Inbreeding and self-fertilization

Introduction

Remember that long list of assumptions associated with derivation of the Hardy-Weinberg principle
that I went over a couple of lectures ago — http://darwin.eeb.uconn.edu/eeb348/lecture-notes/hardy-weinberg/node2.html? Well, we're about to begin violating assumptions to explore the consequences, but we're not going to violate them in order. We're first going to violate Assumption #2:

> Genotypes mate at random with respect to their genotype at this particular locus.

There are many ways in which this assumption might be violated:

- Certain genotypes may be more successful in mating than others — sexual selection.
- Genotypes that are different from one another may mate more often than expected — disassortative mating, e.g., self-incompatibility alleles in flowering plants, MHC loci in humans (the smelly t-shirt experiment).
- Genotypes that are similar to one another may mate more often than expected — assortative mating.
- Some fraction of the offspring produced may be produced asexually.
- Individuals may mate with relatives — inbreeding.
  - self-fertilization
  - sib-mating
  - first-cousin mating
  - parent-offspring mating
  - etc.

©2001-2006 Kent E. Holsinger

When there is sexual selection or disassortative mating genotypes differ in their chances of being included in the breeding population. As a result, allele and genotype frequencies will tend to change from one generation to the next. We'll talk a little about these types of departures from random mating when we discuss the genetics of natural selection in a few weeks, but we'll ignore them for now.

Self-fertilization

Self-fertilization is the most extreme form of inbreeding possible, and it is characteristic of many flowering plants and some hermaphroditic animals, including freshwater snails.[1] It's not too hard to figure out what the consequences of self-fertilization will be without doing any algebra.

- All progeny of homozygotes are themselves homozygous.
- Half of the progeny of heterozygotes are heterozygous and half are homozygous.

So you might expect that the frequency of heterozygotes would be halved every generation, and you'd be right. To see why, consider the following mating table:

		Offsrping genotype		
Mating	frequency	A_1A_1	A_1A_2	A_2A_2
$A_1A_1 \times A_1A_1$	x_{11}	1	0	0
$A_1A_2 \times A_1A_2$	x_{12}	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
$A_2A_2 \times A_2A_2$	x_{22}	0	0	1

Using the same technique we used to derive the Hardy-Weinberg principle, we can calculate the frequency of the different offspring genotypes from the above table.

$$x'_{11} = x_{11} + x_{12}/4 \quad (1)$$

$$x'_{12} = x_{12}/2 \quad (2)$$

$$x'_{22} = x_{22} + x_{12}/4 \quad (3)$$

[1]It may well be characteristic of many hermaphroditic animal parasites. You should also know that I just lied. The form of self-fertilization I'm going to describe actually isn't the most extreme form of selfing possible. That honor belongs to gametophytic self-fertilization in homosporous plants. The offspring of gametophytic self-fertilization are uniformly homozygous at every locus in the genome. For more information, if you're interested, see: Holsinger, K. E. 1990. The population genetics of mating system evolution in homosporous plants. *Amer. Fern J.* 80:153–160.

I use the $'$ to indicate the next generation. We can also calculate the frequency of the A_1 allele among offspring, namely

$$\begin{aligned}
p' &= x'_{11} + x'_{12}/2 & (4)\\
&= x_{11} + x_{12}/4 + x_{12}/4 & (5)\\
&= x_{11} + x_{12}/2 & (6)\\
&= p & (7)
\end{aligned}$$

These equations illustrate two very important principles that are true with any system of strict inbreeding:

1. Inbreeding does not cause allele frequencies to change, but it will generally cause genotype frequencies to change.

2. Inbreeding reduces the frequency of heterozygotes relative to Hardy-Weinberg expectations. It need not eliminate heterozygotes entirely, but it is guaranteed to reduce their frequency.

   - Suppose we have a population of hermaphrodites in which $x_{12} = 0.5$ and we subject it to strict self-fertilization. Assuming that inbred progeny are as likely to survive and reproduce as outbred progeny, $x_{12} < 0.01$ in six generations and $x_{12} < 0.0005$ in ten generations.

Partial self-fertilization

Many plants reproduce by a mixture of outcrossing and self-fertilization. To a population geneticist that means that they reproduce by a mixture of selfing and random mating. Now I'm going to pull a fast one and derive the equations that determine how allele frequencies change from one generation to the next without using a mating table. To do so, I'm going to imagine that our population consists of a mixture of two populations. In one part of the population all of the reproduction occurs through self-fertilization and in the other part all of the reproduction occurs through random mating. If you think about it for a while, you'll realize that this is equivalnt to imagining that each plant reproduces some fraction of the time through self-fertilization and some fraction of the time through random mating. Let σ be the fraction of progeny produced through self-fertilization, then

$$\begin{aligned}
x'_{11} &= p^2(1-\sigma) + (x_{11} + x_{12}/4)\sigma && (8) \\
x'_{12} &= 2pq(1-\sigma) + (x_{12}/2)\sigma && (9) \\
x'_{22} &= q^2(1-\sigma) + (x_{22} + x_{12}/4)\sigma && (10)
\end{aligned}$$

