Conference on Applied Statistics in Agriculture 1990 - 2nd Annual Conference Proceedings

ESTIMATING MIXTURE FRACTION AND MAP DISTANCE IN A MIXED F_2, BC_1 POPULATION

Dennis L. Clason

N. Scott Urquhart

Joe Corgan

Catherine M. Cryder

See next page for additional authors

Recommended Citation

Clason, Dennis L.; Urquhart, N. Scott; Corgan, Joe; and Cryder, Catherine M. (1990). "ESTIMATING MIXTURE FRACTION AND MAP DISTANCE IN A MIXED F_2, BC_1 POPULATION," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1434

Author Information

Dennis L. Clason, N. Scott Urquhart, Joe Corgan, and Catherine M. Cryder

ESTIMATING MIXTURE FRACTION AND MAP DISTANCE IN A MIXED F_2, BC_1 POPULATION

Dennis L. Clason, N. Scott Urquhart
Department of Experimental Statistics
Joe Corgan
Department of Agronomy and Horticulture
New Mexico State University
Catherine M. Cryder
Shamrock Seed Company, Las Cruces NM

ABSTRACT

An F_1 interspecific hybrid onion (Allium cepa x A. fistulosum) was backcrossed to the A. cepa parent line under field conditions. The progeny of this cross were shown by electrophoretic protein analysis to be a mixture of BC_1 (the desired backcross) and F_2 (A. cepa x A. fistulosum) x (A. cepa x A. fistulosum)individuals. This mixture of populations among the progeny render the usual χ^2 test for independent segregation of loci invalid. F_2 is used to denote progeny derived from either selfing of the F_1 or from sib-crosses between two F_1 individuals. A model for this mixed population has been developed; from this model the mixture fraction and crossover frequencies can be estimated using maximum likelihood.

1. Introduction

Because the Japanese bunching onion (Allium fistulosum L.) has many desirable traits, particularly disease resistance, plant breeders have sought to introduce its genetic material into the gene pool for the cultivated bulb onion (A. cepa L.). F_1 hybrids from this cross exhibit a high degree of sterility, but small numbers of backcross (BC_1) and F_2 progeny have been recovered from greenhouse experiments (Cryder, 1988). In 1985 a large population of these F_1 hybrids was open pollinated, but with other pollen sources restricted to A. cepa, the recurrent parent. This resulted in a large seed population, which permitted an extensive study of heritable traits expressed by this backcross. The seed population actually was a mixture of F_2 and first generation backcross (BC_1) genetic populations. F_2 is used here to denote progeny derived either from selfing of the F_1 or from sib-crosses between two F_1 individuals.

In 1985 Peffley, et al. showed the existence of alternate alleles in the A. fistulosum genome coding for several enzymes (including Alcohol dehydrogenase, Isocitrate dehydrogenase, and Phosphoglucoisomerase -- henceforth ADH, IDH, and PGI respectively). Analysis of the isozyme patterns can provide an estimate of the F_2 proportion in the seed population. This paper presents and illustrates a method for estimating the proportion of the mixed population arising from the F_2 and BC_1 crosses, which in turn is an estimate of the proportion of selfing. With this information it is possible to test independent segregation of the ADH, IDH and PGI alleles.

2. Materials and Methods

Samples of the seed populations and the reference populations were analyzed electrophoretically for the enzymes ADH, IDH, and PGI. [For a particular enzyme, an individual bulb was classified displaying:

cepa if it produced only A. cepa isozymes,
fist if it produced only A. fistulosum isozymes,
het if it produced both A. cepa and A. fistulosum isozymes.]

The isozyme analysis was performed on root tissue. Bulbs were allowed to root in moist sand for seven to ten days. Roots were collected in a vial of water and held at 4°C for at least twelve hours and no more than twenty-four hours. Because ADH is inducible (by anaerobic conditions) this holding period enhanced the electrophoretic pattern for ADH.

3. Statistical Models

Under ordinary circumstances (i.e. the observed population is the same as the genetic population), the usual chi-square test of independence provides a test of the null hypothesis that two alleles (here, any two of ADH, IDH and PGI) segregate independently. Because the population observed here is a mixture of two genetic populations (F_2 and BC_1) a large value of the independence chi-square statistic admits either of two explanations. First, it may be that the alleles are linked (i.e., they are carried on the same chromosome), or it may be that the subpopulations have different probability structures. If a gene has two forms or alleles say "a" and "b", an individual will display one of three genotypes being homozygous for allele "a", homozygous for allele "b" or heterozygous - carrying one "a" gene and one "b" gene. If two genetic traits are observed, then plants from a random sample of the population can be classified into 9 cells of a multinomial which can be organized as a 3 × 3 table. In the problem at hand there are two alleles for each gene - one form from A. cepa and one from A. fistalosum. An individual carrying only A. cepa alleles will be referred to as a "cepa" type, similarly, an individual carrying only A. fistulosum alleles is called a "fist" type, and an individual which having one A. cepa allele and one A. fistulosum allele at a particular locus will be referred to as a "het" (i.e., heterozygous) type. The nature of this table differs for F_2 and BC_1 populations. The probabilities over these nine cells for a F_2 population has the following structure (except for divisors of 16 omitted for clarity of pattern):

Genotype at Locus 1	Genotype at Locus 2		
	cepa	het	fist
cepa	1/16	2	1
het	2	4	2
fist	1	2	1 ,

(1)

while a BC_1 population has this structure:

Genotype at Locus 1	Genotype at Locus 2		
	cepa	het	fist
cepa	4/16	4	0
het	4	4	0
fist	0	0	0 .

(2)

Let α be a mixture parameter, $0 \leq \alpha \leq 1$. By this, we mean that the population under study is a mixture of 100 α % F_2 individuals and $100(1-\alpha)$% BC_1 individuals. Such a population has the following structure:

Genotype at Locus 1	Genotype at Locus 2		
	cepa	het	fist
cepa	$[(1-\alpha)4 + \alpha]/16$	$[(1-\alpha)4 + 2\alpha]$	α
het	$[(1-\alpha)4 + 2\alpha]$	4	2α
fist	α	2α	α

(3)

Now we use the cell probabilities given above to obtain the multinomial probability mass function for a data vector

$$\underline{n}' = (n_{11}, n_{12}, n_{13}, n_{21}, n_{22}, n_{23}, n_{31}, n_{32}, n_{33}),\ n = \sum_{ij} n_{ij}.$$

The indices 1,2,3, correspond to "cepa", "het", and "fist" types, respectively.

$$M(\underline{n}|\alpha) = \frac{n!}{\prod_{i,j} n_{ij}!} \{((4-3\alpha)/16\}^{n_{11}} \{((4-2\alpha)/16\}^{n_{12}} \{\alpha/16\}^{n_{13}}$$

$$\times \{((4-2\alpha)/16\}^{n_{21}} \{4/16\}^{n_{22}} \{2\alpha/16\}^{n_{23}} \{\alpha/16\}^{n_{31}} \{2\alpha/16\}^{n_{32}} \{\alpha/16\}^{n_{33}}$$

$$= \frac{n!}{\prod_{i,j} n_{ij}!} \left\{ \frac{2^{n_{23} + n_{32} + n_{12} + n_{21} + 2n_{32}}}{16^{\sum_{i,j} n_{ij}}} \right\} \times$$

$$(4-3\alpha)^{n_{11}} (2-\alpha)^{n_{12} + n_{21}} \times \alpha^{n_{13} + n_{23} + n_{31} + n_{32} + n_{33}} .$$

If we define

$$\alpha = n_{13} + n_{23} + n_{31} + n_{32} + n_{33}$$
$$b = n_{12} + n_{21}$$
$$c = n_{11} \quad ,$$

for notional convenience, the mass function can be interpreted as a likelihood function

$$L(\alpha|\underline{n}) = K\,\alpha^{a}\,(2-\alpha)^{b}\,(4-3\alpha)^{c}. \qquad (4)$$

Here K is a constant of proportionality involving the multinomial combinatorial coefficient and some powers of 2.

If we differentiate L with respect to alpha and set the result equal to zero, the maximum likelihood estimate of alpha must satisfy the following quadratic equation:

$$\partial L/\partial\alpha = 3(a+b+c)\hat{\alpha}^2 - 2(5a+2b+3c)\hat{\alpha} + 8\hat{\alpha} = 0. \qquad (5)$$

One solution always lies in the interval [0,1], while the other always lies outside the interval; the root inside [0,1] maximizes the likelihood function. A chi-square test for independent segregation can be based on this multinomial distribution and the associated maximum likelihood estimate. Note that this chi-square statistic has 7 degrees of freedom because the unconstrained parameter space for the nine cells has dimension 8 while the null hypothesis space has one dimension (for α).

The above material was developed assuming the null hypothesis of independence was true. If the loci fail the test for independence (i.e. the chi-square statistic calculated above is large) it may be that the loci lie on the same chromosome i.e., the loci are linked). During gamete formation, a process called 'crossing over' can occur. Crossing over involves the exchange of genetic material between two homologous chromosomes. The probability that a cross over will occur depends on the spatial separation between the two loci. Let $0 < m \leq 0.5$ denote the probability of cross over. This also is called the 'map distance'. An individual which has linked loci (AB) on one chromosome and (ab) on the homologous chromosome can produce gametes having the following genotypes:

Haploid Genotype	Population fraction (Linked loci)	(Independent loci)
(AB)	(1-m)/2	1/4
(Ab)	m/2	1/4
(aB)	m/2	1/4
(ab)	(1-m)/2	1/4

Note that $m = 0.5$ gives the independent segregation case, hence the restriction on m. Under (assumed) random mating, the following genotypes will be present:

Diploid Genotype	Population fraction
(AB)(AB)	$(1-m)^2/4$
(AB)(Ab)	$m(1-m)/4$
(AB)(aB)	$m(1-m)/4$
(AB)(ab)	$(1-m)^2/4$
.	.
.	.
.	.
(ab)(AB)	$(1-m)^2/4$
.	.
.	.
.	.
(ab)(ab)	$(1-m)^2/4$.

The diploid genotypes can be separated in heterozygous and homozygous types, resulting in the following multinomial distributions for F_2 and BC_1 populations, again omitting the divisors of 16.

F_2 populations:

Genotype at Locus 1	Genotype at Locus 2			
	cepa	het	fist	
cepa	$(1-m)^2$	$2m(1-m)$	m^2	
het	$2m(1-m)$	$2(1-m)^2 + m^2$	$2m(1-m)$	(6)
fist	m^2	$2m(1-m)$	$(1-m)^2$	

BC_1 populations:

Genotype at Locus 1	Genotype at Locus 2			
	cepa	het	fist	
cepa	$2(1-m)$	$2m$	0	
het	$2m$	$2(1-m)$	0	(7)
fist	0	0	0 .	

Assuming the F_2 mixture fraction to be α, and following the same development as for display (3) we obtain the following cell probabilities for the multinomial distribution (divisors of 16 again omitted for clarity):

Genotype at Locus 1	Genotype at Locus 2		
	cepa	het	fist
cepa	$2\alpha(1-m) + (1-\alpha)(1-m)^2$	$2\alpha m + 2(1-\alpha)m(1-m)$	m^2
het	$2\alpha m + 2(1-\alpha)-m(1-m)$	$2\alpha\{(1-m)^2 + m^2\} + 2(1-\alpha)m^2$	$2m(1-m)$
fist	m^2	$2m(1-m)$	$(1-m)^2$.

(8)

The likelihood function can be obtained from table (8) in exactly the same fashion as equation (4) was derived from table (3):

$$L(\alpha,m:\underline{n}) = K \prod_{i=1}^{3}\prod_{j=1}^{3} (p_{ij})^{n_{ij}}$$

$$= K\{2\alpha(1-m)+(1-\alpha)(1-m)^2\}^{n_{11}}\ldots\{\alpha(1-m)^2\}^{n_{33}}. \qquad (9)$$

This function does not submit nicely to algebraic maximization, but is very amenable to numerical maximization using a brute force grid search. This is feasible because of the relatively small number of sufficient statistics (only 8) and because the parameter space is a finite region in two-space. Once again, a chi-square test statistic with 8 - 2 = 6 degrees of freedom can be formed using the maximum likelihood estimators of α and m. If the model is adequate to explain the data, a small value of the chi-square statistic should occur. If the calculated chi-square is large, we should question the assumptions in the model.

4. Results

The cell counts for population 1034[1] are presented in Table 1. Recall that cepa refers to an individual homozygous for A. cepa alleles for the isozyme, fist to an individual homozygous for A. fistulosum alleles, and het to an individual with both A. cepa and A. fistulsom isozymes.

Table 2 presents the estimated mixture fractions for the three allele combinations. The χ^2 test statistic provides the test of model adequacy, as suggested in the text. Figure 1 shows the value of the likelihood function for the mixed population models as a function of alpha.

[1]A mixed (A. fistulosum x A. cepa) x A. cepa and (A. fist. x A. cepa) x (A. fist. x A. cepa) population mentioned in the introduction), one of several available for use as an illustration from Cryder, 1988.

5. Discussion and Conclusions

As Figure 1 shows, the likelihood function remains relatively flat over a fairly broad range. This implies a range values of α which are nearly equally appropriate. The magnitude of the estimated F_2 admixture in the population is between 7.5% and 10.8%, over all combinations of alleles. This is the range of estimates for α, not a confidence interval. Work leading to interval estimates for α is not yet complete. However, these models do not fit the data very well, as measured by the chi-square test statistic, suggesting that model assumptions are unsatisfied. Further work on biological dimensions of this problem suggested meiotic barriers of some type in the F_1 parent.

The general methods used in this paper can be used to estimate self-pollination rates for any population which is known to be a mixture of "hybrids" and "pure" types. These techniques may prove to be of use to population ecologists working with hybridization complexes as well as to plant breeders interested in tracing the heritage of interspecific hybrids.

We realize that the mixing problem could also be approached using a $3 \times 3 \times 3$ table rather than the three 3×3 tables used here. The counts are low there, but the one estimate of alpha coming from that approach is consistent with the results described above. We have also investigated minimum chi-square estimation, but will defer consideration of that to another paper.

