

4 EVOLUTIONARY PROCESSES THAT ERODE GENETIC VARIATION

4.1 NONRANDOM MATING

4.1.a *Assortative mating affects heterozygosity of the genes that control the phenotype*

Assortative mating is non-random mating by *phenotype*. In **positive assortative mating**, mating is more likely between individuals with similar phenotypes than between individuals with dissimilar phenotypes. Examples: assortative mating by height in humans; assortative mating by size in many insects; assortative mating by flowering time in angiosperms. Effects: Decreases heterozygosity, but only for the genes controlling the trait. In **negative assortative mating**, mating is more likely between individuals with dissimilar phenotypes. Examples: self-incompatibility in some plants. S locus controls mating compatibility in some self-incompatible angiosperms. Pollen cannot fertilize ovules of a plant with which it shares an allele at the S locus. Effects: increases heterozygosity for genes controlling the trait.

4.1.b *Inbreeding Reduces Heterozygosity for all genes in the genome*

The most well-known form of non-random mating is **inbreeding**. Inbreeding is mating between relatives. The most extreme form of inbreeding is **self-fertilization**, which occurs naturally in many plants and some invertebrate animals. Because it is relatively simple to treat mathematically, and illustrates most of the important features of less extreme forms of inbreeding, we will consider the population-genetic effects of selfing in some detail.

If we begin by considering a population that consists exclusively of self-fertile heterozygotes of genotype **Aa**. The initial frequency of the A allele (p) is therefore 0.5. After one generation of selfing, the population would consist of **1/4 AA, 1/2 Aa, and 1/4 aa** individuals. So the proportion of heterozygotes (the **heterozygosity** wrt this one locus) has gone from 1 to 1/2 in a single generation of self-fertilization. However, the allele frequency remains $p=0.5$. In the next generation, only heterozygotes can produce heterozygous offspring, and only half the offspring of any heterozygote will also be heterozygotes. So the heterozygosity after two generations of selfing is now $1/2 \cdot 1/2 = 1/4$, and the frequency of the two homozygous genotypes is $3/8$. Three generations of selfing reduce the heterozygosity to $1/8$, and so forth.

So, the primary effect of selfing is that it reduces the heterozygosity by 1/2 every generation. In particular, note that the allele frequency does not change, only the proportion of heterozygotes. Non-random mating therefore causes deviations from Hardy-Weinberg proportions, but does not change allele frequencies.

In weaker forms of inbreeding, such as matings between cousins, the same thing happens, only more slowly. A convenient measure of the effect of inbreeding is based on the reduction in heterozygosity that occurs.

$F = 1 - (\text{observed Het}/\text{expected Het})$, or (Note that your book designates observed Heterozygosity as H and expected Heterozygosity as H_0)
 $F = (1 - H_{\text{Obs}}/2pq)$

F is called the **inbreeding coefficient**, and ranges from 0 in a purely outcrossing population, to 1 in a purely selfing population. For a gene with two alleles, if you know F , you can calculate the expected proportion of heterozygotes in an inbred population:

$$H_{\text{Obs}} = 2pq(1-F) = 2pq - 2pqF.$$

The “missing” $2pqF$ individuals are added to the number of homozygotes (half to each type of homozygote), so the expected proportion of homozygotes for AA is:

$$p^2 + pqF = p^2 + p(1-p)F = p^2(1-F) + pF$$

and the expected proportion of aa homozygotes is:

$$q^2 + pqF = q^2 + q(1-q)F = q^2(1-F) + qF$$

You can use these relationships to infer the amount of inbreeding that occurs in a population. For example, consider a population of plants (e.g., monkey flowers, *Mimulus guttatus*) in which some individuals self-fertilize, and others out-cross. A sample of 140 plants produced the following genotypic frequencies: 60 **AA**, 24 **Aa**, and 56 **aa**. What level of inbreeding is implied by these frequencies?

First, calculate p and q . Since there are 280 alleles in a sample of 140 diploid individuals, $p = \text{freq of A allele} = (60+60+24)/280 = 0.51$. $q = 1-p = 0.49$.

$$2pq = 0.51*0.49 = .4998$$

Next calculate the heterozygosity (H). $H = \text{the observed proportion of heterozygotes} = 24/140=0.17$.

