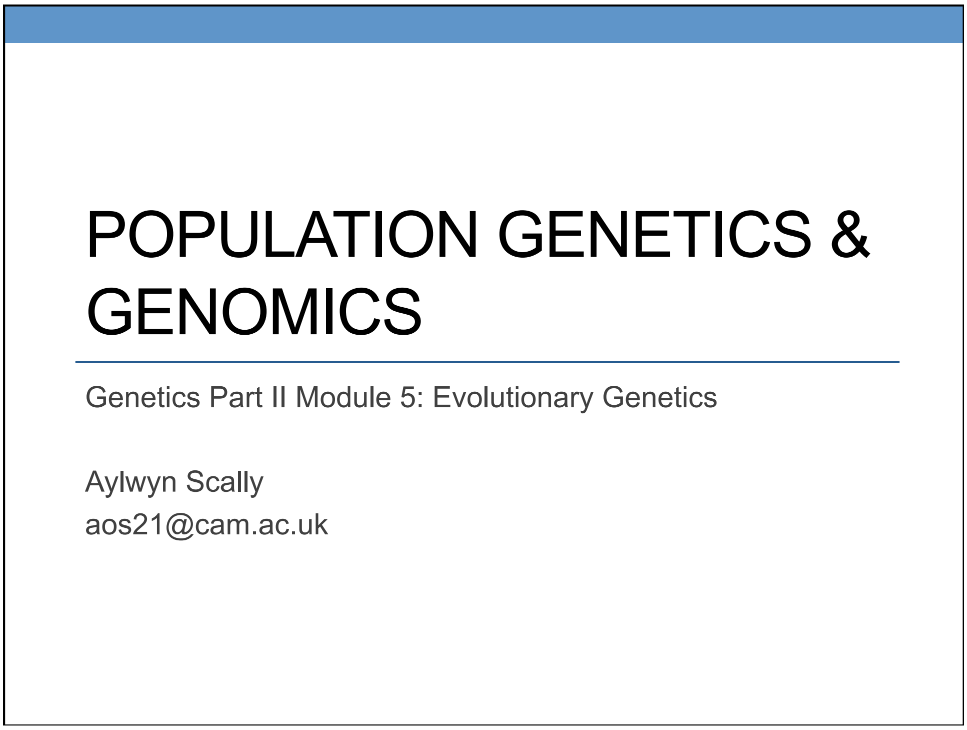
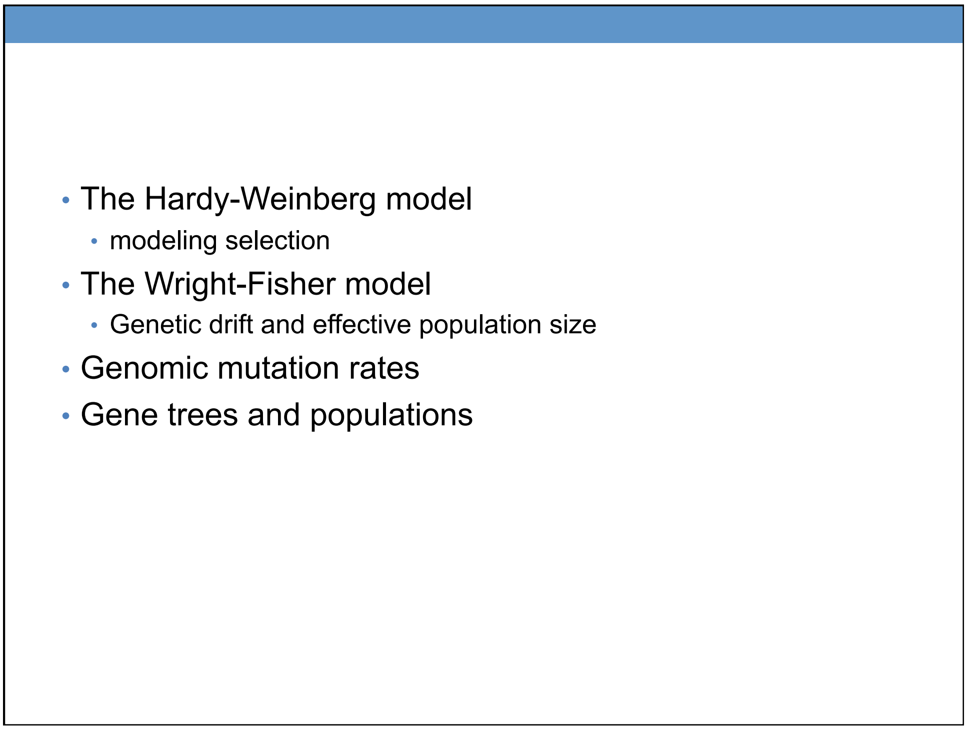


POPULATION GENETICS & GENOMICS

Genetics Part II Module 5: Evolutionary Genetics

Aylwyn Scally
aos21@cam.ac.uk



- The Hardy-Weinberg model
  - modeling selection
- The Wright-Fisher model
  - Genetic drift and effective population size
- Genomic mutation rates
- Gene trees and populations

The Hardy-Weinberg model

- Two alleles at a single locus: A_1, A_2
- Genotypes

	A_1A_1	A_1A_2	A_2A_2	total
number	n_{11}	n_{12}	n_{22}	N
frequency	$P = n_{11}/N$	$H = n_{12}/N$	$Q = n_{22}/N$	1

