Heredity **71** (1993) 591–595
©The Genetical Society of Great Britain

Received 26 March 1993

The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas*. VI. Estimation of the overlap between the allelic complements of a pair of populations

S. O'DONNELL, M. D. LANE & M. J. LAWRENCE*
Wolfson Laboratory for Plant Molecular Biology, School of Biological Sciences, University of Birmingham, Birmingham B15 2TT, U.K.

The data obtained by cross-classifying the self-incompatibility (S−) alleles of samples taken at random from three natural populations of *Papaver rhoeas*, presented in the previous paper (Lawrence *et al.*, 1993), are used here to estimate the extent of the overlap between the complements of alleles that pairs of these populations contain. These estimates indicate that this overlap is very great, so that these populations appear to contain essentially the same set of S-alleles. Three possible explanations of these results, which are not expected on the theory of the self-incompatibility polymorphism possessed by this species, are proposed and discussed. It is argued that the most likely of these alternative explanations is that the number of S-alleles in the species is not very much greater than the number of S-alleles that these natural populations contain, a hypothesis which is put to the test in the following paper.

Keywords: overlap analysis, *Papaver rhoeas*, population genetics, self-incompatibility polymorphism.

Introduction

The results obtained from the cross-classification of the self-incompatibility (S−) alleles of samples taken at random from each of three natural populations of *Papaver rhoeas* against those of the others revealed that 23 of a total of 45 S-alleles occurred in more than one sample and that 15 of these alleles occurred in all three (Lawrence *et al.*, 1993). The obvious implication of this finding is that the overlap between the complements of alleles that these populations contain is extensive. On the other hand, 22 of these alleles appeared to be confined to just one or other of these samples. The question with which we are concerned in the present paper is whether these results indicate that these populations are at least partially differentiated with respect to the S-alleles they contain or whether, alternatively, the overlap between their complement of alleles is so extensive as to suggest that they contain essentially the same set of alleles.

Theory

Consider two populations, A and B, which contain N_1 and N_2 S-alleles, respectively, and suppose that x of these alleles occur in both. Suppose, further, that a sample of n_1 alleles are taken from A and a second sample of n_2 alleles from B, sampling being without replacement in both cases. The samples are now cross-classified and y alleles are found to occur in both. A maximum likelihood estimate of x can be obtained from N_1, N_2, n_1, n_2 and y in the following way.

Consider, first, the sample of alleles taken from A. Let J be a random variable representing the number of alleles in this sample which belong to the set of x alleles common to both A and B. Then J is expected to have the following hypergeometric distribution (see, for example, Spiegel, 1975)

$$P(J=j)=\frac{\binom{x}{j}\binom{N_1-x}{n_1-j}}{\binom{N_1}{n_1}},\ P(j>x)=0,\ P(j>n_1)=0,$$

*Correspondence.

where

$$\binom{x}{j} = \frac{x!}{j!(x-j)!}$$

and similarly for the remaining terms of this expression.

Now consider population B. Within this population is a subset of the J alleles which have occurred in the sample taken from A. For y identities to occur, the sample from B must contain exactly y alleles from this subset, the probability of which is

$$\frac{\binom{j}{y}\binom{N_2-j}{n_2-y}}{\binom{N_2}{n_2}}, \text{ if } j<y, \text{ then this} = 0\ .$$

Hence, the probability of finding y identities between two completely cross-classified samples containing n_1 and n_2 alleles, given that x alleles occur in both of the populations from which these samples have been taken is

$$P(y; x, N_1, N_2, n_1, n_2) =$$

$$\sum_{j=\max(y, n_1+x-N_1)}^{\min(x, n_1, N_2+y-n_2)} \frac{\binom{x}{j}\binom{N_1-x}{n_1-j}\binom{j}{y}\binom{N_2-j}{n_2-y}}{\binom{N_1}{n_1}\binom{N_2}{n_2}}\ .$$

(We are indebted to P. Davies for pointing out the interval within which the terms of this summation are non-zero.)

This formula is the probability function of y. For any cross-classified pair of samples, however, y is fixed so that the formula can be used as a likelihood function for x. The value of x which maximizes this function, given y, N_1, N_2, n_1 and n_2, is the maximum likelihood estimate of x. As there are only a limited number of values that x can take, namely between y and either N_1 or N_2, whichever is the smaller, the maximum likelihood estimate of x can be found by substitution of successive values within this range.