It takes a little more algebra than it did before, but it's not difficult to verify that the allele frequencies don't change between parents and offspring.

$$\begin{aligned}
p' &= p^2(1-\sigma) + (x_{11} + x_{12}/4)\sigma + pq(1-\sigma) + (x_{12}/4)\sigma && (11) \\
&= p(p+q)(1-\sigma) + (x_{11} + x_{12}/2)\sigma && (12) \\
&= p(1-\sigma) + p\sigma && (13) \\
&= p && (14)
\end{aligned}$$

Because homozygous parents can always have heterozygous offspring (when they outcross), heterozygotes are never completely eliminated from the population as they are with complete self-fertilization. In fact, we can solve for the *equilibrium* frequency of heterozygotes, i.e., the frequency of heterozygotes reached when genotype frequencies stop changing.[2] By definition, an equilibrium for x_{12} is a value such that if we put it in on the right side of equation 9 we get it back on the left side, or in equations

$$\begin{aligned}
\hat{x}_{12} &= 2pq(1-\sigma) + (x_{12}/2)\sigma && (15) \\
\hat{x}_{12}(1-\sigma/2) &= 2pq(1-\sigma) && (16) \\
\hat{x}_{12} &= \frac{2pq(1-\sigma)}{(1-\sigma/2)} && (17)
\end{aligned}$$

It's worth noting several things about this set of equations:

1. I'm using $\hat{x}_{12}$ to refer to the equilibrium frequency of heterozygotes. I'll be using hats over variables to denote equilibrium properties throughout the course.[3]

[2] This is analogous to stopping the calculation and re-calculation of allele frequencies in the EM algorithm when the allele frequency estimates stop changing.

[3] Unfortunately, I'll also be using hats to denote estimates of unknown parameters, as I did when discussing maximum-likelihood estimates of allele frequencies. I apologize for using the same notation to mean different things, but I'm afraid you'll have to get used to figuring out the meaning from the context. Believe me. Things are about to get a lot worse. Wait until I tell you how many different ways population geneticists use a parameter f that is commonly called the inbreeding coefficient.

2. I can solve for $\hat{x}_{12}$ in terms of p because I know that p doesn't change. If p changed, the calculations wouldn't be nearly this simple.

3. The equilibrium is approached gradually (or asymptotically as mathematicians would say). A single generation of random mating will put genotypes in Hardy-Weinberg proportions (assuming all the other conditions are satisfied), but many generations may be required for genotypes to approach their equilibrium frequency with partial self-fertilization.

Inbreeding coefficients

Now that we've found an expression for $\hat{x}_{12}$ we can also find expressions for $\hat{x}_{11}$ and $\hat{x}_{22}$. The complete set of equations for the genotype frequencies with partial selfing are:

$$\hat{x}_{11} = p^2 + \frac{\sigma pq}{2(1-\sigma/2)} \tag{18}$$

$$\hat{x}_{12} = 2pq - 2\left(\frac{\sigma pq}{2(1-\sigma/2)}\right) \tag{19}$$

$$\hat{x}_{22} = q^2 + \frac{\sigma pq}{2(1-\sigma/2)} \tag{20}$$

Notice that all of those equations have a term $\sigma/(2(1-\sigma/2))$. Let's call that f. Then we can save ourselves a little hassle by rewriting the above equations as:

$$\hat{x}_{11} = p^2 + fpq \tag{21}$$

$$\hat{x}_{12} = 2pq(1-f) \tag{22}$$

$$\hat{x}_{22} = q^2 + fpq \tag{23}$$

Now you're going to have to stare at this a little longer, but notice that $\hat{x}_{12}$ is the frequency of heterozygotes that we'd observe and $2pq$ is the frequency of heterozygotes we'd expect under Hardy-Weinberg in this population if we were able to observe the genotype and allele frequencies without error. So

$$1 - f = \frac{\hat{x}_{12}}{2pq} \tag{24}$$

$$f = 1 - \frac{\hat{x}_{12}}{2pq} \tag{25}$$

$$= 1 - \frac{\text{observed heterozygosity}}{\text{expected heterozygosity}} \tag{26}$$

f is the inbreeding coefficient. When defined as 1 - (observed heterozygosity)/(expected heterozygosity) it can be used to measure the extent to which a particular population departs from Hardy-Weinberg expectations.[4] When f is defined in this way, I refer to it as the *population inbreeding coefficient*.

But f can also be regarded as a function of a particular system of mating. With partial self-fertilization the population inbreeding coefficient when the population has reached equilibrium is $\sigma/(2(1-\sigma/2))$. When regarded as the inbreeding coefficient predicted by a particular system of mating, I refer to it as the *equilibrium inbreeding coefficient*.

