 & -p_3p_2 & p_3(1-p_3) \end{bmatrix}\begin{bmatrix} \frac{\partial\overline{w}}{\partial p_1} \\ \frac{\partial\overline{w}}{\partial p_2} \\ \frac{\partial\overline{w}}{\partial p_3} \end{bmatrix}$$

variance

$$\Delta\vec{p} = \frac{G}{\overline{w}}\nabla\overline{w}$$

grad mean fitness
wrt $p_i$

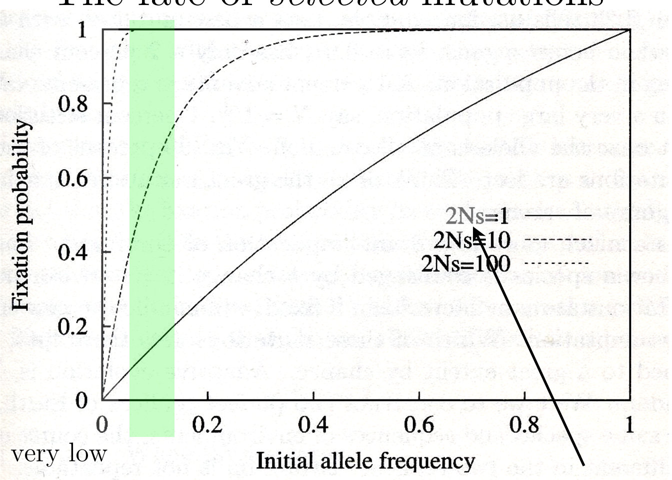
But...



## Climb every mountain? Some surprising results

- The power of selection: what is the fixation probability for a new mutation?
- If no selection, the *pr* **of loss in a** *single generation* **is 1/*e* or 0.3679**
- In particular: suppose new mutation has 1% selection advantage as heterozygote – this is a *huge* difference
- Yet this will have only a 2% chance of ultimate fixation, starting from 1 copy (in a *finite* population a Poisson # of offspring, mean 1+*s*/2, the Pr of extinction in a *single generation* is $e^{-1}$(1-*s*/2), e.g., **0.3642** for *s*= 0.01)
- Specifically, to be 99% certain a new mutation will fix, for *s*= 0.001, we need about 4605 allele copies (independent of population size *N* !!)
- Also very possible for a *deleterious* mutation to fix, if 2*Ns* is close to 1
- Why? Intuition: look at the shape of the selection curve – flat at the start, strongest at the middle
- To understand this, we'll have to dig into how variation changes from generation to generation, in finite populations

## The fate of *selected* mutations

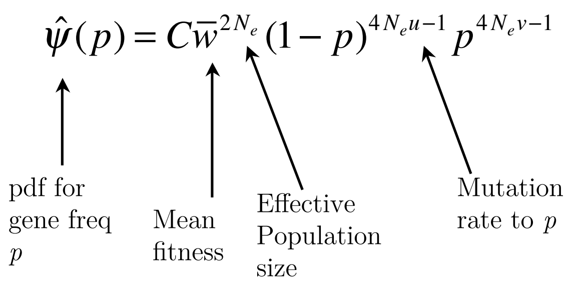


Regime 1: very low copy #

Regime 2: Frequency matters

$2Ns$ (compare to $Nu$ factor)

## Time to fixation for *selected* genes: can we find this in face of population size, mutation, drift?



$$\hat{\psi}(p) = C\overline{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$

pdf for gene freq *p*

Mean fitness

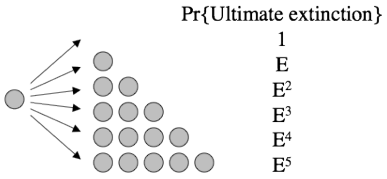
Effective Population size

Mutation rate to *p*

Pr{Fixation}= 1–Pr{extinction} = 2*s*

The fixation probability of *selected* alleles – large population (no effects from ‘demographic stochastics’)

- Branching process argument (Haldane 1927)