You can now solve for F , which is a measure of the amount of inbreeding in the population:

$$F = 1 - H/2pq = 1 - (0.17/.4998)$$

$$F = 1 - 0.34 = 0.66.$$

So, on average, 66% of the population is selfing, and 34% is outcrossing.

4.1.c *Inbreeding is Generally Harmful*

We saw above that inbreeding decreases heterozygosity. The obvious corollary is that inbreeding also increases **homozygosity**, the proportion of individuals within a population that are homozygous with respect to specified loci. In most species, **inbreeding is harmful**, and much of the effect is due to homozygosity for **rare deleterious recessive alleles**. Remember that a recessive allele does not affect the phenotype unless it is homozygous, and that rare alleles are

usually present in heterozygotes, not in homozygotes, since the frequency of the rare (q is small) homozygote is given by q^2 in a random-mating population. However, with inbreeding, the proportion of homozygotes increases. In a completely inbred population, the proportion of homozygotes for the rare allele would approach q , instead of q^2 . If a rare recessive deleterious allele has a frequency of $q=0.1$ in a random-mating population, the proportion of individuals homozygous for the deleterious allele is $q^2=0.01$. In completely inbred population ($F=1$), the proportion of individuals homozygous for the deleterious allele is $q=0.1$. **So there are 10 times as many individuals expressing the deleterious phenotype in a completely inbred population.**

In humans, the most common type of close inbreeding is **between first cousins** ($F=1/16$). Such matings result in an increase in frequency of genotypes that are homozygous for rare, recessive alleles.

The increase in frequency of homozygotes for such an allele can be calculated by noting that, in inbreeding populations, the frequency of homozygotes for the less common allele is given by the formula:

$$q^2 + pqF$$

[Since, in an inbred population, the proportion of heterozygotes is reduced by $2pqF$ compared to a non-inbred population. Of this "missing" $2pqF$, half are homozygotes for one allele and half are homozygotes for the other. Therefore the proportion of homozygotes for the rarer allele becomes $q^2 + pqF$. Can also be expressed only in terms of q : substituting $(1-q)$ for p , we get $q^2 + (1-q)qF = q^2 + qF - q^2F = q^2(1-F) + qF$.

Example: The frequency of **albinism** (homozygous for **aa**) in American Caucasians is 1 in 20,000 ($q^2=0.00005$). But the frequency among offspring of first-cousin marriages is **[0.00005*(15/16)+0.007*(1/16)] = 0.0005] or 1 in 2000.**

Inbreeding also has implications for species with small population sizes, for species in zoos, or for domestic species of plants and animals that are often inbred to "fix" some desirable trait. Highly inbred strains are usually all fixed for the same allele at most loci. This is because, as inbreeding eliminates heterozygotes, genetic drift within a single strain will tend to eliminate alternate alleles. The result is a strain where each individual is highly homozygous, and different individuals are genetically identical. Such populations suffer from inbreeding depression, and suffer the long-term deleterious consequences of the lack of genetic variation within a population--little ability to evolve disease resistance to new pathogens, or to adapt to changing environmental conditions.

4.2 GENETIC DRIFT

Since populations are not infinite in size, allele frequencies can change due to chance. If only two individuals in a population reproduce, one may by chance carry a rare allele. So the allele frequency would go from very rare in one generations to 1/4 in the next. A process of **sampling** takes place every generation, in which the very large number of gametes produced by individuals in a population result in the production of a much smaller number of actual offspring.

Random genetic drift causes random changes in allele frequencies over time. That is, drift leads to nonadaptive evolution.

Figures 3.5 and 3.6 in your text illustrate the effects of genetic drift in real populations.

Q: What is the main difference between these two examples?

A:

If allele frequencies are not equal to start with, then the rarest allele is the most likely to be lost, but even rare alleles will occasionally displace more common alleles by random genetic drift. Demonstrate this to yourself by running the "AlleleA1" simulation program with the following parameters (see the first HW assignment for instructions on how to get this program). You can run it multiple times to simulate drift in multiple populations. To retain all the previous graphs, choose "Auto" from the "Graph lines" section of the window.

AlleleA1 (N=100, G=500, p=0.2)

Even large populations, over a long period of time, can experience drastic changes in gene frequencies due to random genetic drift.