We'll encounter at least two more definitions for f once I've introduced ideas of identity by descent, but that's another lecture. Next we're going to use our new-found knowledge of inbreeding coefficients to go back and talk about how to determine when genotypes in a sample are not in Hardy-Weinberg proportions.[5]

Identity by descent

Self-fertilization is, of course, only one example of the general phenomenon of inbreeding— non-random mating in which individuals mate with close relatives more often than expected at random. We've already seen that the consequences of inbreeding can be described in terms of the inbreeding coefficient, f and I've introduced you to two ways in which f can be defined.[6] I'm about to introduce you to one more.

> Two alleles at a single locus are *identical by descent* if the are identical copies of the same allele in some earlier generation, i.e., both are copies that arose by DNA replication from the same ancestral sequence without any intervening mutation.

We're more used to classifying alleles by type than by descent. All though we don't usually say it explicitly, we regard two alleles as the "same," i.e., identical by type, if they have the same phenotypic effects. Whether or not two alleles are identical by descent, however, is a property of their genealogical history. Consider the following two scenarios:

[4] f can be negative if there are more heterozygotes than expected, as might be the case if cross-homozygote matings are more frequent than expected at random.

[5] And why are we interested in this? Not just because this type of question is likely to be involved in the first problem set, but because we know that if genotypes are *not* in Hardy-Weinberg proportions, then one or more of the assumptions involved in their derivation *must* have been violated, i.e., there are some evolutionary forces at work.

[6] See paragraphs above describing the population and equilibrium inbreeding coefficient.

Identity by descent

$$
\begin{array}{ccccc}
 & & A_1 & \rightarrow & A_1 \\
 & \nearrow & & & \\
A_1 & & & & \\
 & \searrow & & & \\
 & & A_1 & \rightarrow & A_1
\end{array}
$$

Identity by type

$$
\begin{array}{ccccc}
 & & A_1 & \rightarrow & A_1 \\
 & \nearrow & & & \\
A_1 & & & & \\
 & \searrow & & & \\
 & & A_2 & \rightarrow & A_1 \\
 & \uparrow & & \uparrow & \\
 & \text{mutation} & & \text{mutation} &
\end{array}
$$

In both scenarios, the alleles at the end of the process are identical in type, i.e., they're both A_1 alleles. In the second scenario, however, they are identical in type only because one of the alleles has two mutations in its history.[7] So alleles that are identical by descent will also be identical by type, but alleles that are identical by type need not be identical by descent.[8]

A third definition for f is the probability that two alleles *chosen at random* are identical by descent.[9] Of course, there are several aspects to this definition that need to be spelled out more explicitly.

- In what sense are the alleles chosen at random, within an individual, within a particular population, within a particular set of populations?

- How far back do we trace the ancestry of alleles to determine whether they're identical by descent? Two alleles that are identical by type may not share a common ancestor if we trace their ancestry only 20 generations, but they may share a common ancestor if we trace their ancestry back 1000 generations and neither may have undergone any mutations since they diverged from one another.

[7] Notice that we could have had each allele mutate independently to A_2.

[8] Systematists in the audience will recognize this as the problem of homoplasy.

[9] Notice that if we adopt this definition for f it can only take on values between 0 and 1. When used in the sense of a population or equilibrium inbreeding coefficient, however, f can be negative.

Let's imagine for a moment, however, that we've traced back the ancestry of all alleles in a particular population far enough to be able to say that if they're identical by type they're also identical by descent. Then we can write down the genotype frequencies in this population once we know f, where we define f as the probability that two alleles chosen at random in this population are identical by descent:

$$x_{11} = p^2(1-f) + fp \quad (27)$$
$$x_{12} = 2pq(1-f) \quad (28)$$
$$x_{22} = q^2(1-f) + fq \quad . \quad (29)$$

It may not be immediately apparent, but you've actually seen these equations before in a different form. Since $p - p^2 = p(1-p) = pq$ and $q - q^2 = q(1-q) = pq$ these equations can be rewritten as

$$x_{11} = p^2 + fpq \quad (30)$$
$$x_{12} = 2pq(1-f) \quad (31)$$
$$x_{22} = q^2 + fpq \quad . \quad (32)$$

You can probably see why population geneticists tend to play fast and loose with the definitions. *If* we ignore the distinction between identity by type and identity by descent, then the equations we used earlier to show the relationship between genotype frequencies, allele frequencies, and f (defined as a measure of departure from Hardy-Weinberg expectations) are identical to those used to show the relationship between genotype frequencies, allele frequencies, and f (defined as a the probability that two randomly chosen alleles in the population are identical by descent).

Creative Commons License

These notes are licensed under the Creative Commons
Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit
http://creativecommons.org/licenses/by-nc-sa/2.5/ or send a letter to Creative Commons,
559 Nathan Abbott Way, Stanford, California 94305, USA.