Pr{Ultimate extinction}

1

E

$E^2$

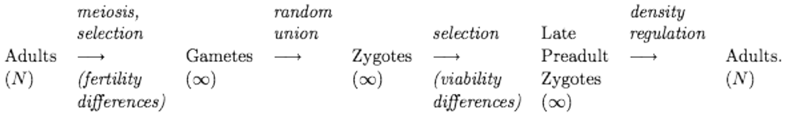
$E^3$

$E^4$

$E^5$

Pr{Extinction} =

$$\lambda \ = \ p_0 + p_1\lambda + p_2\lambda^2 + p_3\lambda^3 + ... + p_k\lambda^k + ...,$$



Adults $(N)$ $\xrightarrow{\text{meiosis, selection}}$ *(fertility differences)* Gametes $(\infty)$ $\xrightarrow{\text{random union}}$ Zygotes $(\infty)$ $\xrightarrow{\text{selection}}$ *(viability differences)* Late Preadult Zygotes $(\infty)$ $\xrightarrow{\text{density regulation}}$ Adults. $(N)$

Assume binomial draw with $N$ trials, pr of success on each trial is $(1+s)/N$

For $N$ large, this is Poisson with mean $1+s$, so the # of Aa with $k$ *surviving* offspring has probability:

$$p_k \;=\; e^{-(1+s)}(1+s)^k/k!,$$

Substitute back for $p_k$

$$\begin{aligned}\lambda &= e^{-(1+s)} + e^{-(1+s)}(1+s)\lambda + e^{-(1+s)}(1+s)^2\lambda^2/2 + \dots + e^{-(1+s)}(1+s)^k/k! + \dots \\ &= e^{-(1+s)}[1 + (1+s)\lambda + (1+s)^2\lambda^2/2 + \dots + (1+s)^k\lambda^k/k! + \dots].\end{aligned}$$

This is a Taylor series expansion of $e^x$ so we can rewrite as:

$$\begin{aligned}\lambda &= e^{-(1+s)}e^{\lambda(1+s)} \\ &= e^{(\lambda-1)(1+s)}.\end{aligned}$$

If $s$ is small, then expand RHS as power series in lambda, dropping terms beyond square, ie, lambda is near 1

$$\lambda \;\simeq\; 1 + (\lambda-1)(1+s) + (\lambda-1)^2(1+s)^2/2$$

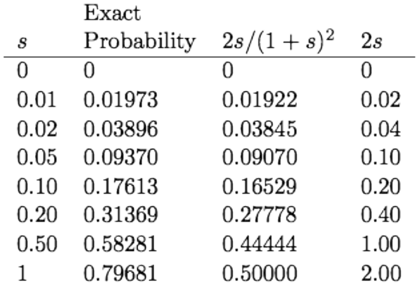
$$(\lambda-1)[1-(1+s)-(\lambda-1)(1+s)^2/2] \;\simeq\; 0$$

Solved when either $\lambda = 1$ or

$$1-\lambda \;\simeq\; \frac{2s}{(1+s)^2}$$

So when $s$ is small, pr of survival of new mutant is either very nearly $2s$ or else 0 (if $s$ less than 0)
When $s$=0.01, only 1 new mutant in 50 will succeed in spreading, despite that all are advantageous; if $s$=0.1, which fixes very rapidly in deterministic case, only 1 in 6 will win

So, $2s$ turns out to be a good approximation to the exact fixation probability for small $s$



| $s$ | Exact Probability | $2s/(1+s)^2$ | $2s$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0.01 | 0.01973 | 0.01922 | 0.02 |
| 0.02 | 0.03896 | 0.03845 | 0.04 |
| 0.05 | 0.09370 | 0.09070 | 0.10 |
| 0.10 | 0.17613 | 0.16529 | 0.20 |
| 0.20 | 0.31369 | 0.27778 | 0.40 |
| 0.50 | 0.58281 | 0.44444 | 1.00 |
| 1 | 0.79681 | 0.50000 | 2.00 |