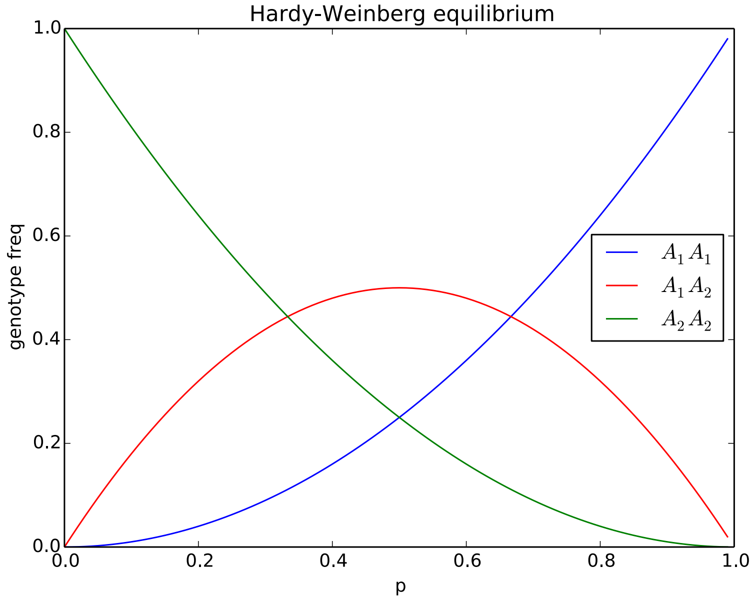
- Alleles

	A_1	A_2	total
frequency	p	q	1
number	$2n_{11} + n_{12}$	$2n_{22} + n_{12}$	2N
	$p = P + H/2$	$q = Q + H/2$	

Hardy-Weinberg distribution

- Mating table

M	F	prob	A_1A_1	A_1A_2	A_2A_2
A_1A_1	A_1A_1	P^2	1	-	-
A_1A_1	A_1A_2	HP	1/2	1/2	-
A_1A_1	A_2A_2	PQ	-	1	-
A_1A_2	A_1A_1	HP	1/2	1/2	-
A_1A_2	A_1A_2	H^2	1/4	1/2	1/4
A_1A_2	A_2A_2	HQ	-	1/2	1/2
A_2A_2	A_1A_1	PQ	-	1	-
A_2A_2	A_1A_2	HQ	-	1/2	1/2
A_2A_2	A_2A_2	Q^2	-	-	1
			$(P + H/2)^2$	$2(P + H/2)(Q + H/2)$	$(Q + H/2)^2$
			p^2	2pq	q^2



Assumptions of the Hardy-Weinberg model

- Population size is large
- Random mating (no substructure)
- Allele frequencies are equal in both sexes
- No selection
- Also
  - discrete generations
  - no migration into or out of the population
  - no mutation
  - no asexual reproduction

> *'Suppose that the numbers are fairly large, so that mating may be regarded as random, that the sexes are evenly distributed among the three varieties, and that all are equally fertile. A little mathematics of the multiplication-table type is enough to show …'*
>
> *Hardy (1908)*

How well does a model fit some data?

- Models make predictions about relationships in data
  - e.g. Hardy-Weinberg equilibrium
- But the real world is noisy, so even if all the assumptions held, we might not expect these relationships to hold *exactly*
- How do we decide whether a departure from model predictions means model assumptions are wrong?
  - Work out the probability of the observed data under the model
  - If this probability is low, reject the model
  - Otherwise, attribute differences between model predictions and data to chance

Chi-squared test for HWE

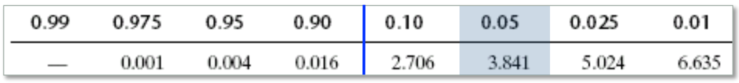
- Calculate allele frequencies from genotype counts
- For each genotype, calculate observed value *O* and expected value *E* under HWE given allele frequencies
- Compute the following statistic, which measures the difference between observed and expected values

$$x = \sum_{\text{genotypes}} \frac{(O-E)^2}{E}$$

- The chi-squared distribution gives the probability (*p-value)* of seeing this value or greater
  - `pchisq(x, 1, lower.tail=FALSE)` in R
  - a low p-value suggests the HW model is a poor fit to the data
    - p-value < 0.05, departure from HWE is ‘significant’ at 5% level

Example

- Genotype data: 7 TT, 5 TG, 0 GG
  - 12 individuals, 24 alleles in total
  - $p = (2 \times 7 + 5) / 24 = 0.79$ $q = 5 / 24 = 0.21$
  - $E(TT) = 12p^2 = 7.49$
  - $E(TG) = 12 \times 2pq = 3.98$
  - $E(GG) = 12q^2 = 0.53$
  - $x = (7 – 7.49)^2 / 7.49 + (5 – 3.98)^2 / 3.98 + (0 – 0.53)^2 / 0.53 = 0.831$
  - By computation (e.g. from R) this gives a p-value = 0.362
  - Hence no significant departure from HWE (at 5% significance level)



p-value	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01
x	—	0.001	0.004	0.016	2.706	3.841	5.024	6.635

0.831

Example: haemoglobin and Malarial resistance

- Two alleles: normal *HbA* and sickle-cell *HbS*
- Genotype data: 25,374 *HbA/HbA*, 5,482 *HbA/HbS*, 67 *HbS/HbS*
  - 30,923 individuals, 61,846 alleles in total
  - $p = 0.909$ $q = 0.091$
  - E(*HbA/HbA*) $= 30{,}923p^2 = 25{,}551.09$
  - E(*HbA/HbS*) $= 30{,}923 \times 2pq = 5{,}115.83$
  - E(*HbS/HbS*) $= 30{,}923q^2 = 256.07$
  - $x = 167.04$
  - p-value << 0.01
  - A significant departure from HWE (at 5% significance level)

