Theoretical and Applied Genetics 42, 327–330 (1972)

The Effect of a Fusion of Subpopulations on the Total Fixation Index

I. HEUCH

Department of Human Genetics, University of Michigan, Ann Arbor, Michigan (USA)

Summary. A general mathematical expression is found for the decrease in the fixation index of a population where subpopulations with different gene frequencies fuse. It is shown that the use of Wright's formulas for a hierarchic structure will not necessarily give the correct result in this situation, since the conditions for their application are usually not satisfied. Two examples are given, one with fusions among subpopulations with a continuously distributed gene frequency, and one with data from real observations producing a discrete distribution.

Introduction

Yasuda (1968, p. 4) has extended the usual Wahlund principle (Wahlund, 1928) to a population divided into continuously distributed subpopulations, or subpopulations having a mixed distribution which is neither discrete nor continuous. He also considers the effect of fusion of some of the subpopulations, but only in the ordinary case with a finite number of such subpopulations (Yasuda, 1968, p. 3 and appendix I). It is the purpose of this paper to establish the results of such fusions in the general case. However, we will also allow local inbreeding within each subpopulation.

The Decrease in F_{IT}

We consider one autosomal locus with only two alleles A and a in diploid organisms. The results of subdivision will be compared with those of inbreeding, and, as pointed out by Li (1969), this cannot easily be done with multiple alleles. We imagine that we have any number of subpopulations, countable or not, each with a certain degree of local inbreeding. Each subpopulation may be regarded as an element ω in a sample space, over which there is defined a probability measure with respect to an appropriate σ-algebra. The value assigned by this measure to a set of subpopulations is assumed to be the relative size of the set. In real applications the number of subpopulations would always be finite, but in spite of this the general model should be useful in some situations. The case treated in Yasuda 1968, appendix I will be obtained with a probability distribution assigning probabilities $w_1, w_2, \ldots, w_n$ to the n subpopulations. It should be noted that all probability distributions are introduced for the purpose of describing the variation of certain quantities in nature, and that they have no connection with actual random sampling from the populations.

The gene frequency p of A and the fixation index (or inbreeding coefficient) F may now be considered as random variables, that is, as (measurable) functions $p = p(\omega)$ und $F = F(\omega)$ of the sample point ω, since it is assumed that each subpopulation has its specific values of p and F. The frequency $f(AA)$ of genotype AA will be a random variable given by

$$f(AA) = (1 - F)\, p^2 + F\, p = p^2 + F\, p\, q\,,$$

with $q = 1 - p$. The frequency of AA in the total population is then found by taking the mean,

$$E(f(AA)) = E(p^2) + E(F\, p\, q)\,,$$

and the frequency of the A gene is $E(p)$. If we had a fixation index F_{IT} in the total population, then the frequency of AA would be $(E(p))^2 + F_{IT}\, E(p)\, E(q)$. Thus the effect of subdivision and local inbreeding is the same as that of total inbreeding F_{IT} given by

$$F_{IT} = [Var\,(p) + E(F\, p\, q)]/[E(p)\, E(q)]\,, \qquad (1)$$

which may also be written as

$$F_{IT} = [(1 - E(F))\, Var\,(p) + Cov\,(F, p\, q)]/ [E(p)\, E(q)] + E(F)\,. \qquad (2)$$

For a finite number of subpopulations this expression is equivalent to eq. (15) in Nei (1965).

Now consider fusion of the original subpopulations into new greater subpopulations. This fusion will be specified by a (measurable) function $s = s(\omega)$, with the convention that each new subpopulation should consist of original subpopulations ω having the same value $s(\omega)$. In this new situation we assume that the frequency of A in a subpopulation will be the mean of the frequencies in the original subpopulations forming the new one. In a new subpopulation with $s(\omega) = s$ this value is the conditional mean $E(p|s)$. We also make the assumption that the level of inbreeding will be stabilized to the mean $E(F|s)$, although this is probably less realistic in some cases. (It is assumed that we can construct actual conditional probability measures, and that the conditional means are actual means with respect to these measures.) The total fixation index F'_{IT} corresponding

to the new situation is now found by replacing p with $E(p|s)$, q with $E(q|s)$ and F with $E(F|s)$ in (1) or (2). Using the rules $E(E(p|s)) = E(p)$, $E(E(F|s)) = E(F)$, we then obtain from (2)

$$F'_{IT} = [(1 - E(F))\ Var\ E(p|s) + Cov\ (E(F|s),\ E(p|s)\ E(q|s))]/[E(p)\ E(q)] + E(F)\ . \quad (3)$$

