Theor Appl Genet (1993) 86:991–1000

Estimation of the number of full sibling families at an oviposition site using RAPD-PCR markers: applications to the mosquito *Aedes aegypti*

B. L. Apostol[1], **W. C. Black IV**[1], **B. R. Miller**[2], **P. Reiter**[3], **B. J. Beaty**[1]

[1] Arthropod-borne and Infectious Diseases Laboratory, Department of Microbiology, Colorado State University, Fort Collins, CO 80523, USA
[2] Medical Entomology and Ecology Branch, Division of Vector-Borne Infectious Diseases, Centers for Disease Control and Prevention, Public Health Service, U.S. Department of Health and Human Services, P.O. Box 2087, Fort Collins, CO 80522 USA
[3] Dengue Branch, Division of Vector-Borne Infectious Diseases, Center for Infectious Diseases, Centers for Disease Control and Prevention, Public Health Service, U.S. Department of Health and Human Services, 2 Calle Casia, San Juan, Puerto Rico 00921-3200

Received: 17 November 1992 / Accepted: 2 February 1993

Abstract. There are many species in which groups of individuals encountered in the field are known to consist of mixtures of full-sibling families. We describe a statistical technique, based on the use of random amplified polymorphic DNA-polymerase chain reaction (RAPD-PCR) markers, that allows for the estimation of the number of families contained in these groups. We test the technique on full-sibling families of the mosquito *Aedes aegypti*, a species that distributes its eggs among several locations. Mixtures of 10 families with 15 individuals per family were analyzed using 40 RAPD-PCR loci amplified by 5 primers. Our analysis accurately estimated the number of families. The technique was accurate when the number of families was small or when family sizes were small and variable.

Key words: RAPD-PCR – DNA Fingerprinting – Sibling analysis – Oviposition – *Aedes aegypti*

Introduction

Genetic fingerprinting has been used to study social organization and interactions among individuals in a wide variety of species. It is widely used in many forensic applications including paternity analysis (Kasai et al. 1990; Chakraborty and Kidd 1990). Fingerprinting has been used to study levels of inbreeding in field populations of snails (Jarne et al. 1992) and mole rats (Reeve et al. 1990), to examine the number of siblings per nest in a number of bird species (Wetton et al. 1987; Burke and Bruford 1987; Westneat 1990), or in an aphid species to study the number of apomictic parthenogenetic maternal lineages in field samples (Shufran et al. 1991).

Lynch (1988) warned of the limitations and problems associated with attempting to ascertain relatedness using DNA fingerprinting techniques. He demonstrated that unbiased estimates of relatedness cannot be obtained among individuals without prior knowledge of allele frequencies in the population or among the individuals analyzed. In addition, he showed that there are large sampling variances surrounding estimates of relatedness and that this variance increases as the relatedness among individuals decreases (e.g., among half-siblings or cousins). Because of these problems, he pointed out that fingerprinting studies cannot be realistically extended beyond parent-offspring or full-sibling analysis even when examining multiple alleles or many loci.

While this limitation is recognized, there are many species in which clusters of individuals encountered in the field are known or believed to consist of mixtures of full-siblings. It is often of interest to know the number of full-sibling families contained in such a cluster. In eusocial species of both vertebrate and invertebrates, colonies often consist of full- and half-siblings (Reeve et al. 1990; Wilson 1971). In solitary species, depending on female mating behavior, nest-mates or litter-mates are often mixtures of half-sibling families. Many insects oviposit in a single location, and emerging immatures are limited in mobility such that groups of individuals contain mixtures of full-siblings, half siblings, and unrelated individuals. Lepidoptera larvae in a tent,

Communicated by J. S. F. Barker
Correspondence to: W. C. Black IV

bark beetles in the phloem of a tree, leaf-mining agromyzids in the mesophyll of a leaf, scabies mites beneath the skin of a vertebrate host, fly larvae in a dung pat, or gasterophilid larvae in the stomach of a vertebrate host are examples of this phenomenon. For various reasons it may be of interest to know whether cohorts represent single or multiple full-sibling families.

To date most fingerprinting studies have employed hypervariable loci that shift in size due to unequal crossing-over and gene conversion among tandem repeats. These are known as VNTR loci (Variable number of tandem repeat) (Jeffreys et al. 1985; Pemberton and Amos 1990 for review). However, these approaches may present problems because of the small size of the organism or the lack of adequate DNA probes. Recently, techniques have been developed to detect genetic variability by the amplification of arbitrary segments of genomic DNA using short and random or nonspecific primers in the polymerase chain reaction (PCR) (Welsh and MCClelland 1990; Welsh et al. 1991; Williams et al. 1991). In RAPD-PCR (random amplified polymorphic DNA amplified by PCR), a single, short (10 bp) primer is used in the PCR reaction to amplify fragments located at numerous locations throughout the genome (Williams et al. 1991). The vast majority of polymorphisms at RAPD-PCR loci segregate as dominant alleles. The genotype of a heterozygous individual with amplification products arising from a single chromosome cannot be distinguished from homozygotes with amplification products arising from both chromosomes (Williams et al. 1991). RAPD-PCR has been used in a variety of studies for species diagnosis, population differentiation, and genetic mapping (Williams et al. 1991; Black et al. 1992; Kambhampati et al. 1992; Ballinger-Crabtree et al. 1992; Puterka et al. 1993). The advantages of this approach include: the small amounts of genomic DNA required, the nonradioactive detection of products, a short processing time, and the ability to carry out the procedure without any prior sequence information.

In this study we describe a statistical technique based on RAPD-PCR markers to estimate the number of full-sibling families at an oviposition site. We compare assumptions and statistical approaches for RAPD-PCR markers and VNTR polymorphisms. We then test the technique on full-sibling families of the mosquito *Aedes aegypti*. These were established directly from adults reared from eggs collected in San Juan, Puerto Rico. *Ae. aegypti* is a species that, as with the examples given above, oviposits all or part of its eggs in a single location, and larvae are confined to that location until adult development occurs. An individual female *Ae. aegypti* typically distributes her eggs among several artificial containers (e.g., cisterns, drains, discarded tires, buckets, and flower pots; Buxton and Hopkins 1927; Christophers 1960; Chadee and Corbet 1991). Many females may visit a container so that larvae within a single oviposition site represent a mixture of families. As we recognize the limitations described by Lynch (1988), the purpose of our technique is not to determine the relatedness among each individual in an oviposition site but rather to estimate the number of full-sibling families at the site.

Materials and methods

Relatedness measures using RAPD-PCR polymorphisms

RAPD-PCR polymorphisms segregate independently, suggesting that they represent individual loci (Williams et al. 1991). They are, in this regard, different from the more commonly employed VNTR polymorphisms that arise through slippage and unequal crossing-over *within* a single genetic locus and do not, therefore, segregate independently. The measure commonly used when comparing individuals with VNTR fingerprinting probes is:

$$S = 2N_{AB}/(N_A + N_B) \quad (1)$$

where N_{AB} is the number of bands shared in common between individuals A and B, and N_A and N_B are the total number of bands observed in A and B, respectively.

Because RAPD loci segregate independently and the vast majority (usually >95% (Williams et al. 1991)) of alleles are dominant, a second measure can be used when comparing RAPD patterns between individuals. The dominant phenotype produced by a RAPD locus is expressed, on an agarose gel, as the presence of a band of a specific molecular weight. The recessive phenotype is the absence of that band. Thus, pairs of individuals can be compared phenotypically at any locus, based on the shared presence or absence of a band. The shared absence of a band actually provides more information regarding their similarity (both homozygote recessives) than does the shared presence of that band (heterozygote or homozygote dominant).

We measure the similarity of pairs of individuals by examining both the shared presence and the shared absence of bands to take advantage of the recessive phenotype. We estimate the fraction of matches (M) using the formula:

$$M = N_{AB}/N_T \quad (2)$$

where N_{AB} is the total number of matches in individuals A and B (i.e., both bands absent or present), and N_T is the total number of loci scored in the overall study. Unlike the similarity index, the denominator for M is fixed, and the absence of a band is scored because it represents the recessive phenotype at a locus. An M value of 1 indicates that two individuals have identical patterns; a value of 0 indicates that two individuals had completely different patterns. As with VNTR markers, fragments that comigrate are assumed to arise from identical alleles. However, we also assume that the absence of a band in two individuals arose from the identical ancestral mutation (i.e., recessive alleles are *identical in state*). This may not be true because there are potentially many point mutations at the primer annealing sites that could interrupt annealing. Furthermore, inversions flanking the annealing sites would prevent amplification, and an insertion that separates sites by a greater distance than can be amplified with routine PCR techniques would also produce a recessive allele. The assumption that recessive alleles are identical in state is valid among full-siblings. However, it may be false among non-siblings, and the scoring of the shared recessive phenotype may overestimate relatedness among non-siblings.

Estimating the numbers of full-sibling families using cluster analysis

Values of M are calculated among all $n(n-1)/2$ pairs of n individuals. Values are placed in a symmetrical matrix, and this matrix is collapsed to construct a dendrogram using the "unweighted pair-group method with arithmetic averaging" (UPGMA) (Johnson and Wichern 1982). Siblings should, on average, be more closely related than non-siblings or half-siblings and should share higher M values amongst themselves than with non-siblings. However, as pointed out by Lynch (1988), there are large variances around estimates of relatedness. This variance can create two types of errors. An individual may not be placed in a cluster with full siblings or may be placed in a cluster with non-siblings. Our goal was not to create clusters of pure siblings but rather to estimate the number of full-sibling families in a mixture of siblings, half-siblings, and unrelated individuals.

This approach requires estimation of a value of $1-M$ that will separate clusters of full-siblings. Clusters formed below this value should primarily contain full-siblings, while members of different families should be joined primarily above this value. To discriminate clusters, average values of $1-M$ among siblings must be less than those among unrelated individuals.

The problem with this approach is illustrated in Fig. 1. Assuming random and independent segregation of RAPD alleles, and averaging across loci and families, the expected value of M among siblings (designated M_S) in Fig. 1A is calculated to be 0.847. Full-siblings will share approximately 85% of RAPD markers. In the cluster analysis, M_S values *within* groups of full-siblings are compared with M values *among* other groups of full-siblings. We designate M among families as M_P. M_P is dependent on allele frequencies in the parent population. The expected value of M_P is calculated according to the method in Fig. 1B and by integrating this function over allele frequencies from 0 to 1. When random mating in the parent population is assumed, the average value of M_P is 0.85 (calculations not shown). M_S and M_P are therefore equal. M among siblings of a specific pair of parents, assuming Mendelian inheritance, and M among siblings from any two randomly selected parents are equal, and no discriminating value of M can be derived.

The distribution of M_P is shown in Fig. 2A. From this curve it can be seen that $M_P \geq M_S$ (0.847) at loci for which the frequency of the dominant 'A' allele is <0.1 or >0.6. When M_P is integrated over loci at which allele frequencies occur between 0.1 and 0.6, the expected value of M_P decreases to 0.770 (calculations not shown), and the difference between M_S and M_P increases to 8%. Thus, by restricting analysis to loci with allele frequencies between 0.1 and 0.6, a value of $1-M$ that discriminates among clusters of full-siblings can be derived. The number of clusters formed below this value estimates the number of families contained within that analysis.

This approach is supported by the findings of Lynch (1988) in that the discriminating value of M is dependent on the frequencies of alleles in the population under study and no general value can be derived. The frequency of RAPD alleles in each population must therefore be estimated before applying the technique. Loci at which the frequency of the dominant allele are below 0.1 and above 0.6 are discarded. Allele frequencies at the remaining loci are placed in the equation of Fig. 1B and a discriminating value of M for each locus derived. A single discriminating M is determined by averaging over all loci.

Verification

To test the validity of this approach we established full-sibling families from parents from field-collected eggs of the mosquito, *Ae. aegypti*. RAPD-PCR was done among members of the parent population to estimate the frequency of dominant alleles at each locus. Those loci at which allele frequencies were between 0.1 and 0.6 were then analyzed in 15 siblings from each of 10 families. The phenotype (1 for presence of a band, 0 for absence) of each individual at each locus was entered into a dataset. This dataset was then analyzed with a FORTRAN program, RAPDPLOT (Kambhampati et al. 1992; available from WCB4, provide 3-1/2″ diskette). This program calculates M among all individuals, builds a matrix of $1-M$, and collapses the matrix to derive a dendrogram. Dendrograms were plotted using modifi-

A. Full-sibling families

parent 2 \ parent 1	AA	Aa	aa
AA	1	1	1
Aa	1	* 0.625	0.5
aa	1	0.5	1

A = amplified allele (dominant)
a = unamplified allele (recessive)

*

sibling 2 \ sibling 1	AA	Aa	aA	aa
AA	1	1	1	0
Aa	1	1	1	0
aA	1	1	1	0
aa	0	0	0	1

$$M = \frac{10}{16} = 0.625$$

$$\text{Average } M = \frac{\sum (\text{M values for siblings})}{9} = \frac{7.625}{9} = 0.847$$

B. Among families

individual 2 \ individual 1	AA (p^2)	Aa $(2pq)$	aa (q^2)
AA (p^2)	1 (p^4)	1 $(2p^3q)$	1 (p^2q^2)
Aa $(2pq)$	1 $(2p^3q)$	0.625 $(4p^2q^2)$	0.5 $(2pq^3)$
aa (q^2)	1 (p^2q^2)	0.5 $(2pq^3)$	1 (q^4)

p = freq. of "A" allele
q = freq. of "a" allele = 1 - p

$$p^4 + 4p^3q + 4pq^3 + 6p^2q^2 + q^4 = 1$$

$$p^4 + 4p^3(1-p) + 4p(1-p)^3 + 6p^2(1-p)^2 + (1-p)^4 = 1$$

substituting values of M:

$$\text{Average } M = p^4 + 4p^3(1-p) + 2p(1-p)^3 + 4.5p^2(1-p)^2 + (1-p)^4$$

Fig. 1A, B. Estimation of average M values within and among families. **A** Estimation of locus M values among full-siblings. M values appear in each cell. In any cross involving a homozygous dominant parent, all offspring will display a band, and M will be 1. M is also 1 among offspring of two homozygous recessive parents. Only half of the offspring will have identical patterns in a cross between a heterozygous and homozygous recessive parent, and M will equal 0.5. In a family produced by two heterozygotes (*), there are four offspring genotypes and therefore 16 possible pair-wise comparisons. In 10 of these, RAPD-PCR patterns will match (i.e., bands will be present or absent in both individuals), and M is 10/16. Summing M across all 9 possible parental crosses, the average M value is 0.847. This is designated as M_S. **B** Estimation of M values among siblings of randomly selected parents from a population mating at random. The frequency of the dominant 'A' allele is **p** and the frequency of the recessive 'a' allele is **q**. Average M (designated M_P) values are calculated by multiplying their M values for crosses estimated in Fig. 1A by their respective frequencies in the population

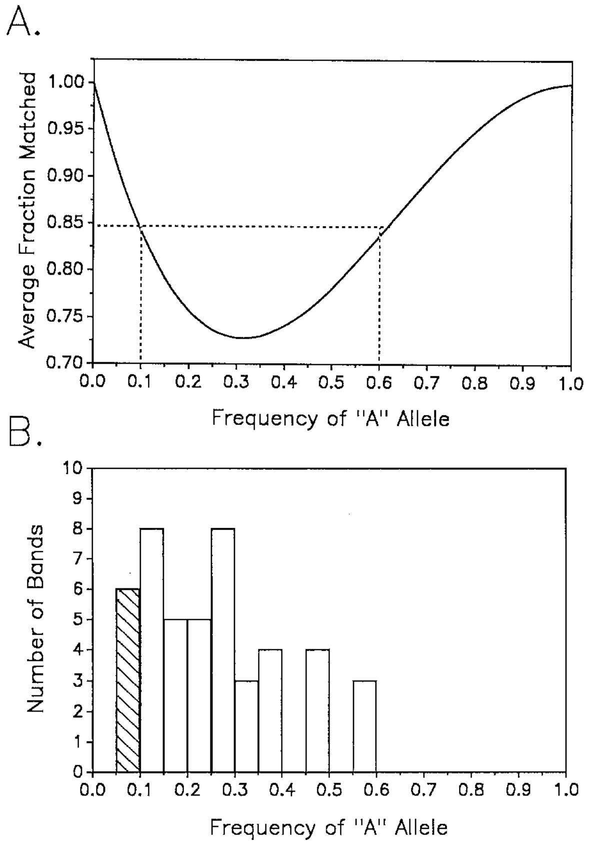

Fig. 2. **A** Distribution of M_P values over frequencies of a dominant 'A' allele. The *dotted line at 0.847* indicates the value of M_S. The *dotted lines at 0.1 and 0.6* are the upper and lower limits of allele frequencies used for discriminating among and within families. **B** Distribution of allele frequencies at the 46 loci scored in the San Juan population. The *hatched box* indicates loci with allele frequencies below 0.1. These were not used in the calculation of M_P

cations of the FORTRAN program of McCammon and Wenninger (1970). The McCammon and Wenninger method estimates both between-group distances (the distance among clusters) and the within-group distances (the distance between adjacent joined groups). With the equation in Fig. 1B a discriminating value of M was derived, and this was used in the dendrogram to estimate the number of full-sibling families.

Mosquitoes

Mosquito eggs were collected in the metropolitan area of San Juan, Puerto Rico (Table 1). Eggs were mailed on oviposition paddles to Fort Collins, Colorado, where they were hatched and reared to adults. Female pupae were removed prior to eclosion and subsequently mated to individual males. Offspring from the F_1 generations were reared to adults and stored at −70 °C prior to extraction of genomic DNA.

Isolation of mosquito DNA

Individual mosquitoes were placed in 1.7-ml microfuge tubes filled with liquid nitrogen and triturated with a microfuge pestle. Lysis buffer (300 µl) (10 mM Tris-HCl, pH 8.0, 5 mM NaCl, 5 mM EDTA, 0.1% SDS, 0.015 mM spermine, 0.05 mM spermidine and 0.33 µg/ml proteinase K) was added, and the preparation vortexed and incubated at 50 °C overnight. After extraction with an equal volume of phenol, 0.2 volumes of 10 M ammonium acetate were added to the aqueous phase. DNA was recovered by precipitation at room temperature with 2.5 volumes of ethanol and centrifugation for 5 min at 15,800 g. Based on OD_{260} measurements, approximately 10 µg of nucleic acid was obtained per mosquito. DNA samples were resuspended in 20 µl water and incubated at 50 °C for 1–2 h to aid in resuspension of high-molecular-weight DNA. For PCR reactions, genomic DNAs were diluted 1:100 in water. The undiluted and diluted samples were stored at 4 °C and remained stable for several months.

RAPD-PCR and gel electrophoresis

With some minor modifications, RAPD-PCR reactions were carried out as previously described (Williams et al. 1991). Reaction mixtures consisted of 2.5 µl of 10X reaction buffer (100 mM Tris-HCl, pH 8.3, 100 mM KCl, 20 mM $MgCl_2$, and 0.01% gelatin), 2.5 µl 1 mM dNTPs, 1 µl primer (see below, 15 ng/µl) and 0.2 µl Amplitaq DNA Polymerase (5 U/µl) in a total volume of 23 µl. Reactions were layered with 50 µl of light mineral oil, and 2 µl (ca. 10 ng) of the diluted genomic DNA was added through the oil into the reaction mixture. Amplifications were performed in a thermocycler (DNA Thermal Cycler 480, Perkin Elmer-Cetus Corp) using the following program: 94 °C for 4 min followed by 45 cycles consisting of 94 °C for 1 min 36 °C for 1 min, and 72 °C for 2 min. A final extension was carried out at 72 °C for 4 min. Upon completion of the amplification, samples were maintained at 4 °C. Amplified products were resolved by electrophoresis on 1.4% TBE agarose gels for 6 h at 4.5 V/cm. DNA fragments were visualized by staining with ethidium bromide (1 µg/ml) for 1 h at room temperature and photographed prior to scoring of bands.

Selection of primers and scoring of DNA fragments

Thirty primers were screened among 18 mosquitoes from nine representative neighborhoods in San Juan. Primers that did not produce well-amplified polymorphic bands that were clearly dis-

Table 1. Origin of parents from San Juan, Puerto Rico used to generate *Aedes aegypti* families

Family[a] designation	Mother	Father
A	PN3[b]	PN2
B	PN6	PN4
C	PN1	PN5
D	BO1	BO1
E	BO2	TA1
F	BO2	BO2
G	V1	V1
H	V1	BO2
I	ER1	BO1
J	OP2	OP1

[a] Letter designation for family

[b] Origin of parents are abbreviated as follows: PN, Puerto Nuevo; BO, Barrio Obrero; V, Virtudes; ER, Eleanor Roosevelt; OP, Ocean Park; TA, Trujillo Alto. The numbers designate individual ovitraps

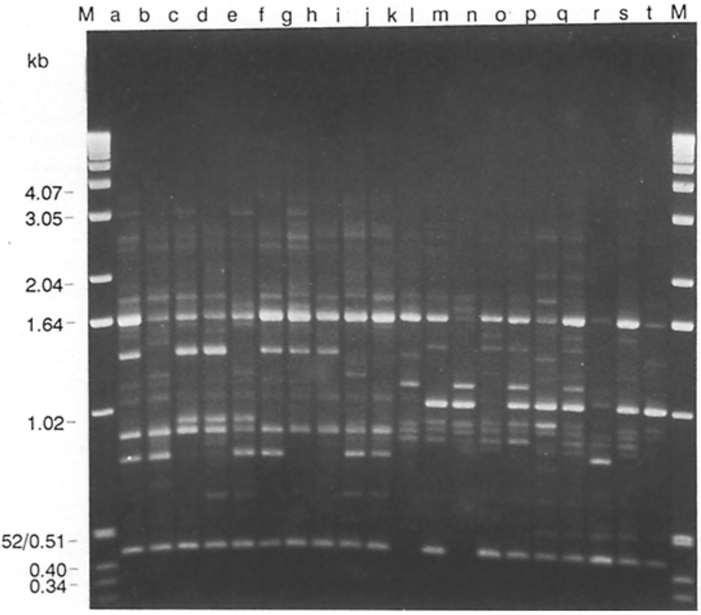

Fig. 3. Examples of RAPD-PCR products with primer C4 and 10 offspring from *Ae. aegypti* families D and F. *Lane M* Molecular weight markers (1-kb DNA ladder, BRL), *lanes a–e* family F females, *lanes f–j* family F males, *lanes k–o* family D females, *lanes p–t* family D males. DNA fragments that were scored are indicated with an *arrow to the right* of the figure

tinguishable from neighboring bands were eliminated to prevent ambiguity in assigning identity among gels. Ultimately, 5 primers: 5′-CCGCATCTAC-3′ [C4], 5′-CTCACCGTCC-3′ [C9], 5′-AAGCCTCGTC-3′ [C13], 5′-CACACTCCAG-3′ [C16], and 5′-GTTGCCAGCC-3′ [C19] ("Kit C", Operon Technologies) were chosen. These primers yielded 46 loci that met the above criteria. Band size ranged from approximately 0.4 to 3.0 kb (e.g. Fig. 3). Preparations from mosquitoes that contained each of the fragments scored were loaded on every gel alongside molecular markers to aid in the scoring of bands. This alleviated the need to account for variation among gels with respect to variation in electrophoresis time, gel concentration, and pH.

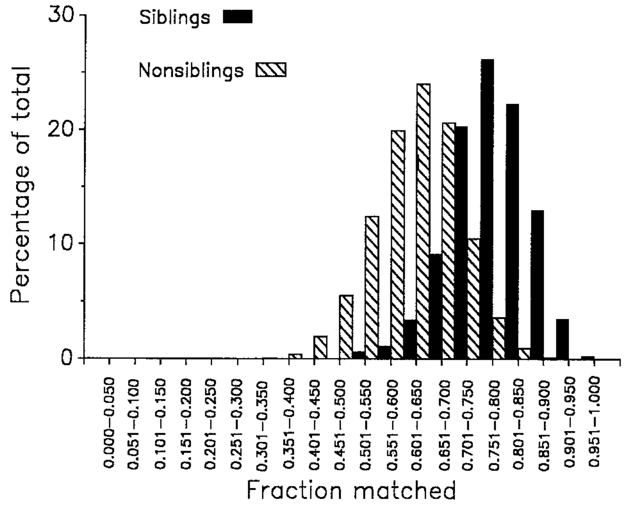

Fig. 4. Distribution of *M* values among siblings and non-siblings. Results are based on 40 bands scored from RAPD-PCR reactions with 5 primers for 10 families, 15 offspring each. A total of 11,175 pair-wise calculations were made and were placed into 5% class intervals

Results

Of the 46 loci analyzed in the overall population, 40 had alleles that occurred in the range of 0.1–0.6. Allele frequencies were not uniformly distributed over the 46 loci selected for study (Fig. 2B). Most were distributed between 0.1 to 0.3, and at no locus were allele frequencies greater than 0.6. This bias occurred because in screening RAPD-PCR loci among 18 individuals, loci that were only absent in 1 or 2 individuals were generally ignored. Therefore, loci with recessive alleles occurring at a frequency from 0–0.33 were not used.

The 40 loci were scored as to their presence or absence in each of 15 siblings in 10 families. Match scores *M* were calculated on a pair-wise basis among all 150 individuals. The distributions of *M* within and among families are shown in Fig. 4. The mean value of *M* among siblings (0.789; range = 0.864–0.715) was greater than that among non-siblings (0.630; range = 0.767–0.507). When analysis was restricted to alleles that occurred in the frequency of 0.1 to 0.6, average relatedness among full-siblings was 16% higher than among non-siblings.

An *M* value of 0.768 was calculated by placing allele frequencies from the 40 loci into the equation of Fig. 1B. This value was in close agreement with that *M* derived by integrating over frequencies from 0.1–0.6 (i.e. 0.77). The theoretical difference between M_S and M_P was therefore 8% (0.85–0.77), although this is half of the observed difference of 16%. Note also that the observed M value (0.789) among siblings was 6% lower than the expected M value (0.847).

Estimation of the number of families

When the expected value of M among non-siblings (0.77) was used, the discriminating value for cluster analysis $(1 - M)$ was 0.23. Only clusters formed below 0.23 were counted as families. A cluster analysis was performed on all 15 offspring from each of the 10 families. The analysis estimated 11 families below 0.23 for $(1 - M)$ (Table 2). In 9 clusters, members of a single family predominated. These clusters were designated as 'family clusters', and non-siblings that fell within these clusters were considered to be misclassified. In family E, 5 individuals clustered together, 5 and 2 individuals were misclassified into families G and I, respectively, and 3 individuals formed an independent cluster. The latter was the source of the eleventh family. The overall misclassification rate was 12.7% (19/150) (Table 2).

We repeated these analyses, first relaxing and then constraining the range of alleles frequencies over which we estimated M. This was done to test whether restriction of analysis to alleles between 0.1–0.6 was optimal in accurately estimating family numbers and reducing misclassification rates. When all of the 46 loci selected in the population were used, the expected value of M among non-siblings was 0.78. The discriminating value in cluster analysis $(1 - M)$ was 0.22 and indicated 14 families, with an overall misclassification rate of 16.8% (24/150). This rate was not significantly greater $[\chi^2(1\,df) = 0.68,\ P = 0.41]$ than when we restricted analyses to those loci at which allele frequencies were between 0.1–0.6, however, the estimated number of families exceeded the actual number by 4. The differences in misclassification rates might have been greater if, in our unrestricted analysis, we had scored loci with frequencies greater than 0.6 (Fig. 2B).

Table 2. Misclassification rates of siblings from ten *Aedes aegypti* families as determined by fraction matched calculations

Family[a]	A[b]	B	C	D	E	F	G	H	I	J	Other[c]
A	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
B	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
C	0.0	0.0	73.3	0.0	0.0	13.3	0.0	0.0	13.3	0.0	0.0
D	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E	0.0	0.0	0.0	0.0	33.3	0.0	33.3	0.0	13.3	0.0	20.0
F	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
G	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0
H	0.0	6.7	13.3	13.3	0.0	0.0	0.0	66.7	0.0	0.0	0.0
I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
J	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0

[a] Families are decribed in Table 1

[b] Percentage classified into family based on analysis of 15 siblings for each family

[c] See Results for definition of "other"

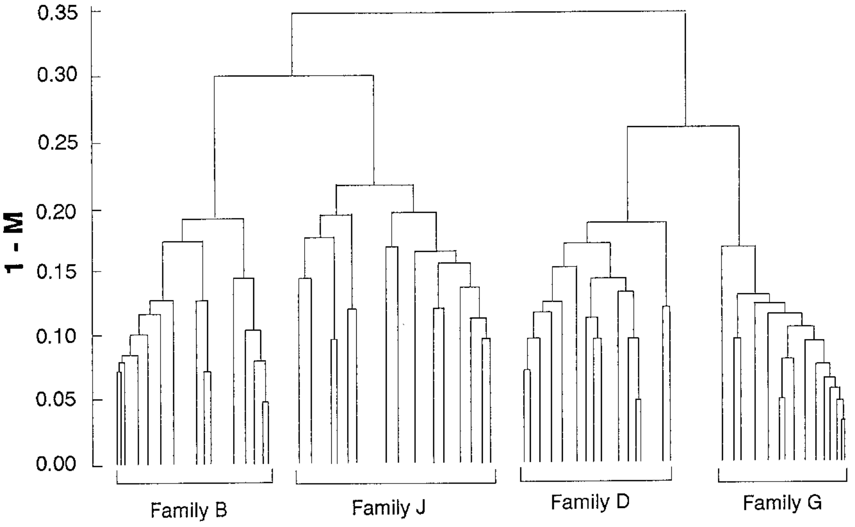

Fig. 5. Dendrograph with four families selected at random. Cluster analysis was performed using $1 - M$ scores and UPGMA (see text). Family labels are shown at the *bottom* of the figure

Table 3. Estimated number of *Aedes aegypti* families determined by cluster analysis of fraction matched values

Actual number of families[a]	Estimated number of families[b]	Percentage misclassified[c]
1	1	0.0
1	1	0.0
1	1	0.0
1	2	20.0
	1.2+/−0.22	5.0+/−4.3
2	2	0.0
2	4	16.7
2	2	0.0
2	2	0.0
	2.5+/−0.43	4.2+/−3.2
3	3	15.6
3	5	17.8
3	4	8.9
3	4	17.8
	4.0+/−0.35	15.0+/−1.8
4	4	6.7
4	5	5.0
4	4	0.0
4	4	13.3
	4.2+/−0.22	6.2+/−2.4
5	6	10.7
5	6	6.7
5	5	0.0
5	5	10.7
	5.5+/−0.22	7.0+/−2.2
6	6	12.2
6	7	12.2
6	6	7.8
6	6	12.2
	6.2+/−0.22	11.1+/−1.0
7	8	16.2
7	7	4.8
7	7	6.7
7	8	10.5
	7.5+/−0.25	9.5+/−2.2
8	8	15.0
8	8	12.5
8	9	15.8
8	9	14.2
	8.5+/−0.25	14.4+/−0.6
9	9	9.6
9	9	11.1
9	9	6.7
9	10	14.8
	9.2+/+0.22	10.6+/−1.5

[a] Families indicated were randomly selected 4 times

[b] Estimated number of families based on a within-group distance of 0.23 (see Results); mean number of groups estimated +/−SE

[c] Percentage of individuals misclassified (see Results); mean percent +/−SE

When the estimation of M was further restricted to loci with allele frequencies between 0.2 and 0.5, the expected M among non-siblings was 0.742, 7 families were estimated, and the overall misclassification rate rose to 33% (50/150). This rate was significantly greater than the rate when using all alleles [χ^2 (1 df) = 12.3, $P < 0.001$] or alleles between 0.1 and 0.6 in frequency [χ^2(1 df) = 18.6, $P < 0.001$]. A value of M that maximizes accurate estimation of family numbers and minimizes the misclassification rate was estimated by the restricted use of alleles with frequencies between 0.1–0.6 in the overall population.

A variable number of families were chosen at random from the original 10 to determine the accuracy with which our analysis predicted family numbers. Families were chosen at random from 1 to 9 at a time. Four replicate random samples were made for each number of families. The actual and predicted numbers of families, and the percent misclassified, are given in

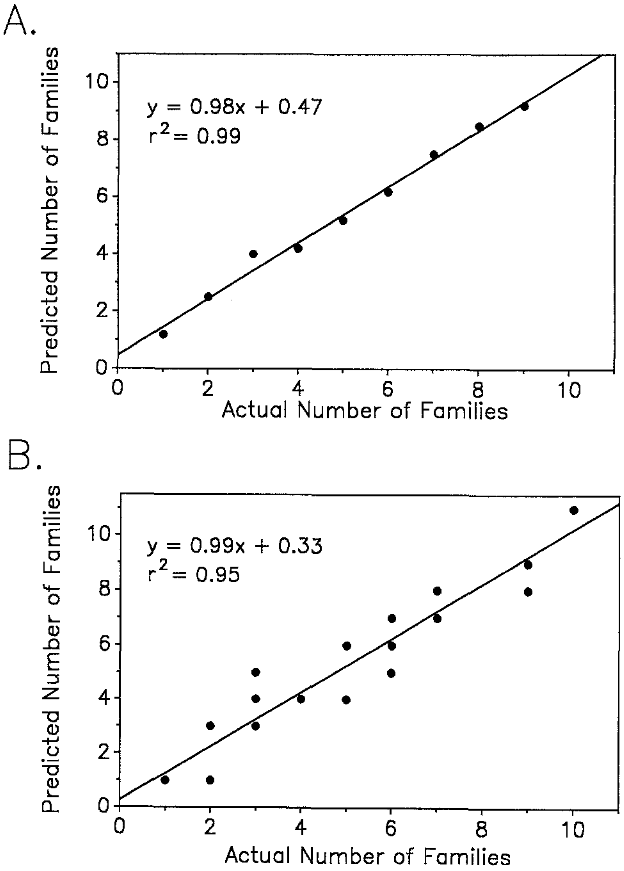

Fig. 6A, B. Linear regression analysis of the actual number of families in an analysis versus the predicted number. **A** Analysis of a variable number of families at a constant family size. **B** Analysis with a variable number of families and a variable number of individuals per family

Table 4. Estimated number of *Aedes aegypti* families from simulations in which the numbers of families and the number of siblings within each family were selected at random

Actual number of families[a]	Estimated number of families[b]	Family size[c]			Percentage individuals misclassified[d]
		Average	Minimum	Maximum	
4	4	10.5	3	14	4.8
10	11	8.9	2	15	20.2
2	1	8.5	2	15	11.8
7	8	8.9	1	14	38.7
6	5	7.4	5	12	15.4
3	4	9.7	3	14	13.8
5	4	9.0	1	15	2.2
3	5	4.7	2	9	14.2
6	6	9.7	1	15	24.4
6	6	10.0	6	15	10.0
5	6	8.8	3	15	31.8
10	11	10.9	5	15	13.7
6	6	4.8	1	9	10.3
9	9	8.1	3	15	26.0
7	7	7.8	1	9	10.2
1	1	15.0	15	15	0.0
3	3	8.0	6	9	20.8
9	8	7.9	3	13	7.0
2	3	7.5	3	12	6.7
6	7	7.0	2	14	7.1

[a] The number of families and the number of siblings per family were selected at random

[b] The estimated number of families based on a within group distance of 0.23 (see Results)

[c] The average number of siblings per family and the minimum and maximum family sizes

[d] The percentage of individuals incorrectly grouped together in the cluster analysis

Table 3. As an example, a dendrograph for one set of 4 families is shown in Fig. 5. For 22 of the 36 different trials, the number of families was correctly estimated. Of the remaining 14 combinations, 12 were off by 1 family, and 2 were off by 2 families. The percentage of individuals misclassified ranged from 0 to 24%; the unweighted average was 8.6%. Linear regression analysis of the true numbers of families regressed on the estimated numbers is shown in Fig. 6A. The slope was not significantly less than 1, and the intercept was not significantly greater than 0.

To determine if the technique is robust for small and unequal family sizes, families and individuals were randomly selected. Family numbers were randomly picked with replacement from 1 through 9; families were then picked at random. The number of siblings in each of these families was randomly selected with replacement from 1 through 15. The members to be removed from each family were randomly selected without replacement. In 20 simulations (Table 4), the misclassification rate averaged 14.4% and varied from 0 to 38.7%. In 19 simulations, the estimated number of families never deviated by more than 1 from the actual number. Regression analysis is shown in Fig. 6B. As above, the slope was not significantly less than 1, and the intercept was not significantly greater than 0.

In both simulations, misclassification rates increased with the number of families in the trial, but rates were independent of the number of misclassified families. This indicated that not all families contained members that were equally well discriminated (Table 2). Some families had members that clustered with other families; others contained members that only clustered amongst themselves. In families D and I, families E, F, and H, and families G and H, one of the parents in each family were reared from the same oviposition trap and therefore may have been siblings. Parents of families A, B, and C were reared from oviposition traps collected in the same neighborhood. We did not, however, observe any preferential misclassification of these individuals into families originating from parents from the same trap or from the same neighborhood.

Discussion

In the majority of simulations, cluster analysis of RAPD-PCR alleles accurately estimated the number of families independent of family number, size, or origin. In regression analysis of predicted and observed values, regression coefficients ranged from 0.95 to 0.99, and slopes and intercepts were not significantly different from 1 and 0, respectively. Across all simulations, the estimated number of families never deviated by more than 2 from the actual number. On the other hand, the average misclassification rate was approximately 10% and ranged from 0 to 38%. The approach we have described cannot, therefore, accurately determine relatedness among 2 individuals selected from an oviposition site. This was anticipated based on the problems indicated by Lynch (1988).

Our measure of similarity (Eq. 2) makes the assumption that recessive alleles are *identical in state.* This assumption is probably false among non-siblings because of the large number of mutations that can interrupt amplification. The scoring of a shared recessive phenotype may therefore overestimate relatedness among non-siblings resulting in the clustering of unrelated individuals. Most fingerprinting procedures use Eq. 1, which uses only the shared presence of bands and therefore makes no assumptions concerning the identity of recessive alleles. We initially attempted to use Eq. 1 in estimating family numbers. The difference in S values among full-siblings and S values among non-siblings was 18%, slightly greater than the 16% difference observed using Eq. 2. However, full siblings were ofter placed in separate clusters such that misclassification rates were large and the number of families was consistently overestimated. We discovered two problems with Eq. 1 with regards to RAPD polymorphisms and the general procedures that we have described. First, the number of loci used when calculating Eq. 1 varies among pair-wise comparisons. When 2 individuals share the recessive phenotype at a locus, that locus drops out of the estimate, thus reducing the number of discriminating loci between individuals. In our data set often half of the loci in siblings were scored as the shared absence of a band. By discarding data on shared absence, the ability to discriminate siblings from non-siblings is greatly reduced. Secondly, because each pair-wise comparison has a different denominator, we were unable to develop a procedure similar to that described in Fig. 1. The S values among full siblings varies by the number of bands within a family, and no single discriminating value can be derived.

It is likely that our method overestimates relatedness among siblings and leads to misclassifications, however our goal was to estimate family numbers, and the method is accurate in that regard. Misclassification rates might be reduced by increasing the number of discriminating loci. This has to be weighed against the time and expense associated with the analysis of more primers.

We emphasize that, in agreement with Lynch (1988), an accurate estimation of families is dependent upon an accurate estimation of allele frequencies in the population. Caution must be exercised in using our approach. If gene flow is restricted, then allele frequencies will vary among subpopulations. A population estimate of allele frequencies may not accurately estimate allele frequencies in every subpopulation. If local restricted gene flow is observed, allele frequencies should be estimated in the subpopulations in which the technique is to be applied. In selecting loci to be used in a study, those that show a great deal of variation outside the range of 0.1–0.6 among subpopulations should be avoided. In general, before this technique can be applied, the researcher should understand the breeding structure of the population under study.

The predicted difference in M_S and M_P was 8%, the observed difference was 16%. Several factors could have accounted for this discrepancy. M_S (Fig. 1A) was calculated assuming RAPD-PCR loci are unlinked. Because of the large number of fragments examined and the relatively small recombinational size of the *Ae. aegypti* genome (220–240 cM, Munstermann and Craig 1979), this assumption is probably false. Linkage would increase observed values of M among siblings and decrease values of M among non-sibling. Linkage of RAPD loci would explain the discrepancy between the observed and expected differences in M between siblings and non-siblings. However, we have no explanation for the observed M among siblings (0.789) being 6% lower than that predicted (M_S) assuming Mendelian inheritance (Fig. 1A).

While RAPD-PCR does not detect the amount of variation that is typical for hypervariable loci, the method can potentially provide an unlimited number of fragments for scoring. Furthermore, it does not require the development of special probes or primers for each species that is to be analyzed. Because the technique uses PCR, it can be applied to even the smallest of organisms and life stages. RAPD patterns similar to those shown from adults have been obtained with mosquito eggs, larvae, and pupae. A single mosquito egg yields sufficient DNA for approximately 50 reactions. The technique we have described can address questions regarding egg laying behavior, and it should be applicable to any species in which clusters of individuals encountered in a field are known or suspected to consist of mixtures of full siblings.

Acknowledgements. We thank Manual Amador for help with collections of *Ae. aegypti* and Mary Ballinger-Crabtree for helpful suggestions. This work was supported by the John D. and Catherine T. MacArthur Foundation. BLA is a postdoctoral fellow supported by training grant AI 07352 from the National Institutes of Health.

References

Ballinger-Crabtree ME, Black WC IV, Miller BR (1992) Use of genetic polymorphisms detected by RAPD-PCR for differentiation and identification of *Aedes aegypti* subspecies and populations. Am J Trop Med Hyg 47:893–901

Black WC IV, DuTeau NM, Puterka GJ, Nechols JR, Pettorini JM (1992) Use of random amplified polymorphic DNA polymerase chain reaction (RAPD-PCR) to detect DNA polymorphisms in aphids. Bull Entomol Res 82:151–159

Burke T, Bruford MW (1987) DNA fingerprinting in birds. Nature 327:149–152

Buxton PA, Hopkins GHE (1927) Researches in Polynesia and Melanesia, an account of investigations in Samoa, Tonga, the Ellice Group and the New Hebrides, in 1924 and 1925. Mem London Sch Hyg Trop Med 1:125–158

Chadee DD, Corbet PS (1991) The gonotrophic status of female *Aedes aegypti* (L.) overnight at the oviposition site (Diptera: Culicidae). Ann Trop Med Parasit 85:461–466

Chakraborty R, Kidd KK (1990) The utility of DNA typing in forensic work. Science 254:1735–1739

Christophers SR (1960) *Aedes aegypti* (L.). The yellow fever mosquito. Cambridge University Press, Cambridge

Jarne P, Delay B, Bellec C, Roizes G, Cuny G (1992) Analysis of mating systems in the schistosome-vector hermaphrodite snail *Bulinus globosus* by DNA fingerprinting. Heredity 68:141–146

Jeffreys AJ, Wilson V, Thein SL (1985) Hypervariable 'minisatellite' regions in human DNA. Nature 314:67–73

Johnson RA, Wichern DW (1982) Applied multivariate statistical analysis. Prentice-Hall, Englewood Cliffs, NJ

Kambhampati S, Black WC IV, Rai KS (1992) RAPD-PCR of mosquito species and population: techniques, statistical analysis and applications. J Med Entomol 29:939–945

Kasai K, Nakamura Y, White R (1990) Amplification of a variable number of tandem repeats (VNTR) locus (pMCT118) by the polymerase chain reaction (PCR) and its applications to forensic science. J Forensic Sci 35:1196–1200

Lynch M (1988) Estimation of relatedness by DNA fingerprinting. Mol Biol Evol 5:584–599

McCammon RB, Wenninger G (1970) The dendrograph. Kans Geol Surv Comput Contrib No. 48

Munstermann LE, Craig GB (1979) Genetics of *Aedes aegypti*: updating the linkage map. J Hered 70:291–296

Pemberton J, Amos B (1990) DNA fingerprinting: a new dimension. Trends Genet 60:101–103

Puterka GJ, Black WC IV, Steiner WM, Burton RL (1993) Genetic variation and phylogenetic relationships among worldwide collections of the Russian Wheat Aphid, *Diuraphis noxia* (Mordvilko), inferred from allozyme and RAPD-PCR markers. Heredity 70:604–618

Reeve HK, Westneat DF, Noon WA, Sherman PW, Aquadro CF (1990) DNA "fingerprinting" reveals high levels of inbreeding in colonies of the eusocial naked mole-rat. Proc Natl Acad Sci USA 87:2496–2500

Shufran KA, Black WC IV, Margolies DC (1991) DNA fingerprinting to study spatial and temporal distributions of an aphid, *Schizaphis graminum* (Homoptera: Aphididae). Bull Entomol Res 81:303–313

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218

Welsh J, Petersen CP, McClelland M (1991) Polymorphisms generated by arbitrarily primed PCR in the mouse: application to strain identification and genetic mapping. Nucleic Acids Res 19:303–306

Westneat DF (1990) Genetic parentage in the indigo bunting: a study using DNA fingerprinting. Behav Ecol Socibiol 27:67–76

Wetton JH, Royston EC, Parkin DT, Walters D (1987) Demographic study of a wild house sparrow population by DNA fingerprinting. Nature 327:147–149

Williams JK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1991) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Wilson EO (1971) The insect societies. Belknap Press, Cambridge, Mass