Theor Appl Genet (1985) 70: 484–490

A FORTRAN program for analysis of genotypic frequencies and description of the breeding structure of populations *

W. C. Black, IV and E. S. Krafsur

Department of Entomology, Iowa State University, Ames, IA 50011, USA

Received February 20, 1984; Accepted March 18, 1985
Communicated by A. Robertson

Summary. A FORTRAN program, "Genestats" was designed to analyse genotypic and allelic frequencies in subpopulations. The genotypes of individuals gathered from electrophoretic analysis at one or more loci are submitted. The program subsequently calculates allele frequencies, determines if significant heterogeneity exists among subpopulations, tests for departures from random mating in subpopulations and calculates F-statistics. A description of the statistical methods is provided. Printout from analysis of allozyme data collected from field subpopulations of the house fly (*Musca domestica* L.) is provided to illustrate and evaluate the computational methods.

Key words: FORTRAN – F-statistics – Electrophoresis – Breeding structure

Introduction

Electrophoretic resolution of allozymes has provided population biologists the means for determining genotypic frequencies among individuals. The breeding structure of field and laboratory populations can be monitored through the periodic censusing of allelic and genotypic frequencies. Statistical analysis of these frequencies allows objective conclusions to·be reached concerning mating patterns within subpopulations and the degree of reproductive isolation among subpopulations.

Extracting allele frequency data from the genotypes of many individuals can be a tedious and inaccurate process when done by hand. Subsequent tests for random mating in subpopulations and analysis of variance of allele frequencies among subpopulations are similiarly laborious and susceptible to error. To overcome these problems, a FORTRAN program "Genestats" was written which calculates allele frequencies from genotype data extracted from individuals. The program subsequently performs any combination of several statistical operations on the data. A more extensive program, BIOSYS-1 (Swofford and Selander 1981), is available which performs some of the operations listed here and a broad array of others helpful in phylogenetic studies. Our program concentrates solely on the analysis of population structure. "Genestats" consists of 850 executable statements, is written in IBM FORTRAN IV and can be run on FORTRAN G and WATFIV compilers. A printed copy of the program, a test data set, and instructions for its use are available from the authors at cost.

The statistical methods used to analyse the breeding structure of populations are scattered throughout the literature. The present report compiles these methods. As each method is described, the corresponding table of computer output is discussed. The data upon which analysis was performed represent a series of house fly (*Musca domestica* L.) populations sampled at different locations in central Iowa.

The program

Data input

The user lists the desired options (Table 1) and provides the names of the subpopulations, the names and number of loci examined, and the number of alleles at each locus. The genotype of each individual is then entered. A key to the subpopulations and the options requested are printed each time the program is run (Table 1).

Allele frequency analysis

The program calculates allele frequencies in subpopulations. Upon request, these frequencies and the corresponding sample

* Journal Paper No. J-11282 of the Iowa Agriculture and Home Economics Experiment Station, Ames, Iowa. Project No. 2411

Table 1. Printout of subpopulation key and options that may be requested by the user

Key to subpopulations

Subpopulation no.	Subpopulation name
1	Nutrition 10/12/82
2	Sheep 10/12/82
3	Pork 10/14/82
4	Swine 10/14/82
5	Dairy 10/14/82

Options requested: Allele frequency calculations
Chi-square analysis on allele frequencies
Test for departures from random mating
- Chi-square analysis on each genotypic class
- Chi-square analysis on observed and expected proportions of heterozygotes

