

Population genetics lecture notes

- 1 - Genetic drift.**
 - 1.1 - Wright-Fisher model.**
 - 1.2 - Effective population size.**
 - 1.3 - Coalescent theory.**
 - 1.4 - Time to fixation and loss.**
- 2 - Mutation.**
 - 2.1 - Infinite alleles model.**
 - 2.2 - Infinite sites model.**
- 3 - Selection.**
 - 3.1 - Frequency spectrum.**
 - 3.2 - Effects on linked variation.**
- 4 - Mutation selection balance.**
- 5 - Measuring polymorphism.**
 - 5.1 - Estimating theta.**
 - 5.2 - Frequency spectrum.**
 - 5.3 - Statistical tests.**
- 6 - Temporal and spatial demographics.**
 - 6.1 - Changes in population size.**
 - 6.2 - Population subdivision and migration.**
- 7 - Polymorphism and divergence.**
 - 7.1 - Hudson-Kreitman-Aguade (HKA) test.**
 - 7.2 - McDonald-Kreitman (MK) test.**
- 8 - Interpreting data.**
- 9 - Reading.**

1 - Genetic drift.

Genetic drift is the stochastic fluctuations in allele frequency due to random sampling in a finite population.

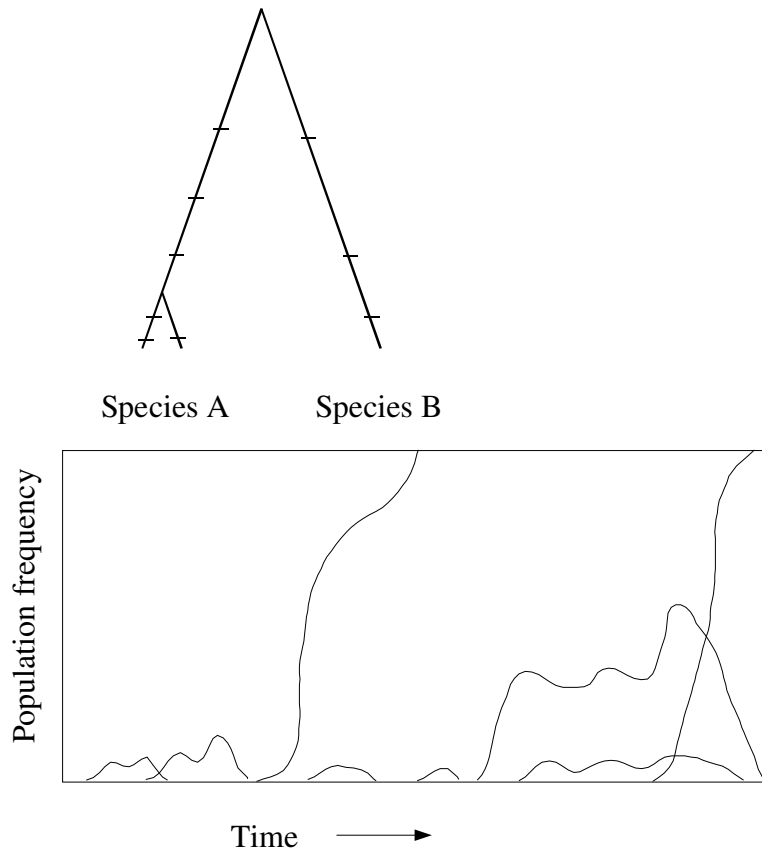
A *polymorphic* site is a position where the common allele frequency > 1%. An outgroup can be used to identify which allele is *derived* and which is *ancestral*. *Allele* refers to either a single site or group of sites.

Polymorphism in *S.cerevisiae* and divergence to *S.paradoxus*.

	10	20	30	40	50	Haplotype
M5	TGGCTCGGTGTCAAAACATAGTTTGC	GTGATAACAGCGTGTGCTCTC				1
M8	1
M13	1
M14C.....A.....	2
M16	1
M22	1
M32T.....	3
M34	1
YPS6C.....T.....	4
S288CC.....T.....T.G..	5
Spar	.C.....C.....	.G.T...G.....G..	