Hence the decrease F_B in the total fixation index due to creation of larger subpopulations is

$$F_B = F_{IT} - F'_{IT} = [(1 - E(F))\ E\ Var\ (p|s) + Cov\ (F, p\ q) - Cov\ (E(F|s),\ E(p|s)\ E(q|s))]/[E(p)\ E(q)]\ , \quad (4)$$

where we have applied

$$Var\ (p) = E\ Var\ (p|s) + Var\ E(p|s)\ . \quad (5)$$

F_B may also be expressed in various other ways, for instance as

$$F_B = [E\ Var\ (p|s) + E(F\ p\ q) - E(E(F|s)\ E(p|s)\ E(q|s))]/[E(p)\ E(q)]\ , \quad (6)$$

derived from (1).

Equation (4) is particularly instructive when both F and $p\ q$ are uncorrelated and $E(F|s)$ and $E(p|s) \times E(q|s)$ are uncorrelated. Then

$$F_B = (1 - E(F))\ E\ Var\ (p|s)/(E(p)\ E(q))\ , \quad (7)$$

and since $F_{IT} = (1 - E(F))\ Var\ (p)/(E(p)\ E(q)) + E(F)$ and $F'_{IT} = (1 - E(F))\ Var\ E(p|s)/(E(p)\ E(q)) + E(F)$, the relation $F_{IT} = F'_{IT} + F_B$ in this case simply reflects the general rule (5), expressing the total variance of p as the mean of the variances of p within the new subpopulations plus the variance of the new subpopulation gene frequencies. The conditions for (7) to hold true are satisfied when F and p are stochastically independent and at the same time $E(F|s)$ and $E(p|s)$ are independent. In many practical situations this would be correct to a high degree of approximation. It is, however, not sufficient that only F and p are stochastically independent.

The expression given by Yasuda (1968, p. 18) for the effect of fusion among certain of a finite number of subpopulations having relative sizes w_i and gene frequencies p_i (with no internal inbreeding) is

$$F_B = \Sigma\Sigma_{i>j}\ w_i\ w_j\ (p_i - p_j)^2/(\bar{p}\ (1 - \bar{p})\ W)\ . \quad (8)$$

Here the first summation is over all new subpopulations, and W is the sum over w_i for old subpopulations absorbed in the particular new subpopulation considered. $\bar{p}$ is the total mean over all p_i. With our notation we find in this case

$$Var\ (p|s) = \Sigma_i\ w_i\ (p_i - \bar{p}_s)^2/W = \Sigma_{i>j}\ w_i\ w_j\ (p_i - p_j)^2/W^2\ ,$$

where the summations are to be taken over subpopulations in the group given by the particular value of s. $\bar{p}_s$ is the mean of p_i among these subpopulations. Then

$$E\ Var\ (p|s) = \Sigma\ W\ \Sigma_{i>j}\ w_i\ w_j\ (p_i - p_j)^2/W^2\ ,$$

proving that (8) is a special case of (7).

Yasuda calls the subpopulations "isolates" and considers the fusion to be a result of breakdown of barriers. In our model the various subpopulations need not be isolated, since we are not concerned about the variation of p and F over any interval of time. If all values p and F are correct immediately before the fusion and F'_{IT} is to be the fixation at once afterwards, then there is no need for p and F to be constant in time.

Comparison with Wright's Hierarchic Structure

It might be supposed that the decrease $F_{IT} - F'_{IT}$ also could be found from Wright's relations for a hierarchic structure. We have (Wright, 1943, 1965)

$$1 - F_{IT} = (1 - F_{ST})\ (1 - F_{IS})\ , \quad (9)$$

where F_{IS} is the mean of local fixation indices in subdivisions, and F_{ST} is the correlation, relative to the total population, between gametes drawn at random from the same subdivision. With both primary (S) and secondary (R) subdivisions

