Heredity **76** (1996) 325–334 *Received 6 April 1995*

Population genetics with RAPD–PCR markers: the breeding structure of *Aedes aegypti* in Puerto Rico

BARBARA L. APOSTOL, WILLIAM C. BLACK IV*†, PAUL REITER‡ & BARRY R. MILLER

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, PO Box 2987, Fort Collins, CO 80522, †Department of Microbiology, Colorado State University, Fort Collins, CO 80523, U.S.A. and ‡Dengue Branch, Division of Vector-Borne Infectious Diseases, Center of 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

RAPD–PCR polymorphisms at 57 presumptive loci were used to examine the breeding structure of the mosquito *Aedes aegypti* in Puerto Rico. Mosquitoes were sampled from 16 locations in six cities and samples were located in a nested spatial design to examine local patterns of gene flow. Allele frequencies were estimated assuming (1) that genomic regions amplified by RAPD–PCR segregate as dominant alleles, (2) that genotypes at RAPD loci are in Hardy–Weinberg proportions, (3) identity in state (*iis*) among dominant amplified alleles and (4) *iis* among null alleles. The average genic heterozygosity was 0.354, more than twice the level detected in earlier allozyme surveys. Nested analysis of variance indicated extensive genetic differentiation among locations within cities. Effective migration rates (Nm) among cities were estimated from F_{ST} assuming an island model of migration. Estimates of Nm ranged from 9.7 to 12.2 indicating a high dispersal rate. The large number of polymorphisms revealed by RAPD–PCR allowed the distribution of F_{ST} and linkage disequilibrium to be examined among loci and demonstrated that small samples inflate F_{ST} and linkage disequilibrium. No linkage disequilibrium maintained through epistasis was detected among alleles at the 57 loci.

Keywords: *Aedes aegypti*, effective migration rate, population genetics, RAPD–PCR.

Introduction

Recently two techniques have been developed that amplify simultaneously many arbitrary regions of a genome, using a single primer annealed at low temperatures in the polymerase chain reaction (PCR) (Williams *et al.*, 1990; Welsh & McClelland, 1990). Random amplified polymorphic DNA (RAPD)–PCR uses a 10 oligonucleotide primer, with a minimum GC content of 60 per cent, that is annealed at $\leq$37°C during PCR (Williams *et al.*, 1990). By using several different primers, polymorphisms at many RAPD loci can be detected among individuals. RAPD–PCR can therefore potentially increase the resolution of genetic differences among individuals in population genetic studies. Furthermore, increased portions of individual genomes can be monitored simultaneously to test for linkage disequilibrium.

The disadvantage of RAPD polymorphisms in population genetics is that the majority of alleles ($>$90 per cent; Williams *et al.*, 1990) segregate as dominant markers. RAPD–PCR produces a fragment with template DNA from individuals that are either homozygous or heterozygous for an amplifiable allele. No fragment is produced in homozygous recessive individuals because amplification is disrupted in both alleles. Dominance prevents tests of random mating within populations because individual genotypes cannot be discerned.

The purpose of this study is to examine the usefulness of RAPD markers in defining the local breeding structure of the mosquito *Aedes aegypti* in

*Correspondence.

Puerto Rico. *Aedes aegypti* is the principal vector of a large number of important arboviruses, including yellow fever and dengue, and has been the subject of a number of population genetics studies using allozymes (see Tabachnick, 1991 for a review). We explore the use of RAPDs in estimating Wright's F_{ST} and θ (Weir & Cockerham, 1984). These are compared with the Lynch & Milligan (1994) method for estimating F_{ST} from RAPD markers. We test for linkage disequilibrium by adapting statistical techniques used in analysing codominant markers. The large numbers of loci revealed by RAPD–PCR permit examination of the distribution of F_{ST}, θ and linkage disequilibrium coefficients among loci.

Materials and methods

Analysis of field collections

The size, number and location of *A. aegypti* egg collections in each city are listed in Table 1. Eggs were collected in enhanced CDC oviposition traps ('ovitraps') (Reiter *et al.*, 1991). Ovitrap pairs, one containing a 10 per cent and the other a 100 per cent hay infusion, were placed under the eaves of houses between 09.00 and 12.00 and collected 1–3 days later. Traps were placed in a nested spatial design (Table 1). The first two collections in a city were located within 0.5 miles of one another in adjacent *barrios* (neighbourhoods) and were designated a 'group'. The third collection was in a nonadjacent *barrio* located 2–5 miles away from the first group. This design allowed us to analyse gene flow at three levels; among *barrios* within a group, between groups within a city and among cities.

Table 1 Sampling locations for *Aedes aegypti* in Puerto Rico. The first two *barrios* (A and B) listed under a city were ≈0.5 miles apart while the third (C) *barrio* was 2–5 miles from A and B

City	*Barrios*	No. individuals analysed
San Juan		150
	A = Reparto Metropolitano	50
	B = Puerto Nuevo	50
	C = Toa Baja	50
Mayaguez		150
	A = Paris	50
	B = Balboa	50
	C = Soledad	50
Arecibo		50
	A = Miramar	17
	B = Garcia	17
	C = Las Brisas	16
Ponce		50
	A = Cirio del Sur	17
	B = Boa Caribe	17
	C = Los Caobos	16
Fajardo		50
	A = Florencio	25
	B = Monte Brisa	25
Caguas		50
	A = Villa del Rey	25
	B = Villa Borinquen	25

Collections were obtained from three *barrios* in San Juan between late August and mid-September in 1992. Collections were made from three *barrios* each in Mayaguez, Ponce and Arecibo and two *barrios* in Fajardo and Caguas in July 1993. Eggs were reared to adults and stored at −70°C until needed. Two adults per ovitrap were analysed by RAPD–PCR. Previous studies of oviposition behaviour in San Juan suggested a low probability that a pair of individuals selected at random from a trap would be siblings (Apostol *et al.*, 1994). RAPD–PCR was performed following the conditions and primers described by Apostol *et al.* (1993). A total of 57 amplified products were scored.

Analysis of RAPD–PCR markers as alleles

We analysed RAPD–PCR polymorphisms as alleles by making four assumptions. First, RAPD products segregate as dominant alleles in a Mendelian fashion. Mendelian segregation was observed in an earlier genetic fingerprinting study of *A. aegypti* (Apostol *et al.*, 1993) and we are using these markers to construct an *A. aegypti* linkage map. Secondly, genotype frequencies at RAPD loci are in Hardy–Weinberg proportions. Thirdly, alleles in a homozygous recessive individual are *identical in state* (*iis*) (i.e. that they arose from identical mutations) among and within individuals. Fourthly, dominant, amplified alleles are similarly *iis*.

Statistical methods and equations are given in the Appendix and are presented in braces throughout the text. We estimate the frequency of a recessive allele a as the square root of the frequency of homozygous recessive individuals {1}. If q is the frequency of the a allele then $1-q=p$ is the frequency of the dominant allele A. Lynch & Milligan (1994) show that this approach underestimates q when sample sizes are small. We compared their less biased estimator and found that our sample sizes were sufficiently large that their correction made no difference for three significant figures. A FORTRAN program (RAPDBIOS) written by W.C.B.4

estimated the frequencies of *A* and *a* using {1} to produce a BIOSYS-1 type 3 dataset (Swofford & Selander, 1981). The expected heterozygosity among individuals at each locus in a *barrio* was estimated with BIOSYS-1. A nested analysis of variance following Wright (1978) was also performed with the WRIGHT78 option in BIOSYS-1. Variance in allele frequencies was partitioned among *barrios* within groups, among groups within a city and among cities.

Calculation of F_{ST} *and estimation of the effective migration rate (*Nm*)*

Wright (1951) developed F_{ST} as a means of estimating 'the correlation between random gametes within a subpopulation relative to gametes within the entire population'. F_{ST} is greater than zero when subpopulations are reproductively isolated because random gametes from a subpopulation carry alleles more often derived from a common ancestor than gametes from the total population. F_{ST} was estimated from {2} (Wright, 1951) and as θ (Weir & Cockerham, 1984). These were compared with Lynch & Milligan's (1994) method of F_{ST} estimation from RAPD markers. Equations to calculate F_{ST} for individual loci do not appear in the paper by Lynch & Milligan (1994) and so were derived by W.C.B.4 {6–11}. *Nm* among populations can be estimated from F_{ST} with equation {12} assuming an island model of migration among populations.

Linkage disequilibrium

Di-locus linkage disequilibrium was estimated from {14–24}. We also used Ohta's (1982a,b) method for partitioning the variance in disequilibrium in the total population into within and among subpopulation components. The within subpopulation component measures the proportion of the total variance attributable to epistasis. The among subpopulation component estimates the proportion attributable to genetic differentiation (Wahlund's effect). The sum of these measures the variance in disequilibrium in the total population.

Results

Breeding structure

The distribution of expected heterozygosities, averaged over locations, is shown in Fig. 1 and had an average of 0.354 across the 57 RAPD loci. The variance in allele frequencies was partitioned into variance among *barrios* within a group, among groups within cities, and among cities, following the method of Wright (1978). These variance components sum to the total variance among all samples so that the relative contribution of each component to the overall variance can be estimated. The variance among *barrios* within groups accounted for the majority (66 per cent) of the total variance. The next largest component occurred among groups within cities and accounted for 31 per cent of the total. Only 3 per cent of total variation was accounted for among cities. We repeated the nested analysis using only San Juan and Mayaguez, to determine whether this pattern of extreme variation among local collections resulted from small sample sizes. Variation among *barrios* within groups still accounted for the majority of the total variance (65 per cent).

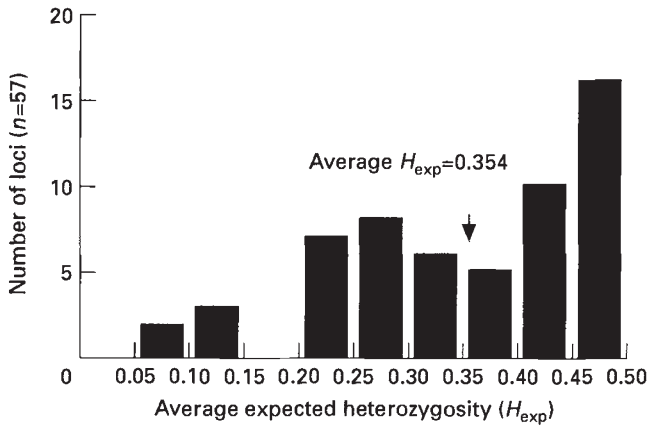

Fig. 1 Distribution for Puerto Rican *Aedes aegypti* of average expected heterozygosity values among 57 RAPD loci.

Estimation of F_{ST} *and* Nm

Estimation of *Nm* from F_{ST} {12} assumes an island model of migration among populations. In our study this would imply a large population (Puerto Rico) that is split into many geographically dispersed subpopulations (cities). Each subpopulation is assumed to be large enough for genetic drift to be negligible. The frequency of alleles in migrants among subpopulations is assumed to be equal to average allele frequencies in the overall population. These assumptions may be valid for city-wide *A. aegypti* populations but the nested analysis of variance suggested genetic drift among local *barrios*. This violates assumptions of the island model; thus *Nm* was only estimated among cities.

Estimates of F_{ST} and θ averaged over all loci are listed in Table 2. The average estimates were similar with all three methods but the standard deviation of estimates among loci was largest with Lynch & Milligan's (1994) F_{ST}. The correlation across loci between Wright's (1951) F_{ST} and θ was high ($r \geq 0.99$, $P \leq 0.0001$) but was slightly lower between Wright's F_{ST} or θ and Lynch & Milligan's F_{ST} (range: $r = 0.87$–0.97, $P \leq 0.0001$).

Frequency histograms of F_{ST} estimates at individual loci among the six cities are shown in Fig. 2. The distributions of Wright's F_{ST} and θ are similar and skewed towards lower values. Lynch & Milligan's F_{ST} is more widely distributed in both the low and high ends of the distribution. Frequency histograms were also plotted for the three F_{ST} estimates within the six cities (analyses not shown) and a similar pattern was detected. Whereas the average F_{ST} estimates obtained by the three methods are similar, estimates for individual loci are not.

The distribution of Wright's F_{ST} and θ in individual cities are shown in Fig. 3. The means of the two parameters are almost identical but the range of θ is always larger. The number of large F_{ST} or θ estimates was inversely proportional to the sample size. The mean and variance of F_{ST} or θ were largest in Arecibo and Ponce, which had average sample sizes of 17, and were smallest in San Juan and Mayaguez, with sample sizes of 50. These results all suggest that F_{ST} and θ are inflated by small samples.

Table 2 Estimates for *Aedes aegypti* of F_{ST} and θ within and among six cities in Puerto Rico. *Nm* was not estimated among *barrios* within cities because the observation of genetic drift among *barrios* violates assumptions of the island model

	Wright's F_{ST}	θ	Lynch & Milligan's F_{ST}
San Juan	0.013	0.009	0.003
	(0.015)	(0.022)	(0.030)
Mayaguez	0.020	0.020	0.016
	(0.021)	(0.031)	(0.039)
Arecibo	0.063	0.060	0.058
	(0.077)	(0.106)	(0.125)
Ponce	0.063	0.062	0.058
	(0.063)	(0.089)	(0.103)
Fajardo	0.026	0.029	0.018
	(0.037)	(0.067)	(0.055)
Caguas	0.029	0.031	0.027
	(0.058)	(0.094)	(0.080)
Cities	0.021	0.020	0.025
	(0.024)	(0.030)	(0.055)
Nm	11.7	12.2	9.7

Standard deviations are in parentheses beneath each entry.

Linkage disequilibrium

With 57 loci there are 1596 [$(57 \times 56)/2$] pair-wise comparisons. The proportion of tests for linkage disequilibrium that were significant at the $P = 0.05$

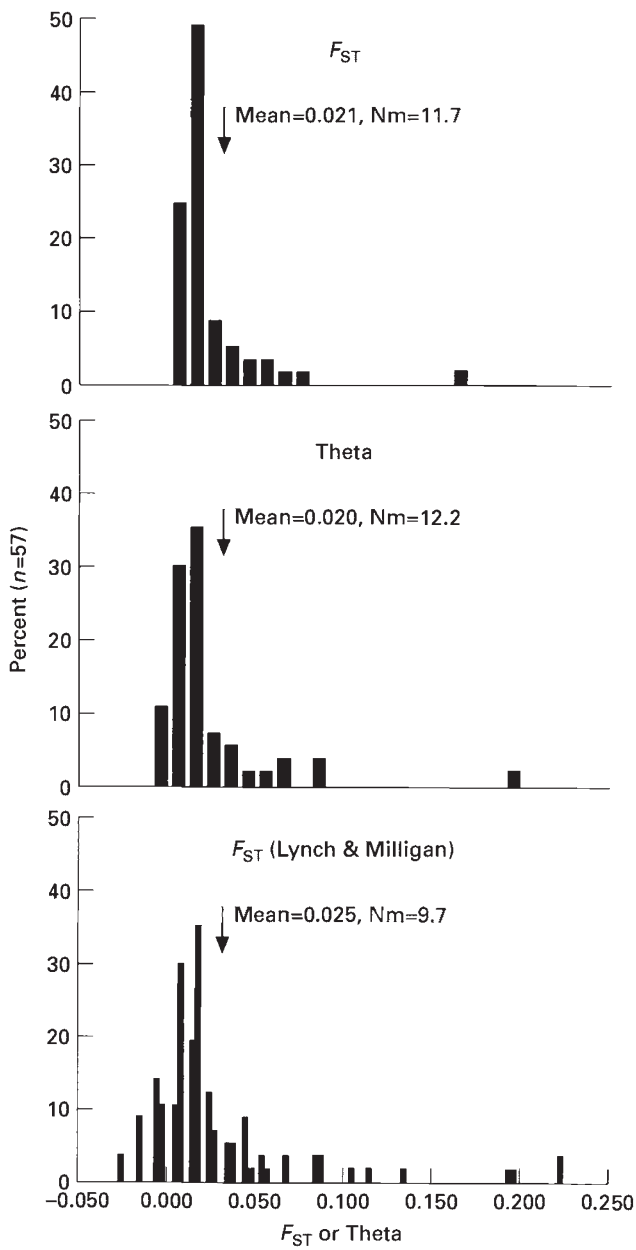

Fig. 2 Distribution for *Aedes aegypti* of F_{ST}- and θ-values at 57 RAPD loci *among* six cities sampled in Puerto Rico.

level are shown in Table 3. The denominator is often less than 1596 because alleles were fixed in some collections. In general, the rate with which significant linkage disequilibrium was detected was consistent with a type I error rate of $\alpha = 0.05$. The single exception was in Mayaguez (Soledad) where 11.4 per cent of analyses were significant.

Values of D_{ij} were transformed to correlation coefficients (R) {19}. Average z-values {20} are shown in Table 3 as are the standard deviations and ranges. The average correlations, their standard deviations and ranges were largest in the smallest collections (Arecibo and Ponce), were intermediate in Caguas and Fajardo and were smallest in San Juan and Mayaguez which had the largest sample sizes. These results demonstrate that small samples inflate linkage disequilibrium estimates.

A large proportion (10.4 per cent) of the linkage disequilibrium coefficients were significant in the overall population. However, linkage disequilibrium in structured populations can arise from drift rather than epistasis. Ohta's analysis of linkage disequilibrium in structured populations was used to test this possibility. The average D_{IT}^2 among 1596 comparisons was 0.03590. The drift component D_{IS}^2 was 0.03498 and accounted for 97.4 per cent of the total whereas the epistasis component D_{ST}^2 was 0.00092 and only accounted for 2.6 per cent. In no comparisons was $D_{ST}^2 \geq D_{IS}^2$.

Epistasis accounted for very little of the total disequilibrium found in this study.

The distribution of significant disequilibrium coefficients among loci was examined to determine if certain pairs of loci appeared in disequilibrium

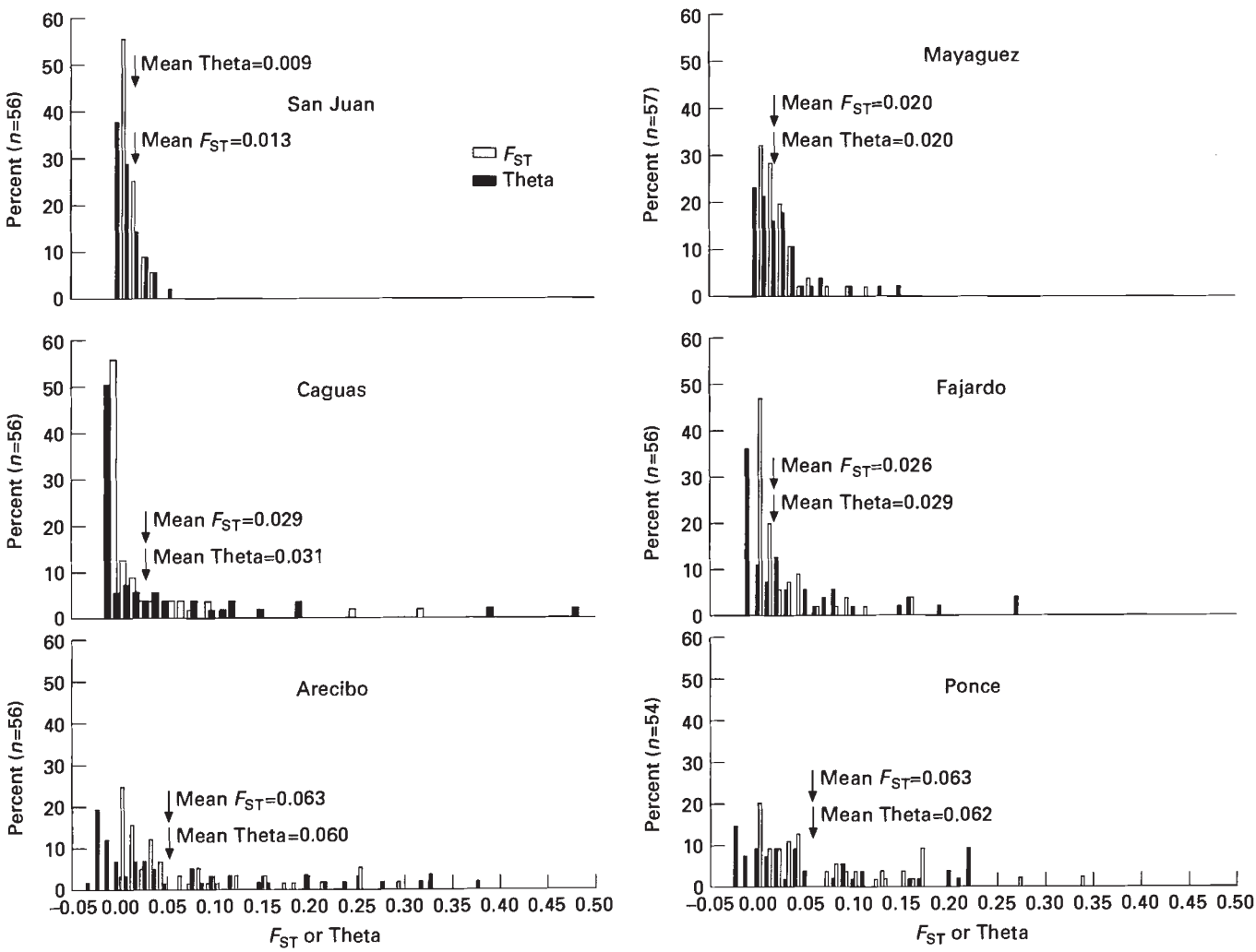

Fig. 3 Distribution for *Aedes aegypti* of F_{ST}- and θ-values among 57 RAPD loci in each of the six cities sampled in Puerto Rico.

Table 3 Analysis for Puerto Rican *Aedes aegypti* of linkage disequilibrium by *barrio* and in all samples combined. The percentage of two-locus linkage disequilibrium analyses that were significant at $P = 0.05$ is indicated. The average size, standard deviation and range of the *z*-transformed correlation is also shown

City	*Barrio*	Per cent of $D_{ij} > 0$ ($P = 0.05$)	Average correlation	SD	Range
San Juan					
	A = Reparto Metropolitano	78/1431 = 5.5%	0.1200	0.0960	0.8522
	B = Puerto Nuevo	95/1540 = 6.2%	0.1257	0.1007	0.8670
	C = Toa Baja	72/1540 = 4.7%	0.1209	0.0920	1.1211
Mayaguez					
	A = Paris	84/1485 = 5.7%	0.1251	0.0912	0.6498
	B = Balboa	95/1540 = 6.2%	0.1281	0.0997	0.8670
	C = Soledad	182/1596 = 11.4%	0.1547	0.2237	7.6246
Arecibo					
	A = Miramar	53/1225 = 4.3%	0.2243	0.2267	6.0870
	B = Garcia	55/1176 = 4.7%	0.2301	0.2881	8.3024
	C = Las Brisas	70/1176 = 6.0%	0.2414	0.2423	6.1024
Ponce					
	A = Cirio del Sur	53/990 = 5.4%	0.2211	0.1593	1.0529
	B = Boa Caribe	75/1378 = 5.4%	0.2375	0.3370	8.3024
	C = Los Caobos	40/946 = 4.2%	0.2265	0.1657	1.1948
Fajardo					
	A = Florencio	103/1378 = 7.5%	0.1963	0.2656	6.1024
	B = Monte Brisa	70/1378 = 5.1%	0.1709	0.1281	0.9473
Caguas					
	A = Villa del Rey	87/1485 = 5.9%	0.1782	0.1300	0.8522
	B = Villa Borinquen	86/1378 = 6.2%	0.1934	0.2121	6.1024
All samples combined		166/1596 = 10.4%	0.0437	0.0389	0.5906

more frequently than expected by chance alone. The 16 collections were treated as independent Bernoulli trials with the probability of any two loci being found in significant disequilibrium equal to 0.05. If alleles at RAPD loci associate independently then the expected frequency with which a pair of loci appear to be in disequilibrium will follow a binomial distribution in which the probability that a pair of loci appears x times in 16 trials is

$$\binom{16}{x} p^x q^{16-x}.$$

The distributions of observed and expected values of x are shown in Table 4. Observed and expected frequencies were similar, suggesting no consistent linkage disequilibrium among alleles at any pairs of RAPD loci.

Discussion

RAPD–PCR reveals large numbers of genetic polymorphisms. The average genic heterozygosity among RAPD loci ($H = 0.354$) was over twice that among 11 allozyme loci in an earlier survey in Puerto Rico ($H = 0.163$; Wallis *et al.*, 1984) or in a survey of 23 allozyme loci in *A. aegypti* populations world-wide ($H = 0.152$; Tabachnick *et al.*, 1979).

Estimates of F_{ST} were upwardly biased and *Nm* underestimated when sample sizes were small. Studies of *Nm* which have used small samples may have overestimated F_{ST} and underestimated migration rates. Slatkin & Barton (1989) found that Weir & Cockerham's (1984) θ tended to overestimate *Nm*. The largest *Nm* was estimated from Weir & Cockerham's θ in our study. Lynch & Milligan's (1994) F_{ST} was smaller than Wright's (1951) F_{ST} or θ in all studies within cities but was largest when estimating F_{ST} among cities or among all *barrios*. This effect arose primarily from corrections for the variance and covariance among H_B and H_W terms. The correction term (by which H_B/H_T is multiplied by in {4}) was close to 1 for within cities estimates but was 1.5 among cities or 1.6 for estimates among all *barrios*.

The mean and variance of transformed disequi-

Table 4 The number of subpopulations of Puerto Rican *Aedes aegypti* in which a significant linkage disequilibrium was observed at a pair of loci compared with the number expected assuming independent segregation of alleles at RAPD loci. Binomial expectations were computed assuming that the probability that alleles at any pair of loci would be in linkage disequilibrium (at the $P = 0.05$ level) would be 5% and treating each of the 16 subpopulations as independent sampling events. The critical value for χ^2 goodness-of-fit at $\alpha = 0.05$ for 15 d.f. is 25.00

No. of subpopulations	Binomial expected frequency	Expected no. (= exp. freq. × 1596)	Observed number	χ^2
0	0.440127	702.4422	716	0.26
1	0.370633	591.5302	563	1.38
2	0.146302	233.4988	237	0.05
3	0.035934	57.35058	66	1.30
4	0.006147	9.809967	8	0.33
5	0.000776	1.239154	5	11.42
6	0.000075	0.119567	1	6.48
7	5.6×10^{-6}	0.00899	0	0.01
8–16	3.5×10^{-7}	0.000558	0	0.00
Total	1.000000	1596	1596	21.23

librium correlation coefficients increased with decreasing sample size. The size of the recombinational map in *A. aegypti* is between 180 and 230 cM (Munstermann & Craig, 1979; Severson *et al.*, 1993). If the RAPD loci were uniformly distributed then this would provide a resolution of one marker every 3–4 cM. Despite this high level of resolution, no linkage disequilibrium maintained through epistasis was detected.

If the island model is valid, the *Nm* values for *A. aegypti* among cities (9.7–12.2) are among the largest reported for insects (see Slatkin, 1985 for comparison) and suggest that *A. aegypti* rapidly disperses. *Aedes aegypti* is a container breeding species that reproduces continuously in tropical or subtropical regions. Mosquitoes emerge from an oviposition site and mate. The female takes a blood meal, matures a batch of eggs and then visits containers (e.g. cisterns, discarded tyres, cans, flowerpots) to oviposit. Sibling analysis (Apostol *et al.*, 1994) and analysis of rubidium-marked eggs from released females (Reiter *et al.*, 1995) showed that a female oviposits only a few of her eggs in individual containers. Reiter *et al.* (1995) monitored dispersal over a distance of 440 m and Hausermann *et al.* (1971) measured dispersal at 580 m with genetically marked strains. These probably underestimate lifetime dispersal but suggest that females deposit few eggs at individual sites as they disperse over long distances.

The large amount of variation detected among proximate collections in the nested analysis of variance supports a model of genetic drift among *A. aegypti* populations in cities. However, genetic drift is counterintuitive given the high *Nm*. Similar patterns were detected with a closely related container breeding mosquito, *A. albopictus*, in North America and Malaysia (Black *et al.*, 1988a,b). Strong genetic differentiation was observed at a local level but allele frequencies were more homogeneous at greater geographical scales. One explanation for the pattern in both species could be the high larval mortality that occurs in container breeding mosquitoes (Service, 1993). Many offspring from different females are oviposited in a container but local populations are maintained by only a few adults that survive to reproduce. This would lead to random genetic differentiation at a local scale while maintaining genetic homogeneity at larger geographical scales.

Acknowledgements

We thank Jose Santo Domingo and Juan Valentin for mosquito collections and the Puerto Rico residents for their cheerful co-operation.

References

APOSTOL, B. L., BLACK, W. C., IV, MILLER, B. R., REITER, P. AND BEATY, B. J. 1993. Estimation of the number of full sibling families at an oviposition site using RAPD–PCR markers: applications to the mosquito *Aedes aegypti*. *Theor. Appl. Gen.*, **86**, 991–1000.

APOSTOL, B. L., BLACK, W. C., IV, REITER, P. AND MILLER, B. R. 1994. Use of RAPD–PCR markers to estimate the number of *Aedes aegypti* families at oviposition sites in

San Juan, Puerto Rico. *Am. J. Trop. Med. Hyg.*, **51**, 89–97.

BLACK, W. C., IV AND KRAFSUR, E. S. 1985. A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. *Theor. Appl. Genet.*, **70**, 491–496.

BLACK, W. C., IV, FERRARI, J. A., RAI, K. S. AND SPRENGER, D. A. 1988a. Breeding structure of a colonizing species: *Aedes albopictus* in the United States. *Heredity*, **60**, 173–181.

BLACK, W. C., IV, HAWLEY, W. A., RAI, K. S. AND CRAIG, G. B., JR 1988b. Breeding structure of a colonizing species: *Aedes albopictus* (Skuse) in peninsular Malaysia and Borneo. *Heredity*, **61**, 439–446.

HAUSERMANN, W., FAY, R. W. AND HACKER, C. S. 1971. Dispersal of genetically marked female *Aedes aegypti* in Mississippi. *Mosq. News*, **31**, 37–51.

HILL, W. G. 1981. Estimation of effective population size from data on linkage disequilibrium. *Genet. Res.*, **38**, 209–216.

LYNCH, M. AND MILLIGAN, B. G. 1994. Analysis of population genetic structure with RAPD markers. *Mol. Ecol.*, **3**, 91–99.

MUNSTERMANN, L. E. AND CRAIG, G. B., JR 1979. Genetics of *Aedes aegypti*: updating the linkage map. *J. Hered.*, **70**, 291–296.

OHTA, T. 1982a. Linkage disequilibrium due to random genetic drift in finite subdivided populations. *Proc. Natl. Acad. Sci. U.S.A.*, **79**, 1940–1944.

OHTA, T. 1982b. Linkage disequilibrium with the island model. *Genetics*, **101**, 139–155.

REITER, P., AMADOR, M. A. AND COLON, N. 1991. Enhancement of the CDC ovitrap with hay infusions for daily monitoring of *Aedes aegypti* populations. *J. Am. Mosq. Control Assoc.*, **7**, 52–55.

REITER, P., AMADOR, M. A., ANDERSON, R. A. AND CLARK, G. G. 1995. Dispersal of *Aedes aegypti* in an urban area after blood feeding as demonstrated by rubidium marked eggs. *Am. J. Trop. Med. Hyg.*, **52**, 177–179.

SERVICE, M. W. 1993. *Mosquito Ecology: Field Sampling Methods*, 2nd edn. Elsevier Applied Science, New York.

SEVERSON, D. W., MORI, A., ZHANG, Y. AND CHRISTENSEN, B. M. 1993. Linkage map for *Aedes aegypti* using restriction fragment length polymorphisms. *J. Hered.*, **84**, 241–247.

SLATKIN, M. 1985. Rare alleles as indicators of gene flow. *Evolution*, **39**, 53–65.

SLATKIN, M. AND BARTON, N. H. 1989. A comparison of the three indirect methods for estimating average levels of gene flow. *Evolution*, **43**, 1349–1368.

SOKAL, R. R. AND ROHLF, F. J. 1981. *Biometry*, 2nd edn. W. H. Freeman and Co., New York.

SWOFFORD, D. L. AND SELANDER, R. B. 1981. BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. *J. Hered.*, **72**, 281–283.

TABACHNICK, W. J. 1991. The yellow fever mosquito: evolutionary genetics and arthropod-borne disease. *Am. Entomol.*, **37**, 14–24.

TABACHNICK, W. J., MUNSTERMANN, L. E. AND POWELL, J. R. 1979. Genetic distinctness of sympatric forms of *Aedes aegypti* in east Africa. *Evolution*, **33**, 287–295.

WALLIS, G. P., TABACHNICK, W. J. AND POWELL, J. R. 1984. Genetic heterogeneity among Caribbean populations of *Aedes aegypti*. *Am. J. Trop. Med. Hyg.*, **3**, 492–498.

WEIR, B. S. 1979. Inferences about linkage disequilibrium. *Biometrics*, **35**, 235–254.

WEIR, B. S. AND COCKERHAM, C. C. 1984. Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

WELSH, J. AND MCCLELLAND, M. 1990. Fingerprinting genomes using PCR with arbitrary primers. *Nucl. Acids Res.*, **18**, 7213–7218.

WILLIAMS, J. K., KUBELIK, A. R., LIVAK, K. J., RAFALSKI, J. A. AND TINGEY, S. V. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucl. Acids Res.*, **18**, 6531–6535.

WORKMAN, P. L. AND NISWANDER, J. D. 1970. Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. *Am. J. Hum. Genet.*, **22**, 24–49.

WRIGHT, S. 1931. Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

WRIGHT, S. 1951. The genetical structure of populations. *Ann. Eugen.*, **15**, 323–354.

WRIGHT, S. 1978. *Evolution and the Genetics of Populations*, vol. 4, *Variability Within and Among Natural Populations*. University of Chicago Press, Chicago, IL.

Appendix

Estimation of allele frequencies and average heterozygosities

Using the four assumptions described in the text, we estimated $q_j(i)$, the frequency of the null allele a at locus i ($i = 1,\ldots,L$) in population j ($j = 1,\ldots,r$), as:

$$q_j(i) = \sqrt{x_j(i)} \qquad (1)$$

where $x_j(i)$ is the frequency of null recessive homozygotes in subpopulation j at locus i.

Estimation of F_{ST} *and* Nm

There are a variety of methods used to estimate F_{ST}; however, only three can be used without information on heterozygote frequencies. These are the methods of Wright (1951), Weir & Cockerham (1984) and Lynch & Milligan (1994). With Wright's method:

$$F_{ST} = s_q^2/(\bar{q}(1-\bar{q})) \qquad (2)$$

where s_q^2 is the variance in the frequency of a RAPD allele among subpopulations and $\bar{q}$ is the weighted average frequency among all subpopulations. This F_{ST} is the ratio of the observed variance in the frequency of an allele among subpopulations relative

to its maximum variance in the total population. Multiplying this estimate of F_{ST} by $2N$ is equivalent to performing a χ^2 contingency test (Workman & Niswander, 1970) using the formula:

$$\chi^2 = 2NF_{ST} \text{ where } N = \sum_{j}^{r} n_j \quad (3)$$

with d.f. = r − 1.

BIOSYS-1 cannot estimate F_{ST} without heterozygote frequencies. A FORTRAN program, RAPDFST, was written by W.C.B.4 that uses equation {2} to compute F_{ST} values for each RAPD locus. It then calculates χ^2 contingency values with {3} and computes probability values to test for significance. Weir & Cockerham (1984) introduced θ as an estimator of F_{ST} that incorporates small and unequal sample sizes in subpopulations as well as small numbers of subpopulations. RAPDFST estimates θ assuming that genotypes are in Hardy–Weinberg proportions using the formula of Weir & Cockerham (1984; top of p.1364). Lynch & Milligan (1994) estimate F_{ST} as a ratio of the excess heterozygosity among genetically differentiated subpopulations (H_B) relative to heterozygosity in the total population (H_T) using the equation:

$$F_{ST} = \frac{H_B}{H_T} \times \left(1 + \frac{H_B \text{Var}(H_W) - H_W \text{Var}(H_B) + (H_B - H_W)\text{Cov}(H_B, H_W)}{H_B H_T^2} \right)^{-1} \quad (4)$$

where

$$H_T = H_W + H_B. \quad (5)$$

Equations appearing in Lynch & Milligan (1994) are indicated as LM[#] throughout this appendix, where # is the reference number of the equation in their paper. H_W estimates the average expected heterozygosity within subpopulations. It is estimated in RAPDFST using LM[4a, 5 and 7]. The variance of H_W is calculated with LM[1]. H_B estimates the excess heterozygosity arising among genetically differentiated subpopulations. It is estimated in RAPDFST from LM[9a, 10a, 12, and 13]. The variance of H_B is calculated with LM[13b]. The covariance term in {4} is LM[14b].

V_B and C_B are the variance and covariance, respectively, among estimates of H_{jk} (from LM[12]). However, V_B is estimated from **nonoverlapping** values of H_{jk} whereas C_B is estimated from **overlapping** values of H_{jk}.

$$V_B = \frac{1}{(novl - 1)} \left[\sum_{j<k}^{r} H_{jk}^2 - \left(\frac{\sum_{j<k}^{r} H_{jk}}{novl} \right)^2 \right] \quad (6)$$

for **nonoverlapping** values of H_{jk} (e.g. H_{12}, H_{34}, H_{56}, etc.) where *novl* = number of nonoverlapping values of H_{jk}.

$$C_B = \frac{1}{(ovl - 1)} \left[\sum_{\substack{j<k \\ j<l}}^{r} H_{jk} H_{jl} - \left(\frac{\sum_{j<k}^{r} H_{jk} \sum_{j<k}^{r} H_{jl}}{ovl} \right) \right] \quad (7)$$

for **overlapping** values of $H_{jk}H_{jlk}$ (e.g. [H_{12}, H_{13}] [H_{12}, H_{14}] [H_{34}, H_{35}]) where *ovl* = number of overlapping values of H_{jk}. Lynch & Milligan (1994) do not provide a method to estimate F_{ST} at individual loci. In RAPDFST we calculate $H_W(i)$, $H_B(i)$, $\text{Var}(H_W(i))$, $\text{Var}(H_B(i))$ and $\text{Cov}(H_W(i), H_B(i))$ for individual loci. These values are then used in {4} to estimate $F_{ST}(i)$. We calculate $H_W(i)$ as:

$$H_W(i) = \frac{1}{r} \sum_{j=1}^{r} H_j(i) \quad \text{(8) from LM[7]}$$

and $\text{Var}(H_W(i))$ as:

$$\text{Var}(H_W(i)) = \frac{1}{r(r-1)} \sum_{j=1}^{r} (H_j(i) - H_W(i))^2. \quad \text{(9) from LM[1]}$$

$H_B(i)$ is:

$$H_B(i) = \frac{2}{r(r-1)} \sum_{j<k}^{r} H_{jk}(i) \quad \text{(10) from LM[13a]}$$

and $\text{Var}(H_B(i))$ is:

$$\text{Var}(H_B(i)) = \frac{2[V_B(i) + 2(n-2)C_B(i)]}{r(r-1)} \quad \text{(11) from LM[13b]}$$

where $V_B(i)$ is calculated from {6}, substituting $H_{jk}(i)$ for H_{jk}, and $C_B(i)$ is calculated from {7}, substituting $H_{jk}(i)$ and $H_{jl}(i)$ for H_{jk} and H_{jl}, respectively. The $\text{Cov}(H_W(i), H_B(i))$ is estimated from LM[14b] by substituting $H_j(i)$ for H_j and $H_{jk}(i)$ for H_{jk}. With all three methods RAPDFST calculates Nm from:

$$Nm = (1 - F_{ST})/4F_{ST} \quad (12)$$

substituting θ for F_{ST}. This comes from the formula originally derived by Wright (1931):

$$F_{ST} = 1/(4Nm+1). \quad (13)$$

Linkage disequilibrium analysis

The frequencies p_{xy} of the four types of two-locus genotypes at loci X and Y are estimated, where p_{11} is the observed frequency of individuals with a band at both loci, p_{10} is the observed frequency of individuals with a band at locus X but not locus Y, p_{01} is the observed frequency of individuals with no band at locus X but one at locus Y, and p_{00} is the observed frequency of individuals with bands at neither locus.

Next, we estimate expected di-locus genotype frequencies. The expected frequency of individuals that produce a band at locus X is $B_X = p^2 + 2pq$ {14} and the frequency of individuals that produce no band is $b_X = q^2$ {15}. With n individuals in a population, the expected frequencies of genotypes at a pair of loci are then:

$$P_{11} = B_X B_Y n, \quad (16a)$$

$$P_{10} = B_X b_Y n, \quad (16b)$$

$$P_{01} = b_X B_Y n \text{ and} \quad (16c)$$

$$P_{00} = b_X b_Y n. \quad (16d)$$

If loci are in linkage equilibrium then:

$$p_{xy} = P_{xy}. \quad (17)$$

Otherwise linkage disequilibrium occurs and is computed as:

$$D_{xy} = p_{xy} - P_{xy}. \quad (18)$$

It can be shown that D_{xy} is equivalent over all values of x and y. The magnitude of D_{xy} is affected by allele frequencies but can be standardized across all allele frequencies by dividing D_{xy} by:

$$\sqrt{B_X b_X B_Y b_Y}. \quad (19)$$

Disequilibrium calculated in this way is a correlation R (Weir, 1979; Hill, 1981) and can vary from 0 to 1. The distribution of correlations is markedly asymmetrical but can be transformed to z-values, which have a normal distribution, using the equation (Sokal & Rohlf, 1981):

$$z = 0.5 \ln [(1+R)/(1-R)]. \quad (20)$$

When z is small it is very similar to R but diverges as the absolute value of R approaches 1.

A χ^2 goodness-of-fit test with one degree of freedom can be used to test for significant disequilibrium by:

$$\chi^2 = \sum_x \sum_y (p_{xy} - P_{xy})^2 / P_{xy} \quad (21)$$

over all four combinations of x and y (Weir, 1979). These values were calculated for all alleles in all subpopulations with a FORTRAN program RAPDLD written by W.C.B.4.

The derivations of these formulae using Weir & Cockerham's (1984) composite disequilibrium statistic is discussed briefly by Black & Krafsur (1985). Ohta (1982a,b) developed a nomenclature for these statistics that was meant to parallel Wright's F-statistics (i.e., I = individuals, S = subpopulations, T = total population). Intuitively, D_{IS}^2 should measure disequilibrium arising through epistasis and D_{ST}^2 should measure disequilibrium arising from drift. Ohta labels the additive components of this model D'^2_{IS} and D'^2_{ST} opposite to this intuition. D'^2_{IS} is

$$\sum_x \sum_y (p_{xy,j} - p_{xy})^2 \quad (22)$$

which is the variance in p_{xy} among r populations summed over all values of x and y. D'^2_{ST} is:

$$\sum_x \sum_y (p_{xy} - P_{xy})^2 \quad (23)$$

which is the variance in D_{xy} among r populations summed over all values of x and y. D'^2_{IS} and D'^2_{ST} sum to D'^2_{IT} which is:

$$\sum_x \sum_y (p_{xy,j} - P_{xy})^2 \quad (24)$$

and is the total variance of disequilibrium in a subdivided population summed over all values of x and y. This is the number of times a pair of alleles occurs in an individual relative to the frequencies of those alleles in the total population. If there is no reproductive isolation among subpopulations then disequilibrium can only arise through epistasis and $D'^2_{IT} = D'^2_{ST}$.

The program RAPDLD estimates D'^2_{IS}, D'^2_{ST} and D'^2_{IT} and determines what proportion of D'^2_{IT} is attributable to D'^2_{IS} or D'^2_{ST}. The variance arising from drift usually accounts for most of the variance. The program flags cases in which $D'^2_{ST} \geq D'^2_{IS}$.

All of the FORTRAN programs described above are available by anonymous FTP from lamar.colostate.edu. The programs are located in the pub/wcb4 directory. Follow the instructions in the READ.ME file to acquire the programs, instructions and sample input and output data files.