Here there are 5 polymorphic sites and 5 fixed differences between *S.cer* and *S.par*. The 5 polymorphic sites produce 5 haplotypes in *S.cer*.

Polymorphism describes sites variable within a species. *Divergence* describes sites variable between species.



1.1 - Wright-Fisher model.

The Wright-Fisher model describes the process of genetic drift in a finite population. The model assumes:

1. N diploid organisms
2. Monoecious reproduction with an infinite number of gametes
3. Non-overlapping generations
4. Random mating
5. No mutation
6. No selection

Given two alleles in a population, A_1 and A_2 , *genetic drift* is their change in frequency over time.

Because there are an infinite number of gametes, the transition probability of going from i copies of A_1 to j copies of A_1 in the next generation is given by the binomial probability distribution:

$$P_{ij} = \binom{2N}{j} \binom{i}{2N} \left(1 - \frac{i}{2N}\right)^{2N-j}$$

From the Hardy-Weinberg principle, if p is the frequency of allele A_1 and q is the frequency of allele A_2 , and if there is random mating, the frequency of A_1A_1 , A_1A_2 and A_2A_2 individuals in the next generation is given by p^2 , $2pq$ and q^2 . Thus, the proportion of homozygous individuals, $F = p^2 + q^2$, and the fraction of heterozygous individuals, $H = 2pq$.

The probability of an individual being homozygous in the next generation is:

$$F_{t+1} = 1/2N + (1 - 1/2N)F_t$$

The probability of an individual being heterozygous, $H_t = 1 - F_t$. Thus,

$$H_{t+1} = (1 - 1/2N)H_t$$

$$H_t = (1 - 1/2N)^t H_0$$

When $x \ll 1$, then $(1 - x)^t \approx e^{-xt}$

$$H_t \approx H_0 e^{-t/2N}$$

As $t \rightarrow \infty$ $H_t \rightarrow 0$

Fluctuations in population size:

If the population size is $N_1, N_2 \dots N_t$ in generation 1, 2 . . . t:

$$H_1 = (1 - 1/2N_1)H_0$$

$$H_2 = (1 - 1/2N_2)H_1 = (1 - 1/2N_2)(1 - 1/2N_1)H_0$$

$$H_t = (1 - 1/2N_t)H_{t-1} = (1 - 1/2N_t) \dots (1 - 1/2N_2)(1 - 1/2N_1)H_0$$

Let N_e be the effective size of the population, i.e. the size of a population that has the same rate of loss of heterozygosity as the one with fluctuating population sizes. Thus, we want to find the value of N_e that satisfies:

$$(1 - 1/2N_e)^t H_0 = (1 - 1/2N_t) \dots (1 - 1/2N_2)(1 - 1/2N_1)H_0$$

$$e^{-t/2N_e} = (e^{-1/2N_t}) \dots (e^{-1/2N_2})(e^{-1/2N_1})$$

$$\frac{1}{N_e} = \left(\frac{1}{t}\right) \left(\frac{1}{N_t} \dots \frac{1}{N_2} + \frac{1}{N_1}\right)$$

Thus, N_e is the harmonic mean of the actual population size. Since the harmonic mean is dominated by the smallest terms, population bottlenecks, or brief reductions in actual population size, can have a strong influence on the effective population size and heterozygosity.

1.2 - Effective population size.

The *effective population size* is the size of an idealized population, such as that described by the Wright-Fisher model, that has the same magnitude of genetic drift. The effective population size is always less than the actual population size due to:

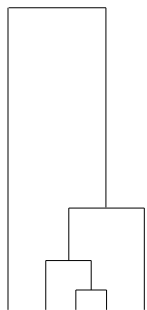
1. Unequal numbers of males and females. If N_m and N_f are the number of males and female:

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

2. Overlapping generations.
3. Non-poisson distribution of fecundity.
4. Non-random mating, e.g. population structure.

1.3 - Coalescent theory.