$$1 - F_{IT} = (1 - F_{ST})\ (1 - F_{RS})\ (1 - F_{IR}) \quad (10)$$

(Wright, 1951, 1965). In our situation we would use (10) with the original subpopulations ω as the secondary subdivisions and the groups of subpopulations that are going to fuse as the primary ones. After the fusion, (9) might be applied with the same F_{ST}, with F'_{IT} instead of F_{IT}, and, under the assumption of maintenance of local inbreeding, with $F_{IS} = F_{IR}$. Thus (9) and (10) would give

$$F_{IT} - F'_{IT} = (1 - F_{ST})\ F_{RS}\ (1 - F_{IR})\ . \quad (11)$$

We will now show how the relations (9) and (10) could be derived in our model. The necessary conditions will however be rather restrictive, so (11) will have much less generality than (4) or (7). Consider first the case with only one set of subdivisions. At the moment these may be taken as the subpopulations ω, and then (2) implies that

$$F_{IT} = (1 - E(F))\ Var\ (p)/(E(p)\ E(q)) + E(F)$$

when $Cov\ (F, p\ q) = 0$. Substituting $F_{IS} = E(F)$ and $F_{ST} = Var\ (p)/(E(p)\ E(q))$ we readily obtain (9). But the condition for (9) to hold true is that F and $p\ q$ are uncorrelated, which is particularly correct if F and p are stochastically independent. The latter assumption is made in the derivation in Wright (1965), and the derivation given above is actually only a formalization of that one. A corresponding relation could be constructed in the general case without any such condition, as done by Barrai (1971), but then the simplicity is lost. Crow and Kimura (1970, section 3.12) have derived (9) in a different manner, interpreting F_{IS}, F_{ST}, and F_{IT} as probabilities. However, they define F_{IS} and F_{ST} with respect to only one particular subpopulation, making the situation somewhat different.

We now pass to the case with both primary and secondary subdivisions, where the subpopulations ω

constitute the secondary divisions, and the sets of ω with the same value $s(\omega)$ the primary ones. For any particular primary subdivision s we now find from (9)

$$1 - F_{IS} = (1 - F_{RS})\,(1 - F_{IRS})\,,$$

where $F_{IRS} = E(F|s)$ and $F_{RS} = Var\,(p|s)/\big(E(p|s) \times \times E(q|s)\big)$, under the condition that

$$Cov\,(F,\, p\,q|s) = 0 \qquad (12)$$

for this s. Furthermore, (9) applied to the total population with primary subdivisions gives

$$1 - F_{IT} = (1 - F_{ST})\,(1 - F_{IS})\,, \qquad (13)$$

where we now must define $F_{IS} = E(F_{IS})$ and $F_{ST} = Var\,E(p|s)/[E\,E(p|s)\;E\,E(q|s)] = Var\,E(p|s)/\big(E(p)E(q)\big)$. This is valid under the condition

$$Cov\,\big(F_{IS},\, E(p|s)\,E(q|s)\big) = 0\,. \qquad (14)$$

When (12) is correct, then F_{IS} may here according to (2) be expressed as

$$\big(1 - E(F|s)\big)\;Var\,(p|s)/[E(p|s)\;E(q|s)] + E(F|s)\,.$$

We now find

$$\begin{aligned} 1 - F_{IS} &= E\,(1 - F_{IS}) \\ &= \big(1 - E(F_{RS})\big)\,\big(1 - E(F_{IRS})\big)\,, \end{aligned} \qquad (15)$$

and if we introduce

$$F_{IR} = E(F_{IRS}) = E\,E(F|s) = E(F)$$

and

$$F_{RS} = E(F_{RS}) = E\,[Var\,(p|s)/\big(E(p|s)\;E(q|s)\big)]\,,$$

substitution of (15) in (13) produces (10). A necessary (and sufficient) condition for (15) to hold true is, however, that

$$Cov\,(F_{RS},\, F_{IRS}) = 0\,. \qquad (16)$$

To obtain (11) we must apply (9) to the situation after the fusion, and this is permitted only if

$$Cov\,\big(E(F|s),\; E(p|s)\,E(q|s)\big) = 0\,. \qquad (17)$$

Thus finally we get the expression (11) for $F_{IT} - F'_{IT}$, but only under the conditions (12) (for almost all s), (14), (16) and (17).

