

## SHORT COMMUNICATION

**A new approach to detecting mixed families**

TIMOTHY H. VINES\* and NICHOLAS H. BARTON

*Institute of Cell, Animal and Population Biology, University of Edinburgh, King's Buildings, Edinburgh, EH9 3JT, UK***Abstract**

**There are several analyses in evolutionary ecology which assume that a family of offspring has come from only two parents. Here, we present a simple test for detecting when a batch involves two or more subfamilies. It is based on the fact that the mixing of families generates associations amongst unlinked marker loci. We also present simulations illustrating the power of our method for varying numbers of loci, alleles per locus and genotyped individuals.**

*Keywords:* linkage disequilibrium, matrix algebra, mixed families, multiple mating, parentage, randomization test

*Received 16 January 2003; revision received 14 March 2003; accepted 14 March 2003*

**Introduction**

There are several situations in evolutionary biology when one must be sure that a field-collected set of offspring has come from only one pair of parents. For example, the heritability of a quantitative trait cannot reliably be inferred from sets of siblings if there is uncertainty over how many families are present (Falconer & Mackay 1996). Second, if the family is not the product of a single mating, inferences about the identity of the parents will be misleading (e.g. Vines 2002). Third, valuable information about the mating system can be gleaned from the contributions of different males to the offspring of a single female (e.g. deWoody *et al.* 2000; Emery *et al.* 2001). In this paper we present a novel technique for detecting mixed families that is based on the fact that mixing of families generates associations amongst unlinked loci.

Traditionally, the detection of mixing has relied on the presence of alleles in the offspring that were in neither of the putative parents. When the parents are unknown, finding five or more alleles in a batch indicates that more than two parents were involved. Detecting mixed families using the number of alleles is therefore only effective when the probability of observing the extra alleles is high. This requires high levels of polymorphism at each of the marker loci, especially when allele frequencies are skewed (Neff & Pitcher 2002).

By contrast, the method presented here can be applied to offspring genotyped at loci with fewer than five alleles. In addition, our approach is more effective when allele frequencies are skewed. The underlying principle is similar to the generation of linkage disequilibrium by the mixing of distinct populations: alleles at different loci will tend to be associated across subfamilies (Barton 2000). These associations are not expected in families produced from a single segregation, as alleles are distributed between individuals independently at each locus, provided there is no linkage. This fact is used to test for the presence of mixed families: the associations between loci should drop dramatically when we shuffle alleles between individuals when more than two parents have contributed to the batch.

This paper presents a method for quantifying the associations between unlinked loci, and a randomization test for detecting mixed families based on this statistic. Lastly, we test the power of this approach using simulations.

**Method**

To accommodate varying numbers of alleles per locus, we represent the genotype of an individual at a  $t$ -allelic locus as a column vector with  $t$  elements: e.g. the genotype vector  $\mathbf{G}$  for individual  $i$  at tri-allelic locus  $j$  would be

$$\mathbf{G}_{i,j} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

Correspondence: T. H. Vines. \*Present address: Zoology Department, University of British Columbia, Vancouver, BC, Canada V6T 1Z4. Fax: + 1604 8222416; E-mail: vines@zoology.ubc.ca

where  $x, y$  and  $z = 0, 1$ , or  $2$  and  $x + y + z = 2$ . For example,

$$\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

represents a heterozygote for the first and second alleles, and

$$\begin{pmatrix} 1 & 2 & 0 & 1 \\ 1 & 0 & 2 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

represents a batch with four individuals genotyped at a single tri-allelic locus. The genotypes at other loci are given by similar matrices. The list of deviations in allelic state  $\zeta$  for allele  $x$  at locus  $j$  is given by

$$\zeta_{x,j} = n(x)_{i,j} - p_{x,j}$$

where  $p_{x,j}$  is the frequency of  $x$  alleles at locus  $j$ , and  $n(x)_{i,j}$  is the number of  $x$  alleles possessed by individual  $i$  at locus  $j$ . The deviations in allelic state for the example above are

$$\begin{pmatrix} 0.5 & 1.5 & -0.5 & 0.5 \\ 0.5 & -0.5 & 1.5 & 0.5 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The squared covariance between allele  $x$  at locus  $j$  and allele  $c$  at locus  $k$  is given by

$$[(\zeta_{x,j} \cdot \zeta_{c,k})/N]^2$$

where  $N$  is the number of individuals. The squared covariance is summed over all the possible between-locus pairings of alleles and averaged over all  $m$  possible pairs of loci, as shown in equation 1.

$$CV^2 = m^{-1} \sum_{j \neq k}^m \sum_x \sum_c [(\zeta_{x,j} \cdot \zeta_{c,k})/N]^2 \quad (1)$$

We denote this statistic  $CV^2$ . Individuals with missing data are best left out of the analysis, as the calculation assumes there are the same numbers of data points per locus.