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



Let T_{21} be the time in generations until the most recent common ancestor of two genes chosen at random from a population of size N , here $N = N_e$.

$$P(T_{21} > t) = (1-1/2N)^t \text{ for } t = 0, 1, 2, 3, \dots \infty$$

$$P(T_{21} = t+1) = P(T_{21} > t) - P(T_{21} > t+1) = 1/2N(1-1/2N)^t \text{ for } t = 0, 1, 2, 3, \dots \infty$$

This is the geometric distribution so the expectation and variance of T_{21} is:

$$E(T_{21}) = 2N$$

$$\text{Var}(T_{21}) = 2N(2N-1)$$

The coefficient of variation:

$$\frac{\sqrt{\text{Var}(T_{21})}}{E(T_{21})} = \sqrt{\frac{2N-1}{2N}} \rightarrow 1 \text{ as } N \rightarrow \infty$$

For a sample of k genes from a population:

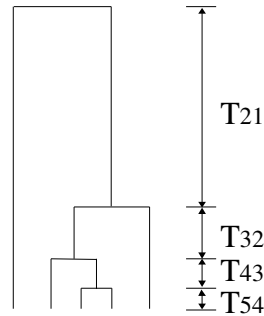
$$P(T_{k, k-1} > t) \approx \left(1 - \binom{k}{2} \frac{1}{2N}\right)^t \rightarrow e^{-\binom{k}{2} \frac{t}{2N}} \text{ as } N \rightarrow \infty$$

If time is measure in $2N$ generations by $\tau = t/2N$ generations:

$$P(T_{k, k-1} > \tau) = e^{-\binom{k}{2} \tau}$$

$$E(T_{k, k-1}) = \binom{k}{2}^{-1}$$

$$\text{Var}(T_{k, k-1}) = \binom{k}{2}^{-2}$$



The time to the most recent common ancestor (MRCA) of n genes is given by:

$$T_{n1} = \sum_{k=2}^n T_{k, k-1}$$

$$E(T_{n1}) = 2\left(1 - \frac{1}{n}\right)$$

Where the unit is $2N$ generations.

Incorporating mutations into the coalescence process is straightforward since the number of mutations that have occurred over t generations is Poisson distributed with rate, $\lambda = \mu t$.

Recombination can also be incorporated into the coalescence. ***In the absence of recombination a single genealogy can be generated. However, in the presence of recombination different genealogies are found at different sites.

1.4 - Time to fixation and loss.

The expected time to fixation, T_{Fix} , and loss, T_{Loss} , as a function of initial frequency, p , is:

$$T_{Fix}(p) = -4N_e \frac{(1-p)}{p} \ln(1-p)$$

$$T_{Loss}(p) = -4N_e \frac{p}{(1-p)} \ln(p)$$

For $p = 1/2N$:

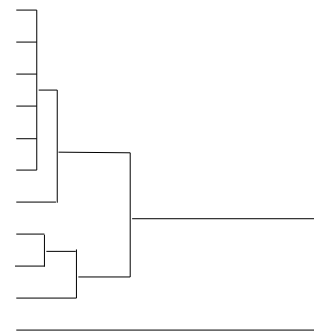
$$T_{Fix} \approx 4N_e$$

$$T_{Loss} \approx 2N_e \ln(2N)/N$$

where N is the breeding population size, i.e. the current number of breeding individuals. For $N_e = N = 10^6$, $T_{Loss} = 29$ generations.

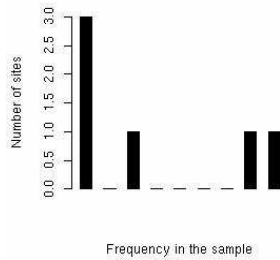
2 - Mutation.

M5	TGGCTCGGTGTCAAAACATAGTTTGC	GTGATAACAGCGTGTGTGCTCTC
M8
M13
M14
M16
M22
M32C.....A.....
M34T.....
YPS6C.....T.....
S288CC.....T.....T.G..
Spar	.C.....C.....G.T...G.....G..



Two measures of polymorphism are heterozygosity and the frequency spectrum. Homozygosity is the probability two alleles are identical and heterozygosity is one minus homozygosity. The frequency spectrum is the frequency distribution of alleles in a sample.

2.1 - Infinite alleles model.