Degrees of freedom in a hypothesis test

$$\text{DOF} = \begin{array}{c}\text{number of values}\\\text{needed to fully}\\\text{describe the data}\end{array} - \begin{array}{c}\text{number of parameters}\\\text{in the model}\end{array}$$

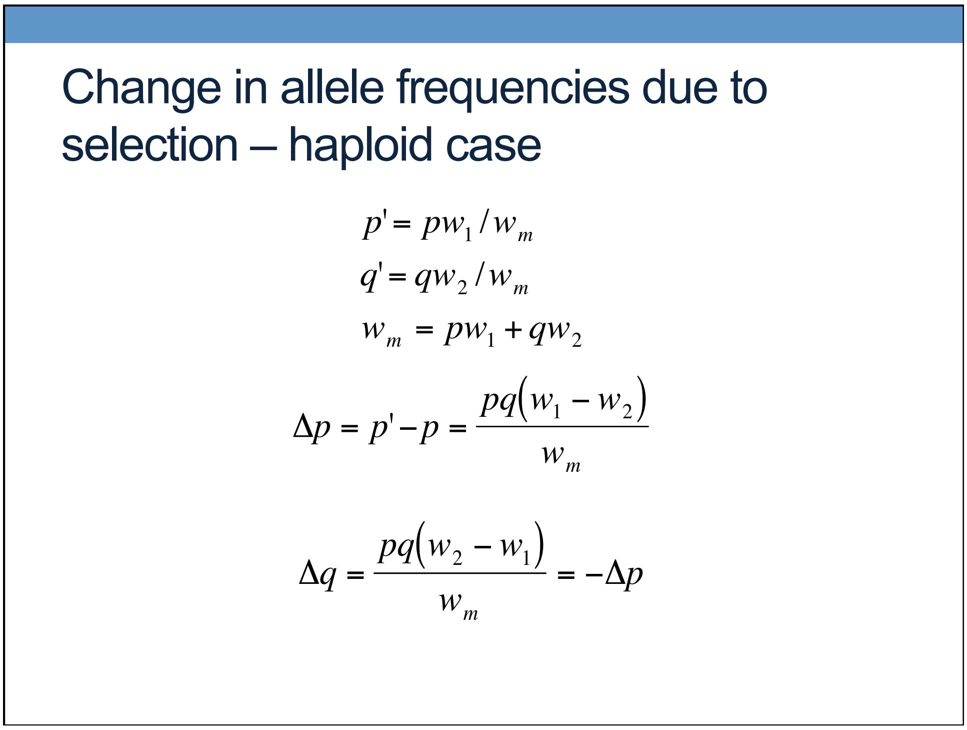
- For a biallelic locus the data has 3 independent values (e.g. the three genotypes), while the model has two parameters (allele frequency *p* and the total number of individuals). Hence DOF = 3 – 2 = 1.

Adding selection to the model: haploid case

- Define the **fitness** of a genotype as the relative contribution to the next generation by an individual of that type

genotype	**A_1**	**A_2**
fitness	w_1	w_2
Frequency in generation 0	p	q
Frequency in generation 1	pw_1 / w_m	qw_2 / w_m

- The **mean fitness** in generation 0 is $w_m = pw_1 + qw_2$



Change in allele frequencies due to selection – haploid case

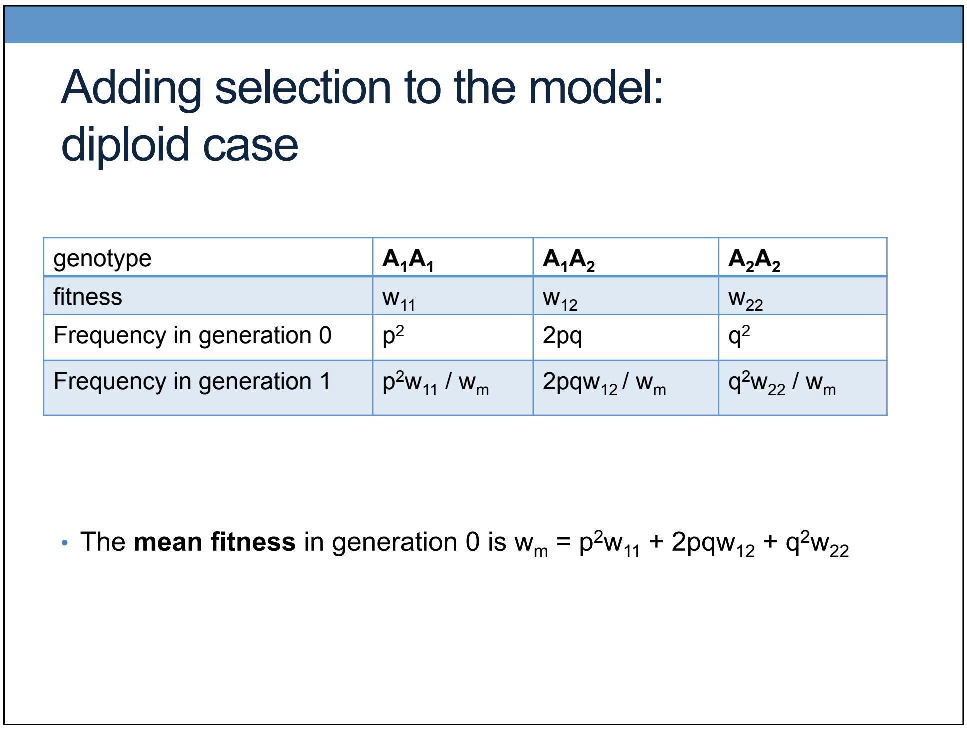
$$p' = pw_1 / w_m$$

$$q' = qw_2 / w_m$$

$$w_m = pw_1 + qw_2$$

$$\Delta p = p' - p = \frac{pq(w_1 - w_2)}{w_m}$$

$$\Delta q = \frac{pq(w_2 - w_1)}{w_m} = -\Delta p$$



Adding selection to the model: diploid case

genotype	**A_1A_1**	**A_1A_2**	**A_2A_2**
fitness	w_{11}	w_{12}	w_{22}
Frequency in generation 0	p^2	2pq	q^2
Frequency in generation 1	p^2w_{11} / w_m	$2pqw_{12} / w_m$	q^2w_{22} / w_m

