1 The Theoretical Principles of Population Genetics

Before turning to an analysis of genetic processes in populations, we will review in general terms the main theoretical principles of population genetics. These are of importance for understanding and evaluating the contents of this and following chapters. Population genetics, perhaps the most theoretically advanced field of research, occupies a special place in modern genetics and biology as a whole.

In population genetics, quantitative theory has existed for a long time and continues to improve. It is a theory involving natural factors, under pressure of which a population either remains constant or changes in successive generations with a concomitant change in biologically important traits. In other words, population genetics has mathematical models. The models may or may not correspond closely to nature, but they are important since they permit one to plan research in a certain way. Furthermore, if we observe the conformity with the natural situation of a model, it becomes possible to evaluate numerically the changes in populations and to predict the possible consequences. Because a large number of diversified papers and books have been published on this theme in recent years, we shall examine only the main population genetic terms, models, and approaches. We base our further discussion on works by Sewall Wright (1931, 1951, etc.), Neel and Schull (1958), Ehrlich and Holm (1963), Dobzhansky (1970), Kimura (1983), Kimura and Ohta (1971), Cavalli–Sforza and Bodmer (1971), Nei (1975, 1987), Li (1976), and several others to be mentioned in the text.

1.1 Estimation of Gene Frequencies

To a first approximation, a population may be defined (as did Dobzhansky) as *an aggregate of freely interbreeding individuals that share a common gene pool.* Because the number of segregating loci in the genome is large, one can understand the difficulties that confront the researcher in attempting to give an adequate description of this pool of inherited information. But at the same time it is evident that, however great these difficulties may be, there is only one way of obtaining such a description: by defining the frequencies of allelic genes at each single locus. Knowledge of the spatial and

temporal distribution of gene frequencies enable a quantitative assessment of how genetic processes in populations are influenced by given external and internal factors.

There are several carefully formulated methods for evaluating gene frequencies in populations. We shall examine two of them, the first applied to a situation without dominance and the second to inheritance with dominance. The situation relates to a pair of alleles at a single autosomal locus.

Absence of Dominance. Let us assume that of N diploid individuals N_1, carrying only the allele A, that is, are AA; N_2 carrying the heterozygotes AB; and N_3 carrying the homozygotes BB, so that $N_1 + N_2 + N_3 = N$, and the total number of genes is $2N$. Each AA homozygote has two A genes, and each AB heterozygote has only one gene of this kind. Consequently, the total number of A genes in the group under study equals $2N_1 + N_2$ and the fraction (frequency) of this gene is

$$pA = \frac{2N_1 + N_2}{2N} = \frac{N_1 + \left(\frac{1}{2}\right) N_2}{N} .$$

The frequency of gene B is defined in exactly the same way:

$$qB = \frac{2N_3 + N_2}{2N} = \frac{N_3 + \left(\frac{1}{2}\right) N_2}{N} ,$$

so that $p + q = 1$. We can also apply the same method to loci with more than two alleles.

Of course, a complete examination of natural populations is generally impossible, which in practical terms means that sampling is necessary. Hence, the reliability of estimates of gene frequency depends very much on the numbers sampled. Such estimates should be characterized by the least possible error or dispersion factor, that is, they should satisfy the so-called *criterion of effectiveness*. Thus, the requisite sample size depends on the genetic population structure, which can be established by preliminary research. The most dependable results come from the method of "directly calculating" genes, which was formulated by Fisher and which we have used in the above example. However, if there is **dominance** of one allele (A) over the other (a), then only two distinguishable phenotypes are present in a population, and of them only one phenotype – the homozygote (aa) for the recessive allele – corresponds to only one genotype.

The method of directly determining the allelic frequency is inapplicable to a genetic situation of this kind, and one must allow the hypothesis that the Hardy–Weinberg equation (see the next section) holds in the population; namely, that the distribution of genotypes in random mating conforms to the coefficients of the binomial expansion $p^2 + 2pq + p^2 = 1$.

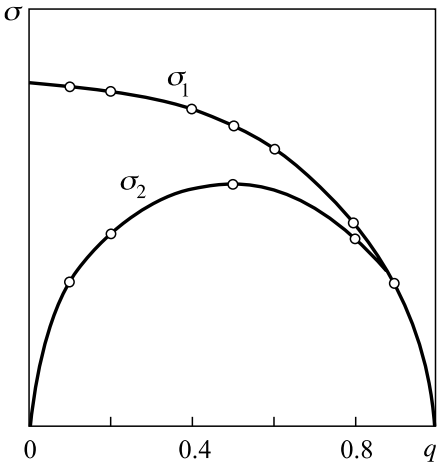

Fig. 1.1. Distributions of the standard errors of the two values of a gene frequency for a single pair of alleles without dominance (based on Neel and Schull 1958). $\sigma_1 = \sqrt{(1-q^2)/4N}$ (the gene frequency found by extracting the square root of the fraction of homozygous genotype); $\sigma_2 = \sqrt{q(1-q)/2N}$ (the same allele frequency determined by the direct calculation method)

It follows that in order to obtain an effective estimate of the frequency (q) of a recessive gene (a) we should extract the square root of the fraction of aa individuals in our sampling: $q = \sqrt{aa}$. Accordingly, $pA = 1 - q$. In principle, this method could also be used to estimate allelic frequency in the absence of dominance, but it only gives least biased estimates at high frequency values (Fig. 1.1).

The reader can find more general examples of estimating gene frequencies (for example, for multiple alleles or linked loci) in Li's book (1978) and other publication on theoretical population genetics (e.g., Zhivotovsky 1991; Weir 1995; Hedrick 1999).

A number of software packages used in population genetics for estimating gene frequencies, intra- and intergroup components of gene diversity, genetic distances, for constructing phylograms, etc., can be found on the Internet at the following addresses:

Software package	Site address
Analysis of Molecular Variance – Amova	http://anthropologie.unique.ch/ftp/com/win/amova/
Arlequin	http://anthropologie.unige.ch/arlequin
Genetic Data Analysis (GDA)	http://lewis.eeb.uconn.edu/lewishome/software.html
GENEPOP	ftp://ftp.cefe.cnrs-mop.fr/PC/MSDOS/GENEPOP/
Molecular Evolutionary Genetics Analysis (MEGA)	http://www/megasoftware.net/

Software package	Site address
Phylogeny Inference Package (PHYLIP)	http://evolution.genetics.washington.edu/phylip.html
POPGENE	http://www.ualberta.ca/$\sim$fyeh/
FSTAT	http://www.unil.ch/izea/softwares/fstat.html
Genetic Analysis in Excel (GenAlEx V5)	http://www.anu.edu.au/BoZo/GenAlEx/
Multilocus Mating System Program (MLTR)	http://genetics.forestry.ubc.ca/ritland/programs.html
Structure	http://pritch.bsd.uchicago.edu/
PCAGEN	http://www2.unil.ch/popgen/softwares/pcagen.htm

1.2 The Hardy–Weinberg Rule

The field of population genetics examines the principles regulating the maintenance and dynamics of population genotypic structure in time and space. The Hardy–Weinberg principle provides a theoretical basis for this view. It reflects the invariability of the genetic composition of a randomly mating (panmictic) population, unlimited in number and existing extra-environmentally, (i.e., in the absence of environmental pressure). In this structureless community, the genotype frequencies and, hence, also the gene frequencies at an autosomal locus having a pair of alleles A and a, reach equilibrium in the generation that follows random mating.

Inasmuch as random matings signify merely a random association of gametes, it can easily be verified that this combination of $p(A)$ and $q(a)$ male gametes and $p(A)$ and $q(A)$ female gametes, when $p + q = 1$, produces an invariable distribution of $p^2AA + 2pqAa + q^2aa = 1$.

		Male gametes	
		$p(A)$	$q(a)$
Female gametes	$p(A)$	p^2	pq
	$q(a)$	pq	q^2

The algebraic calculations supporting the Hardy–Weinberg equilibrium are given in Table 1.1. As we are concerned here with autosomal genes, the reciprocal crosses (that is of the type *male Aa*×*female AA* or *male AA*×*female Aa*, etc.) may be combined, and consequently the nine possible crosses reduce to six, given that $p + q = 1$.

It is clear that this equilibrium ratio of genotypes is provided by the symmetry of the distribution of allelic genes into male and female gametes and by the free combination of these into the zygotes formed in the process

Table 1.1. Types of matings and proportions of genotypes of population progeny at genetic equilibrium

Type of mating	Frequency of mating	Proportions of genotypes among progeny		
		AA	*Aa*	*aa*
$AA \times AA \quad (p^2 \times p^2)$	p^4	p^4	0	0
$AA \times Aa \quad 2(p^2 \times 2pq)$	$4p^3q$	$2p^3q$	$2p^3q$	0
$Aa \times Aa \quad (2pp \times 2pq)$	$4p^2q^2$	$2p^2q^2$	$2p^2q^2$	p^2q^2
$AA \times aa \quad 2(p^2 \times q^2)$	$2p^2q^2$	0	$2p^2q^2$	0
$Aa \times aa \quad 2(2pq \times q^2)$	$4pq^3$	0	$2pq^3$	$2pq^3$
$aa \times aa \quad (q^2 \times q^2)$	q^4	0	0	q^4
Totals for population	1,00	p^2	$2pq$	q^2

of reproduction. From this it follows that when there are no disturbances affecting a population of unlimited numbers, the frequencies of the genotypes and genes that characterize it remain unchanged in an infinitely long series of generations – the so-called absolute zero of genetic dynamics.

In Table 1.2 an elementary example is presented of testing an empirical distribution of genotypes at a diallele locus of the MN blood type (codominant expression) for goodness-of-fit to Hardy–Weinberg proportions in a British population. The gene frequencies are $pM = 0.542$ and $qN = 458$. These data show virtually ideal goodness-of-fit of the observed genotype distribution to that expected: the total chi-square value at one degree of freedom (the number of genotypes minus the number of allelic genes) is as low as 0.22, which is much lower than the threshold level of 3.84.

But ideal populations of this kind are virtually never encountered in nature; there are always natural factors that shift them from the point of equilibrium, disturbing their stability – random genetic drift, mutations, migration, and natural selection. These are the "factors of evolution" or microevolutionary forces that we shall now examine.

Table 1.2. Empirical distribution of genotypes at a diallele locus of the MN blood type (codominant expression) for goodness-of-fit to Hardy–Weinberg proportions in a British population

Genotype	Observed number (obs)	Expected number (exp)	$\chi_i^2 = \frac{(\text{obs}-\text{exp})^2}{\text{exp}}$
MM	298	$p^2n = 294.3$	0.05
MN	489	$2pqn = 496.4$	0.11
NN	213	$q^2n = 209.3$	0.06
Total n	1,000	1,000	$\sum \chi_i^2 = 0.22$

The gene frequencies are $pM = 0.542$ and $qN = 458$

http://www.springer.com/978-3-540-25490-4

Intraspecific Genetic Diversity
Monitoring, Conservation, and Management
Altukhov, Y.P.
2006, XV, 438 p., Hardcover
ISBN: 978-3-540-25490-4