F-statistics: Nei
F-statistics: Weir and Cockerham

Table 2. Allele frequencies detected in 5 house fly subpopulations

Locus	Allele frequencies in subpopulations				
	1.	2.	3.	4.	5.
AMY					
(*N*)	50	50	48	50	50
1	0.010	0.040	0.000	0.010	0.010
2	0.080	0.030	0.010	0.020	0.040
3	0.130	0.110	0.094	0.100	0.180
4	0.740	0.710	0.844	0.830	0.610
5	0.040	0.110	0.031	0:030	0.120
6	0.000	0.000	0.021	0.010	0.040
ADH					
(*N*)	50	50	49	50	50
1	0.000	0.010	0.000	0.010	0.010
2	0.740	0.690	0.714	0.750	0.660
3	0.150	0.160	0.092	0.170	0.160
4	0.110	0.130	0.184	0.070	0.170
5	0.000	0.010	0.010	0.000	0.000
PGM					
(*N*)	50	50	50	50	50
1	0.010	0.000	0.000	0.000	0.010
2	0.000	0.000	0.030	0.010	0.000
3	0.950	1.000	0.950	0.980	0.990
4	0.040	0.000	0.020	0.010	0.000
SOD					
(*N*)	50	50	50	50	50
1	0.980	0.960	0.940	0.950	0.950
2	0.020	0.040	0.060	0.050	0.050

sizes are printed (Table 2). In the following discussion, the calculated frequency of an allele x at locus k in subpopulation i will be denoted p_{ikx}. The next option calculates the weighted mean allele frequency $\bar{p}_{kx}$,

$$\bar{p}_{kx} = \sum_i w_i p_{ikx} \quad \text{(Table 3)} \tag{1}$$

where

$$w_i = n_i/N \tag{1a}$$

n_i is the sample size from subpopulation i, and N is the sum of all the samples. An estimate of the weighted variance uncorrected for sample number (Workman and Niswander 1970) is next calculated,

$$s_{kx}^2 = \left(\sum_i w_i p_{ikx}^2\right) - \bar{p}_{kx}^2 . \tag{2}$$

A chi-square test for homogeneity of allele frequencies among subpopulations is performed for each allele (Workman and Niswander 1970):

$$\chi^2 = 2N\, s_{kx}^2/\bar{p}_{kx}(1-\bar{p}_{kx}) \quad \text{(Table 3)} \tag{3}$$

with $(r-1)$ degrees of freedom for r subpopulations. A heterogeneity chi-square value is subsequently calculated for all alleles at a locus. The formula (Workman and Niswander 1970) is:

$$\chi^2 = 2N \sum_x (s_{kx}^2/\bar{p}_{kx}) \quad \text{(Table 3)} . \tag{4}$$

This test employs $(r-1)$ $(m-1)$ degrees of freedom, where m is the number of alleles at the locus. These heterogeneity chi-square values are summed over all loci to give an overall comparison of subpopulations. The significance of each chi-square value is tested and printed in a subroutine which approximates the cumulative chi-square distribution.

If the chi-square tests on specific alleles proved significant at the 10% level, a table is printed (Table 4) indicating the relative contributions of the different subpopulations to the chi-square value.

Tests for random mating

"Genestats" tabulates and lists the number of individuals observed in each genotypic class (Table 5). It then calculates and prints the number of individuals expected by the Hardy-Weinberg rule. The chi-square statistic calculated for each genotypic class is then printed. The total chi-square accumulated over all genotypes is given with the degrees of freedom (number of genotypes minus the number of alleles) and its significance. This method is suspect when the expected numbers in classes are less than one (e.g. *ADH*, Table 5). The user is provided the option of suppressing this lengthy output and printing the second half of Table 5. The observed and expected numbers of heterozygotes are printed along with a conservative chi-square test for the homogeneity of observed $H_0(ik)$ and expected $H_S(ik)$ heterozygote frequencies. The formula (Weir and Cockerham 1984) is,

$$\chi^2 = n_i(H_S(ik) - H_0(ik))^2/\left(\sum_x p_{ikx}^2 + \left(\sum_x p_{ikx}^2\right)^2 - 2\sum_x p_{ikx}^3\right) \quad \text{(Table 5)} \tag{6}$$

with one degree of freedom.