- The **mean fitness** in generation 0 is $w_m = p^2w_{11} + 2pqw_{12} + q^2w_{22}$

Change in allele frequencies due to selection

$$p' = (p^2 w_{11} + pqw_{12}) / w_m$$

$$q' = (q^2 w_{22} + pqw_{12}) / w_m$$

$$w_m = p^2 w_{11} + 2pqw_{12} + q^2 w_{22}$$

$$\Delta p = p' - p = \frac{pq\left(p\left(w_{11} - w_{12}\right) + q\left(w_{12} - w_{22}\right)\right)}{w_m}$$

$$\Delta q = \frac{pq\left(p\left(w_{12} - w_{11}\right) + q\left(w_{22} - w_{12}\right)\right)}{w_m} = -\Delta p$$

Representing different types of selection

genotype	**A_1A_1**	**A_1A_2**	**A_2A_2**
fitness	w_{11} = 1	w_{12} = 1 + hs	w_{22} = 1 + s

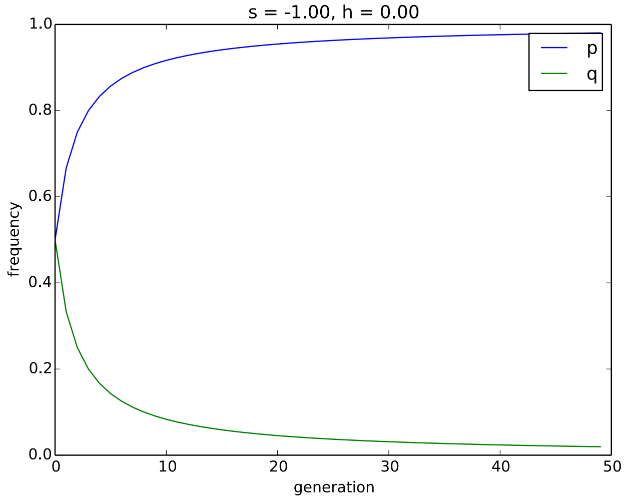
- s: selection coefficient
  - purifying selection for A_1: s < 0
  - positive selection for A_2: s > 0
- h: dominance parameter
  - recessive: h = 0
  - additive: h = 1/2
  - dominant: h = 1

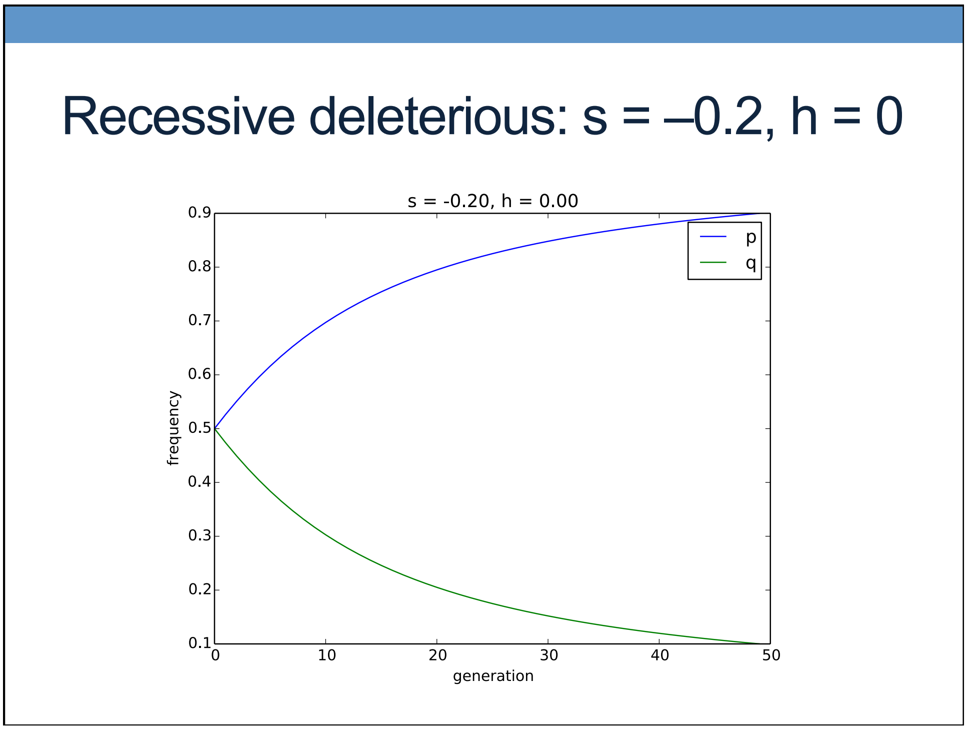
Simulating the selection model

Algorithm

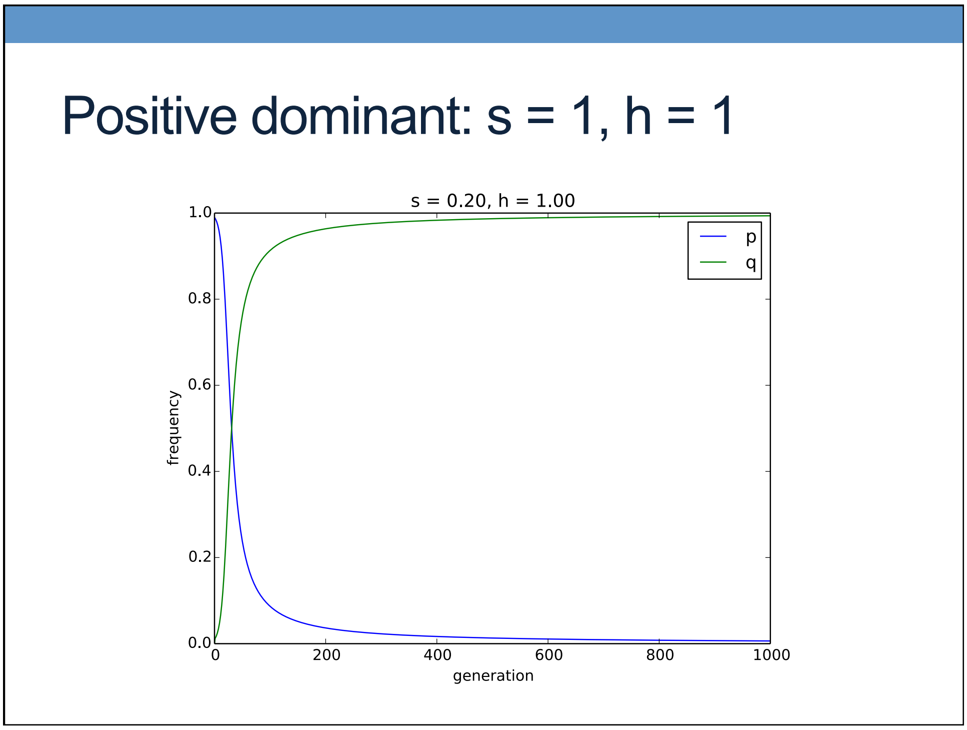
1. user input parameters:
   1. dominance h and selection coefficient s
   2. initial allele frequency q
   3. number of generations to simulate: n_{gen}
2. calculate fitnesses w_{11}, w_{12}, w_{22} using h and s
3. for each generation n from 0 to n_{gen}:
   1. plot a point at (n, q)
   2. calculate Δq using q, w_{11}, w_{12}, and w_{22}
   3. set $q = q + \Delta q$