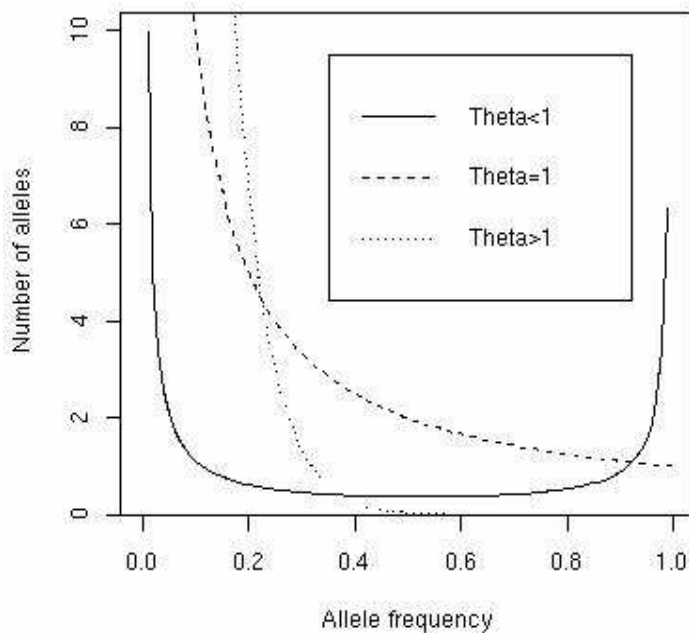
An allele mutates with rate μ per generation and produces a novel allele. There is no recombination within an allele. This is reasonable for proteins with 100 codons since there are 100^{60} possible alleles.

If $\mu \ll 1$ the expected heterozygosity is:

$$E(H) \approx \theta / (1 + \theta), \text{ where } \theta = 4N_e\mu.$$

The frequency spectrum $\Phi(x)$ is the expected number of alleles segregating in a population as a function of their frequency, x . The frequency spectrum can also be obtained for the distribution of alleles in a population *sample*. The frequency spectrum is given by (Kimura and Crow 1964):

$$\phi(x) \approx \frac{\theta(1-x)^{\theta-1}}{x}$$



2.2 - Infinite sites model.

Sites mutate with rate μ per generation and each mutation occurs at a novel site, i.e. a site monomorphic in the population. There can either be no recombination between sites or free recombination since recombination does not affect the expectation but only the variance in the parameters of interest. If the proportion of variable sites $\ll 1$, then the infinite sites model is appropriate.

The expected heterozygosity is:

$$E(H) \approx \theta, \text{ where } \theta = 4N_e\mu.$$

The frequency spectrum of segregating sites is given by:

$$\phi(x) \approx \frac{\theta}{x}$$

where x refers to the frequency of the derived rather than ancestral state (see 3.1 for curve).

The expected number of polymorphic sites can be found by integration of $\phi(x)$ from $1/2N$ to $1-1/2N$.

$$\int \phi(x) dx = \theta \ln(2N)$$

3 - 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, or fecundity. Natural selection can drive mutations to fixation, due positive selection for an adaptive mutation, or to loss, due to negative selection against a deleterious mutation, or to intermediate frequency in a population, due to balancing selection (see below).

The *selection coefficient* is the fitness effect of a mutation across genetic backgrounds and environments. In a haploid population with two alleles A_1 and A_2 with fitness values of w_1 and w_2 , the selection coefficient is w_1-w_2 . Fitness values typically take on arbitrary units since they are measured relative to a populations mean fitness, which is set to 1.

If w_{11} , w_{12} and w_{22} are the fitness values associated with genotypes A_1A_1 , A_1A_2 and A_2A_2 :

