Estimation of inbreeding coefficients from genotypic data on multiple alleles, and application to estimation of clonality in malaria parasites

WILLIAM G. HILL*, HAMZA A. BABIKER, LISA C. RANFORD-CARTWRIGHT AND DAVID WALLIKER

Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh, EH9 3JT

(*Received 5 July 1994 and in revised version 1 November 1994*)

Summary

Methods for estimating probability of identity by descent (f) are derived for data on numbers of genotypes at single loci and at pairs of loci with many alleles at each locus. The methods are general, but are specifically applied to data on genotype frequencies in zygotes of the malaria parasite sampled from its mosquito host in order to find the extent of outcrossing in the parasite and the degree of clonality in populations. It is assumed that zygotes are the outcome either of gametes of the same clone, in which they are identical at all loci, or are products of two random, unrelated clones. From the estimate of f an effective number of clones per human host can also be derived. For *Plasmodium falciparum* from a Tanzanian village, estimates of f are 0·33 from data on zygote frequencies at two multiallelic loci, indicating that two-thirds of zygotes produce recombinant types.

1. Introduction

Inbreeding as a consequence of selfing or of mating among relatives leads to an increase in homozygosity (Wright, 1921; e.g. Hartl & Clark, 1989). From the homozygote excess, the amount of inbreeding and, for example, the selfing rate can be estimated (Weir, 1990). Such information is important in understanding the population biology of infectious agents such as parasitic protozoa, which undergo gamete formation and fertilization during their life cycle.

In the human malaria parasite, *Plasmodium falciparum*, sexual reproduction occurs in the mosquito host. Gametes are derived from gametocytes ingested in a single blood meal from a human host, and the insect then transmits haploids to man when it subsequently feeds. Many people in malaria endemic areas are infected with mixtures of genetically distinct clones (Thaithong *et al.* 1984; Creasey *et al.* 1990). As a person can be infected by mosquitoes on many occasions, or by different haploid clones from one mosquito which has itself ingested a mixture of clones, he or she may be infected by one or more clones. If the mosquito in its meal ingests gametocytes derived from a single clone, then the diploids formed are completely homozygous. Further, if more than one clone is ingested and mating is random among clones, the probability that diploids are homozygous is a function of the frequency of the ingested clones. Most simply, with random pairing and two clones of equal frequency, half of the diploids are inbred and half are crosses (Ranford-Cartwright *et al.* 1993). For further description of the biology and assumptions, see Babiker *et al.* (1994).

Meiosis in zygotes formed from gametes of two clones probably produces new variants. The extent of crossing is important in understanding the level of variability in this parasite and has important implications for planning control measures based on vaccines and chemotherapy. The extent of cross-mating is controversial: one view, based on the genotype diversity and frequency and on the evidence of meiosis, is that it is quite frequent in nature (Walliker, 1989; Conway & McBride, 1991; Babiker *et al.* 1991); the alternative view is that natural infections are derived from one or a few genotypes which undergo mating infrequently, so that reproduction is mainly clonal (Tibayrenc *et al.* 1991). The genotypes of oocysts, each of which is the product of a single zygote in an infected mosquito, can be obtained for marker loci, and the homozygosity used to infer the frequency with which they have derived from gametes from a single clone (Ranford-Cartwright *et al.* 1993; Babiker *et al.* 1994). Using this information Babiker *et al.* (1994) were then able to estimate the average number of clones harboured by individual human hosts in a

* Corresponding author.

https://doi.org/10.1017/S0016672300033000 Published online by Cambridge University Press

population in a Tanzanian village. They obtained data on two highly polymorphic single copy genes, MSP-1 and MSP-2, both varying in numbers of tandem DNA repeats (i.e. as VNTR loci) and in sequence differences identified by hybridization to specific probes. Most of their analysis was based only on the sequence differences, giving only three and two alleles at the respective loci, whereas if both fragment length and sequence differences had been taken into account, there would have been 15 or more alleles identified at each locus. It is reasonable to assume that more information could be obtained by using the multiple allelic data.

The aims of this paper are therefore twofold: to extend methods for estimation of inbreeding coefficients at loci with multiple alleles, and to use these to analyse the data of Babiker *et al.* (1994) to obtain estimates of the clonality rate of *P. falciparum* malaria in a Tanzanian village population.

2. Estimation of inbreeding coefficient

(i) *Single locus*

Consider a locus A at which the alleles A_i have frequencies p_i. If the inbreeding coefficient, i.e. the probability that the parental gametes are identical by descent, is f and there is random association among non-identical allele types, then the expected genotype frequencies, Q_{ij}, are:

$$\text{homozygotes:} \quad Q_{ii} = fp_i + (1-f)\,p_i^2, \qquad (1a)$$

$$\text{heterozygotes:} \quad Q_{ij} = 2(1-f)\,p_i p_j, \quad \text{where} \quad i<j. \qquad (1b)$$

Assume a sample of n zygotes are taken, each from a different mosquito such that they can be considered independently sampled genotypes, and that all genotypes can be identified (i.e. no dominance or nulls). Of these n_{ii} are A_{ii} homozygotes and n_{ij} are A_{ij} heterozygotes, and the total number of A_i alleles is $n_i = 2n_{ii} + \Sigma_{i \neq j} n_{ij}$. There are k alleles represented in the sample. The inbreeding coefficient and allele frequencies can be estimated by maximum likelihood (ML). With two alleles, the ML estimates are obtained by equating observation to expectation, giving estimates

$$p_1 = n_1/2n, \quad p_2 = n_2/2n, \quad f = 1 - 2n_{12}/(n_1 n_2). \qquad (2)$$