Wright's F-statistics

The statistical methods discussed heterofore provide the inferential tools necessary to detect significant departures from random mating in a field population. Chi-square analysis of genotype frequencies within subpopulations (Table 5) indicates significant departures from random mating. Chi-square tests of heterogeneity in allele frequencies among subpopulations (Tables 3 and 4) indicates subpopulation differentia-

Table 3. Weighted mean allele frequencies and analysis of variation in allele frequencies

Chi-square analysis of allele frequencies

Locus	Weighted mean	Chi-square	*P*	Heterogeneity Chi-square	d.f.	*P*
AMY						
(*N*)	248.000		(4 d.f.)			
1	0.014	6.56	0.1613			
2	0.036	8.21	0.0841			
3	0.123	4.47	0.3466			
4	0.746	19.03	0.0008			
5	0.067	12.85	0.0120			
6	0.014	8.11	0.0875			
				43.12	20	0.0020
ADH						
(*N*)	249.000		(4 d.f.)			
1	0.006	1.99	0.7372			
2	0.711	2.64	0.6205			
3	0.147	3.08	0.5439			
4	0.133	7.30	0.1209			
5	0.004	3.04	0.5506			
				14.74	16	0.5441
PGM						
(*N*)	250.000		(4 d.f.)			
1	0.004	3.01	0.5558			
2	0.008	8.57	0.0728			
3	0.974	8.38	0.0787			
4	0.014	8.11	0.0875			
				19.72	12	0.0726
SOD						
(*N*)	250.000		(4 d.f.)			
1	0.956	2.19	0.7001			
2	0.044	2.19	0.7014			
				2.19	4	0.7014
			Total	79.76	52	0.0079

Table 4. Contribution of individual subpopulations to significant ($P \leq 0.1$) variation in allele frequencies

Contribution of subpopulations to structuring

Allele	Chi-square	*P*	Relative contributions to total Chi-square Subpopulation no. 1	2	3	4	5
AMY 2	8.212	0.0841	0.665	0.014	0.224	0.092	0.005
AMY 4	19.029	0.0008	0.001	0.036	0.255	0.196	0.513
AMY 5	12.852	0.0120	0.088	0.237	0.150	0.167	0.358
AMY 6	8.113	0.0875	0.176	0.176	0.038	0.015	0.594
PGM 2	8.569	0.0728	0.094	0.094	0.712	0.006	0.094
PGM 3	8.378	0.0787	0.272	0.319	0.272	0.017	0.121
PGM 4	8.114	0.0875	0.604	0.175	0.032	0.014	0.175

tion. Departures from random mating within and among subpopulations create hierarchical structuring in the total population. Wright (1951) introduced F-statistics as a convenient means of describing the breeding structure of natural populations.

To describe nonrandom mating among individuals in subpopulations, Wright (1951) defined the coefficient F_{IS} as "the average over all subpopulations of the correlation between uniting gametes relative to those of their own subpopulation".

To describe nonrandom mating among individuals from different subpopulations, Wright (1951) defined F_{ST} as "the correlation between random gametes within a subpopulation relative to the gametes within the entire population". F_{ST} is positive when subpopulations are reproductively isolated because random gametes from a subpopulation bear alleles more often derived from a common ancestor than gametes from the total population. A catastrophic reduction in subpopulation size also increases the correlation among random gametes.