AlleleA1 (N=10000, G=10000, p=0.5)

Probability of fixation of an allele is equal to its starting frequency

$p_1=0.2$, then Probability of fixation of the A1 allele =0.2, and the probability of fixation of the alternate allele (A2) = $1-0.2=0.8$.

AlleleA1 (N=50, G=500, p=0.2)

Profound ramifications of this simple fact:

1. In a diploid population of size N , a brand new mutation has a frequency $1/2N$, and this is the probability of that it will eventually replace the original allele. *A new mutation has a higher probability of fixation in a small population.*
2. If the same mutation arises in many populations of size N , it will become fixed in $1/2N$ of them.
3. Of all the new mutations that arise in a single generation, $1/2N$ of them will become fixed.
4. If many populations of the same size start out with 2 alleles at a locus, at frequencies p and q , a proportion p of them will become fixed for 1 allele, and a proportion q will become fixed for the other. Q: What are the average allele frequencies over all populations after they have all become fixed for one of the alleles?
5. Over time, all genes will become fixed for one allele. Therefore, both heterozygosity and polymorphism go to zero, although within a generation, genotypes are in approximate Hardy-Weinberg proportions.
6. If unopposed, random genetic drift eliminates genetic variation.

If random genetic drift is the only force operating within a population, all loci will eventually become fixed for a single allele. This means there is no genetic variation at all left within such a population. This is why small population sizes in species of organisms that are threatened or endangered is a cause of concern. Without a reserve of genetic variation, a population will be unable to respond genetically to changing environmental conditions, to new diseases or parasites, or to the introduction of new competitors, predators or resources. Without genetic variation, no adaptation can occur. The relatively new discipline of **Conservation Genetics** has grown from the need for biologists and wildlife managers to understand the forces affecting genetic variation in small populations.

Genetic drift does not generally cause much deviation from HW equilibrium, as does inbreeding. So the main effect of drift is changes in allele frequencies.

Several factors can **retard the rate of loss of genetic variation** due to genetic drift:

1. Large population size
2. migration between subpopulations
3. mutation
4. natural selection ('balancing' selection such as frequency-dependent selection, environment-dependent fitness, and heterozygote superiority)

4.3 EFFECTIVE POPULATION SIZE

In fact, the loss of variation seen in the simulations is an underestimate of what happens in real populations, because the simulations assume that equal numbers of males and females participate in producing the next generation, that generations do not overlap, and that population size is constant. That is, the population is an ideal population. Most populations are not ideal, and deviations from any of these assumptions leads to genetic drift acting faster than it would in an ideal population. For example, many populations have skewed sex ratios (there may be fewer males than females). In many species, mortality is higher for one sex than for the other, perhaps because only one sex disperses. For example, sex ratios in baboons are often female-biased. So, a new generation of baboons may be produced by 100 mothers, but only 50 fathers. Although the census size of the breeding population is 150, the so-called effective size (N_e) is equal to 133 ($N_e = [4N_f N_m] / [N_f + N_m]$). **133 is the size of the ideal population that would lose genetic variation due to drift at the same rate as the real population.** So our population of 150 breeding baboons would lose genetic variation due to genetic drift at the same rate as an ideal population of only 133 individuals. In general, the effective size of a population is less than the census size.

Sexual or natural selection will also cause N_e to be less than the census population size. Greater than random variation in reproductive success means that fewer adults than expected will contribute gametes to the next generation, and this will increase the rate of genetic drift. In red deer (*Cervus elaphus*), males can have four times greater variation in reproductive success than do females. In one study, there were 33 males, and 35 females, but the variance in reproductive success among males was 4 times greater than for females $V_m=42$, $V_f=9$. When there is greater than random variation in reproductive success $N_e = 8(N_f+N_m)/(V_f+V_m+4) = 9.9$. So this population will experience drift at the same rate as an ideal population consisting only of 10 individuals!

Variation in population size over generations also causes N_e to be less than the census size (see text and lecture slides). N_e when population size fluctuates is equal to the harmonic mean population size, which is dominated by smaller numbers (see text and slides).

4.4 POPULATION DIFFERENTIATION

In isolated populations, genetic drift will lead to genetic divergence between populations. (Text Fig. 3.6). If a population consists of subpopulations with limited gene flow between them, it is called a “metapopulation”. The entire set of populations in Fig 3.6 could be considered a metapopulation, although in this case gene flow is zero. So consider this the extreme case of population subdivision.