Recessive lethal: s = –1, h = 0

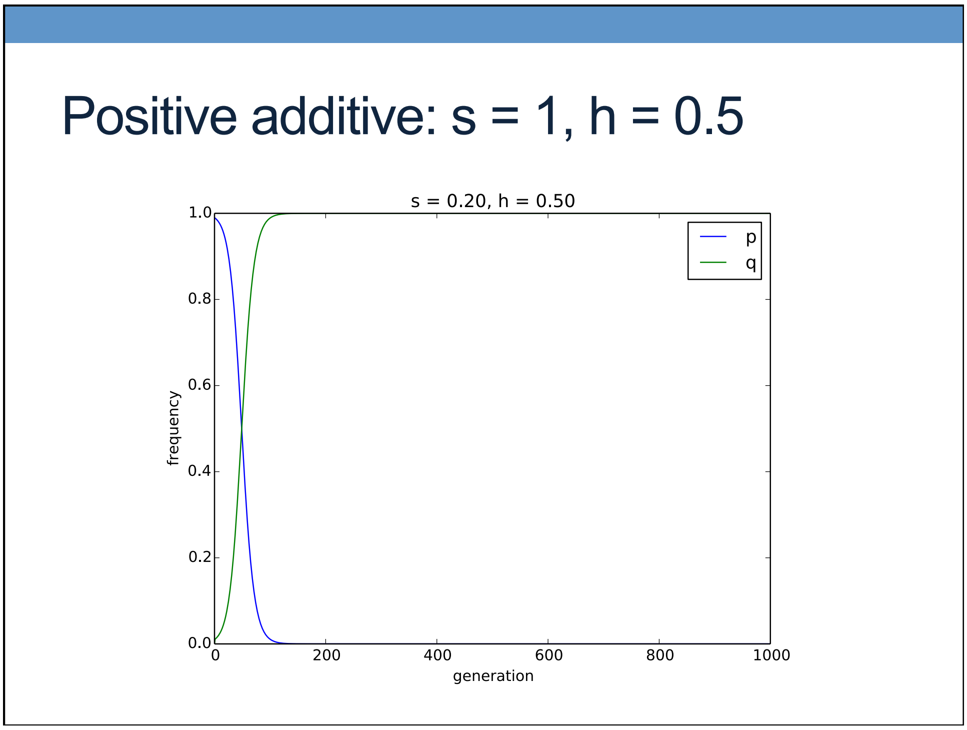




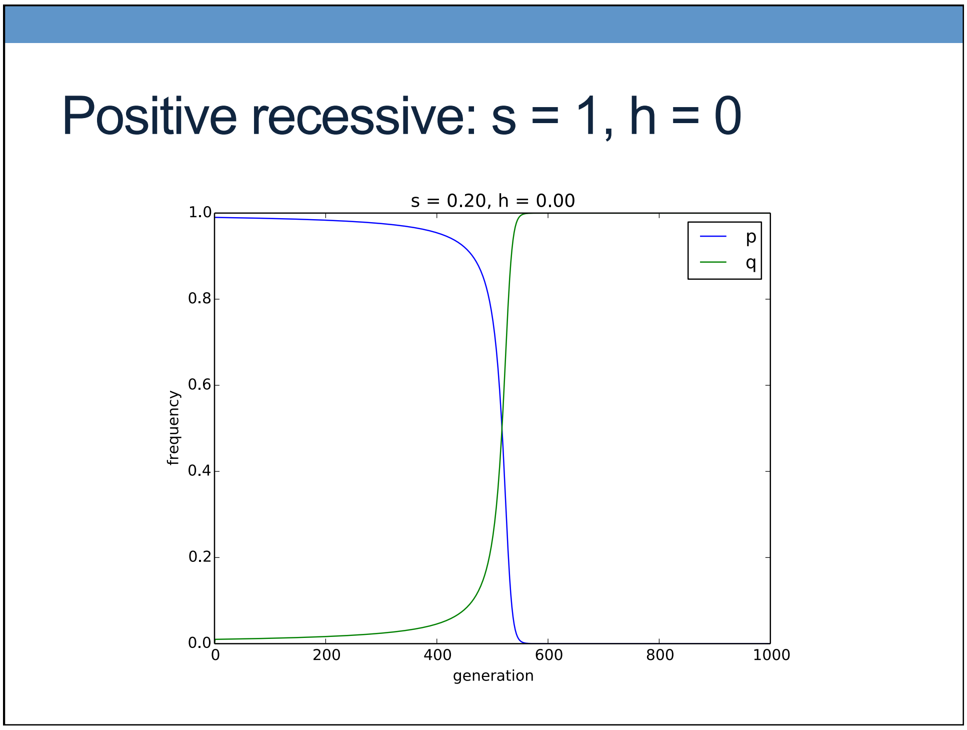
Recessive deleterious: s = –0.2, h = 0



Positive dominant: s = 1, h = 1



Positive additive: s = 1, h = 0.5



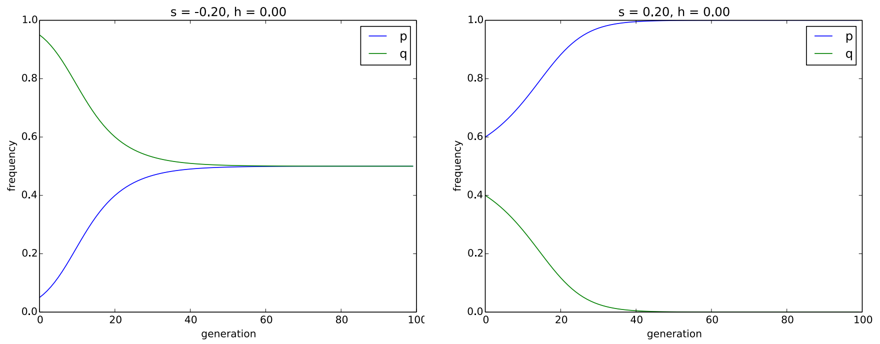
Positive recessive: s = 1, h = 0

Heterozygote advantage/disadvantage

genotype	A_1A_1	A_1A_2	A_2A_2
fitness	w_{11} = 1 + s1	w_{12} = 1	w_{22} = 1 + s2

- s_1, s_2: selection coefficients for homozygous genotypes
  - heterozygote advantage: $s_1 < 0$, $s_2 < 0$
  - heterozygote disadvantage: $s_1 > 0$, $s_2 > 0$

Heterozygote advantage/disadvantage



See Hedrick Ch 3 for some interesting experimental examples and a fuller discussion

Adding relatedness to the Hardy-Weinberg model

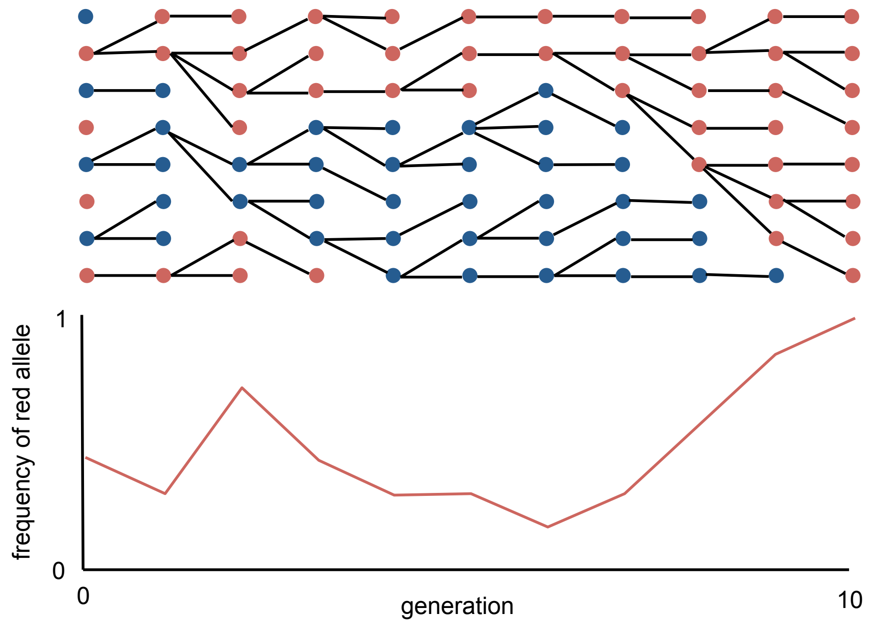
Consider selecting 2 alleles from a population where there is a probability *f* that they are identical by descent (IBD)

Allele 1	Allele 2		
	Unrelated (probability 1 – *f*)		IBD (probability *f*)
	A_1 (probability *p*)	A_2 (probability *q*)	same as allele 1
A_1 (probability *p*)	A_1A_1	A_1A_2	A_1A_1
A_2 (probability *q*)	A_2A_1	A_2A_2	A_2A_2