The significance of the  $CV^2$  value is assessed by comparing the original value to a null distribution generated by repeatedly randomizing the alleles at each locus between individuals and recalculating  $CV^2$  each time. In families arising from a single segregation, alleles are distributed at random at each locus, whereas they will be grouped within subfamilies in a mixed batch. Randomizing alleles removes this structure, and hence the associations between loci (measured by  $CV^2$ ) will disappear. The test statistic is therefore the proportion of randomization  $CV^2$  values that exceed the  $CV^2$  of the original family. We denote this proportion  $M$ .

## Simulations

We assessed the power of our approach with simulations, and concentrate on two main cases: batches containing offspring from two unrelated pairs of parents ('four-parent batches'), and those where the two subfamilies share one parent ('three-parent batches'). In both cases the parents of each subfamily are drawn from a population in Hardy-Weinberg equilibrium in which alleles are equiproportional within loci. A pair of parents then produces the desired number of offspring according to Mendelian inheritance, and the two subfamilies are melded together to produce a mixed batch.

For both three- and four-parent batches, we explored the effect of the number of individuals per batch, the number of marker loci and the number of alleles per locus on our ability to detect mixed families. For simplicity, two of these variables were held constant, and the third took on the values 2, 4, 6, 10 and 20. Batch size was kept at 10 individuals, allele number at four alleles and the number of marker loci at six when they were not being varied. Each simulated batch was randomized only 100 times, as more required prohibitive amounts of computer time. A batch was taken as mixed when the proportion of randomizations exceeding the original test statistic ( $M$ ) was less than 0.05. We generated 200 simulated batches for each combination of variables.

Since these simulations cannot cover all allele frequency distributions or numbers of loci, we recommend that the power of our method be assessed separately for each dataset. To this end, C++ code and Mathematica notebooks (Wolfram 1999) implementing both the randomization test and the power simulations are available from <http://helios.bto.ed.ac.uk/evolgen/>.

## Results and Discussion

The results of the simulations for the four-parent batches are presented in Table 1. Our method works well when there are more than six loci or six alleles per locus, except when the contributions of each subfamily are very skewed (90 : 10 columns). When one subfamily is represented by one or two offspring there is less opportunity to generate large values of  $CV^2$ , and hence less power to detect mixtures. Since one does not know the relative contributions of the subfamilies *a priori*, we recommend simulating a wide range of contributions for each batch in a data set. If there is adequate power when one subfamily is very small, there is a high probability that any type of mixture will be detected.

The simulation results for the three-parent batches are presented in Table 2. When the two subfamilies share a parent they will have more alleles in common, and this reduces the covariance in allelic state across the

**Table 1** Simulation results for the four-parent families

	% contribution 50 : 50					% contribution 70 : 30					% contribution 90 : 10				
	2	4	6	10	20	2	4	6	10	20	2	4	6	10	20
No. of loci	0.28	0.63	<b>0.83</b>	1	1	0.21	0.5	<b>0.83</b>	0.97	1	0.04	0.21	<b>0.31</b>	0.39	0.79
No. of alleles	0.49	<b>0.87</b>	0.95	1	1	0.35	<b>0.83</b>	0.96	0.98	1	0.07	<b>0.28</b>	0.29	0.35	0.48
No. of offspring	0	0.25	0.63	<b>0.88</b>	0.99	—	0.29	0.46	<b>0.82</b>	0.96	—	—	0.31	<b>0.38</b>	0.52

% contribution refers to the percentage of the batch contributed by each subfamily. Each row of data represents that variable taking the values 2, 4, ..., 20, the other two being held constant. When not being varied, number of loci was set at six, number of alleles at four and number of offspring at 10. The numbers are the proportion of simulated mixed batches detected as mixed when  $M < 0.05$ ; those in bold are replicates of the case where number of loci was six, number of alleles was four and number of offspring was 10 for each type of mixed family.

**Table 2** Simulation results for the three-parent families

	% contribution 50 : 50					% contribution 70 : 30					% contribution 90 : 10				
	2	4	6	10	20	2	4	6	10	20	2	4	6	10	20
No. of loci	0.14	0.25	<b>0.53</b>	0.72	0.96	0.11	0.2	<b>0.24</b>	0.57	0.92	0.03	0.06	<b>0.07</b>	0.19	0.27
No. of alleles	0.17	<b>0.51</b>	0.57	0.68	0.89	0.24	<b>0.35</b>	0.42	0.46	0.57	0.07	<b>0.1</b>	0.08	0.15	0.1
No. of offspring	0	0.06	0.3	<b>0.45</b>	0.78	—	0.05	0.2	<b>0.38</b>	0.54	—	—	0.13	<b>0.1</b>	0.15

% contribution refers to the percentage of the batch contributed by each subfamily. Each row of data represents that variable taking the values 2, 4, ..., 20, the other two being held constant. When not being varied, number of loci was set at six, number of alleles at four and number of offspring at 10. The numbers are the proportion of simulated mixed batches detected as mixed when  $M < 0.05$ ; those in bold are replicates of the case where number of loci was six, number of alleles was four and number of offspring was 10 for each type of mixed family.

batch as a whole. Nonetheless, with 50 : 50 contribution by each subfamily and more than 10 loci or 10 alleles per locus, the majority of three-parent mixed batches are detected. Even if a mixed family is missed, the consequences for the conclusions of subsequent analyses may not be serious: the subfamilies may be so similar that keeping them together has little impact on the outcome. However, this is best confirmed for each analysis by simulating the effect of including such batches. In general, retaining mixed batches will increase the frequency of heterozygous parents, and this may generate considerable bias in some cases.