Results

Estimates of x, the number of alleles that two populations have in common, have been obtained by applying the procedure described in the previous section to the data shown in Figs 1–3 of the previous paper (Lawrence *et al.*, 1993). This procedure requires that the alleles of one sample be completely cross-classified against those of the second, a requirement that is met in the case of the R102 × R104 matrix (Fig. 1 of the previous paper). Hence, for this part of the data, $n_1 = 27$, $n_2 = 25$ and $y = 19$. Estimates of the number of alleles in these populations are given in Table 3 of the previous paper, from which $\hat{N}_1 = 35$ and $\hat{N}_2 = 32$.

The data shown in the other two matrices, however, cannot be used as they stand because in both cases cross-classification is not complete. It is necessary, therefore, to derive submatrices which are complete or nearly so by omitting poorly explored arrays from the original matrices. A submatrix which is 96 per cent complete can be derived from the R104 × R106 matrix (Fig. 3 of previous paper) by omitting the row of the R104 allele S_{18} and the columns of the R106 alleles, S_6, S_{23} and S_{31}. Although the columns of the R106 alleles S_{16} and S_{26} are as incomplete as those excluded, they have been retained because these alleles are probably identical to the R104 alleles S_{14} and S_{16}, respectively (or *vice versa*; see Table 4 of the previous paper) which, although they have not been fully located, have nevertheless been included in the total number of identities found for this pair of populations. The omission of four arrays from the original matrix yields a submatrix for which $n_1 = 24$, $n_2 = 28$, $y = 17$, $\hat{N}_1 = 32$ and $\hat{N}_2 = 38$.

It is not possible to derive a completely cross-classified submatrix from the R102 × R106 data (Fig. 2 of previous paper) without discarding an unacceptably large amount of information. The rows of the R102 alleles S_{23} and S_{29} can be safely discarded, however, as can the column of the R106 allele S_{31}, for these contain no direct information. The array of the R102 allele S_{26} has been retained because this allele is probably the same as either S_{16}, S_{26} or S_i of the R106 sample (see Table 4 of previous paper) and this identity, although not completely located, has been included in the total number of identities found for this pair of populations. The omission of three arrays from the original R102 × R106 matrix leaves a submatrix which is 89 per cent complete for which $n_1 = 25$, $n_2 = 30$, $y = 16$, $\hat{N}_1 = 35$ and $\hat{N}_2 = 38$.

The estimates of the number of alleles which pairs of these populations have in common are shown in Table 1. In terms of the S-alleles they contain, R102 and R104 appear to be identical and R104 and R106 are very nearly so which suggests, of course, that the same should be true of R102 and R106. The fact that the estimated overlap between the latter pair of populations is only 83 per cent is almost certainly due to incomplete cross-classification. Thus, in estimating the overlap between these populations we have had to assume that all of the alleles that are common to the samples taken from these have been found, which is equivalent to the assumption that none of the pair-wise comparisons between the alleles of these samples that have not been made, either directly or indirectly, involve identities. As 11 per cent of the R102 × R106 comparisons fall into this category this is clearly a rather questionable assumption. Furthermore, an examination of the distribution of S-alleles between the three samples (Table 6 of the previous paper) reveals an apparent excess of alleles that occur only in R106 and a corresponding deficiency of alleles that occur in this population and also in R102 and R104. As was pointed out in the previous paper, this rather peculiar distribution is almost certainly indicative of incomplete cross-classification and the same point holds here with regard to the estimate of x for these pairs of populations. Hence, this analysis of the overlap between the allelic complements of these three natural populations strongly suggests that this is at least very great and probably complete.

Discussion

Because the self-incompatibility polymorphisms are maintained by frequency dependent selection, the selective advantage of any allele being negatively related to its frequency in the population, the number of alleles in a species with a homomorphic system of self-incompatibility is theoretically expected to be very large. The number of alleles that can be maintained in a population at equilibrium of such a species, however, depends on its effective size and the mutation rate of the S-gene (Wright, 1939). Unless this size is very large and the mutation rate very high, any population is expected to contain, at any one time, only a subset of the total number of alleles of the species. On this argument, different populations, provided they are isolated, are expected to contain different subsets of alleles. Hence, if the number of alleles in the species is large, these populations are expected to be differentiated with respect to the S-alleles they contain.