- P = $p((1 – f)p + f) = p^2 + fpq$
- Q = $q^2 + fpq$
- H = freq(A_1A_2) + freq(A_2A_1) = $2pq(1 – f)$
  - Relatedness or inbreeding reduces heterozygosity
  - *f* is called the **inbreeding coefficient**

Wright-Fisher model

- Finite population size
  - N individuals
  - 2N alleles (i.e. chromosomes) at any given locus
- Alleles in each generation are sampled randomly from those in the previous one
  - leads to stochastic variation in allele frequencies: **genetic drift**



Simulating the Wright-Fisher model

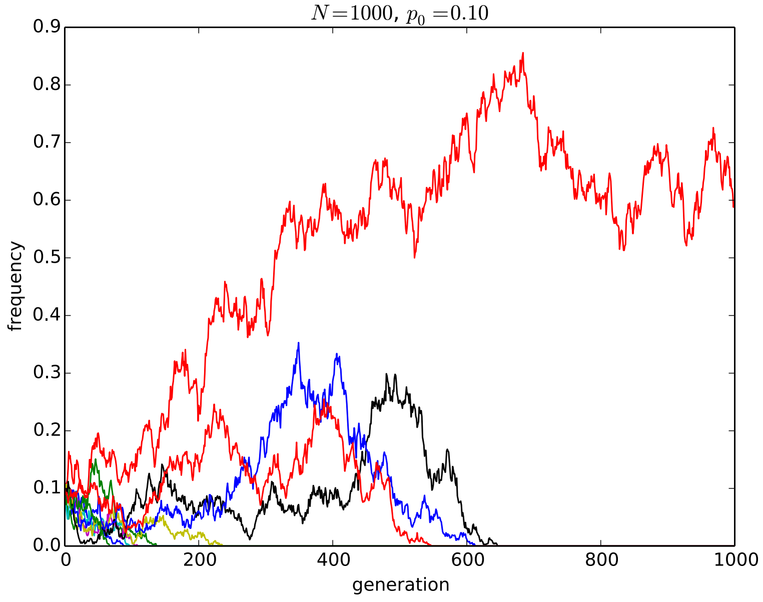
Algorithm

```
1. input parameters:
    1. initial allele frequency $p$
    2. population size $N$
    3. number of generations to simulate: $T$
2. for each generation $t$ from 0 to $T$:
    1. plot a point at $(t, p)$
    2. generate number $n$ of $A_1$ alleles in generation $t + 1$
       by sampling from Binom(N, p)
    3. set $p = n / N$
```

Simulating the Wright-Fisher model with selection

Algorithm

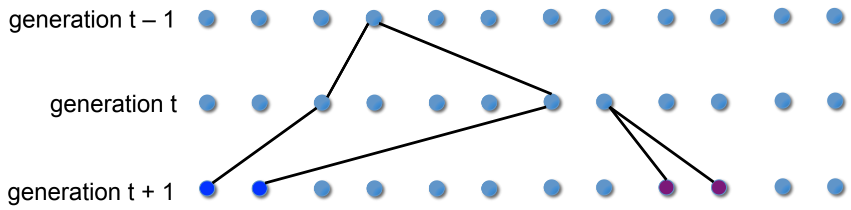
```
1. input parameters:
   1. initial allele frequency $p$
   2. population size $N$
   3. number of generations to simulate: $T$
2. for each generation $t$ from 0 to $T$:
   1. plot a point at $(t, p)$
   2. generate number $n$ of $A_1$ alleles in generation $t + 1$
      by sampling from Binom(N, p)
   3. set $p = n (1 + s) / N$
```



Can show that the (long-term) probability of allele fixation is equal to p_0

Relatedness in the Wright-Fisher model

- Chance of any pair descending from same allele in previous generation is 1 / 2N
- In generation *t* + 1, two alleles are IBD if either:
  - they have the same parental allele in generation t
  - their parental alleles in generation t are IBD



IBD because parental alleles IBD

IBD because parental alleles same

Increase of relatedness due to drift

- If f_t is the fraction of alleles related IBD in generation *t*, we expect:

$$f_t = \underbrace{\frac{1}{2N}}_{\text{probability that parental alleles are same}} + \underbrace{\left(1-\frac{1}{2N}\right)f_{t-1}}_{\text{probability that parental alleles are different but IBD}}$$

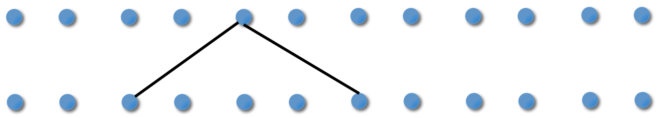
  - Define $h_t = 1 - f_t$, so

$$h_t = \left(1-\frac{1}{2N}\right)h_{t-1} = \left(1-\frac{1}{2N}\right)^t h_0$$

- For large *N*, we can approximate this as $h_t = h_0 e^{-t/2N}$
- Thus relatedness increases (diversity decreases) over time:
  - If $f_0 = 0$ then $f_t = 1 - e^{-t/2N}$

Effective population size

- We can model the evolution of real population using a Wright-Fisher population of size N_e individuals
- Various ways to define N_e
  - classically: defined so that model population has same rate of increase of inbreeding or variance in allele frequencies
  - coalescent approach: defined in terms of P_2, the probability of two alleles deriving from same copy (**coalescing**) in the previous generation
  - $P_2 = 1 / N_e$ for a haploid population
  - $P_2 = 1 / 2N_e$ for a diploid population



- N_e is affected by several factors
  - population structure, non-random mating, sex ratio, selection, locus type
  - usually N_e is smaller than census population size – sometimes by a large factor