For more than two alleles an iterative procedure is necessary in order to maximize the likelihood and obtain the ML estimates of the parameters. One method is that of Robertson & Hill (1984), but it is more simple to apply the Expectation–Maximization (EM) algorithm (Dempster *et al.* 1977), initially used in similar genetic applications as 'gene counting' (Smith, 1957). See Weir (1990) for further explanation. Basically the EM method as applied to the present problem involves assigning homozygous genotypes into identical and non-identical types according to the current estimates of allele frequencies and f, and then repeating the process until estimates converge. The choice of starting values for this EM iteration is not critical as convergence is rapid. Convenient starting values are:

$$p_i = n_i/2n, \quad i = 1, \ldots, k \qquad (3)$$

and, equating homozygote numbers to expectation,

$$f = (\Sigma_i n_{ii} - n\Sigma_i p_i^2)/[n(1-\Sigma_i p_i^2)]. \qquad (4)$$

Iteration then proceeds, taking steps (5a), (5b), (5c), (5a), … in turn, in each case replacing previously by newly computed values, until convergence is reached:

$$x_i = fn_{ii}/[f + p_i(1-f)], \quad i = 1, \ldots, k \qquad (5a)$$

where x_i is the expected number of homozygotes for A_i generated from identical alleles and thus $n_i - x_i$ is the expected number of independent A_i alleles. Thus

$$p_i = (n_i - x_i)/(2n - \Sigma_i x_i), \quad i = 1, \ldots, k, \qquad (5b)$$

$$f = \Sigma_i x_i/n. \qquad (5c)$$

The (natural) log likelihood, $L = \Sigma\Sigma_{i \leqslant j} n_{ij} \ln(Q_{ij}) +$ constant terms involving binomial coefficients, when f is fitted, is given by

$$\begin{aligned} L_f &= \Sigma_i n_{ii} \ln[fp_i + (1-f)\,p_i^2] \\ &\qquad + \Sigma\Sigma_{i<j} n_{ij} \ln[2(1-f)p_i p_j)] + \text{const.} \\ &= \Sigma_i n_{ii} \ln\{1 + f/[p_i(1-f)]\} + n \ln[2(1-f)] \\ &\qquad + \Sigma_i n_i \ln p_i + \text{const.} \end{aligned} \qquad (6)$$

and the reduction in L_f as f is changed about its ML value (and the allele frequencies re-estimated to give a profile likelihood) can be used to construct a support range or confidence interval for f. Here, invoking the χ^2 approximation for natural log likelihoods, a reduction of two is used to give a range equivalent to a 95% confidence interval. A test for $f > 0$ can be obtained from the likelihood ratio, by comparing $2(L_f - L_0)$ to χ^2 with 1 D.F., where L_0 is the log likelihood with $f = 0$. For the case of multiple alleles where some or many genotypes have small expected numbers, an exact test is preferable, and this can be performed by permutation (Guo & Thompson, 1992).

While it is straightforward to estimate f and to test whether it departs significantly from zero, a much harder problem in the case of multiple alleles is to test whether the model with $f > 0$ fits the data adequately. If L_n denotes the log likelihood where genotype numbers are equated to expectations as for a perfect fit, i.e. $Q_{ij} = n_{ij}/n$, then $2(L_n - L_f)$ is expected to have a χ^2 distribution with $(k-2)(k+1)/2$ D.F. if the χ^2 approximation to likelihood ratio holds. With 60 observations and 15 alleles, as in one example in the malaria data to be analysed here, there are therefore 104 residual D.F., greatly in excess of the number of observations. Simulation was undertaken for some examples to investigate the properties of this residual likelihood statistic. Samples of $n = 60$, the size of the malaria data set, were sampled with k alleles either

https://doi.org/10.1017/S0016672300033000 Published online by Cambridge University Press

Table 1. *Simulation to examine properties of likelihood test of fit of model to data. Replicate samples of $n = 60$ with k alleles were taken with either equal or unequal population frequencies and specified f, and analysed by maximum likelihood*

k		Equal frequency			Unequal frequency		
(D.F.)	f	$E(f)$	$E[2(L_n-L_f)]$	$V[2(L_n-L_f)]$	$E(f)$	$E[2(L_n-L_f)]$	$V[2(L_n-L_f)]$
4	0·2	0·206	5·4	13·5	0·183	4·6	7·9
(5)	0·4	0·386	5·0	10·0	0·386	4·6	7·0
	0·6	0·593	5·4	12·5	0·584	4·6	5·4
6	0·2	0·191	15·4	35·7	0·189	11·7	17·6
(14)	0·4	0·394	16·9	38·2	0·400	11·6	15·8
	0·6	0·598	16·5	25·2	0·589	11·2	15·2
10	0·2	0·190	51·7	66·1	0·190	33·1	43·1
(44)	0·4	0·395	49·6	52·9	0·395	32·7	43·5
	0·6	0·590	44·4	36·2	0·590	29·5	40·5
15	0·2	0·195	102	79·8	0·186	67·3	67·7
(104)	0·4	0·395	92·0	74·8	0·392	64·2	71·3
	0·6	0·593	76·2	77·2	0·590	56·4	70·8
24	0·2	0·191	177	78·5	0·183	145	124
(275)	0·4	0·387	155	136	0·378	130	133
	0·6	0·588	123	197	0·584	108	134