It is clear that the results we have obtained from a survey of the distribution of alleles between three natural populations of *P. rhoeas* are quite inconsistent with this expectation. Indeed, the most striking feature of these results is not merely that these populations fail to show any evidence of differentiation but that they appear to contain virtually the same set of alleles.

There are three possible explanations of these unexpected results. First, we have inadvertently drawn our samples from the same population and that each is large enough to have captured most of the alleles that this super-population contains; second, only a subset of the alleles of the species is present in this country because the British Isles is on the western periphery of the geographical distribution of the species; third, despite theoretical expectation, the number of S-alleles in the species is not very large, so that its constituent populations, provided they are long established and large, come to contain, through mutation, most if not quite all of these alleles.

The first explanation is unlikely for three chief reasons. First, the distance between the closest pair of populations (R102 and R106) is no less than 43 km and each is separated from the others by extensive tracts of predominantly pastoral agriculture in which *P. rhoeas* is conspicuous by its absence; indeed, this was the chief reason why these populations were chosen for investigation. Second, these populations are known to differ, on average, for a number of metrical characters (Ooi, 1970), so there is no question that they are at least partially differentiated for the genes determining these characters at least. Third, though they may contain essentially the same set of alleles, those which occur at a relatively high frequency in one do not, in general, occur at a relatively high frequency in the others (Lawrence *et ál.*, 1993). Taken together, these facts suggest quite strongly that these populations are independent. On the other hand, we cannot rule

Table 1 Estimates of the number of alleles that occur in pairs of populations and of the percentage overlap between their complements of alleles, which is calculated as $(100 \times x)/N$ where $N = \min(N_1, N_2)$. R102 = Wellesbourne, Warwickshire, R104 = Broad Oak, Herefordshire and R106 = Hackmans Gate, Worcestershire

Comparison	$\hat{N}_1$	$\hat{N}_2$	n_1	n_2	y	$\hat{x}$	% Cross-classification	% Overlap
R102 × R104	35	32	27	25	19	32	100	100
R102 × R106	35	38	25	30	16	29	89	83
R104 × R106	32	38	24	28	17	31	96	97

out the possibility that before the introduction of efficient seed cleaning machinery into agriculture there was a significant amount of migration of seed between these populations either directly or *via* intermediate populations. On this view, we would have to suppose that the inertia of the self-incompatibility polymorphism to change is sufficiently great to have prevented the divergence of these populations, under the joint influence of mutation and drift, with respect to the S-alleles they contain. Very little is known about the dynamics of this polymorphism, so it is impossible to say how many generations are required before three daughter populations have diverged to the point where their differentiation could be detected in an experiment of realistic size, although in principle this problem could be investigated by computer simulation.

The second explanation supposes that these populations contain the same set of alleles because only a limited number occur in the British Isles. For this explanation to hold, we should also have to suppose that the mutation rate from one functional allele to another was so low that there had been insufficient time since the species first appeared in Britain for many, if any, new alleles to have occurred. While estimates of the mutation rate of this locus suggest that this is not high (Emerson, 1939; Lewis, 1948, 1951), poppies are known to have been present in the British Isles for at least 3000 years because seed remains have been found in a Bronze Age settlement in Wiltshire (Robinson, 1989). The species may not, of course, always have been as abundant as it is at present. Furthermore, the age in generations of any one of its constituent populations is almost certainly less than its age in years, both because a stand of flowering plants is expected only when seed germinates on recently disturbed ground and because of overlap between generations in the bank of seed in the soil by which a population persists between one flowering episode and the next. Nevertheless, it is difficult to believe that no new alleles have occurred by mutation since the time when the species first appeared in Britain.

The third explanation appears, at first sight, to be the least likely because whereas the first two explanations attempt to accommodate our results to the theory, the third challenges the notion that the effect of frequency dependent selection is unconditionally to generate a very large number of alleles at the S-locus. There are two reasons for doubting whether this notion is true. Firstly, while the number of alleles in populations of *Trifolium pratense* (Williams & Williams, 1947; O'Donnell & Lawrence, 1984) and *T. repens* (Atwood, 1944) appear to be large, this number appears to be much smaller in the two most thoroughly investigated species with a one-locus, multi-allelic, gametophytic system of self-incompatibility. Thus Emerson (1940) found 45 different S-alleles in *Oenothera organensis* and we have found, quite coincidentally, the same number in *P. rhoeas*. Furthermore, only 49 different S-alleles have been found in *Brassica oleracea* (Ockendon, 1985), which is the most extensively investigated species with a one-locus, multi-allelic, sporophytic system of self-incompatibility. That three quite different species appear to possess a similar number of alleles is surely not just coincidence. Secondly, while the selective advantage of an allele is negatively related to its frequency in a population, it is also negatively related to the number of alleles present. Because of the latter, the selective advantage of a new allele appearing in a population previously containing only three is much greater than in one previously containing, say, thirteen. The consequences of this attenuation of the frequency dependent effect as the number of alleles in a population rises is an aspect of the theory which appears to have received less attention than it deserves. Although we shall return to this subject in a future paper, it is worth pointing out here that as the cross-compatibility between the individuals of a population containing as few as 20 equally frequent alleles is no less than 99.47 per cent, it is not obvious why any population needs many more than this number to accomplish a full set of seed. In short, on this argument, the question we have been discussing changes from asking why we find so few alleles to asking why, on the contrary, we find so many.

Now, whereas we have no way of distinguishing between these explanations on the evidence presently available, in principle, such evidence is obtainable. Thus, if the third explanation is correct, we expect no population of *P. rhoeas* to contain many more than 45 alleles and that different long-established and large populations will contain most of these alleles and, hence, contain very similar sets of alleles. The second explanation, on the other hand, leads to the expectation that populations found elsewhere within the geographical range of the species will contain different subsets of alleles from those contained by British populations; and the first explanation leads to a similar expectation with regard to populations found elsewhere in the British Isles. In the following paper we consider the results obtained from a partial cross-classification of the alleles of a sample taken from a Spanish population against those of the R104 sample which goes some way towards distinguishing between these explanations.

Acknowledgements

In carrying out the work, one of us (S.O'D.) was supported by a SERC Research Studentship, a second (M.D.L.) by a NERC Research Studentship and the third (M.J.L.) by a NERC Research Grant, all of which are gratefully acknowledged. We are grateful also to P. Davies, for advice on mathematical matters and to J. Greig for archeological information.

References

ATWOOD, S. S. 1944. Oppositional alleles in natural populations of *Trifolium repens. Genetics*, **29**, 428–435.

EMERSON, S. 1939. A preliminary survey of the *Oenothera organensis* population. *Genetics*, **24**, 528–537.

EMERSON, S. 1940. Growth of incompatible pollen tubes in *Oenothera organensis. Bot. Gaz.*, **101**, 890–911.

LAWRENCE, M. J., LANE, M. D., O'DONNELL, S. AND FRANKLIN-TONG, V. E. 1993. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas.* V. The cross-classification of the S-alleles of samples from three natural populations. *Heredity*, **71**, 581–590.

LEWIS, D. 1948. Structure of the incompatibility gene. I. Spontaneous mutation rate. *Heredity*, **2**, 219–236.

LEWIS, D. 1951. Structure of the incompatibility gene. III. Types of spontaneous and induced mutation. *Heredity*, **5**, 399–414.

OCKENDEN, D. J. 1985. Genetics and physiology of self-incompatibility in *Brassica. Current Communications in Molecular Biology*, Cold Spring Harbor Laboratory, NY, pp. 1–6.

O'DONNELL, S. AND LAWRENCE, M. J. 1984. The population genetics of the self-incompatibility polymorphism in *Papaver rhoeas.* IV. The estimation of the number of alleles in a population. *Heredity*, **53**, 495–507.

OOI, S. C. 1970. *Variation in wild populations of Papaver rhoeas* L. Ph.D. Thesis, University of Birmingham.

ROBINSON, M. 1989. Seeds and other plant macrofossils. In: Bell, M. and Proudfoot, E. (eds), *Wilsford Shaft: Excavations 1960–62*, English Heritage Archeological Report, **11**, pp. 78–90.

SPIEGEL, M. R. 1975. *Theory and Problems of Probability and Statistics.* McGraw-Hill, New York.

WILLIAMS, R. D. AND WILLIAMS, W. 1947. Genetics of red clover (*Trifolium pratense* L.) compatibility. III. The frequency of incompatibility S-alleles in two non-pedigree populations of red clover. *J. Genet.*, **48**, 64–79.

WRIGHT, S. 1939. The distribution of self-sterility alleles in populations. *Genetics*, **24**, 538–552.