Hardy–Weinberg Equilibrium and the Foundations of Evolutionary Genetics

2. The Inertial State in Population Genetics

Amitabh Joshi

The year 2008 marks the 100th anniversary of the enunciation of the principle defining the inertial state of populations from a genetic viewpoint through the independent publication of papers on the topic by G H Hardy and W Weinberg. In this part, we formally build up the model that underlies the Hardy-Weinberg Equilibrium, as well as its mathematical representation.

We saw in the example considered in Part 1[1] of this article that genotypic frequencies did not change from one generation to the next. We will now examine this situation in greater detail and also develop a somewhat idealized, but very useful, depiction of what happens to genes in a population over generations, at least under certain simplifying assumptions. We start by building up a picture of an *ideal large population*, a hypothetical construct that, nevertheless, captures the essential genetic aspects of what happens during the transition from one generation to the next in real diploid, sexually reproducing, populations. It is more amenable to simple depiction, mathematical representation and algebraic manipulation.

The Ideal Large Population

Consider a very large population of *N* individuals of a diploid, sexually reproducing, hermaphroditic species (if you prefer to think of a species with separate sexes, assume a sex ratio of 1:1) whose genetic composition at some locus *A*, at the breeding adult stage of a given generation *t*, is described by an array of genotypic frequencies P_{ij}, $i,j = 1 \dots m$ among the breeding adults. As we saw in Part 1, such a population can also be described in terms of an array of allele frequencies p_i, $i = 1 \dots m$, that are related to the

Amitabh Joshi studies and teaches evolutionary genetics and population ecology at the Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore. His current research interests are in life-history, evolution, the evolutionary genetics of biological clocks, the evolution of ecological specialization dynamics. He also enjoys music (especially traditional qawali in Braj, Farsi, Punjabi and Urdu), history, philosophy, and reading and writing poetry in Urdu, Hindi and English.
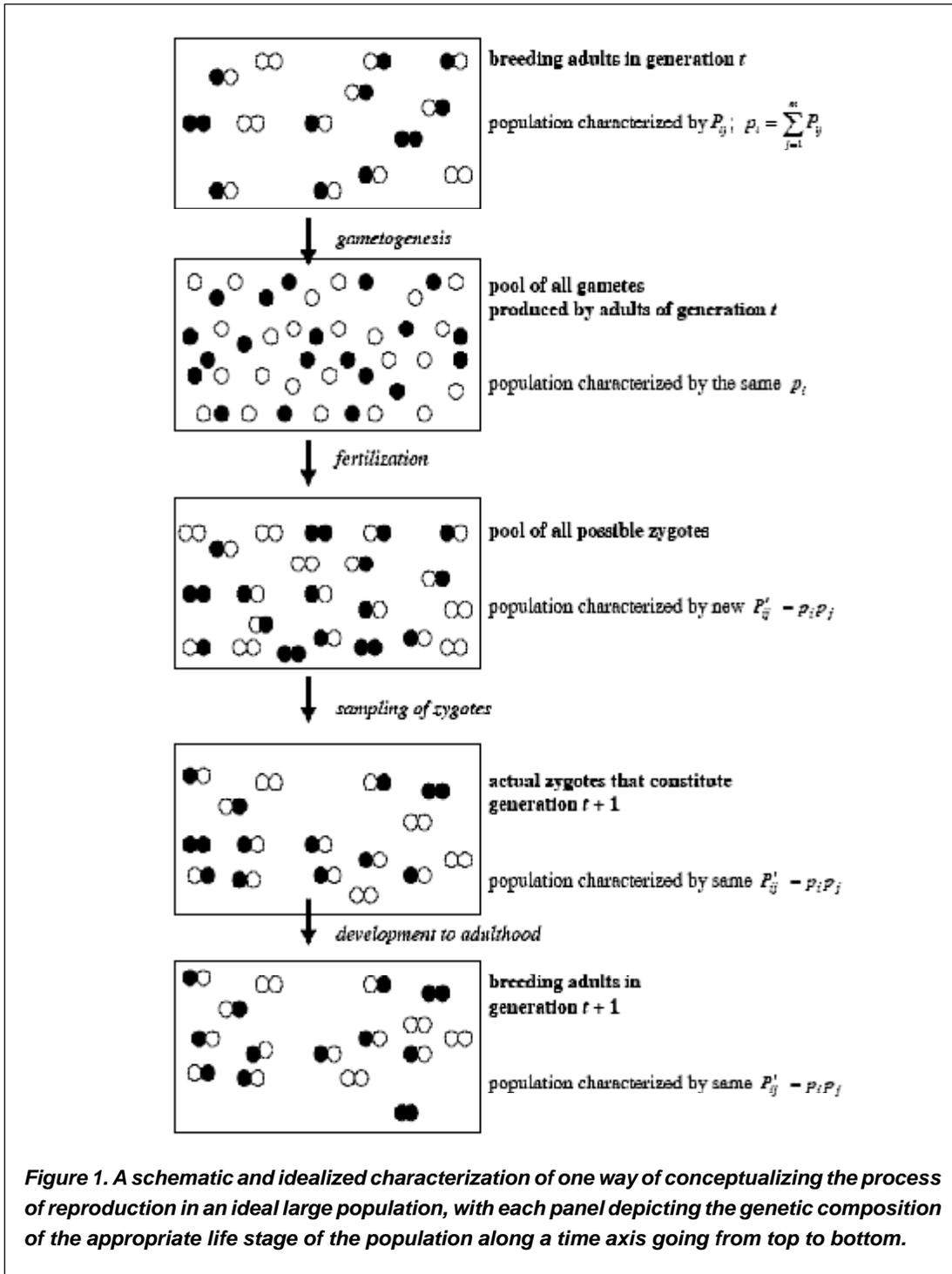
genotypic frequencies by $p_i = \sum_{j=1}^{m} P_{ij}$. This is the situation represented in panel 1 of *Figure* 1, with the paired black or white circles representing different alleles. For simplicity, only two alleles are depicted in the figure; the argument we are developing, nevertheless, applies to any number of alleles. Next, these breeding adults will produce gametes and we shall imagine that we can collect all these gametes into a large gamete pool. Of course, this is not what really happens during sexual reproduction (at least not in species with internal fertilization). If we assume that no net mutation from A_i to A_j takes place in the cells that will give rise to gametes, and that fertility is not correlated with A locus genotypes (such that individuals of each A locus genotype produce the same number of gametes on average), it follows that the allele frequencies p_i will not change in the transition from breeding adults to gamete pool. This is because, under the assumptions we have made, all that happens when we go from the breeding adults to the gamete pool is that alleles paired up into diploid individuals separate and are copied many times into haploid gametes, and the number of copies made is the same for different alleles A_i, $i = 1...m$ (*Figure* 1, panel 2).

We now assume that all these gametes in the gamete pool fuse in pairs at random (see *Box* 1) to give rise to a pool of all possible diploid zygotes (*Figure* 1, panel 3)[1]. What we are saying is that, conceptually, the process of fertilization occurs in two separate steps: (a) the fusion of all gametes to produce a large pool of potential zygotes, and (b) a sampling of a sub-set of these potential zygotes to represent the actual zygotes that become the next generation.

Given that millions of gametes are typically produced by each individual, and that we have assumed that all of them fuse to form zygotes, we now have millions of zygotes in our zygote pool. However, our population consists of N individuals, and we will assume that the population tends to maintain constant numbers from one generation to the next. Once again, this statement actually contains two assumptions, namely that (a) the number of breeding adults N is constant, and (b) generations are discrete

[1] This is an important step and, once again, is not an accurate representation of what really happens during sexual reproduction. In reality, large numbers of the gametes formed never undergo fertilization and do not, therefore end up as zygotes at all.

Figure 1. A schematic and idealized characterization of one way of conceptualizing the process of reproduction in an ideal large population, with each panel depicting the genetic composition of the appropriate life stage of the population along a time axis going from top to bottom.

Box 1. Random Fusion of Gametes

What exactly is meant by the assumptions that gametes fuse at random? This phrase, which we shall encounter often in this article, actually covers two distinct assumptions. The first is the assumption of randomness of gametic fusion with regard to the A alleles carried by the gametes. This simply means that the probability of a fusion between an A_i bearing gamete and an A_j bearing gamete is equal to the product of the frequencies of these two types of gamete in the gamete pool (i.e., $p_i p_j$). Note that this does not imply that gametes need to fuse at random in an absolute sense; gamete fusion may be very non-random with regard to another locus. For example, suppose that there is one more locus B, with two alleles B_1 and B_2, and that B_1 gametes fuse only with B_1 gametes and B_2 gametes fuse only with B_2 gametes. This is clearly an extreme case of non-random fusion but, as long as there is no particular pattern of association of specific A alleles with B alleles in gametes (i.e. every A_i has probability equal to its frequency p_i of being associated with any B_i), fusion will still be random with regard to A alleles. The second assumption implicit in the phrase "gametes fuse at random" is that there are no differences in the intrinsic abilities of gametes carrying particular A alleles to successfully achieve fusion. Once again, such differences may exist with regard to alleles at other loci, but as long as these differences are unrelated to the particular A allele present in the gamete, it does not invalidate the assumption. These assumptions of randomness of gamete fusion, and equal fusion capacities of all gametes are equivalent to assuming, under a more realistic visualization, that diploid individuals choose mates at random with regard to their A locus genotypes, and that differences in fertility among mating individuals are unrelated to their A locus genotypes. It is clear, from the very manner in which randomness is defined, that under these assumptions, the genotypic frequencies in the pool of all possible zygotes are related to allele frequencies in the gamete pool as $P_{ij} = p_i p_j$. The use of P'_{ij} rather than P_{ij} stresses that the genotypic frequencies in the pool of possible zygotes that will give rise to generation $t+1$ need not necessarily be identical to the genotypic frequencies among breeding adults in generation t. It is also clear that the allele frequencies p_i in the pool of possible zygotes will be the same as those in the gamete pool, because all that has happened is that the gametes have now been paired up; no gametes have been added to or removed from the pool, and no mutations are allowed which could change the identity of an allele carried by a gamete.

[2] Since we started with the assumption that N was very large, we are dealing here with a very large sample and, therefore, we can reasonably expect the representation of different genotypes $A_i A_j$ in the sample of N individuals to be an accurate reflection of their relative proportions in the pool of possible zygotes.

such that offspring do not coexist with their parents. Now, we need to conceptualize how we get N actual zygotes that form generation $t + 1$ from these millions of zygotes in the pool of all possible zygotes. We will assume that this is a purely random process with regard to A locus genotype, and that, therefore, what we are doing is sampling N zygotes at random from this very large pool. Since we are sampling from a very large pool, we can assume the sampling, for all practical purposes, to be '*with replacement*'[2]. In other words, we do not expect the genotypic frequencies P'_{ij} to change as we go from the pool of all possible zygotes to the pool of actual zygotes in generation $t + 1$ (*Figure* 1, panel 4).

The only process now remaining in our conceptualization of what happens genetically from one generation to the next is the transition from zygotes of generation $t + 1$ to breeding adults of generation $t + 1$. We assume that any differences in viability (survivorship) among individuals are unrelated to their A locus genotypes. Consequently, on average, we expect individuals of the various genotypes A_iA_j to have equal viability. We should add here that if the overall survival is less than 100%, all we need to do to account for that is to assume an appropriately larger sample of actual zygotes. For example, if only 50% of zygotes survive to adulthood (reproductive age), then to maintain N constant, we will assume that a sample of $2N$ zygotes was taken from the pool of all possible zygotes. Since the viability of different genotypes A_iA_j is not different, we again do not expect the genotypic frequencies P'_{ij} to change as we go from zygotes to breeding adults in generation $t + 1$ (*Figure* 1, panel 5).

The Hardy–Weinberg Law

Let us now summarize the results obtained from our conceptualization of the dance of the genes across generations. We made many assumptions, which together defined the ideal large population (see *Box* 2).

What we obtained from our conceptual exercise was the result that in an ideal large population, (a) allele frequencies will not change from one generation to the next, and (b) genotypic frequencies will be related to allele frequencies as $P'_{ij} = p_i p_j$ by random mating. The corollary to point (b) is that regardless of what genotypic frequencies we start with, one generation of random mating in an ideal large population will alter them to values determined by the allele frequencies (following the relationship $P'_{ij} = p_i p_j$) and the genotypic frequencies, too, will remain constant over generations thereafter. This principle was discovered independently by G H Hardy, an English mathematician, and W Weinberg, a German physician, in 1908, and is often called the Hardy–Weinberg Law. Populations fulfilling the relationship $P'_{ij} = p_i p_j$ for any locus are said to be in Hardy–Weinberg

---

Box 2. The Ideal Population

(a)   Consists of a species that is hermaphroditic, or if dioecious, has a 1:1 sex ratio,

(b)   Discrete generations, and constant large N,

(c)   Random mating with regard to the locus under consideration (for a hermaphroditic species, this includes selfing too),

(d)   No differential mutation or migration, with regard to different genotypes at the locus under consideration,

(e)   No differences among gametes carrying different alleles at the locus under consideration in ability to undergo fertilization, and

(f)   No differences among genotypes at the locus under consideration with regard to fertility and survival to breeding age (points (e) and (f) can be summarized as the assumption that no natural selection is operating on the locus under consideration).

---

equilibrium with regard to that locus.

Hardy and Weinberg actually arrived at this result through the much cleaner route of algebra, which we shall now describe. The reason I preferred to first develop the model of the ideal large population verbally was to emphasize and exemplify that in constructing a model one is primarily trying to build up an idealized version of some real situation, a connection that is often lost when one goes straight to a mathematical depiction of the model. Moreover, I hope that being able to compare the algebraic and verbal formulations of the model will leave you with a better feel for what exactly it is that lies hidden behind the veil of algebra.

What Hardy, with a mathematician's insight had clearly seen was that under certain simplifying assumptions (random mating, no mutation, selection, etc., at that locus), the proportion of genotypes in a generation could be obtained by a binomial expansion of the allelic frequencies. This is of course a piece of high school algebra we are all familiar with, namely that $(a + b)^2 = a^2 + 2ab + b^2$. Thus, if the allele frequencies of the two alleles in a one-locus biallelic system are p_1 and p_2, the frequencies of the three genotypes will be given by $(p_1 + p_2)^2$, as $P_{11} = p_1^2$, $P_{12} = 2p_1p_2$, and $P_{22} = p_2^2$ (*Table* 1). This is, of course, exactly the same logic as

---

Box 3. G H Hardy's Letter to *Science* (1908)

"*I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists.*" Thus starts a letter to the editor of the journal *Science*, written in 1908 by Hardy [1], which can formally be said to mark the beginning of population genetics as a scientific discipline. The backdrop to this letter is an interesting one. At the time, many people were arguing that Mendel's Laws suggested that the proportion of individuals in a population exhibiting the recessive phenotype for a single locus trait should be 25 %. In retrospect, we can see that this is a clear misapplication of the Mendelian 3:1 ratio among the F_2 progeny of a cross between an A_1A_1 and an A_2A_2 individual to a situation where one is dealing with the transmission of traits in a whole population rather than a family with a particular combination of parental genotypes. At that time, though, many such misconceptions existed among biologists who were trying to extrapolate population characteristics from Mendel's Laws. One of Hardy's colleagues, Punnett, was uncomfortable with this argument for 25% of a population showing the recessive phenotype but was unable to clearly refute it. He happened to discuss the problem with Hardy, and after asking him a few questions about inheritance, Hardy provided him not only with a clear refutation of this argument but also with a system with which one could, under certain assumptions, deduce the genotypic frequencies in a population from the allelic frequencies (the Hardy–Weinberg Law). Punnett impressed upon Hardy the importance of publishing the result, which Hardy was hesitant to do because the mathematics involved was clearly trivial and Hardy was mindful of his reputation as one of the outstanding mathematicians of his time. Mercifully for many of us, Punnett succeeded and Hardy finally wrote up his results and sent them for publication.

---

that used by Mendel to predict the 3:1 ratio among progeny resulting from a cross between two F_1 heterozygotes. The difference is that Mendel was dealing with a specific case where the allele frequencies among the uniting gametes were 0.5 for both alleles, whereas Hardy's algebraic formulation is much more generally applicable to populations with any combination of allele frequencies p_1, p_2 ($0 \leq p_i \leq 1$; $i = 1,2$).

Now let us return to the assumptions we have made in describing what happens in an ideal large population between generations. The assumptions regarding mutation, migration, selection and large population size are, of course, to be taken as axioms for the present, one can always examine, within the same general framework, the consequences of relaxing these assumptions in some detail. But what of the manner in which we chose to model the process of reproduction in which we substituted a gamete pool

In fact one great advantage of a mathematical approach to a problem is that mathematical formulations are typically generalizable and this generalization often leads to a better understanding even of specific situations than a more restrictive treatment would have yielded.

Table 1. The derivation of genotypic frequencies in the next generation resulting from random mating in an ideal large population with allele frequencies p_1 *and* p_2 *at a given locus* **A** *with two alleles. Gametic and genotypic frequencies are at the top left of each cell in the table.*

♂ gametes →	p_1	p_2
♀ gametes ↓	A_1	A_2
p_2 A_1	$p_1 \times p_1 = p_1^2$ $A_1 A_1$	$p_1 \times p_2 = p_1 p_2$ $A_1 A_2$
p_2 A_2	$p_1 \times p_2 = p_1 p_2$ $A_2 A_1$	$p_2 \times p_2 = p_2^2$ $A_2 A_2$

and a pool of all possible zygotes for the more natural process of mating between individuals. Clearly, it would be desirable to convince ourselves that this representation does in fact mirror what happens as a consequence of random mating.

Hardy–Weinberg Equilibrium through Mating Frequencies

Consider an ideal large population whose genetic composition at a biallelic locus, A, is described by the genotypic frequencies P_{11}, $2P_{12}$, and P_{22}, and that the allele frequencies are $p_1 = P_{11} + P_{12}$, $p_2 = 1 - p_1 = P_{22} + P_{12}$. Now assume that the individuals in this population mate at random, and each type of mating produces the same number of offspring. The frequencies of the various matings, and the genotypes and frequencies of the resulting offspring can, as we saw in Part 1, easily be described (*Table* 2). If the representation in *Figure* 1 is an accurate depiction of random mating in an ideal large population, then the conclusions reached from that representation should be in accordance with those reached through examining mating and offspring frequencies. We should be, therefore, able to show that (i) the allele frequencies among offspring are unchanged, and (ii) the genotypic frequencies among the offspring follow the relationship $P_{ij} = p_i p_j$. To do this, let us first sum up the frequencies in column three of *Table* 2, to obtain the new genotypic frequencies (P_{ij}') in the offspring.

Type of mating	Frequency	Offspring Frequencies		
$A_1A_1 \times A_1A_1$	P_{11}^2	A_1A_1 P_{11}^2	A_1A_2 0.0	A_2A_1 0.0
$A_1A_1 \times A_1A_2$	$4P_{11}P_{12}$	A_1A_1 $2P_{11}P_{12}$	A_1A_2 $2P_{11}P_{12}$	A_2A_1 0.0
$A_1A_1 \times A_2A_2$	$2P_{11}P_{22}$	A_1A_1 0.0	A_1A_2 $2P_{11}P_{12}$	A_2A_1 0.0
$A_1A_2 \times A_1A_2$	$4P_{12}^2$	A_1A_1 P_{12}^2	A_1A_2 $2P_{12}^2$	A_2A_1 P_{12}^2
$A_2A_2 \times A_1A_2$	$4P_{11}P_{22}$	A_1A_1 0.0	A_1A_2 $2P_{12}P_{22}$	A_2A_1 $2P_{12}P_{22}$
$A_2A_2 \times A_1A_2$	P_{22}^2	A_1A_1 0.0	A_1A_2 0.0	A_2A_1 P_{22}^2

$$P'_{11} = P_{11}^2 + 2P_{11}P_{12} + P_{12}^2$$

$$2P'_{12} = 2P_{11}P_{12} + 2P_{11}P_{22} + 2P_{12}P_{22} + 2P_{12}^2$$

$$P'_{22} = P_{12}^2 + 2P_{12}P_{22} + P_{22}^2$$

Let us now examine the new allele frequencies (p'_i) in the offspring. By definition, $p'_1 = P'_{11} + P'_{12}$ and, therefore, $p'_1 = P_{11}^2 + 2P_{11}P_{12} + P_{12}^2 + P_{11}P_{12} + P_{11}P_{22} + P_{12}P_{22} + P_{12}^2$. A little algebraic rearrangement and simplification shows that, in fact $p'_1 = p_1$.

$$p'_1 = P_{11}^2 + 2P_{11}P_{12} + P_{12}^2 + P_{11}P_{12} + P_{11}P_{22} + P_{12}P_{22} + P_{12}^2$$
$$= P_{11}(P_{11} + 2P_{12} + P_{12} + P_{22}) + P_{12}(P_{12} + P_{22} + P_{12})$$

(collecting like terms outside)

$$= P_{11}(1 + P_{12}) + P_{12}(1 - P_{11})$$

(making use of the fact that $P_{11} + 2P_{12} + P_{22} = 1$)

$$= P_{11} + P_{12} = p_1$$

And since $p_1 + p_2 = 1$, and $p'_1 + p'_2 = 1$, if $p'_1 = p_1$, then $p'_2 = p_2$. The model we used appears to work well because the offspring genotypic frequencies, calculated from mating frequencies, are also in Hardy–Weinberg Equilibrium, as shown below.

Table 2. The frequencies of different possible types of matings between individuals with different genotypes at a biallelic locus, along with the types of offspring that each mating will produce, and the proportion of each type of offspring with respect to the total offspring population resulting from all matings. For example, an $A_1A_1 \times A_1A_2$ mating will produce 50% A_1A_1 offspring and 50% A_1A_2 offspring, and this type of mating occurs with a frequency of $2 \times P_{11} \times 2P_{12} = 4P_{11}P_{12}$ in the population. Therefore, the proportion of each of the two genotypes of offspring resulting from this mating among the total number of offspring in the next generation will be $2P_{11}P_{12}$.

Our conceptualization seems a reasonable characterization of reality: allele frequencies do not change after one round of random mating in an ideal large population, regardless of whether we model using mating frequencies or the gamete pool.

$$P'_{11} = P_{11}^2 + 2P_{11}P_{12} + P_{12}^2$$
$$= (P_{11} + P_{12})^2$$
$$= p_1^2 \quad \text{(using the fact that } p_1 = P_{11} + P_{12})$$

It can similarly be shown that $P'_{22} = p_2^2$; and since $\sum_{i,j=1}^{2} P_{ij} = 1$, if $P'_{11} = p_1^2$ and $P'_{22} = p_2^2$, then $2P'_{12} = 1 - p_1^2 - p_2^2 = 2p_1p_2$.

It is, thus, clear that our formulation of reproduction via a gamete pool and fusion at random among the gametes yields the same results as a more realistic tracking of offspring frequencies for all the different possible matings in a random mating population. We can, therefore, be confident about using the gamete pool representation for other scenarios too.

Box 4. Extending the Hardy–Weinberg Law: Multiple Alleles at a Locus

If we contemplate a triallelic locus, with the frequencies of the three alleles being p_1, p_2 and p_3, we can obtain the array of genotypic frequencies P_{ij} $(i,j = 1,2,3)$ by expanding the expression $(p_1 + p_2 + p_3)^2$. The frequencies of the six unordered genotypes are

A_1A_1:- $P_{11} = p_1^2$; A_2A_2:- $P_{22} = p_2^2$; A_3A_3:- $P_{33} = p_3^2$; A_1A_2:- $2P_{12} = 2p_1p_2$;
A_1A_3:- $2P_{13} = 2p_1p_3$; A_2A_3:- $2P_{23} = 2p_2p_3$.

So far we have considered situations where we know the frequencies of all the genotypes. However, when one or more of the alleles at a locus are dominant, all we can empirically detect are the frequencies of the dominant and recessive phenotypes. Thus, for a bialleleic locus where A_1 is dominant over A_2, we cannot phenotypically distinguish between A_1A_1 and A_1A_2 individuals. Nevertheless, on the assumption of Hardy–Weinberg Equilibrium, we can predict the proportion of heterozygotes among the individuals who show the dominant phenotype, and this can be important in genetic counseling in order to assess the likelihood of an individual being a carrier for a recessive genetic disorder.

For example, the frequency of the recessive disorder 'cystic fibrosis' is about 1 in 1800 births in some human populations. In other words, $P_{22} = 1/1800 \approx 0.000556$. Assuming the population to be in Hardy–Weinberg equilibrium at this locus, the frequency of the recessive allele $p_2 = \sqrt{P_{22}} = 0.02357$. The frequency of the

Box 4. continued...

Box 4. continued...

dominant allele, therefore, is $p_1 = 1 - p_2 = 0.97643$, and the expected frequency of carriers (heterozygotes who do not suffer from the disorder but can transmit it to their offspring) is $2 p_1 p_2 = 0.04603$. The expected proportion of carriers among phenotypically normal individuals, therefore, is $\dfrac{2 p_1 p_2}{p_1^2 + 2 p_1 p_2} = 0.04606$ (in other words, approximately four and a half percent of normal individuals are carriers).

A similar approach can be applied to the ABO blood group system in humans, where there are three alleles – i^A, i^B and i^O - and, consequently, six possible genotypes at this locus – $i^A i^A$, $i^A i^B$, $i^A i^O$, $i^B i^B$, $i^B i^O$, and $i^O i^O$. The dominance relations between these alleles are that i^A and i^B are codominant among themselves, and both are dominant to i^O. Thus, an individual with phenotype (blood group) A can be genotypically $i^A i^A$ or $i^A i^O$, one with phenotype B can be $i^B i^B$ or $i^B i^O$, whereas individuals with AB and O phenotypes are genotypically $i^A i^B$ and $i^O i^O$, respectively. Empirically, all that can be determined is the frequency of the four different phenotypes. Let the phenotypic frequencies of A, B, AB and O individuals be given by P_A, P_B, P_{AB}, P_O, respectively. Assuming the population to be in Hardy–Weinberg Equilibrium, we can estimate the frequency of the i^O allele as $p_O = \sqrt{P_O}$. We can further estimate the frequency of the i^A allele as follows:

$$P_A = p_A^2 + 2 p_A p_O$$
$$\therefore\ P_A + P_O = p_A^2 + 2 p_A p_O + p_O^2 = \left(p_A + p_O\right)^2$$
$$\therefore\ p_A + p_O = \sqrt{P_A + P_O}$$
$$\therefore\ p_A = \sqrt{P_A + P_O} - p_O = \sqrt{P_A + P_O} - \sqrt{P_O}\ .$$

In this manner, the frequency of the i^A allele can be determined from the empirically observable frequencies of individuals with the A and O blood groups. Of course, once we know two of the allele frequencies, we can get the third as $p_B = 1 - p_A - p_O$, and then we can calculate the expected frequencies of all genotypes.

Extending the Hardy–Weinberg Law: X-linked Loci

So far, we have been dealing with autosomal loci, or loci present on chromosomes other than the sex (X and Y) chromosomes. As we saw in the case of linkage in Part 1, Mendel's Laws often need to be modified to account for genetic scenarios different from the ones he happened to work with: Mendel, fortuitously had worked with unlinked autosomal loci. What happens in the case of loci on the X or Y chromosomes is a straightforward consequence of the way in which X and Y chromosomes are inherited (in the follow-

Mendel's Laws often need to be modified to account for genetic scenarios different from the ones he happened to work with: Mendel, fortuitously had worked with unlinked autosomal loci.

Figure 2. Schematic depiction of the inheritance of alleles at X- and Y-linked loci (squares: Y-linked; circles: X-linked) in an ideal population with a 1:1 sex ratio. At generation t, all males have the 'light grey' allele on their X chromosome, half have a 'dark grey' allele on the Y chromosome and half a 'white' allele. The females in generation t have two other X-linked alleles, a 'white' and a 'black' allele. Note that at any generation, all the Y-linked alleles will be in the male part of the population, whereas 2/3 of the X-linked alleles will be in the female and 1/3 in the male part of the population (for autosomal loci, the male and female parts of the population contain half the alleles each). During the transition to the next generation (t+1), all Y-linked alleles are passed from males to males (solid arrow), and all the X-linked alleles of males in generation t will be passed to females in generation t+1 (dotted arrow). Of the X-linked alleles in females, half will be passed to females and half to males in the next generation (dashed arrows).

ing discussion I assume that we are dealing with a species in which the females, as in humans, are XX and males XY). The basic point about the inheritance of loci on the Y chromosome is that they are inherited from males to males: sons get their Y chromosomes exclusively from their fathers, and fathers pass on Y chromosomes only to sons. This results in very simple inheritance at the populational level. If the frequency of an allele at a locus on the Y chromosome (a Y-linked locus) in an ideal large population is p_i in generation t, it will remain p_i in generation $t+1$, because all these alleles continue to remain in the male population in exactly the same frequency as before (*Figure* 2). In the case of X-linked loci, the situation is a little different. First, the distribution of alleles at X-linked loci is uneven among the sexes in a population: two-thirds of all X chromosomes are in the females and only one-third in the males. Moreover, males pass on their X chromosomes only to their daughters, whereas females pass on their X chromosomes equally to sons and daughters, resulting in a lopsided criss-cross pattern of across-generation inheritance of X chromosomes (*Figure* 2). We will now examine the consequences of this pattern of inheritance on the behaviour of allele frequencies across generations.
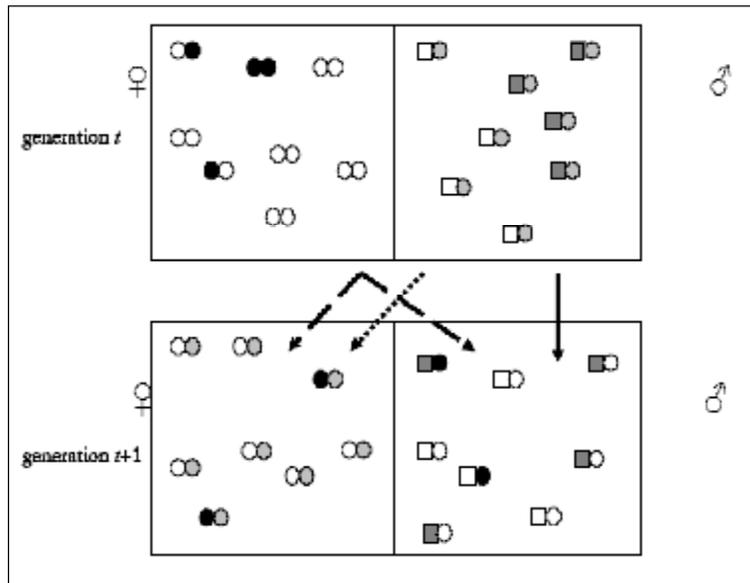
Imagine an X-linked locus with two alleles, A_1 and A_2. Males can only have two genotypes, A_1Y and A_2Y, whereas in females all three genotypes, A_1A_1, A_1A_2 and A_2A_2 are possible. Let the frequencies of these five sex-specific genotypes at generation t be $P_1, P_2, Q_{11}, 2Q_{12}$ and Q_{22}, respectively. Then, the allele frequencies in generation t can be written separately for the two sexes as (a) in males $p_1 = P_1$, $p_2 = P_2$, and (b) in females $q_1 = Q_{11} + Q_{12}$, $q_2 = Q_{22} + Q_{12}$. (Note that we now have to keep track of allele frequencies separately in males and females, unlike with autosomal loci.) Now, all males in generation $t + 1$ will get their X chromosomes from their mothers. The allele frequencies at the A locus in males of the next generation will, therefore, simply be the allele frequencies in the females in the previous generation. So, using primes (′) for the frequencies in generation $t + 1$, we can write $p_1' = q_1$, and $p_2' = q_2$. The females, of course, get one X chromosome from their father and one from their mother. So, an A_1A_1 female in generation $t+1$, for example must get an A_1 allele from each of her parents. Thus, the genotypic frequencies in females in generation $t + 1$ can be written as $Q_{11}' = p_1q_1$, $2Q_{12}' = p_1q_2 + p_2q_1$, and $Q_{22}' = p_2q_2$. Once we have the genotypic frequencies, we can derive the new allele frequencies in females in generation $t + 1$ as follows.

$$q_1' = Q_{11}' + Q_{12}' = p_1q_1 + \tfrac{1}{2}\left(p_1q_2 + p_2q_1\right)$$

$$= p_1q_1 + \tfrac{1}{2}\left[p_1\left(1 - q_1\right) + \left(1 - p_1\right)q_1\right]$$

$$\text{using } p_2 = 1 - p_1 \text{ and } q_2 = 1 - q_1$$

$$= p_1q_1 + \tfrac{1}{2}\left(p_1 + q_1\right) - \tfrac{1}{2}\left(p_1q_1 + p_1q_1\right)$$

$$= p_1q_1 + \tfrac{1}{2}\left(p_1 + q_1\right) - p_1q_1$$

$$= \tfrac{1}{2}\left(p_1 + q_1\right).$$

In exactly the same way we can show that $q_2' = \tfrac{1}{2}\left(p_2 + q_2\right)$. In other words, the frequency of an X-linked allele in females is the average of what it was in males and in females in the previous generation. This is a clear consequence of the fact that females get half their chromosomes from their fathers and half from their mothers.

Let us now examine the behaviour of the difference between the male and female allele frequencies for an allele. Let the difference between the frequency of A_1 in males and females be given by y. Then, $y = p_1 - q_1$, and similarly, in the next generation, $y' = p_1' - q_1'$. But, we can now substitute the earlier derived values for p_1' and q_1' into the expression for y', as follows.

$$y' = p_1' - q_1' = q_1 - \tfrac{1}{2}(p_1 + q_1)$$
$$= \tfrac{1}{2}q_1 - \tfrac{1}{2}p_1 = -\tfrac{1}{2}(p_1 - q_1)$$
$$= -\tfrac{1}{2}y.$$

In other words, in each generation the difference between male and female frequencies for an allele is halved in magnitude and reversed in sign. Thus over a long period of time, the difference in male and female frequencies should vanish altogether and p_i and q_i should converge to a common value.

The question then is: What is the common value to which the allele frequencies in males and females converge? Intuitively, we can guess that it should be the initial frequency in the whole population. After all, from one generation to the next, X-linked alleles are not entering or leaving the population; they are merely being reallocated between the male and female parts of the population. We can now confirm this guess algebraically. Let x be the frequency of A_1 in the whole population. Since the male and female parts of the population contain different proportions of all X-linked alleles, the average of male and female frequencies must be weighted by these proportions. Thus, the average frequency can be written as $x = \tfrac{1}{3}p_1 + \tfrac{2}{3}q_1 = \tfrac{1}{3}(p_1 + 2q_1)$. Similarly, the average in the next generation can be written as $x' = \tfrac{1}{3}(p_1' + 2q_1')$. Substituting the values for p_1' and q_1' into this expression yields

$$x' = \tfrac{1}{3}\left[q_1 + 2\left(\tfrac{1}{2}(p_1 + q_1)\right)\right]$$
$$= \tfrac{1}{3}(q_1 + p_1 + q_1) = \tfrac{1}{3}(p_1 + 2q_1) = x.$$

The allele frequency in the population, therefore, does not change

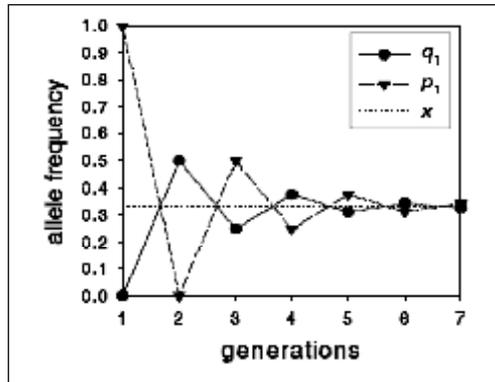over generations. What happens is that the male and female frequencies oscillate around the overall frequency (because of $y' = -\frac{1}{2}y$), eventually converging to it (*Figure* 3).

Thus, equilibrium for sex-specific allele frequencies is not reached for several generations in the case of X-linked loci, unlike in the case of autosomal loci, where neither overall allele frequencies nor sex-specific allele frequencies change in an ideal large population. I hope that this discussion has also served to highlight the fact that sometimes algebra can lead us to conclusions that would not, perhaps, be immediately intuitively obvious.

Extending the Hardy–Weinberg Law: Two Biallelic Loci

As we briefly discussed in Part I, genetic scenarios rapidly become complex and cumbersome as one moves beyond the simple case of one biallelic locus. If we consider two biallelic loci, there is a whole range of complications with which we must deal. We need to consider the possibility of linkage and crossing over while making the transition from diploid genotypes to haploid gametes. Moreover, we have to develop a notion of what we mean by genetic equilibrium for two loci. We start by describing the genotypic composition of a population at two loci, A and B, each with two alleles subscripted as 1 and 2, respectively. There are now ten possible genotypes, which are listed below:

$$\frac{A_1B_1}{A_1B_1}, \quad \frac{A_1B_1}{A_1B_2}, \quad \frac{A_1B_1}{A_2B_1}, \quad \frac{A_1B_1}{A_2B_2}, \quad \frac{A_1B_2}{A_1B_2},$$

$$\frac{A_1B_2}{A_2B_1}, \quad \frac{A_1B_2}{A_2B_2}, \quad \frac{A_2B_1}{A_2B_1}, \quad \frac{A_2B_1}{A_2B_2}, \quad \frac{A_2B_2}{A_2B_2}.$$

There are two things to note about this way of depicting genotpes[3], which could also be written as A_iB_j/A_kB_l $(i, j, k, l = 1, 2)$.

Consider now the gametes that these various individuals can produce. The four possible gametic types are A_1B_1, A_1B_2, A_2B_1, and A_2B_2. Note also that these are not only the four possible kinds of gametes but also the four possible kinds of haploid biallelic genomes that individuals, can get from their parents. The gametic types are, thus, analogous to alleles in the one-locus formulation. We will designate the frequency of the diploid genotype A_iB_b/A_kB_l as $P_{ij/kl}$, and the frequency of the gametic or haploid genotype A_iB_j as p_{ij}. We further designate the allele frequencies of the i-th and j-th alleles at the A and B loci as p_i and q_j, respectively. We can now establish a criterion for genetic equilibrium at this two-locus system in a manner analogous to that for the Hardy–Weinberg Equilibrium[4] in the one-locus case. For the two-locus system, the criterion for between-locus equilbrum is that the coming together of a specific A allele and a specific B allele to form a two-locus haploid genome should be a random event. The gametic frequency p_{ij} should, therefore, be the product of the relevant allele frequencies p_i and q_j. If the equality $p_{ij} = p_iq_j$ is satisfied for all i and j, the system is said to be in gametic phase equilibrium (also called 'linkage equilibrium') at these two loci A and B. If, on the other hand, $p_{ij} \neq p_iq_j$, then the deviation from the equilibrium expectation $D_{ij} = p_{ij} - p_iq_j$ can be used as a measure of disequilibrium, and we can say that gametic phase equilibrium would exist if $D_{ij} = 0$ for all i, j.

We will again take recourse to our notion of the idealized large population to explore how gametic frequencies change from one

generation to the next when we are simultaneously considering two biallelic loci. Once two loci are under consideration, we have to consider the possibility of crossing over between them[6]. For example, if the frequency of crossing over between the two loci is 0.3 (30% of all gametes produced result from a crossing over event), an individual of genotype A_1B_1/A_2B_2 will produce non-recombinant gametes bearing haploid genotypes A_1B_1 and A_2B_2, corresponding to the haploid genomes carried by gametes in the previous generation that fused to give rise to that individual. These non-recombinant gametes will together constitute 0.7 of all gametes produced by this individual; the remaining 0.3 of the gametes will be, in equal proportions, A_1B_2 and A_2B_1, representing gametes resulting from a crossing over between the two loci. One way of working out the recursion equation linking gametic frequencies in subsequent generations (p'_{ij} and p_{ij}) would be to list all genotypes and their frequencies, work out the proportions of all kinds of gametes produced by each genotype, and then sum up the frequencies of the various gametes across parental genotypes to get the new gametic frequencies. This would be the equivalent of working out the one-locus case through mating frequencies (*Table* 2). We will, instead, use a simpler argument, based once again on an idealized version of what really happens, which can be shown to yield the same result as the more cumbersome method of going through the individual genotypes and mating frequencies[7].

Imagine we are dealing with two biallelic loci with a recombination frequency of c $(0 \leq c \leq 0.5)$ between them. Let us focus on the haploid genomes present at any generation $t + 1$. Consider, for example, the haploid genome A_2B_2. All these haploid genomes present at generation $t + 1$ would definitely have arisen in one of two ways: either they are copies of A_2B_2 genomes that were present in generation t and did not undergo crossing over during meiosis, or they are products of recombination. Now imagine a 'haploid genome pool' which consists of all the haploid genomes present in the diploid individuals of generation t, and let the gametic (haploid genome) frequencies be p_{ij}. Now imagine

[6] Crossing over, of course, takes place during meiosis when gametes are being formed, and each individual can, in principle, produce both recombinant (resulting from cross-overs) and non-recombinant gametes.

[7] I strongly encourage you to actually work this out the long way to confirm for yourself that you get the same result as that obtained by the more elegant simplification described below. When I first studied this topic, I could not believe this simplification would actually work and the understanding gained by going through all the crossovers and matings was well worth the time spent on it.

dividing this pool into two parts: a recombinant part which will contain a fraction c of all the haploid genomes, and a non-recombinant part containing the remaining $1 - c$ of all haploid genomes. Individual haploid genomes are assigned to these two parts of the pool at random. Now, the haploid genomes from the non-recombinant part will become gametes as they are, without undergoing any crossing over. Since allotment of haploid genomes to the two parts of the pool was random, the frequencies p_{ij} still hold in both parts of the pool. Therefore, a fraction $1 - c$ of gametes (and, thus, haploid genomes in the generation $t + 1$) will continue to have the same frequencies p_{ij}.

Now consider the recombinant part of the pool. To represent the process of crossing over, we assume that all the two-locus haploid genomes present here are broken into constituent one locus parts, containing a single A allele or B allele. Gametes are then produced by putting together pairs of A and B alleles at random within the recombinant part of the pool. The frequency of gametes containing the alleles A_i and B_j, formed in this recombinant part of the pool will, therefore, be the product of the respective allele frequencies, $p_i q_j$. Therefore, the remaining fraction c of gametes will continue frequencies $p_i q_j$. Putting these two results from the recombinant and non-recombinant parts of the haploid genome pool together, we can construct the recursion for gametic frequencies between subsequent generation as

$$p'_{ij} = (1 - c)p_{ij} + cp_i q_j$$

Note that allele frequencies at the two loci will not change from one generation to the next. In fact, each locus itself will follow the Hardy–Weinberg Equilibrium expectations. What we would like to investigate, using this recursion, is the behaviour of the between locus disequilibrium.

Recall the measure of disequilibrium, $D_{ij} = p_{ij} - p_i q_j$. We will now derive a recursion for this measure of gametic phase disequilibrium. Let the disequilibrium in generation t be D_{ij}, and in generation $t + 1$ be D'_{ij} .

$$D'_{ij} = p'_{ij} - p'_i \; q'_j$$

but, since allele frequencies do not change,

$$= p'_{ij} - p_i q_j$$

$$= (1-c)p_{ij} + cp_i q_j - p_i q_j = (1-c)p_{ij} - (1-c)p_i q_j$$

$$= (1-c)D_{ij} .$$

In other words, in each generation the disequilibrium is reduced by a factor of c, the recombination frequency. In general, if disequilibrium at an arbitrary starting generation 0 is $D_{ij}(0)$, the disequilibrium after an arbitrary number of generations t, will be given by
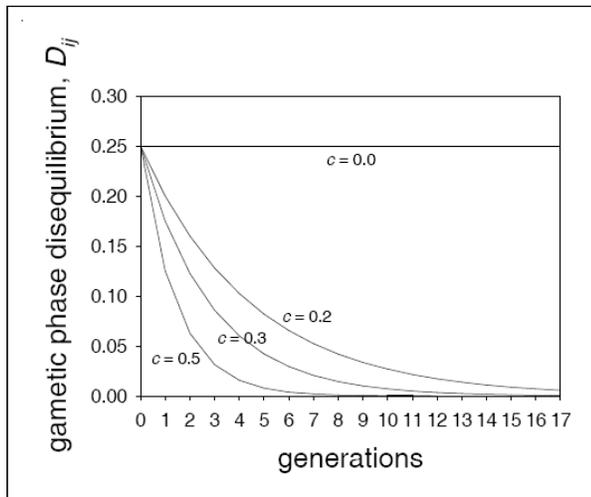
$$D_{ij}(t) = (1-c)^t D_{ij}(0)$$

It is clear that in the presence of recombination, disequilibrium will decay with time and the rate of this decay will depend upon the recombination frequency (*Figure* 4). For loci that are un-linked ($c = 0.5$), gametic phase disequilibrium will be halved each generation. There are many further refinements and technicalities to the issue of gametic phase disequilibrium, such as devising measures of disequilibrium that would be independent of allele frequencies so that comparisons could be made across popula-tions, or extending the arguments described here to situations with more than two alleles/loci.

> In the presence of recombination, disequilibrium will decay with time and the rate of this decay will depend upon the recombination frequency.

Figure 4. The decay of ga-metic phase disequlibrium over time at different re-combination frequencies 'c'. $D_{ij}(0)$ has been chosen as 0.25 which is the maxi-mal magnitude of disequi-librium possible in a sys-tem of two biallelic loci.

Implications of the Hardy–Weinberg Law

We have seen that under a set of assump-tions, which together describe the ideal large population, the allelic composition of a population does not change from one generation to the next. If we consider genotypic composition, equilibrium is at-tained in one generation of random mat-ing if we consider only one autosomal locus. For X-linked loci, equilibrium for

Suggested Reading

[1]   G H Hardy, *Mendelian proportions in a mixed population, Science,* Vol.28, pp.49–50, 1908.

[2]   D L Hartl and A G Clark, *Principles of Population Genetics,* 2nd Edition, pp.45–57, 1989. Sinauer, Sunderland. (pp.95–106 in the 3rd Edition, Sinauer, Sunderland, 1997).

Address for Correspondence
Amitabh Joshi
Evolutionary and Organismal Biology
Jawaharlal Nehru Centre for Advanced Scientific Research
Jakkur P.O.
Bangalore 560 064, India.
Email:ajoshi@jncasr.ac.in

allele and genotype frequencies in the males and females may take several generations to attain. This is also the case for two loci, especially if they are linked. But in all these cases, equilibrium, or a point so close to equilibrium that for all practical purposes genetic composition can be assumed to be unchanging, is attained within a relatively small number (< 15) generations. This tells us that the process of sexual reproduction in diploid organisms is inherently a conservative one and, by itself, tends to maintain whatever genetic variation is present in the population. Thus, the ideal large population allows us to describe the inertial state in population genetics. Unless an evolutionary force (such as mutation, migration or selection) is acting on it, a large population, buffered by its size from random fluctuations, tends to maintain its genetic composition unchanged over generations.

In fact, though the assumptions underlying the ideal large population may seem very restrictive and unrealistic, quite a few populations do appear to be in Hardy–Weinberg Equilibrium with regard to many loci. Recall that the requirement for Hardy–Weinberg Equilibrium to hold at a locus is only that there should be random mating and no net mutation, migration or selection *with regard to the genotypes at that locus*. Even more important than the fact that many populations do appear to be in Hardy–Weinberg Equilibrium for several loci, however, is the fact that the ideal large population gives us a good framework within which we can analyse the consequences of relaxing particular assumptions we made. Another useful point to note is that the framework we have developed here tells us that as long as there is random mating, genotypic frequencies in the zygotes will be the product of allelic (or haploid genomes for more than one locus) frequencies in the gametes. This relationship is often of practical value because it allows us to cast and follow what is happening genetically at the level of haploid frequencies, which means there are less variables to keep track of. In all these senses, the Hardy–Weinberg principle, and the model of genetics of a population that it embodies, truly forms the basic foundation of our understanding of evolutionary genetics.