# of copies of allele matters - must get over the initial ‘hump’

Pr that $n$ copies go extinct (since all lineages are independent):

$$1-\lambda^n \simeq (1-2s)^n.$$

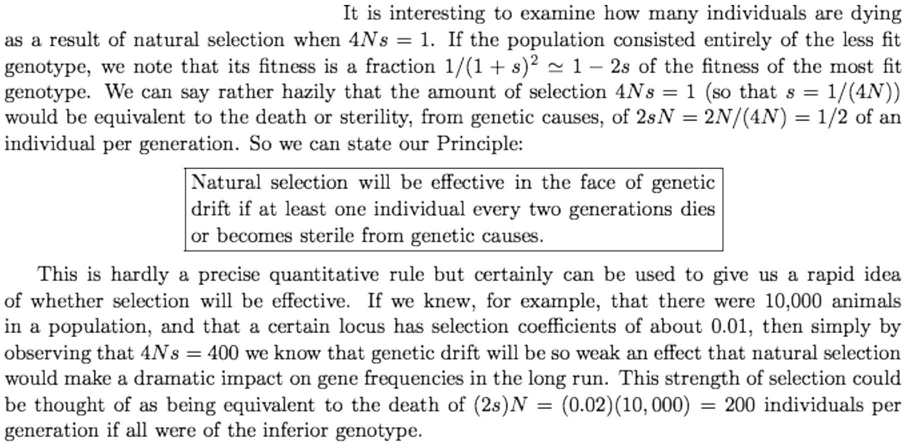
Eg, once 100 copies present, $s$=0.01, pr loss is only 0.14; with 1000 copies, less than 3 x $10^{-9}$

Tells us about time course of selection with new mutation: it *does* follow two regimes…

What about branching process vs. deterministic equations - the difference is between the # of copies and the gene *frequency*

How do we put back drift?

## And a General Rule



It is interesting to examine how many individuals are dying as a result of natural selection when $4Ns = 1$. If the population consisted entirely of the less fit genotype, we note that its fitness is a fraction $1/(1+s)^2 \simeq 1 - 2s$ of the fitness of the most fit genotype. We can say rather hazily that the amount of selection $4Ns = 1$ (so that $s = 1/(4N)$) would be equivalent to the death or sterility, from genetic causes, of $2sN = 2N/(4N) = 1/2$ of an individual per generation. So we can state our Principle:

> Natural selection will be effective in the face of genetic drift if at least one individual every two generations dies or becomes sterile from genetic causes.

This is hardly a precise quantitative rule but certainly can be used to give us a rapid idea of whether selection will be effective. If we knew, for example, that there were 10,000 animals in a population, and that a certain locus has selection coefficients of about 0.01, then simply by observing that $4Ns = 400$ we know that genetic drift will be so weak an effect that natural selection would make a dramatic impact on gene frequencies in the long run. This strength of selection could be thought of as being equivalent to the death of $(2s)N = (0.02)(10,000) = 200$ individuals per generation if all were of the inferior genotype.

But what about the interaction with drift??????

# The whole banana: the diffusion approximation to evolutionary processes

Our goal: how can we model the full interaction of stochastic forces and selection, mutation, migration…?

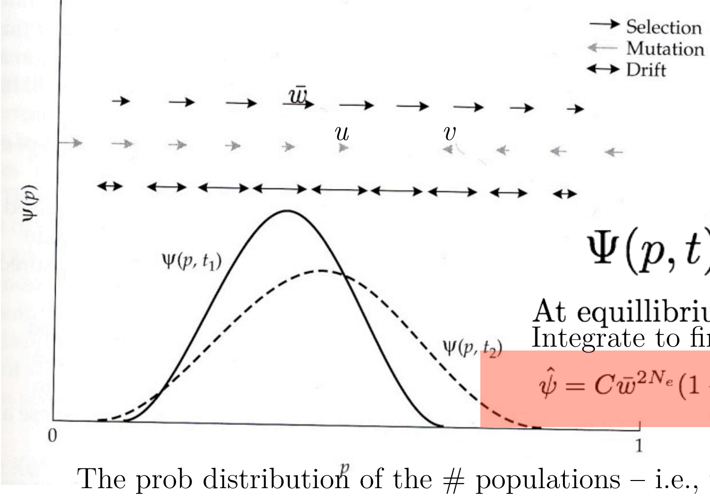
Our answer: write an approximating *differential equation* that involves all these 'forces'

$$\frac{\partial\psi(p,t)}{\partial t} = -\frac{\partial}{\partial p}\left[\psi(p,t)M(p)\right] + \frac{1}{2}\frac{\partial^2}{\partial p^2}\left[\psi(p,t)V(p)\right]$$