In comparison with the conditions $Cov\,(F,\, p\,q) = 0$ and $Cov\,\big(E(F|s),\; E(p|s)\,E(q|s)\big) = 0$ for (7) to be correct, (12), (14), (16) and (17) seem very restrictive. It is for instance not sufficient to assume that at the same time F and p are independent in all conditional distributions given s, $E(p|s)$ and $E(F|s)$ are independent, and F and p are independent. (No single one of these three restrictions, or any pair of them, implies any other.) If we assume that (12) is correct and that the pair $\big(E(p|s),\; Var\,(p|s)\big)$ is independent of $E(F|s)$, then the only condition left for (11) to hold true is the (reduced) equation (14):

$$Cov\,\big(Var\,(p|s)/\big(E(p|s)\;E(q|s)\big),\; E(p|s)\;E(q|s)\big) = 0\,. \qquad (18)$$

In particular, this is the only restriction if F is identical to a constant. But even now (18) imposes conditions on the distribution of p which will frequently not be satisfied. What causes the difficulties is that the "inbreeding coefficient" F_{IS} also includes a component due to secondary subdivision, and reasonable assumptions concerning actual local inbreeding coefficients will not apply to the effect of such subdivision. If we insert the values of F_{ST}, F_{RS} and F_{IR} in (11), the expression found will in the general case be quite different from (4) or (7).

It also follows from the treatment above that one should be cautious when applying (10) in other kinds of situations.

Example 1

Suppose that a population is evenly distributed over the quadratic region given by $0 \leq x \leq 1$, $0 \leq y \leq 1$ with respect to a coordinate system. The collection of individuals at each point $\omega = (x, y)$ is considered as a subpopulation. Since the density is constant, we may regard x and y as independent random variables with a uniform distribution over the interval $[0, 1]$. Let the original frequency of the gene A in the subpopulation at (x, y) be $p = (x + y - x\,y)/2$. The fixation index F is supposed to be identical for all such subpopulations. The fusion is now assumed to create new subpopulations such that all individuals at points (x, y) with identical y will belong to the same new population. Thus we have $s(\omega) = s(x, y) = y$.

This model does not necessarily require a fusion of subpopulations with different geographical positions. The y coordinate might for instance describe the variations among social groups living in the same place, and we would then find the consequences of an elimination of social barriers.

We now find $Var\,(p|y) = (1 - y)^2/48$, $E(p) = 3/8$, giving a decrease in inbreeding due to fusion of subpopulations equal to

$$F_B = 4(1 - F)/135\,,$$

found from (7). We also have $E(p|y) = (1 + y)/4$, $Var\,E(p|y) = 1/192$, giving $F'_{IT} = 3(1 - F)/135 + F$, and $Var\,(p) = 7/576$, giving $F_{IT} = 7\,(1 - F)/135 + F$. For example, when $F = 0$, a little more than half the amount of apparent total inbreeding is lost in the fusion.

The quantities used in (11) are found to be $F_{IR} = F$, $F_{ST} = 1/45$ and

$$\begin{aligned} F_{RS} &= \int_0^1 (1 - y)^2/\big(3\,(1 + y)\,(3 - y)\big)\,dy \\ &= (\log 3 - 1)/3\,. \end{aligned}$$

The result from eq. (11), which is really not applicable, will then be

$$F_{IT} - F'_{IT} = 44\,(\log 3 - 1)\,(1 - F)/135\,,$$

or approximately $0.0321 \cdot (1 - F)$, which is 8.3% greater than the correct value $0.0296 \cdot (1 - F)$ found above.

Example 2

Among the western group of the Yanomama Indians in Southern Venezuela, it is possible to recognize three village clusters, based on historical relationships (Ward, 1972). All of the Shamatari, Namoweitari and Wanaboweitari clusters are known to descend from single villages, and the construction of a genetic network for the total western group shows that, as a general rule, villages within each cluster are closely related. Our model will be applied to the M/N locus (disregarding S/s), with the total population consisting of all three clusters, the primary subpopulations $s(\omega)$ of the individual clusters, and the secondary subpopulations ω of the villages. Thus we are interested in what would happen to F_{IT} if all villages fused within each cluster.