Simulation was undertaken with either equal population gene frequencies, $p_i = 1/k$, or with unequal frequencies with p_i having a geometric distribution with parameter $b = 0{\cdot}25$, 0·42, 0·64, 0·77 and 0·88 for $k = 4$, 6, 10, 15 and 24, the value of b being chosen to give a high chance of having k alleles in the sample. Replicates were included only if there were k alleles in the sample. 400 replicates were taken, but only 100 with 15 and 24 alleles and unequal frequencies (most samples did not have the required k). Standard deviations of estimates of f among replicates were approximately 0·07, depending on allele number and f.

with equal frequencies or with frequencies sampled from a geometric distribution, in which the frequency of the j allele was b times that of the $j-1$ allele. The value of b was chosen to give a high probability that the sample would contain k alleles, for analyses of sample data were undertaken only on samples with k alleles. Results are shown in Table 1. It is seen that the χ^2 distribution, assessed in terms of mean (expectation = D.F. for χ^2) and variance (expectation = 2 D.F.) performs satisfactorily with few alleles, none of them rare, but in the multiple allele situation with unequal frequencies, $2(L_n-L_f)$ tends to be much smaller than the χ^2 expectation. Therefore spurious fits might be inferred. No suitable alternative, preferably exact, test has been devised for testing whether the model gives a satisfactory fit to the data.

(ii) *Two loci*

If data are available on two loci, as in the case of the data of Babiker *et al.* (1994), f can be estimated from data for each locus separately. In the model used here in which matings are either clonal or outcrosses, combining gametes are therefore identical at both loci or neither loci. Estimates of inbreeding coefficient from each locus are therefore correlated and can not readily be combined. The maximum likelihood approach is to use the genotypic data on the two loci in a single analysis. The simplest assumption is that the loci are in linkage equilibrium; this is reasonable when the loci are unlinked as Triglia *et al.* (1992) showed for those used by Babiker *et al.* (1994). It is feasible to avoid or to test this assumption when there are only two or three alleles at each locus, and thus only one (if both loci have two alleles) or a few D.F. for disequilibrium. If, however, there are many alleles at either locus (196 D.F. if two loci each have 15 alleles), there are likely to be more D.F. needed to fit the disequilibrium than there are data points and the estimation and, certainly, testing are not likely to be worthwhile. The analysis assuming linkage equilibrium is given first and that allowing for disequilibrium follows.

(*a*) *Linkage equilibrium*

Assume the two loci A and B have alleles A_i and B_j, with frequencies in the population of p_i and q_j, and k_A and k_B alleles, respectively, present in the sample. In the sample of n individuals there are n_{hijk} of genotype $A_{hi}B_{jk}$, where $i \geqslant h$ and $k \geqslant j$. The total number of alleles of types A_i and B_j carried by individuals in the sample are $n_{i.}$ and $n_{.j}$, respectively. The expected genotype frequencies are:

$$A_{ii}B_{jj}\colon\ Q_{iijj} = fp_iq_j + (1-f)p_i^2q_j^2, \qquad (7a)$$

$$A_{ii}B_{jk}\colon\ Q_{iijk} = 2(1-f)p_i^2q_jq_k, \qquad (7b)$$

$$A_{hi}B_{jj}\colon\ Q_{hijj} = 2(1-f)p_hp_iq_j^2, \qquad (7c)$$

$$A_{hi}B_{jk}\colon\ Q_{hijk} = 4(1-f)p_hp_iq_jq_k. \qquad (7d)$$

The EM algorithm can be used to estimate f and the allele frequencies simultaneously for the two loci by simple extension of the method for one locus. Suitable initial estimates of allele frequencies are as for a single locus (eqn 3) and that for f from the excess of homozygotes over expectation:

$$f = (\Sigma_i \Sigma_j n_{iijj} - n\Sigma_i \Sigma_j p_i^2 q_j^2)/[n(1-\Sigma_i \Sigma_j p_i^2 q_j^2)]. \qquad (8)$$

Iteration proceeds in steps (9*a–d*)

$$x_{ij} = (f n_{iijj})/[f + p_i q_j (1-f)], \qquad (9a)$$

where here and subsequently evaluation and summation are over $i = 1, \ldots, k_A$ and $j = 1, \ldots, k_B$ as appropriate, with sums represented by

$$x_{i.} = \Sigma_j x_{ij}, \quad x_{.j} = \Sigma_i x_{ij}, \quad x_{..} = \Sigma_i x_{i.},$$

$$p_i = (n_{i.} - x_i)/(2n - x_{..}), \qquad (9b)$$

$$q_j = (n_{.j} - x_{.j})/(2n - x_{..}), \qquad (9c)$$

$$f = x_{..}/n. \qquad (9d)$$

The log likelihood is given by $L = \Sigma\Sigma\Sigma\Sigma n_{hijk} \ln(Q_{hijk})$ + constant terms, a simple extension of that for one locus, inserting estimated genotype frequencies. As for fitting single loci, three likelihoods can be considered: L_0 for $f = 0$, L_f for $f \neq 0$, and L_n for a complete fit to the data.