These simulations only address three- and four-parent batches, and clearly more parents could be involved. Increasing the number of contributing parents for a given number of offspring is expected to decrease the proportion detected. This is most easily understood in the extreme case where each individual comes from a different pair of adults: if the adult population is in linkage equilibrium,  $CV^2$  is expected to be zero. Nonetheless, as long as more than three or four offspring represent each subfamily, our method should be able to detect mixed families produced by any number of parents.

To illustrate the relative merits of our approach over methods based on finding five or more alleles in a mixed batch, we simulated 100 mixed batches for a varying number of loci, alleles per locus, and number of offspring. We repeated these for equiprobable and skewed allele frequency distributions. The skewed frequencies were calculated so that each additional allele had 50% of the frequency of the previous allele. The simulation results are shown in Table 3. Similar results for the allele-counting method when one parent is known can be found in Neff & Pitcher (2002).

Our approach has several advantages over the allele-counting method. First, the  $CV^2$  statistic can be calculated for loci with fewer than five common alleles, which broadens the range of usable loci. Second, even if highly polymorphic loci are available, our method can detect mixed batches that were missed by the five-allele approach. This becomes more important when allele frequencies are skewed because the parents are more likely to carry only the common alleles. Nonetheless, since our approach and the allele-counting method use distinct aspects of the data to detect mixed batches, it is sensible to employ both when analysing batch genotypes.

**Table 3** Results for the simulations comparing the proportion of mixed batches detected using our  $CV^2$  method (upper numbers) with the five-allele method (lower numbers) for four-parent families with 50 : 50 contributions

	Equiprequent allele distribution					Skewed allele distribution				
	2	4	6	10	20	2	4	6	10	20
No. of loci	0.28	0.63	0.83	1	1	0.4	0.76	0.96	1	1
	0.80	0.95	0.98	1	1	0.56	0.79	0.89	0.97	1
No. of alleles	0.49	0.87	0.95	1	1	0.42	0.88	0.98	1	1
	—	—	0.99	1	1	—	—	0.91	1	1
No. of offspring	0	0.25	0.63	0.88	0.99	0	0.28	0.72	0.84	1
	—	0.81	0.96	0.99	0.99	—	0.62	0.83	0.9	0.93

As in Tables 1 and 2, each row of data represents that variable taking the values 2, 4, ..., 20, the other two being held constant. When not being varied, number of loci was set at six, number of alleles at six and number of offspring at 10. The skewed allele frequencies were calculated so that each additional allele is half as common as the previous allele.

### Conclusions

This paper presents a novel method for detecting when batches originate from three or more parents. It neatly complements techniques based on the number of alleles present in the batch (e.g. Neff & Pitcher 2002), especially when the available loci are not highly polymorphic. We anticipate that it will be most useful as a preliminary test prior to analyses that assume that each batch represents a single segregation.

### Acknowledgements

We thank Toby Johnson and Beate Nürnberger for many useful insights, and Brad Davis for help with C++ programming. Arianne Albert, Louis Bernatchez and two anonymous reviewers supplied helpful comments on an earlier version of this manuscript. T.V. was supported by a NERC studentship and N.B. was supported by NERC grant GR3/11635.

### References

Barton NH (2000) Estimating multilocus linkage disequilibria. *Heredity*, **84**, 373–389.

Emery AM, Wilson IJ, Craig S, Boyle PR, Noble LR (2001) Assignment of paternity groups without access to parental genotypes: multiple mating and developmental plasticity in squid. *Molecular Ecology*, **10**, 1265–1278.

Falconer DS, Mackay TFC (1996) *An Introduction to Quantitative Genetics*, 4th edn. Longman, Harlow, UK.

Neff BD, Pitcher TE (2002) Assessing the statistical power of genetic analyses to detect multiple mating in fishes. *Journal of Fish Biology*, **61**, 739–750.

Vines TH (2002) Migration, habitat choice and assortative mating in a *Bombina* hybrid zone. PhD Thesis, University of Edinburgh.

Wolfram S (1999) *The Mathematica Book*, 4th edn. Wolfram Media & Cambridge University Press, Cambridge.

deWoody JA, Walker D, Avise JC (2000) Genetic parentage in large half-sib clutches: theoretical estimates and empirical appraisals. *Genetical Research*, **75**, 95–105.

This work was part of T. H. V.'s PhD thesis on assortative mating in the fire-bellied toad hybrid zone in Romania. The test was originally developed to detect multiply parented egg batches, which were common in our study site. T. H. V. is currently studying assortative mating between sympatric stickleback species at the University of British Columbia, and N. H. B. continues to work on population genetics and speciation in Edinburgh.