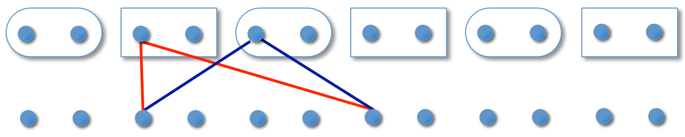
species	N_e
Homo sapiens	~20,000
Drosophila melanogaster	~1,000,000
Caenorhabditis elegans	~80,000
Escherichia coli	~25,000,000

Autosomal N_e for two sexes (dioecy)

- Consider population as comprising two separate populations: males (N_m) and females (N_f)
  - For any two alleles to have same parental allele means either same mother or same father
  - The probability of two alleles coalescing in a mother is:

prob ½ that allele 1 derives from mother × prob ½ that allele 2 derives from mother × prob 1 / N_f that both derive from same mother × prob ½ that both derive from same chromosome within the mother $= \frac{1}{8N_f}$

- Similarly, the probability of coalescing in a father is 1 / 8N_m



Autosomal N_e for two sexes (dioecy)

- Thus the combined probability of coalescing is

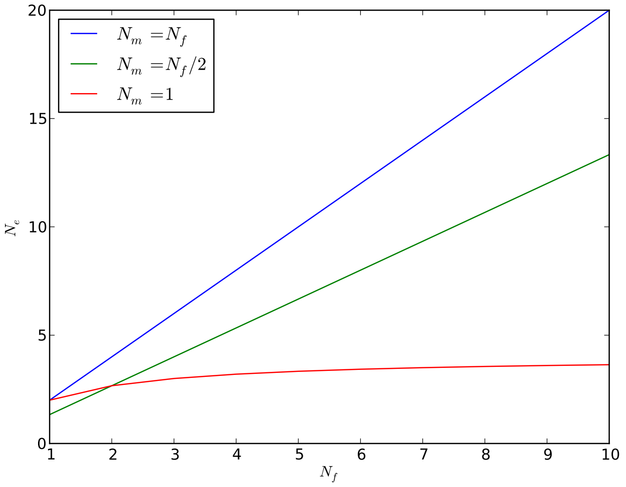
$$\frac{1}{2N_e} = P_2 = \frac{1}{8N_f} + \frac{1}{8N_m}$$

- Hence

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

- Note that if $N_f = N_m$ then $N_e = 2N_f = 2N_m$
  - i.e. the expected result.

N_e with an unequal breeding ratio



N_e on chrX, chrY and mtDNA

- For X, we must account for the fact that X-linked alleles spend 2/3 of their time in females

$$N_{eX} = \frac{9N_f N_m}{2N_f + 4N_m}$$

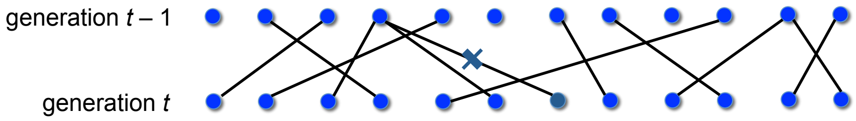
  - if $N_m = N_f$ then $N_{eX} = 9N_f / 6$, and since the autosomal effective population size $N_{eA} = 2N_f = 2N_m$, we can write $N_{eX} = 3N_{eA} / 4$

- Chromosome Y spends all its time in males, and there is only one Y per male, so $P_2 = 1 / N_m$
  - we define effective population size as $1 / 2P_2$, so this means $N_{eY} = N_m / 2$
  - if $N_m = N_f$ then $N_{eY} = N_{eA} / 4$

- For the mitochondrial genome, by a similar argument to that for chrY: $N_{eMT} = N_{eA} / 4$

Adding mutation to the Wright-Fisher model

- A mutation at a locus creates a new allele in one chromosome
  - initial frequency is 1 / 2*N*
- Under neutrality (no selection), the probability of fixation is 1 / 2*N*
  - so when *N* is large, vast majority of new mutations will not persist.



Heterozygosity due to mutations

- Let mutations occur at rate μ_{gen} per parent-child transmission
  - i.e. μ_{gen} mutations per generation on each lineage
- Suppose all alleles match (no heterozygosity) in generation 0.
  - Let m_t be chance of two alleles matching in generation *t*. Then

$$m_{t+1} = (1-\mu_{gen})^2\left(\frac{1}{2N} + \left(1-\frac{1}{2N}\right)m_t\right)$$

probability that neither allele has mutated × probability that parental alleles match

- assumes probability of both mutating to same new allele is negligible: ‘infinite alleles’, or very low mutation rate
- Then assume μ_{gen} is small enough that we can neglect terms in μ_{gen}^2 and μ / 2*N*

$$m_{t+1} \approx \frac{1}{2N} + \left(1 - 2\mu_{gen} - \frac{1}{2N}\right)m_t$$

Equilibrium heterozygosity under mutation and drift

- Locus slowly reaches an equilibrium between heterozygosity loss due to drift and gain due to mutation
  - at equilibrium, $m_{t+1} = m_t = m_{eq}$, so

$$m_{eq} = \frac{1}{1 + 4N\mu_{gen}} = \frac{1}{1+\theta}$$

  - where we define $\theta = 4N\mu_{gen}$
  - $H = 1 - m$, so we also have:

$$H_{eq} = \frac{\theta}{1+\theta}$$

Mutations at a single nucleotide

- For a single nucleotide we can assume θ is << 1
- So at equilibrium, the probability that two chromosomes will differ at a particular site is $H_{eq} = \theta / (1 + \theta) \approx \theta$
- Thus we can estimate θ by evaluating π, the average number of pairwise differences between chromosomes at a site: $E(\pi) = \theta = 4N\mu_{gen}$