The distributions of genotypes in samples from the villages are known (Gershowitz *et al.*, 1972). The villages belonging to the Shamatari cluster are $03D$, $03H$, $11G$ and $11HI$, to the Namoweitari cluster, $03A$, $03B$, $03C$ and $08ABC$, and to the Wanaboweitari cluster, $03E$, $03F$, $03G$, $03I$, $08N$, $08S$ and $08T$. The clusters are assigned s-values 1, 2 and 3, respectively. Our model is concerned with actual parameter values rather than estimates, but since a large proportion of each village was sampled (for most villages over 70%), it seems justified to use the estimated values as our values for p and F. The estimate for the gene frequency p of M for each village is found by gene counting in the usual way, and F is obtained from

$$F = 1 - f(M\,N)/(2\,p\,q)\,,$$

where $f(M\,N)$ is the (observed) relative frequency of heterozygotes. Means (and variances) are computed by giving village i weight N_i/N, where N_i is the total population size of the village, and N is the total size of all three clusters. The value for N_i does not coincide with the sample size, but is "the approximate village size" given in Gershowitz *et al.*, 1972, Table 1. Conditional means are found with weights N_i/n_s, where n_s is the size of cluster s, and means over variables depending on s are found using weights n_s/N.

The quantities needed for each cluster are given in Table 1. Using (6), we then find a decrease in F_{IT} equal to

$$F_B = 0.0225\,.$$

Actually, (1) and (3) give $F_{IT} = 0.0757$ and $F'_{IT} = 0.0532$. Ignoring the original local departures (in different directions) from random mating, (8) gives

$$F_B = 0.0287\,.$$

The indices required in (11) are $F_{IR} = 0.0082$, $F_{RS} = 0.0279$ and $F_{ST} = 0.0422$. Hence (11) would give

$$F_B = 0.0265\,.$$

The values for F_B given by (8) and (11) are 27.6% and 17.8% greater than the correct one given by (6). Now all conditions (12), (14), (16) and (17) are violated. The covariances in (12) are given in Table 1, and the covariances in (14), (16) and (17) are found to be 0.0011, 0.0016 and 0.0007.

Table 1. *The variation of parameters among clusters in example 2*

s	n_s	$E(p\|s)$	$Var(p\|s)$	$E(F\|s)$	F_{RS}	$Cov\,(F, pq\|s)$
1	397	0.793	0.0007	−0.0287	0.0044	−0.00009
2	417	0.562	0.0029	−0.0424	0.0119	−0.00057
3	421	0.614	0.0156	0.0930	0.0659	−0.00231

Acknowledgements

This work has been supported by a research fellowship from the Norwegian Research Council for Science and the Humanities.

Literature

1. Barrai, I.: Subdivision and inbreeding. Amer. J. Hum. Genet. **23**, 95–96 (1971). — 2. Crow, J., Kimura, M.: An introduction to population genetics theory. New York/Evanston/London: Harper and Row 1970. — 3. Gershowitz, H., Layrisse, M., Layrisse, Z., Neel, J., Chagnon, N., Ayres, M.: The genetic structure of a tribal population, the Yanomama Indians. II. Eleven blood-group systems and the ABH-Le secretor traits. Ann. Hum. Genet., Lond. **35**, 261–269 (1972). — 4. Li, C. C.: Population subdivision with respect to multiple alleles. Ann. Hum. Genet. Lond. **33**, 23–29 (1969). — 5. Nei, M.: Variation and covariation of gene frequencies in subdivided populations. Evolution **19**, 256–258 (1965). — 6. Wahlund, S.: Zusammensetzung von Populationen und Korrelationserscheinungen vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas **11**, 65–106 (1928). — 7. Ward, R.: The genetic structure of a tribal population, the Yanomama Indians. V. Comparison of a series of genetic networks. Ann. Hum. Genet., Lond. In press (1972). — 8. Wright, S.: Isolation by distance. Genetics **28**, 114–138 (1943). — 9. Wright, S.: The genetical structure of populations. Ann. Eugenics **15**, 323–354 (1951). — 10. Wright, S.: The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution **19**, 395–420 (1965). — 11. Yasuda, N.: An extension of Wahlund's principle to evaluate mating type frequency. Amer. J. Hum. Genet. **20**, 1–23 (1968).

Received December 17, 1971

Communicated by R. W. Allard

I. Heuch
Institute of General Genetics
University of Oslo
Oslo 3 (Norway)