In the entire metapopulation represented by Fig. 3.6, what are the allele frequencies at generation 19? What are the genotype frequencies? If we sampled across the entire metapopulation, would we observe any deviations from H-W genotypic proportions?

Within sub-populations, we expect H-W to hold, assuming that individuals mate at random within a subpopulation. However, individuals do not mate at random across the entire metapopulation: they have a higher probability of mating with individuals within their own population and a low probability of mating with individuals from a different subpopulation. Therefore, across the entire metapopulation, we will tend to see deviations from H-W that resemble the effects of inbreeding: We will see fewer than expected heterozygotes and more than expected homozygotes.

Therefore genetic drift in a metapopulation produces genotypic effects that are similar to the effects of inbreeding, except that inbreeding would cause deviations from H-W within subpopulations, whereas drift causes deviations from H-W only when the entire metapopulation is sampled. Therefore, to distinguish the effects of inbreeding from the effects of population structure, we need to assess genotypic frequencies within populations and in the metapopulation as a whole.

Sewall Wright (one of the founders of population genetics who spent most of his career at the University of Chicago) invented “F statistics” to understand the genetic effects of population structure. Note: These are not the same as the “F ratio” used in the statistical procedure called Analysis of Variance.

F statistics are based on measures of heterozygosity at different levels of population structure. Please see the very cogent description of these measures in your textbook (Chapter 3).

H_s = observed heterozygosity (i.e, proportion heterozygotes) within subpopulations

H_s = expected heterozygosity within a subpopulation. Say we have i different alleles at a locus in a subpopulation, and p_i is the frequency of the i th allele.

$$H_s = 2 \sum_{i,j} p_i p_j = 1 - \sum_i p_i^2$$

H_T = expected heterozygosity if there was random mating across the entire metapopulation.

$H_T = 1 - \sum_i \bar{p}_i^2$ where \bar{p}_i is the average frequency of the i th allele, calculated as the average of the values of p_i over all subpopulations. The summation is over all alleles at a locus.

The F statistics are as follows:

F_{IS} tells us if there is inbreeding within subpopulations by comparing H_I and H_S :

$$F_{IS} = \frac{\bar{H}_S - \bar{H}_I}{\bar{H}_S}$$

The bars above the H symbols mean that the values are the averages over all the subpopulations that we are considering. So F_{IS} measures whether there is, on average, a deficit of heterozygotes within subpopulations.

In contrast, gene flow between populations will tend to homogenize different populations. So drift and migration have opposite effects on population differentiation.

F_{ST} is the statistic that tells us how differentiated the subpopulations are. Formally, F_{ST} tells us if there is a deficit of heterozygosity in the metapopulation, due to differentiation among subpopulations:

$$F_{ST} = \frac{H_T - \bar{H}_S}{H_T}$$

F_{ST} increases with increasing differentiation among sub-populations. $F_{ST}=0$ indicates that all subpopulations have the same allele frequencies (no differentiation), and $F_{ST}=1$ indicates maximal differentiation (F_{ST} will = 1 when H_S for every subpopulation = 0, so every subpopulation is fixed for one of the alleles.) F_{ST} is sometimes called the fixation index because it increases as more subpopulations become fixed for a single allele. More intuitively, F_{ST} indicates how much of the total genetic variation in the metapopulation is distributed among subpopulations, rather than within subpopulations.

F_{IT} tells us how much population structure has affected the average heterozygosity of individuals within the population:

$$F_{IT} = \frac{H_T - \bar{H}_I}{H_T}$$

The mathematical relationship among the three F statistics is:

$(1-F_{IS})(1-F_{ST}) = (1-F_{IT})$. This relationship is useful for checking your calculations!

Please read Box 3.1 in the text carefully.

Example: *D. melanogaster* has two allozyme alleles at the *Adh* locus (F and S). In a high-elevation population, 100 flies were sampled and the frequency of the F allele was 0.3. In a nearby low-elevation population, 100 flies were sampled, and the frequency of F was 0.03. What is the value of F_{ST} for this locus?