Table 1. Frequency counts for population 1034.

ADH	IDH	PGI Classification Cepa	Het	Fist
cepa	cepa	41	15	2
	het	19	15	0
	fist	0	1	1
het	cepa	14	22	0
	het	17	24	1
	fist	0	3	0
fist	cepa	0	1	0
	het	0	1	0
	fist	0	0	0

Table 2. Maximum likelihood estimates of mixture fractions and map distances.

Model	Alleles	Estimated Mixture Fraction	Estimated Map Distance	Chi-Square
Mixture	All	0.094	--	15.89
	IDH,PGI	0.096	--	10.58
	IDH,ADH	0.075	--	22.45
	ADH,PGI	0.084	--	12.60
Linked	IDH,PGI	0.108	0.420	8.43
	IDH,ADH	0.080	0.380	16.31
	ADH,PGI	0.098	0.430	9.21

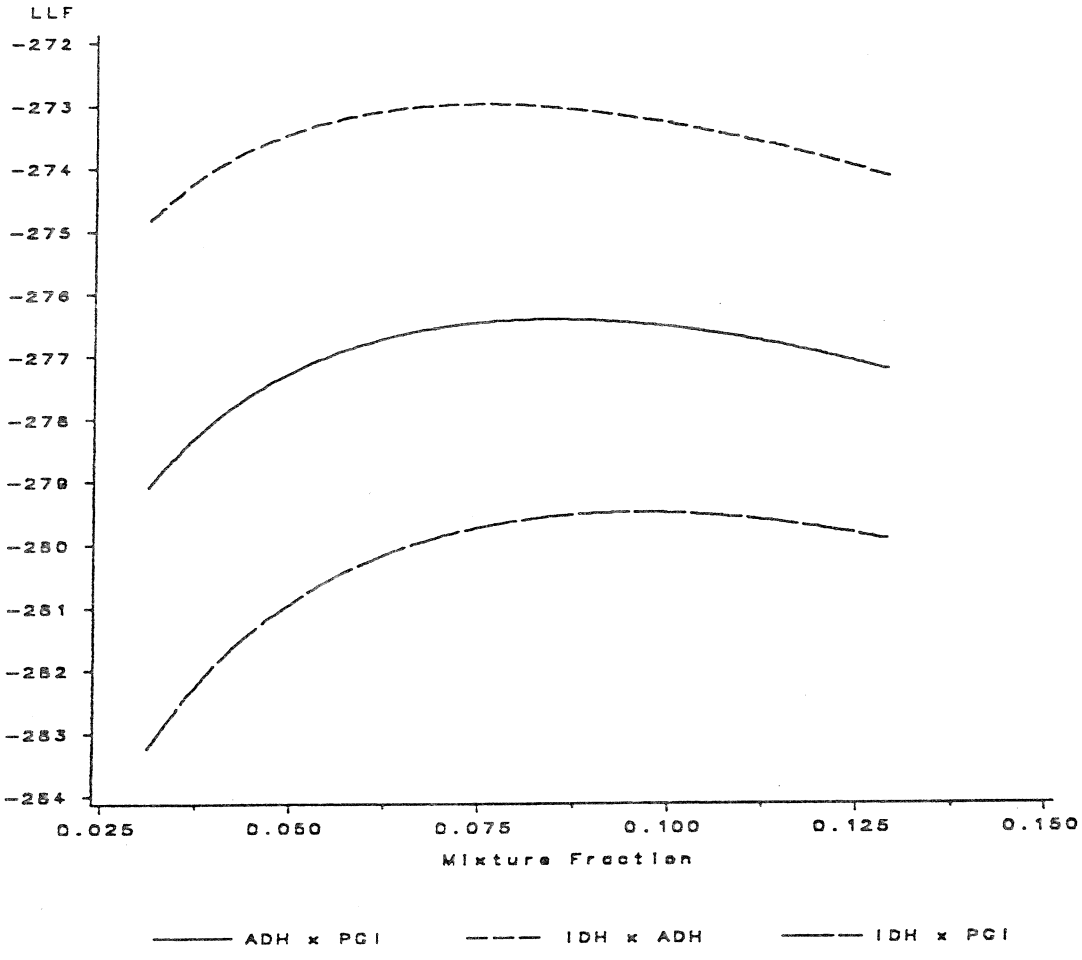

Figure 1. Log-likelihood function for the 3 mixture models.

Literature Cited

Cryder, C. M. 1988, A Study of the Associations of Heritable Traits in Progeny From the Interspecific Backcross (Allium fistulosum x Allium cepa L.) x Allium cepa L. Unpublished Ph.D. dissertation available from UMI.

Peffley, E. B., J. N. Corgan, K. E. Horak and S. D. Tankley 1985, Electrophoretic Analysis of Allium alien addition lines. Theor. Appl. Genet. 71:176-184.