(*b*) *Linkage disequilibrium*

On the assumption that coupling and repulsion heterozygotes can not be distinguished, it is necessary to estimate simultaneously the disequilibrium and allele frequencies or, more simply, to estimate directly the haplotype frequencies, v_{ij} of haplotype $A_i B_j$. This is a simple extension of the method for equilibrium above. The expected genotypic frequencies are

$$Q_{iijj} = f v_{ij} + (1-f) v_{ij}^2, \qquad (10a)$$

$$Q_{iijk} = 2(1-f) v_{ij} v_{ik}, \qquad (10b)$$

$$Q_{hijj} = 2(1-f) v_{hj} v_{ij}, \qquad (10c)$$

$$Q_{hijk} = 2(1-f)(v_{hj} v_{ik} + v_{hk} v_{ij}). \qquad (10d)$$

An EM procedure to estimate f is given in Appendix 1. If the likelihood fitting this model is denoted L_{fD}, and there are sufficiently few D.F. fitted that χ^2 can be assumed under the null hypothesis, a test of whether there is linkage disequilibrium is given by $2(L_{fD} - L_f)$, against χ^2 with $(k_A - 1)(k_B - 1)$ D.F.

(*c*) *Single/two locus identity*

In the previous analysis it has been assumed that the only kinds of mating are between identical haploid clonal members or between unrelated clones. More generally, there are breeding systems in which this does not apply, a simple example being the case of random selfing and outcrossing in a monoecious plant. In that case, there are three kinds of double homozygotes for a pair of loci: identical (by descent) at both loci, or identical at one locus, or identical at neither. In general, the probability that a pair of unlinked or partially linked loci are both identical by descent is not equal to the probability that genes at a single locus are identical, as is assumed in eqn (7) and (10). This could also apply, as discussed later, in the malaria situation, if two clones carried by one human host were the product of a single infection by a mosquito, perhaps itself derived from one human host, or by infections from different mosquitoes, independently infected. The model is therefore generalized to allow for different probabilities of identity by descent: f is the conventional identity by descent at a single locus, and it is convenient to define two new parameters: g, the probability of identity by descent at both loci, and h, the probability of identity by descent at locus A but not at B and, assumed to be the same, the probability of identity at B but not at A. Hence $f = f + h$. In the above models (7) and (10), it is assumed that $h = 0$. In the notation of Weir & Cockerham (e.g. 1973), g is equivalent to F_{11}, h to F_{01} and F_{10}, and the probability of identity at neither locus, $1 - g - 2h$, to F_{00}.

Genotype frequencies, assuming linkage equilibrium, are given by

$$Q_{iijj} = g p_i q_j + h(p_i^2 q_j + p_i q_j^2) + (1-g-2h) p_i^2 q_j^2, \qquad (11a)$$

$$Q_{iijk} = 2h p_i q_j q_k + 2(1-g-2h) p_i^2 q_j q_k, \qquad (11b)$$

$$Q_{hijj} = 2h p_h p_i q_j + 2(1-g-2h) p_h p_i q_j^2, \qquad (11c)$$

$$Q_{hijk} = 4(1-g-2h) p_h p_i q_j q_k. \qquad (11d)$$

A procedure for estimating g and h is given in Appendix 2. Let the log likelihood fitting this model be L_{gh}. Under the assumption of clonality and random mating used above, $g = f$ and $h = 0$ and the statistic $2(L_{gh} - L_f)$ has a χ^2 distribution with 1 D.F., at least in large samples. This therefore provides some test of the model assumptions without the problem of many more D.F. than observations discussed above. Extension of the analysis to include linkage disequilibrium is straightforward, and such might be undertaken if either model (10) or (11) does not fit.

For a model of random selfing and outcrossing with probability of selfing equal to s, then f and g asymptote at $f = s/(2-s)$ and $g = [s(2+s)]/[(2-s)(4-s)]$ (Weir & Cockerham, 1973). A test for this equilibrium is obtained by fitting a single parameter with these respective values, but some of the simple structure of the EM estimates is lost.

(iii) *Clonality*

The inbreeding coefficient, f, gives the probability that the two uniting malaria gametes in the mosquito are identical. Assuming that these gametes come from the same human host, the gametes are identical if the person carries just one malaria clone, or carries more than one but by chance the two gametes are sampled from the same clone. Thus it seems appropriate to define an effective number of parasite clones per host as $n_e = 1/f$. For example, if all individuals carry two clones and these are equally represented, then $n_e = 2$, and if the two clones are in the relative proportions 1/3 and 2/3, then $f = 5/9$ and $n_e = 1{\cdot}8$. Alternatively, if individuals are equally likely to carry one clone or two equally frequent clones, $f = 3/4$ and $n_e = 1{\cdot}33$.

https://doi.org/10.1017/S0016672300033000 Published online by Cambridge University Press

3. Analysis of malaria data

(i) *Fitting few alleles*

Let us consider first the analysis utilizing the data on only the sequence differences identified by PCR and sequence-specific probes, as was done by Babiker *et al.* (1994). The data are shown in Table 2 (from their Table 1). Results of the ML analysis are shown in Table 3, with estimates of f of 0·15 and 0·39 from the two loci separately, and 0·23 from a simultaneous fit, assuming linkage equilibrium, a value which is well within the support limits for the estimate from the separate loci. The presence of inbreeding, or departure from Hardy–Weinberg expectations, is shown by the comparison $L_f - L_0$, which is highly significant ($P < 0{\cdot}01$) for the fit of two loci. There is no evidence that the data are not well fitted by the model including inbreeding, for the value of $2(L_n - L_f)$ is close in value to the D.F., as expected from χ^2. The estimate of effective number of clones, n_e, is 4·26 from the estimate of f obtained by fitting both loci.

Table 2. *Numbers of genotypes in oocysts of alleles identified as sequence differences at two loci, MSP-1 and MSP-2*

	MSP-2 Genotype[a]			
MSP-1 Genotype[a]	1/1	1/2	2/2	Total
1/1	8	4	5	17
1/2	3	5	4	12
1/3	5	7	5	17
2/2	3	0	3	6
2/3	2	1	0	3
3/3	4	1	0	5
Total	25	18	17	60

[a] For: MSP-1 alleles 1, 2 and 3 are K1, MAD20 and RO33 types, respectively; MSP-2 alleles 1 and 2 are ICI and FC27 types, respectively.

For the two locus data, model (10) including linkage disequilibrium was fitted. There was no significant disequilibrium: $2(L_{fD} - L_f) = 1{\cdot}548$ with 1 D.F. The model (11) with different single and two locus identity was also fitted. The likelihood was maximized at $h = 0$, as would be expected if the primary model of only clonal or non-clonal matings applied.

(ii) *Fitting multiple alleles*

Utilizing both fragment length and sequence difference, H. A. Babiker (personal communication) provided data on 15 alleles at MSP-1 and 24 at MSP-2. It is not feasible to tabulate all the genotypic data, but those for MSP-1 are shown in Table 4 so as to give some feel for the data, and a summary is provided in Table 5. It is noteworthy that, despite there being only 60 zygotes, 20 were homozygous at both loci, indicative of substantial inbreeding. Only one two-locus genotype occurred more than once in the sample. Estimates of parameters are given in Table 5, again assuming linkage equilibrium in the two locus case. Estimates of f are higher and of effective number of clones correspondingly lower than when only a few alleles are fitted, whether loci are fitted singly or together. The likelihood ratios, $2(L_f - L_0)$, indicate highly significant inbreeding in each case. This was confirmed ($P < 0{\cdot}001$) by a Monte Carlo permutation test for each of the loci separately (Guo & Thompson, 1992). No equivalent test was performed for the loci together, but since less than one double homozygote would be expected with $f = 0$ and 20 were found, significance can be assumed. The support intervals for f are generally rather smaller with multiple than with two alleles. The difficulty is in testing whether the model fitted: for each locus separately, the residual

Table 3. *Estimation of f and of effective number of clones per host, n_e, from data on sequence differences with limited numbers of alleles on loci separately and jointly*

Locus	MSP-1	MSP-2	Joint[a]
No. of alleles	3	2	—
Estimated allele frequencies	0·5292	0·5667	—
	0·2191	0·4333	—
	0·2517	—	—
ML estimate of f	0·1492	0·3891	0·2351
Interval for f			
Lower[b]	0	0·136	0·093
Upper[b]	0·346	0·608	0·393
$2(L_f - L_0)$ (D.F.)	2·66 (1)	9·27 (1)**	13·00 (1)**
$2(L_n - L_f)$ (D.F.)	3·06 (2)	0 (0)	16·20 (13)
ML estimate of n_e	6·70	2·57	4·26

[a] Assuming linkage equilibrium.
[b] Reduction of 2 in log likelihood.
** $P < 0{\cdot}01$, $P > 0{\cdot}05$ otherwise.

https://doi.org/10.1017/S0016672300033000 Published online by Cambridge University Press

Table 4. *Genotype numbers at the MSP-1 locus of sampled oocysts*

				Allele														
Allele	Seq[a]	Bin[b]	$n_{i.}$[c]	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	K	600	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
2	K	580	1		0	0	0	0	0	0	0	0	0	0	0	1	0	0
3	K	560	3			0	0	0	0	0	0	0	0	1	0	0	0	2
4	K	540	17				4	0	0	0	0	0	0	3	0	0	0	6
5	K	520	19					5	0	0	0	0	1	3	0	0	2	3
6	K	500	5						2	0	0	1	0	0	0	0	0	0
7	K	480	9							3	0	0	0	0	0	1	0	2
8	K	460	7								2	0	0	1	0	0	0	2
9	K	410	1									0	0	0	0	0	0	0
10	M	560	3										1	0	0	0	0	0
11	M	520	21											5	0	0	0	2
12	M	500	1												0	0	0	1
13	M	470	2													0	0	0
14	R	490	2														0	0
15	R	470	28															5

[a] K, M and R denote sequences K1, MAD20 and RO33 types, respectively.
[b] Bin denotes fragment size group.
[c] Number of alleles of type *i*.

Table 5. *Estimation of f from data on sequence and repeat differences with multiple alleles on loci separately and jointly*

Locus	MSP-1	MSP-2	Joint[a]
No. of alleles	15	24	—
No. of homozygotes	27	34	20
Estimated allele frequency range			
Low	0·0102	0·0114	—
High	0·2498	0·1278	—
ML estimate of f	0·3643	0·5404	0·3276
Interval for f			
Lower[b]	0·223	0·404	0·214
Upper[b]	0·512	0·671	0·456
$2(L_f - L_0)$ (D.F.)	41·5 (1)**	117·5 (1)**	114·7 (1)**
$2(L_n - L_f)$ (D.F.)	72·8 (104)	130·4 (275)	538·2 (35971)
ML estimate of n_e	2·74	1·98	3·05

[a] Assuming linkage equilibrium.
[b] Reduction of 2 in log likelihood.
** $P < 0·01$, $P > 0·05$ otherwise.

doubled log likelihood, $2(L_n - L_f)$, is somewhat less than the residual degrees of freedom; for the pair of loci, the residual is 538 with 35971 D.F.! The problems illustrated in Table 1 for single loci with multiple alleles are exacerbated.

4. Discussion

The methods for estimating inbreeding coefficients using data on zygote frequencies at loci with multiple alleles described here are a simple extension of standard methods, but seem not to have been derived previously. Applications are not restricted solely to the estimation of clonality in malaria, but can be used in any population in which there may be non-random mating, for example in a species which reproduces by a mixture of selfing and outcrossing, such as barley or *Phlox cuspidata*, and in which the outcrossing rate is to be estimated (Levin, 1978). Among other parasitic protozoa, trypanosome species, particularly *Trypanosoma brucei*, also appear to undergo selfing and crossing during their cycle in the tsetse-fly vector *Glossina* (Tait & Turner, 1990). Other aspects, such as the sampling errors of estimates of f are discussed by Robertson & Hill (1984), whose methods of estimation were more complicated and less general .

The ML estimates of f are slightly biassed downwards (most easily explained for a sample of one zygote: if the locus is segregating so f can be estimated, then the sample comprises only a heterozygote and the estimate is $f = -\frac{1}{2}$), by an amount inversely proportional to the sample size. For samples of 60, the size of the present data set, the bias is negligible relative to f for values as large as 0·2 or more, as obtained here. There is, however, an unresolved problem of testing whether the model of mixed clonal and non-clonal matings fits the data when there are many alleles, with residual degrees of freedom well in excess of observations. This is not a problem unique to genetics; but in studies of alternative methods for testing goodness-of-fit (Koehler & Larntz, 1980; Simonoff, 1985), much less sparse situations have been considered than those found here, particularly when there are data on two loci. An approach would, perhaps, be to pool cells in a logical manner before testing fit, so as to increase expected numbers in each class; but when there are many thousands of classes and under 100 observations, any pooling is liable to be an arbitrary process. The main novelty in the analysis is the extension to enable simultaneous estimation of probabilities of identity at one and at two loci, which enable some models, such as that of simply clonal or non-clonal mating of the malaria parasite in its mosquito host, to be tested using a single degree of freedom.

The malaria parasite depends entirely on the mosquito vector for transmission between hosts, and fusion of gametes in the mosquito is an obligatory phase of the life-cycle. In order to understand the genetic structure of malaria populations it is essential to study genotypes of individual zygotes in wild-caught mosquitoes, but techniques for examining this stage have only recently been developed (Ranford-Cartwright *et al.* 1991). In previous studies, alleles of polymorphic loci of *P. falciparum* in samples of blood from patients in malaria-endemic areas have been examined (e.g. Carter & McGregor, 1973; Conway & McBride, 1991). Such studies have demonstrated clearly that most patients are infected with mixtures of different parasite genotypes and provided strong circumstantial evidence that random mating between clones occurs, but the actual extent of crossing and selfing could not be quantified.

Among the gametocytes of malaria parasites there is usually an excess of females, typically six to eight times more frequent than males. It has been argued that the evolutionary stable sex ratio is related to the inbreeding coefficient (Read *et al.* 1992; Dye & Godfray, 1993). As the sex ratio has not yet been estimated in the Tanzanian population analysed here and the inbreeding coefficient from oocysts not estimated in others, it is not yet possible to test their predictions. Using Dye & Godfray's predictions, a value of f of 0·33 (Table 5) would correspond to a proportion of male gametocytes of $(1-f)/2 = 0{\cdot}33$.

Natural *P. falciparum* populations are not homogeneously distributed within the infected individuals of a community but substructured to varying extent, dependent on many epidemiological factors such as the frequency with which inhabitants are bitten by infected mosquitoes, the proximity of their houses to mosquito breeding sites, seasonality of malaria transmission, and clone-specific immune responses of the human host. An individual might be infected by a single or by many haploid clones, which would affect mating patterns among the parasites. The oocyst data used here are the first to be obtained in any malarious region. The village of Michenga is in a highly malaria-endemic area of Tanzania, in which individuals may be exposed to as many as 300 infectious bites per year (Smith *et al.* 1993). In areas with lower infection rates, lower heterozygosity and higher inbreeding coefficients of the malaria parasite would be observed, and lower numbers of clones per person inferred.

Estimates of inbreeding were higher when multiple rather than two- or three-allele data were used: the reason is not clear, but most likely it is just due to chance. For example there is no noticeable relationship between allele number and estimate of f shown in the simulations in Table 1. As a consequence, estimates of effective number of clones are lower using multiple alleles (Table 5) than two or three alleles (Table 3). The highest value, close to 3·0, for multiple alleles comes from fitting the loci simultaneously, and this has to be regarded as the best estimate in that it uses all the data available. The relationship between the effective and actual number of clones depends on assumptions made about the distribution of numbers of clones and of the relative abundance of different clones in people who carry more than one.