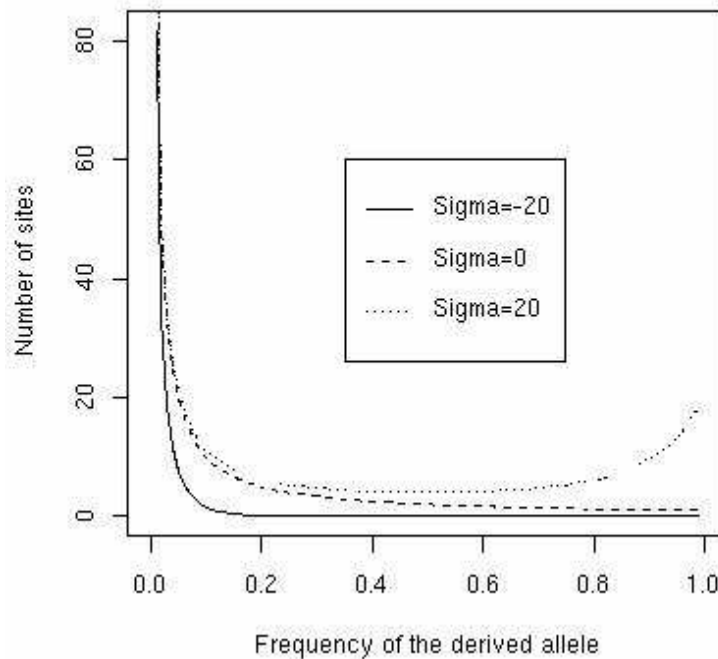
- 1) If $w_{11} < w_{12} < w_{22}$, there is positive, directional selection for A_2 and negative directional selection against A_1 .
- 2) If $w_{11} < w_{12} > w_{22}$, there is heterosis or balancing selection for A_1A_2 . If stable, both A_1 and A_2 be maintained in the population.

3.1 - Frequency spectrum.

Under the infinite sites model, if genotypes A_1A_1 , A_1A_2 and A_2A_2 , have fitness values of $1+s$, 1 and $1-s$:

$$\phi(x) \approx \frac{\theta [1 - e^{\sigma(x-1)}]}{(1 - e^{-\sigma})x(1-x)}$$

where $\sigma = 4Nes$.



3.2 - Effects on linked variation.

If there is no recombination, both kinds of direction selection (positive and negative) as well as balancing selection affect levels and patterns of linked neutral variation. If there is free recombination, e.g. two separate chromosomes, selection does not influence the frequency or fixation probability of neutral polymorphic sites.

Background selection describes the selective removal of rare deleterious mutations from a population. In general, background selection results in a reduction in levels of neutral polymorphism.

Hitchhiking describes the spread of neutral alleles through a population due to their linkage with a mutation under positive selection. In the absence of any recombination, hitchhiking results in a loss of all neutral variation followed by a gradual approach to equilibrium levels of variation.

Both background selection and hitchhiking can explain the general observation that levels of variation are correlated with rates of recombination.

4 - Mutation selection balance (no drift).

If w_{11} , w_{12} and w_{22} are the fitness values associated with genotypes A_1A_1 , A_1A_2 and A_2A_2 , and A_1 mutates to A_2 with rate μ , the frequency of A_1 in the next generation is:

$$p_{t+1} = \frac{p_t(p_t w_{11} + (1 - p_t) w_{12})(1 - \mu)}{\bar{w}}$$

where \bar{w} is the mean population fitness and p_t is the frequency of A_1 in the population. At equilibrium $p_t = p_{t+1}$. If $w_{11} = 1$, $w_{12} = 1 - hs$, and $w_{22} = 1 - s$, at mutation-selection balance:

$$\hat{q} = \sqrt{\frac{\mu}{s}} \quad \text{for } h = 0, \text{ and}$$
$$\hat{q} = \frac{\mu}{hs} \quad \text{when } q \ll 1.$$

9 - Reading.

Lecture 3

Nei M., Fuerst P.A., Chakraborty R. 1976. Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. *Nature*, **262**:491-93.

Begun DJ, Aquadro CF. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature*, **356**:519-20.

Lecture 4

Tajima F. 1989a. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**:585-95.

Cann RL, Stoneking M, Wilson AC. 1987 Mitochondrial DNA and human evolution. *Nature*, **325**:31-6.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature*, **351**:652-4.

Further reading

Kreitman M, Hudson RR. 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics*, **127**:565-82.

Tajima F. 1989b. The effect of change in population size on DNA polymorphism. *Genetics*, **123**:597-601.

Templeton A. Out of Africa again and again. *Nature*. 2002 Mar 7;416(6876):45-51.

Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics*, **116**:153-9.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics*, **155**:1405-13.

Wang W, Thornton K, Berry A, Long M. 2002. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science*, **295**:134-7.