A: Pop.1 (high elevation): $H_S = 1 - [(0.3)^2 + (0.7)^2] = 1 - [0.09 + 0.49] = 0.42$

Pop. 2 (low elevation): $H_S = 1 - [(0.03)^2 + (0.97)^2] = 1 - [0.0009 + 0.9409] = 0.06$

The average expected heterozygosity = $\overline{H_S} = (0.42 + 0.06) / 2 = 0.24$

The average allele frequency of the F allele over both populations is $(0.3 + 0.03) / 2 = 0.165$, so the average frequency of S is $(1 - 0.165) = 0.835$.

$$H_T = 1 - \sum_i p_i^2 = 1 - [(0.165)^2 + (0.835)^2] = 0.276$$

$$F_{ST} = \frac{H_T - \overline{H_S}}{H_T} = (0.276 - 0.24) / 0.276 = 0.13$$

This is a fairly high value of F_{ST} , indicating moderate to high genetic differentiation between the two populations. Any value of F_{ST} over 0.15 is considered a large amount of differentiation.

For multiple loci, the average F_{ST} is:

$$\overline{F}_{ST} = \frac{\overline{H_T} - \overline{\overline{H_S}}}{\overline{H_T}}$$

where $\overline{\overline{H_S}}$ is $\overline{H_S}$, averaged over all loci, subpopulations, and $\overline{H_T}$ is H_T , averaged over loci.

4.4.a The rate at which populations differentiate depends on N_e

The magnitude of genetic drift depends on population size, so the rate at which populations differentiate due to drift also depends on population size. The differences between populations (F_{ST}) will increase every generation as:

$$\Delta F_{ST} = \frac{1}{2N_e}$$

So the smaller the effective population size, the faster differences accumulate due to drift.

F_{ST} will increase with the generations, so, over multiple generations,

$$F_{ST_t} = 1 - (1 - \Delta F_{ST})^t \text{ where } F_{ST_t} \text{ is } F_{ST} \text{ in generation } t.$$

4.5 ESTIMATING MIGRATION RATES

1. Direct measures of the movement of individuals over their lifetime, e.g., mark-recapture methods. Very labor-intensive. Often does not tell you how gametes move. Miss rare events.
2. Release of individuals with a recognizable genetic marker that can be scored in their offspring can provide information on the movement of gametes. Any single study will provide information on short-term gene flow, but may miss important, but rare cases of gene flow that can have long-term effects on population structure.
3. Indirect measurement: infer gene flow from the differences in allele frequencies among populations. Strong divergence among populations indicates little gene flow, while similarity in allele frequencies indicate substantial gene flow. That is, the more similar the allele frequencies among populations, the higher the rate of gene flow. Note: this argument makes a couple of assumptions. A) neutrality; B) allele frequencies must be at an equilibrium between gene flow and genetic drift.

4.5.a Indirect measures

The indirect method of estimating gene flow usually involves the calculation of F_{st} (or related measures of population subdivision). If you know the allele frequencies at several genes in two different populations, you can calculate a value of F_{st} for each gene.

Under different migration models, the value of F_{ST} can be used to estimate the actual amount of migration between subpopulations. For example, under the continent-island model that we discussed previously, $F_{ST} = 1/[4N_e m + 1]$, where N_e is the effective population size, and m is the migration rate.

Q: What is the estimated migration rate between the high and low-elevation *Drosophila* populations from the example above? (Assume that the effective size is equal to the census size: $N_e=100$).

A:

The assumption that genetic markers (the genes we use to evaluate F_{ST}) are neutral wrt to selection is important because if selection favors the same alleles in different populations, those populations will have similar allele frequencies, even if there is little gene flow. Conversely, if selection favors different alleles, allelic frequencies can differ between populations, even if they experience gene flow.

Can we evaluate these assumptions?

A) Neutrality of alleles used as markers. Gene flow and genetic drift affect all genes in the same way. However, natural selection affects different loci independently (it favors alleles at one locus

independently of its effects on alleles at a different, unlinked locus). So, genetic drift and gene flow should lead to similar measures of population substructure in all the different genes that are studied, whereas natural selection will lead to differences among genes.

B) Equilibrium between gene flow and genetic drift. This requires that the populations have been separate long enough so that the allele frequencies have reached the values determined by the balance between drift (which tends to eliminate all but one allele at a locus) and gene flow (which tends to introduce new alleles).

Q: What would happen if an ancestral large population, had recently become divided into two populations that no longer exchanged alleles (no gene flow). What would allele frequencies look like for the first several generations after the separation of two populations? Would you get an accurate measure of gene flow by using the indirect method?

A: