Li, 1961

Chapter 4

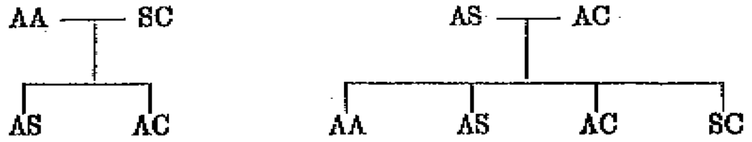
Multiple Alleles and Blood Types

The number of alleles at any given locus is by no means limited to two; there could be three, four, or more and then they are called *multiple* alleles. In fact, the MN locus, the Rh locus, the sickle hemoglobin locus, all of which have been mentioned in previous chapters, involve multiple alleles. In this chapter, we shall extend the previous family and population laws and methods of analysis to cases of multiple alleles.

The Sickle Hemoglobin Locus

The number of alleles at this locus is still a matter for study, but it has been shown definitely that there are at least three. In addition to the normal adult hemoglobin (A) and the sickle hemoglobin (S), there is a third, known as hemoglobin type C and also considered abnormal. A formal notation for the series of three alleles is Hb^A, Hb^S, Hb^C, for good reasons. For our purpose here, we shall informally write only the superscripts A, S, C, with the understanding that they are the alleles of the Hb locus.

One of the main difficulties in studying multiple alleles in man is to demonstrate that they are actually alleles. Not all populations have all three alleles. Allele S may be abundant in an African population that lacks C completely. A South Asian population may have C but no S. To prove allelism, all three must be present in the same family which segregates. The allelism of A, S, C is proved by the rare occurrence of the following types of families:



Proof of allelism is time-consuming when the allele in question is rare in the population, because we just have to wait for the right family with a fairly large number of children. In the hemoglobin case, it is fortunate that all six genotypes (AA, AS, SS, AC, SC, CC) can be distinguished by filter-paper electrophoresis (see, e.g., Ranney, 1954). Dominance greatly complicates the task of proving allelism.

For a population in which all three alleles are present, the genotypic proportions are given by an obvious extension of the Hardy-Weinberg law. Let p, q, r denote the frequencies of the alleles A, S, C, respectively. Then the six genotypic proportions are the terms of $(p + q + r)^2$. For example, if $p = .95$, $q = .04$, $r = .01$, the population will consist of:

Genotype:	AA	AS	AC	SS	SC	CC	Total
Proportion:	p^2	$2pq$	$2pr$	q^2	$2qr$	r^2	1.00
Example:	.9025	.0760	.0190	.0016	.0008	.0001	1.00

This example represents roughly the situation in American Negroes. When there is no dominance, the estimation of the gene frequencies follows the same procedure as that for the MN system, described in Chap. 3. Let n_{AA} be the observed number of AA individuals, etc., and G the total number in a random sample from the population. The estimate of the frequency of the A allele and its variance are:

$$p = \frac{2n_{AA} + n_{AS} + n_{AC}}{2G} \qquad V(p) = \frac{p(1 - p)}{2G}$$

The estimates of q and r take the same form. The formula above applies to all cases without dominance.

The frequency of the various types of mating is calculated in the manner shown in Chap. 2. For instance,

$$\begin{aligned} \text{AS} \times \text{AC:}\quad & \text{freq} = 2(2pq)(2pr) = 2(.076)(.019) && = .002{,}888 \\ \text{AA} \times \text{SC:}\quad & \text{freq} = 2(p^2)(2qr) = 2(.9025)(.0008) && = .001{,}444 \end{aligned}$$

The former type of mating is *always* twice as frequent as the latter, whatever the gene frequency. In our particular example, the total frequency of these two types of families is only about four per thousand.

Suppose that a number of AA × SC families have been observed and their total offspring consist in a individuals with S hemoglobin and b with C hemoglobin. We may wish to test whether these numbers conform with the expected 1:1 ratio. The chi square for this test is, writing $a + b = n$,

$$\chi^2 = \frac{(a - \frac{1}{2}n)^2}{\frac{1}{2}n} + \frac{(b - \frac{1}{2}n)^2}{\frac{1}{2}n} = \frac{(a - b)^2}{n}$$

with one degree of freedom. Similarly, if a number of AS × AC families have been observed and their total offspring consist in a individuals with A hemoglobin (AA or AS or AC) and b without (i.e., SC), the chi square for testing the expected 3:1 ratio is:

$$\chi^2 = \frac{(a - 3b)^2}{3n}$$

with one degree of freedom. These tests for Mendelian ratios are possible only when the genotypes of both parents are known.

The ABO Blood Groups

This is probably the best-known series of multiple alleles in man, and hence we shall not describe the antigen-antibody relationship here because it has been so admirably explained and illustrated in most textbooks of general genetics and human genetics. The ABO locus has also three (major) alleles, a formal notation for which is I^A, I^B, I^O. Again, for the sake of simplicity, we shall only write the superscripts A, B, O, with the locus identification symbol I understood. Now the reader may appreciate why the formal notation is necessary: the A of the ABO series has nothing to do with the A of the ASC series.

The ABO locus is more susceptible of gene frequency analysis on the population level than the hemoglobin locus, because all three alleles are usually present in all populations and none of them is too rare. There are exceptions; for instance, some American Indian and Australian populations have only A and O but no B. There is, however, a complication. Whereas all the six genotypes of the hemoglobin series are distinguishable, this is not the case with ABO. In the latter series, both A and B are dominant over O, but AB individuals have antigen A as well as antigen B (*codominant*). Consequently, there are only four distinguishable phenotypes (i.e., blood groups). To emphasize the symmetrical nature of the four blood groups, the six genotypes can be arranged in a triangular pattern, as shown in the left of Fig. 4-1, in which the three homozygotes are at the bottom and each heterozygote is at the vertex of a triangle with the two corresponding homozygotes at its base. To have a quantitative representation of the genotypic proportions, an equilateral triangle may be drawn with the base subdivided into three segments proportional in length to $p:r:q$. From such points of segmentation, lines parallel to both sides are drawn. The resulting areas of the three equilateral triangles on the base represent the frequencies of the three homozygotes (p^2,r^2,q^2), and the three parallelograms represent the three heterozygotes ($2pr$,$2pq$,$2qr$). The combined area of AA and AO represents the

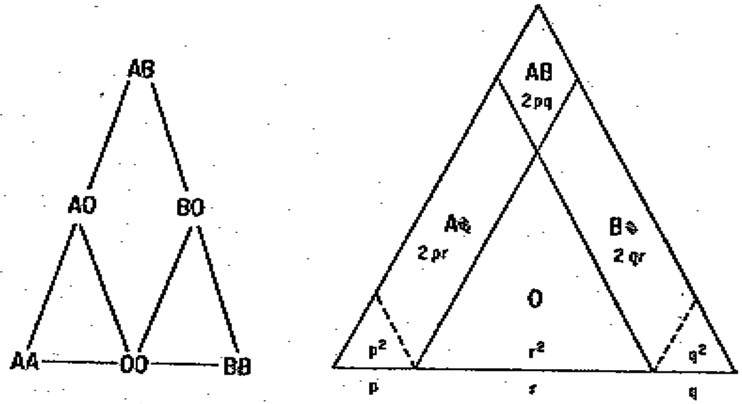


FIG. 4-1. Diagrammatic representation of the ABO system. Left: arrangement of the six genotypes. Right: the gene frequencies are represented by the segments of the base, and the genotypic proportions are represented by the areas of triangles and parallelograms.

frequency of group A and, similarly, BB and BO represent group B.

For ease of drawing, it is assumed in Fig. 4-1 that $p = .20$, $r = .60$, $q = .20$, roughly representing the situation of an Asian population with comparatively high B frequency. In a Western European population the gene frequencies are roughly $p = .30$, $r = .60$, $q = .10$. The difference in the values of p and q is more pronounced than that in r. In either case, a glance at the triangular diagram shows that most of the individuals of groups A and B are heterozygotes on account of the high frequency of allele O.

The symmetrical nature of the four blood groups immediately suggests a method of estimating the gene frequencies from observed number of individuals. Since the area $A = p^2 + 2pr$ in the diagram represents the proportion of group A individuals, and area $B = q^2 + 2qr$ represents that of group B individuals, etc., we may write:

$$\begin{array}{lll} B + O = (q + r)^2 & \sqrt{B + O} = q + r & 1 - \sqrt{B + O} = p \\ A + O = (p + r)^2 & \sqrt{A + O} = p + r & 1 - \sqrt{A + O} = q \\ O = r^2 & \sqrt{O} = r & \sqrt{O} = r \end{array}$$

Numerical estimates are based on these theoretical relations. As an example, the data of an Iowa study (Buckwalter et al., 1956) may be cited:

AB, 61	A, 672	B, 147	O, 743	Total, 1,623

It is desirable to perform the arithmetic systematically according to the algebraic expressions shown above. The details are given in Table 4-1. The first five columns of Table 4-1 need no explanation, except to note that the sum of the first estimates is 1.0024 instead of unity. Apparently each estimate should be diminished slightly so they would add up to unity. Let $d = .0024$, the deviation from unity. The adjusting factor for p' and q' is

TABLE 4-1. THE ESTIMATION OF ABO GENE FREQUENCIES FROM 1,623 INDIVIDUALS IN IOWA

Blood group	*Observed number*	*Proportion*	*$\sqrt{Proportion}$*	*First estimate*	*Adjusting factor*	*Adjusted estimate*	*Standard error*
B + O	890	0.5484	0.7405	$.2595 = p'$	.9988	$.2592 = p$	.0083
A + O	1,415	.8718	.9337	$.0663 = q'$	.9988	$.0662 = q$	.0044
O	743	.4578	.6766	$.6766 = r'$	See text	$.6746 = r$	.0089
Total..	1,623	......	......	1.0024	....	1.0000	

$$1 - \tfrac{1}{2}d = 1 - .0012 = .9988$$

so the adjusted estimate

$$p = p'(1 - \tfrac{1}{2}d) = .2595 \times .9988 = .2592$$

etc. The adjustment for r' is a little different, being:

$$r = (r' - \tfrac{1}{2}d)(1 - \tfrac{1}{2}d) = (.6754)(.9988) = .6746$$

The adjusted estimates now add up to unity. If they do not, a second round of adjustment may be performed but usually this is unnecessary. The adjustment method is due to Bernstein (1930). For all practical purposes, the adjusted estimates may be accepted as the maximum likelihood solutions (Stevens, 1938).

Let us recall that, if allele A were dominant over all other alleles, the variance of the estimate of its frequency would be:

$$V(p) = \frac{p(1-p)}{2G} + \frac{p^2}{4G} = \frac{(.2592)(.7408)}{2(1623)} + \frac{(.2592)^2}{4(1623)} \quad 1.03$$
$$= 69.50 \times 10^{-6}$$

The first component, $p(1 - p)/2G$, represents the variance when there is no dominance and the number of alleles in a sample is known; therefore, it is the minimum variance of any sample estimate. Now, the A in the ABO system is like a complete dominant allele most of the time, except in genotype AB, which consists of 3.76 per cent of the individuals in our sample. The variance of p (maximum likelihood solution) is slightly smaller than the one indicated above, being (DeGroot, 1956; Li, 1956b):

$$V(p) = \frac{p(1-p)}{2G} + \frac{p^2}{8G}\left(1 + \frac{r}{pq + r}\right)$$
$$V(q) = \frac{q(1-q)}{2G} + \frac{q^2}{8G}\left(1 + \frac{r}{pq + r}\right)$$

The factor $r/(pq + r)$ is equal to 0.9752 in our example, so $V(p) = 69.37 \times 10^{-6}$ and $V(q) = 19.71 \times 10^{-6}$. In most human populations the value of $r/(pq + r)$ is larger than .90 and in many cases larger than .95, so the approximate and exact expressions for variance differ only very slightly, as illustrated by our example. The variance of r is:

$$V(r) = \frac{r(1-r)}{2G} + \frac{(1-r)^2}{8G} + \frac{r(p-q)^2}{8G(pq+r)} = 78.58 \times 10^{-6}$$

The square root of the variance is given in the last column of Table 4-1.

The frequencies of various types of mating and their offspring, of mother-child combinations, and of sib-sib pairs have all been given in detail elsewhere (Li, 1955a, Chap. 4). The traditional method of arriving at those results is long and tedious, but they can be achieved at one step by the *ITO* method explained in Chap. 3.

Subgroups of A

The major allele A consists of two alleles A_1 and A_2, the former being dominant to the latter. Let p_1 and p_2 denote the frequency of A_1 and A_2, so that $p = p_1 + p_2$ is the frequency of A. As an exercise, the reader may arrange the ten genotypes into a triangular pattern with the four homozygotes (A_1A_1, A_2A_2, OO, BB) at the base and draw a corresponding equilateral triangle to represent the genotypic proportions in exactly the same manner as shown in Fig. 4-1. He will then find that the six phenotypes have the following frequencies:

A_1B $2p_1q$	A_2B $2p_2q$	B (BB + BO) $q^2 + 2qr$	
A_1 ($A_1A_1 + A_1A_2 + A_1O$) $p_1^2 + 2p_1p_2 + 2p_1r$	A_2 ($A_2A_2 + A_2O$) $p_2^2 + 2p_2r$	O (OO) r^2	$(p + r)^2$

$$\underbrace{p_2^2 + 2p_2(q + r) \qquad (q + r)^2}_{(p_2 + q + r)^2}$$

Note that B and O groups remain the same; AB and A are each split into two subgroups:

$$\begin{aligned} AB &= A_1B + A_2B = 2p_1q + 2p_2q = 2pq \\ A &= A_1 \;\;\; + A_2 \;\; = (p_1 + p_2)^2 + 2(p_1 + p_2)r = p^2 + 2pr \end{aligned}$$

The symmetrical nature still remains; hence the method of estimating the gene frequencies follows the same general procedure as in the case of three alleles. Ignoring the subgroups, we may estimate p; and pooling $A_2B + A_2$ and $B + O$, we may estimate p_1. Then p_2 is obtained by subtraction. In brief, the estimation is based on the following theoretical relations (Mourant, 1954, p. 219):

$$\left.\begin{aligned} p_1 &= 1 - \sqrt{A_2B + A_2 + B + O} \\ p_2 &= \sqrt{A_2B + A_2 + B + O} - \sqrt{B + O} \end{aligned}\right\} \quad \begin{aligned} p &= 1 - \sqrt{B + O} \\ q &= 1 - \sqrt{A + O} \\ r &= \sqrt{O} \end{aligned}$$

Table 4-2 gives a numerical example. The first five columns (up to first estimates) are done according to the algebraic expressions above; for instance, $p_2' = .7931 - .7401 = .0530$. The sum of these first estimates is .9930; its deviation from unity is $d = .9930 - 1 = -.0070$. Each estimate should be increased

TABLE 4-2. THE ESTIMATION OF A_1A_2BO GENE FREQUENCIES

Sample	A_1B,	4	A_2B	2	B	29	35
	A_1	124	A_2	26	O	160	310
		128		28		189	345

Group	Number	Prop.	$\sqrt{Prop.}$	First estimate	Adj. factor	Adj. estimate	Standard error
$A_2B + A_2 + B + O$	217	.6290	.7931	$.2069 = p_1'$	1.0035	$.2076 = p_1$	.0164
$B + O$	189	.5478	.7401	$.0530 = p_2'$	1.0035	$.0532 = p_2$	.0097
$A + O$	310	.8985	.9479	$.0521 = q'$	1.0035	$.0523 = q$	.0086
O	160	.4638	.6810	$.6810 = r'$	......	$.6869 = r$	.0190
Total.............	345	.....	.....	.9930	......	1.0000	

slightly so they will add up to unity. The method of adjustment is the same as before; thus,

$$p_1 = p_1'(1 - \tfrac{1}{2}d) = .2069 \times 1.0035 = .2076$$

etc., and:

$$r = (r' - \tfrac{1}{2}d)(1 - \tfrac{1}{2}d) = .6845 \times 1.0035 = .6869$$

The adjusted estimates now add up to unity and for all practical purposes may be regarded as the maximum likelihood solutions. The variance of p_1 is obtained by substituting p_1 for p in the previous formula:

$$V(p_1) = \frac{p_1(1 - p_1)}{2G} + \frac{p_1^2}{8G}\left(1 + \frac{r}{p_1q + r}\right)$$

while the expressions for $V(q)$ and $V(r)$ remain the same as before with $p = p_1 + p_2$. The only problem is $V(p_2)$, because allele A_2 is recessive to A_1 but behaves like the ordinary A of the three-allele case to B and O. The author finds the following expression quite satisfactory:

$$V(p_2) = \frac{p_2(1 - p_2)}{2G(1 - p_1)} + \frac{p_2^2}{8G(1 - p_1)^2}\left(1 + \frac{r}{p_2q + r}\right)$$

It is the same formula for $V(p)$, except that $2G$ is multiplied by $(1 - p_1)$ in the first component and $8G$ is multiplied by $(1 - p_1)^2$ in the second component. The numerical values are as follows, and the reader is urged to perform the arithmetic himself.

$$V(p_1) = 2.694 \times 10^{-4} \qquad V(p_2) = 0.954 \times 10^{-4}$$
$$V(q) = 0.738 \times 10^{-4} \qquad V(r) = 3.627 \times 10^{-4}$$

The standard errors are given in the last column of Table 4-2. The method of adjustment and the variance formula for the A_1A_2BO system described in this section do not seem to have been fully investigated by biometricians. The estimates and their variances obtained for this example are identical (within the limit of rounding-off errors) with those obtained by Stevens (1938) through the formal procedure of maximum likelihood.

The MNS System

The two-allele MN system has been mentioned several times in earlier chapters. In fact, this locus also has a series of multiple alleles. The three genotypes (MM, MN, NN) are distinguished by anti-M and anti-N sera. If a third type of serum (anti-S) is used, each genotype may be further identified as S positive or S negative. Each of the major alleles M and N consists of two alleles (analogous to $A = A_1 + A_2$), one yielding S-positive reaction and one S-negative. These four alleles may be designated by MS, Ms, NS, Ns and their respective frequencies by m_1, m_2, n_1, n_2, so $m = m_1 + m_2$ is the frequency of the major allele M and $n = n_1 + n_2$ is the frequency of N. Furthermore, the alleles MS and NS are dominant to Ms and Ns with respect to the S reaction (but there is no dominance with respect to the MN reactions). Hence, with respect to the S reaction, $s = m_2 + n_2$ is the total frequency of the recessive alleles Ms and Ns. The situation may be summarized as follows:

MS m_1	NS n_1	$m_1 + n_1 = 1 - s$
Ms m_2	Ns n_2	$m_2 + n_2 = s$
M m	N n	1.00

On account of the dominance in S reaction, there are only six phenotypes, because each of the M, MN, N types is subdivided into S positive and negative. In the upper portion of Table 4-3 are the theoretical proportions of the six phenotypes (MS, etc.) and general notation for observed numbers (D_1, etc.); in the lower portion is a numerical example showing the arithmetic procedure of estimating the gene frequencies. The frequencies of the major alleles M and N are estimated first by the usual method:

$$m = \frac{2G_1 + G_2}{2G} \qquad n = \frac{G_2 + 2G_3}{2G}$$