Babiker *et al.* (1994) fitted a geometric distribution to estimate the number of clones carried per person and assumed equal frequencies of clones within hosts. Using the sequence data only, they obtained an estimate of mean number of clones of 3·2, close to the value for the effective number obtained here from the multiple allele data. An estimate of numbers of (haploid) genotypes per person was also obtained directly by Babiker *et al.* by PCR analysis to detect sequence differences in blood samples from members of the village studied. The mean was estimated to be in the range 2·2 and 3·2, according to whether the minimum or maximum number of two-locus haplotypes were counted: for example, if alleles A_1, A_2, B_1 and B_2 are detected, these could be just haplotypes A_1B_1 and A_2B_2, or could also include A_1B_2 and/or A_2B_1. The number of different types detected in this way is, however, an underestimate of the number of different clones, because there are, in effect, only six different sequence haplotypes for two loci, MSP-1 having three alleles and MSP-2 having two. For the allele frequencies in the present data (Table 2) and assuming linkage equilibrium, the probability that two unrelated clones are of the same type is $\Sigma_i \Sigma_j v_{ij}^2 = \Sigma_i p_i^2 \Sigma_j p_j^2 = 0{\cdot}191$. This assumes that all bloodform parasites are represented in the gametocyte population, which may not be the case. Nevertheless there is reasonable agreement between the observed numbers in the human host and the values inferred from homozygosity of oocysts in the mosquito host.

Such estimates of numbers of clones per host are based on the assumption that the uniting gametocytes (gametes) of the parasite in the mosquito come from a single human host; if they can come from more than one person, who may carry different clones, then the numbers of clones per host will be lower than indicated from f or n_e. Mixed blood-feeding is likely to come from interrupted feeding, but few studies of its frequency have been carried out. Boreham *et al.* (1979) typed blood meals for haptoglobin-type from *A. gambiae* caught in houses in a Nigerian village, and concluded that the incidence of mixed feeding was of the order of 10%. Not all contributions to mixed feeds would be from infected individuals, however, so the contribution of mixed feeding to heterozygosity may be small. In Papua New Guinea, Burkot *et al.* (1988) found that 13% of anophelines captured in houses inhabited by two people with different ABO blood groups had fed on them. The development of DNA profiling of such blood meals (Gokool *et al.* 1993) should provide more precise information on this subject.

The extent of outcrossing $(1-f)$ of over 50% shown by *P. falciparum* in this region of Tanzania has important implications for control of malaria. In particular, new antigenic types may be formed and recombination can occur between drug resistant genes with the risk that multiple resistance genotypes will be produced, particularly in areas where multiple anti-malarial regimes are in common use. Thus there may be benefits in using a mixture of drugs (Curtis & Otoo, 1986) in the short term when resistant types are rare for each drug, although in the long term that may not be the case.

This work was supported by the Medical Research Council. We are grateful to Bruce Weir and Margaret Mackinnon for helpful comments.

Appendix

(i) *Fitting linkage disequilibrium*

Initial allele frequencies can be computed from (3) and haplotype frequencies as $v_{ij} = p_i q_j$, as for linkage equilibrium. A corresponding initial value for f is given by (8). Iteration proceeds straightforwardly, the only addition being that the double heterozygotes, which are of two types, have to be apportioned differently:

(a) $x_{ij} = n_{iijj} f/[f+(1-f)\,v_{ij}]$,

(b) $v_{ij} = [2n_{iijj} - x_{ij} + \Sigma_{k\neq j} n_{iijk} + \Sigma_{h\neq i} n_{hijj}$
$\quad + \Sigma_{h\neq i}\Sigma_{k\neq j}(n_{hijk}\, v_{ij}\, v_{hk})/(v_{hj}\, v_{ik} + v_{hk}\, v_{ij})]/(2n - x_{..})$

(c) $f = x_{..}/n$.

(ii) *Fitting single and two locus identities*

Initial allele frequencies can be computed from (3), and initial values of g and h by, for example, computing f from (4) and taking $g = h = f/2$. Iteration proceeds as follows, where it is convenient to compute a number of intermediate quantities and dots denote summation:

(a) $y_{iijj} = n_{iijj}/[g + h(p_i + q_i) + (1-g-2h)\,p_i\, q_j]$,
$\quad z_{iijk} = n_{iijk}/[h + (1-g-2h)\,p_i]$,
$\quad z_{hijj} = n_{hijj}/[h + (1-g-2h)\,q_j]$,
$\quad Y_{ij} = g y_{iijj}$,
$\quad Z_{Ai} = \Sigma_j\, h q_j\, y_{iijj} + \Sigma\Sigma_{k\neq j}\, h z_{iijk}$,
$\quad Z_{Bj} = \Sigma_i\, h p_i\, y_{iijj} + \Sigma\Sigma_{i\neq h}\, h z_{hijj}$,

(b) $p_i = (n_{i.} - Y_{i.} - Z_{Ai})/(2n - Y_{..} - Z_{A.})$,

(c) $q_j = (n_{.j} - Y_{.j} - Z_{Bj})/(2n - Y_{..} - Z_{B.})$,

(d) $g = Y_{..}/n$,

(e) $h = (Z_{A.} + Z_{B.})/(2n)$.

References

Babiker, H. A., Creasey, A. M., Fenton, B., Bayoumi, R. A. L., Arnot, D. E. & Walliker, D. (1991). Genetic diversity of *Plasmodium falciparum* in a village in eastern Sudan. 1. Diversity of enzymes, 2D-PAGE proteins and antigens. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **85**, 572–577.

Babiker, H. A., Ranford-Cartwright, L. C., Currie, D., Charlwood, J. D., Billingsley, P., Teuscher, T. & Walliker, D. (1994). Random mating in a natural population of the malaria parasite *Plasmodium falciparum*. *Parasitology* **109**, 413–421.

Boreham, P. F. L., Lenahan, J. K., Boulzaguet, R., Storey, J., Ashkar, T. S., Nambiar, R. & Matsushima, T. (1979). Studies on multiple feeding by *Anopheles gambiae* s.l. in a Sudan savanna area of north Nigeria. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **73**, 418–423.

Burkot, T. R., Graves, P. M., Paru, R. & Lagog, M. (1988). Mixed blood feeding by the malaria vectors in the *Anopheles punctulatus* complex (Diptera: Culicidae). *Journal of Medical Entomology* **25**, 205–213.

Carter, R. & McGregor, I. A. (1973). Enzyme variation in *Plasmodium falciparum* in The Gambia. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **67**, 830–837.

Conway, D. J. & McBride, J. S. (1991). Population genetics of *Plasmodium falciparum* within a malaria-hyperendemic area. *Parasitology* **103**, 7–16.

Creasey, A., Fenton, B., Walker, A., Thaithong, S., Oliveira, S., Matambu, S. & Walliker, D. (1990). Genetic diversity of *Plasmodium falciparum* shows geographical variation. *American Journal of Tropical Medicine and Hygiene* **42**, 403–413.

Curtis, C. F. & Otoo, L. N. (1986). A simple model of the build-up of resistance to mixtures of anti-malarial drugs. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **80**, 889–892.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **B39**, 1–38.

Dye, C. & Godfray, H. C. F. (1993). On sex ratio and inbreeding in malaria parasite populations. *Journal of Theoretical Biology* **161**, 131–134.

Gokool, S., Curtis, C. F. & Smith, D. F. (1993). Analysis of mosquito bloodmeals by DNA profiling. *Medical and Veterinary Entomology* **7**, 208–215.

Guo, S. W. & Thompson, E. A. (1992). Performing the

https://doi.org/10.1017/S0016672300033000 Published online by Cambridge University Press

exact test for Hardy–Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.

Hartl, D. L. & Clark, A. G. (1989). *Principles of Population Genetics*. Sunderland, MA: Sinauer.

Koehler, K. J. & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association* **75**, 336–344.

Levin, D. A. (1978). Genetic variation in annual *Phlox*: Self-compatible versus self-incompatible species. *Evolution* **32**, 245–263.

Ranford-Cartwright, L. C., Balfe, P., Carter, R. & Walliker, D. (1991). Genetic hybrids of *Plasmodium falciparum* identified by amplification of genomic DNA from single oocysts. *Molecular & Biochemical Parasitology* **49**, 239–244.

Ranford-Cartwright, L. C., Balfe, P., Carter, R. & Walliker, D. (1993). Frequency of cross-fertilisation in the human malaria parasite *Plasmodium falciparum*. *Parasitology* **107**, 11–18.

Robertson, A. & Hill, W. G. (1984). Deviations from Hardy–Weinberg proportions: sampling variances and use in estimation of inbreeding coefficients. *Genetics* **107**, 703–718.

Simonoff, J. S. (1985). An improved goodness-of-fit statistic for sparse multinomials. *Journal of the American Statistical Association* **80**, 671–677.

Smith, C. A. B. (1957). Counting methods in genetical studies. *Annals of Human Genetics* **21**, 254–276.

Smith, T., Charlwood, J. D., Kihonda, J., Mwankusye, S., Billingsley, P., Meuwissen, J., Lyimo, E., Takken, W., Teuscher, T. & Tanner, M. (1993). Absence of seasonal variation in malaria parasitaemia in an area of intense seasonal transmission. *Acta Tropica* **54**, 55–72.

Read, A. F., Nabara, A., Nee, S., Keymer, A. E. & Day, K. P. (1992). Gametocyte sex ratios as indirect measures of outcrossing rates in malaria. *Parasitology* **104**, 387–395.

Tait, A. & Turner, C. M. R. (1990). Genetic exchange in *Trypanosoma brucei*. *Parasitology Today* **6**, 70–75.

Thaithong, S., Beale, G. H., Fenton, B., McBride, J., Rosario, V., Walker, A. & Walliker, D. (1984). Clonal diversity in a single isolate of the malaria parasite *Plasmodium falciparum*. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **78**, 242–245.

Tibayrenc, M., Kjellberg, F., Arnaud, J., Oury, B., Breniere, S. F., Darde, M.-L. & Ayala, F. J. (1991). Are eukaryotic microorganisms clonal or sexual? A population genetics vantage. *Proceedings of the National Academy of Sciences USA* **88**, 5129–5133.

Triglia, T., Wellems, T. E. & Kemp, D. J. (1992). Towards a high resolution map of the *Plasmodium falciparum* genome. *Parasitology Today* **8**, 225–229.

Walliker, D. (1989). Implications of genetic exchange in the study of protozoan infections. *Parasitology* **99**, S49–S58.

Weir, B. S. (1990). *Genetic Data Analysis*. Sunderland, MA: Sinauer.

Weir, B. S. & Cockerham, C. C. (1973). Mixed self and random mating at two loci. *Genetical Research* **21**, 247–262.

Wright, S. (1921). Systems of mating. *Genetics* **6**, 111–178.

https://doi.org/10.1017/S0016672300033000 Published online by Cambridge University Press