Set to 0 and solve to find equilibrium allele frequency distribution for $p$

$$\hat{\psi} = C\bar{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$

Diffusion Theory: Conceptual Framework



$$\Psi(p,t) = \text{pdf}$$

At equillibrium, $\frac{\partial\psi(p,t)}{\partial t} = 0$
Integrate to find:

$$\hat{\psi} = C\bar{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$

The prob distribution of the # populations – i.e., the prob density – with different allele frequencies *shifts* under the directional effects of selection & mutation (& migration) and *flattens out* under the effects of drift

Two ‘classes’ of evolutionary ‘processes’ pushing a population into and out of a time slice of allele frequencies from $p$ to $p+e$ (think of heat/water diffusing along a pipe)

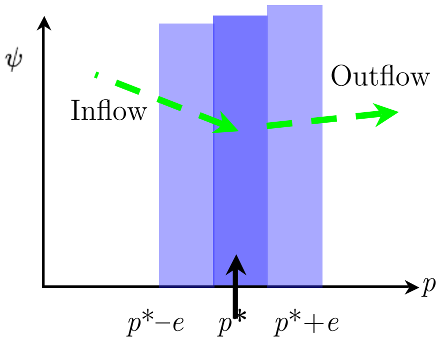
1. **Directional (‘mean’) processes, *M*(*p*):** nonzero expected change in allele frequency within any one population (selection, mutation, migration, recombination) – measured by expected change over one generation.

2. **Nondirectional (‘variance’) processes,** *V*(*p*): produce expected changed of zero but cause distribution to spread – all driftlike processes – measured by expected variance in next generation

Intuitive formulation of this differential equation (the Kolmogorov forward equation)

We want an equation for $\frac{\partial \psi(p,t)}{\partial t}$

Ask: *How* can we figure out the change in density at a region density centered at $p^*$?



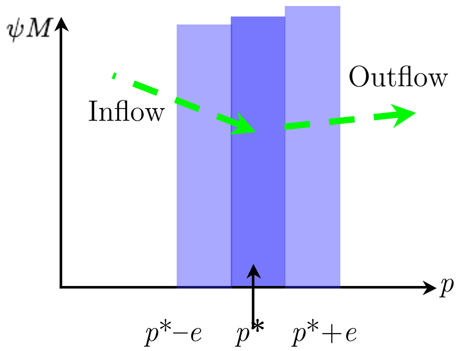
Answer: it can change *either* due to $M(p)$ or $V(p)$ – consider in turn what each can do by figuring out the net **inflow-outflow** that *each* can produce

$$\frac{\partial \psi(p)}{\partial t} = \text{change from } M(p) + \text{ change from } V(p)$$

## Contribution from $M(p)$

Inflow – outflow calculation into slice centered at $p^*$ for *directional* evolutionary process, $M(p)$ – shifts entire density distribution over. Note that $M(p)$ is the *rate* of flow

The *direction* of flow is fixed; what matters is the magnitude or *volume* of the region from which it originates – the difference in volume to the left of $p^*$ and the volume slice at $p^*$



Therefore: (1) flow into this region is given by density centered at $p^*$-$e$ times rate of flow at $p^*$-$e$, or:

$$\psi(p^*-e)M(p^*-e)$$

(2) Flow out of this region is given by density centered at $p^*$ times rate of flow at $p$, or:

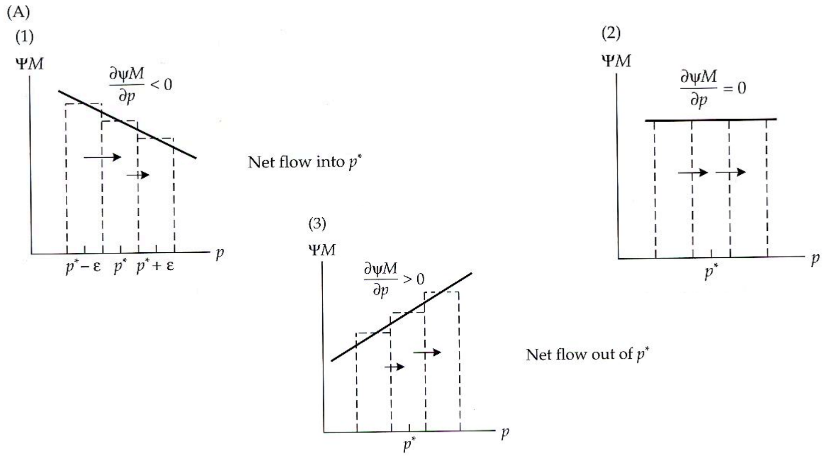
$$\psi(p^*)M(p^*)$$

Putting these together:

$\partial\psi(p^*) = \psi(p^*-e)M(p^*-e) - \psi(p^*)M(p^*)$ and let $dp \to 0$ we get $\frac{\partial}{\partial p}\left[\psi(p,t)M(p)\right]$

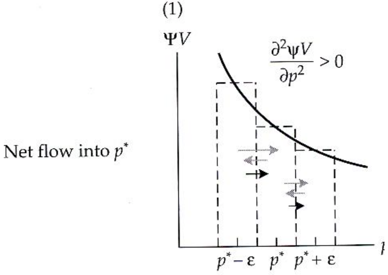
now we flip the sign (why?) to get the contribution to $\frac{\partial\psi(p,t)}{\partial t}$ from $M(p)$:

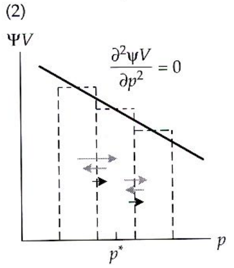
$\frac{\partial\psi(p,t)}{\partial t} = -\frac{\partial}{\partial p}\left[\psi(p,t)M(p)\right]$ +contribution from $V(p)$



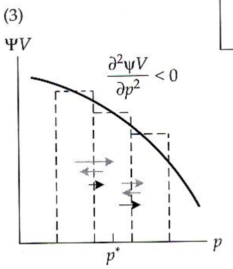
(A)
(1)
ΨM
$\frac{\partial \psi M}{\partial p} < 0$
Net flow into $p^*$
$p^* - \varepsilon$
$p^*$
$p^* + \varepsilon$
$p$
(2)
ΨM
$\frac{\partial \psi M}{\partial p} = 0$
$p^*$
$p$
(3)
ΨM
$\frac{\partial \psi M}{\partial p} > 0$
Net flow out of $p^*$
$p^*$
$p$

For <u>nondirectional</u> processes $V(p)$, populations can move either way – net flow is determined by the *difference* between the *differential* flow to the left of $p^*$ and the *differential* flow to the right of $p^*$, i.e., the 2nd derivative wrt $p$





Further, there's a factor of only 1/2 because: only *half* the populations diffusing out from, say, $p^*$-$e$ go to $p^*$



The Kolmogorov forward equation

$$\frac{\partial \psi(p,t)}{\partial t} = -\frac{\partial}{\partial p}\left[\psi(p,t)M(p)\right] + \frac{1}{2}\frac{\partial^2}{\partial p^2}\left[\psi(p,t)V(p)\right]$$

Now solve for equilibrium by setting this 0....

## Solution for equilibrium frequency

Setting this to 0 and integrating first term over all values of $p$ (since the eqn holds for all values of $p$, we get:

$$\frac{1}{2}\frac{\partial}{\partial p}\left[\hat{\psi}(p,t)V(p)\right] - \left[\hat{\psi}(p,t)M(p)\right] = 0$$

now substitute to get first-order homogenous diffeqn:

$F(p) = \hat{\psi}V(p)$ which gives us:

$$\frac{\partial F}{\partial p} - F\frac{2M(p)}{V(p)} = 0$$

This can be solved by standard means…

## The Grail Quest ends…

$F = Ce^{\int \frac{2M}{V}dp}$ and substituting back for

$F(p) = \hat{\psi}V(p)$ and solving for $\hat{\psi}$ *finally* gives us:

$$\hat{\psi} = \frac{C}{V}e^{\int \frac{2M}{V}dp}$$

Well, almost… let's solve this for particular case of $M$ and $V$

$$M = \frac{p(1-p)}{2\bar{w}}\frac{d\bar{w}}{dp} - up + v(1-p)$$

(selection + mutational change f/back)

$$V = \frac{p(1-p)}{2N_e}$$

(variance in Wright-Fisher model)

Since $1/xdx/dp = d\ln x/dp$, we can find

$$\frac{2M}{V} = 2N_e\frac{d\ln(\bar{w})}{dp} - 4N_e u(1-p)^{-1} + 4N_e v p^{-1}$$

using the fact that $\int p^{-1}dp = \ln p$ and $\int (1-\mathrm{p})^{-1}dp = -\ln(1-p)$

we can integrate this equation to get:

$$\int \frac{2M}{V} = 2N_e\ln(\bar{w}) + 4N_e u\ln(1-p) + 4N_e v\ln p$$

substituting back in the equation for $\hat{\psi}$

$\hat{\psi} = \frac{C}{V}e^{\int \frac{2M}{V}dp}$ and including $2N_e$ in $C$ gives us:

$$\hat{\psi}(p) = C\bar{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$

Mutation vs. drift: set $\bar{w} = 1$ = constant, $u = v$,

$\hat{\psi}(p) \propto [p(1-p)]^{4N_e u-1}$

Drift wins when $4N_e u \ll 1$

Figure removed due to copyright reasons.

$$\hat{\psi}(p) = C\bar{w}^{2N_e}(1-p)^{4N_e u-1}p^{4N_e v-1}$$

$$\hat{\psi}(p) \propto \frac{e^{4N_e sp(1-p)}}{p(1-p)}$$

$$\hat{\psi}(p) \propto \frac{e^{4N_e sp}}{p(1-p)}$$

Figures removed due to copyright reasons.

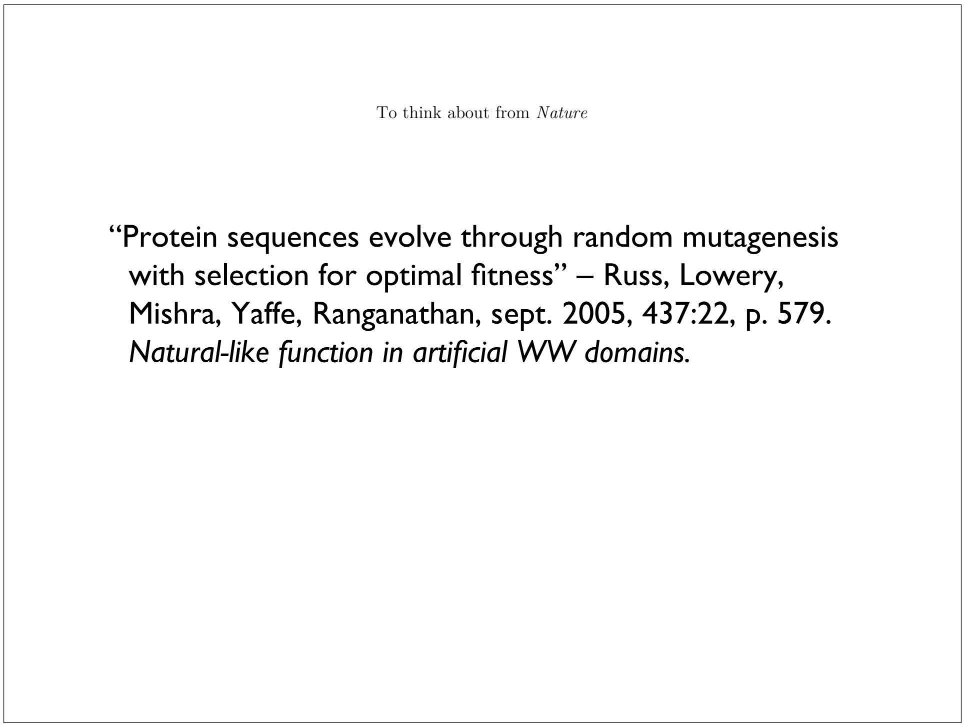
Balancing selection

Directional selection

Drift wins when $4N_e$« 1

Cannot say how effective selection is without knowing effective population size!!!

For next time:
OK, how do we *use* this stuff to figure out whether selection's been at work????



To think about from *Nature*

"Protein sequences evolve through random mutagenesis with selection for optimal fitness" – Russ, Lowery, Mishra, Yaffe, Ranganathan, sept. 2005, 437:22, p. 579. *Natural-like function in artificial WW domains.*