Table 5. Analysis of random mating

Subpopulation: nutrition 10/12/82

Locus	Genotype	Observed no.	Hardy-Weinberg Expectation	Chi-square	d.f.	*P*
AMY						
	1/1	0	0.005	0.005		
	1/2	0	0.080	0.080		
	1/3	0	0.130	0.130		
	1/4	1	0.740	0.091		
	1/5	0	0.040	0.040		
	2/2	0	0.320	0.320		
	2/3	1	1.040	0.002		
	2/4	7	5.920	0.197		
	2/5	0	0.320	0.320		
	3/3	0	0.845	0.845		
	3/4	10	9.620	0.015		
	3/5	2	0.520	4.212		
	4/4	27	27.380	0.005		
	4/5	2	2.960	0.311		
	5/5	0	0.080	0.080		
			Total	6.654	10	0.7577
ADH						
	2/2	31	27.380	0.479		
	2/3	8	11.100	0.866		
	2/4	4	8.140	2.106		
	3/3	2	1.125	0.681		
	3/4	3	1.650	1.105		
	4/4	2	0.605	3.217		
			Total	8.452	3	0.0375
PGM						
	1/1	0	0.005	0.005		
	1/3	1	0.950	0.003		
	1/4	0	0.040	0.040		
	3/3	45	45.125	0.000		
	3/4	4	3.800	0.011		
	4/4	0	0.080	0.080		
			Total	0.139	3	0.9868
SOD						
	1/1	48	48.020	0.000		
	1/2	2	1.960	0.001		
	2/2	0	0.020	0.020		
			Total	0.021	1	0.8853

Locus	Heterozygotes		Chi-square	*P*
	Observed	Expected		
AMY	23.0	21.37	0.63	0.4277
ADH	15.0	20.89	6.85	0.0089
PGM	5.0	4.79	0.13	0.7207
SOD	2.0	1.96	0.02	0.8853

To summarise nonrandom mating in the entire population, Wright (1951) introduced a third coefficient, F_{IT}, defined as "the correlation between gametes that unite the produce individuals, relative to the gametes of the total population". F_{IT} describes nonrandom mating arising among and within subpopulations and so is expressed in terms of F_{IS} and F_{ST},

$$F_{IT} = F_{IS} + F_{ST} - [F_{IS} \cdot F_{ST}] . \qquad (6)$$

Wright (1969) discussed situations which give rise to negative or positive values of F_{IT}.

F-statistics have been subject to many qualifications and refinements since the original definitions (Wright 1951). An approach by Nei (1977) has come into general use but not without objections. Weir and Cockerham (1984) note that Nei's method does not account for bias due to small and unequal sample sizes or small numbers of subpopulations. Thus, F-statistics calculated by Nei's method are a function of the sampling scheme. This is of concern in field studies where sample sizes are often unequal and small. It is also of concern when authors compare F-statistics from different studies. A method prescribed by Weir and Cockerham (1984) incorporates sample size and subpopulation number into the F-statistics, thus making interpretation independent of the sampling scheme. The methods of Nei and Weir and Cockerham are both offered as options in "Genestats".

Nonrandom mating due to inbreeding within subpopulations

F_{IS} was originally defined as an "average correlation" over subpopulations. Nei (1977) extended F_{IS} to describe the correlation in a single subpopulation *i* between gametes bearing an allele *x* at locus *k*. This is:

$$F_{IS}(ikx) = 1 - [H_0(ikx)/H_S(ikx)] \quad \text{(Table 6)} \qquad (7)$$

where $H_0(ikx)$ is the observed frequency of heterozygotes with allele *x* and the frequency expected under random mating is:

$$H_S(ikx) = 2\,p_{ikx}(1 - p_{ikx}) \quad \text{(Table 6)} . \qquad (7a)$$

A second refinement of F_{IS} (Nei 1977) calculates the correlation among all alleles at a locus in a subpopulation,

$$F_{IS}(ik) = 1 - [H_0(ik)/H_S(ik)] \quad \text{(Table 6)} . \qquad (8)$$

Here the expected number of heterozygotes is:

$$H_S(ik) = 1 - \sum_x p^2_{ikx} \quad \text{(Table 6)} . \qquad (8a)$$

Following the original definition, F_{IS} is next calculated as the average over all subpopulations of the correlation between uniting gametes bearing a specific allele. Nei (1977) employs observed and expected homozygosities,

$$F_{IS}(kx) = (P_{kx} - \overline{p^2_{kx}})/(\bar{p}_{kx} - \overline{p^2_{kx}}) \quad \text{(Table 6)} , \qquad (9)$$

where

$$P_{kx} = \sum_i w_i P_{ikx} \qquad (9a)$$

and P_{ikx} is the frequency of *x* homozygotes in a subpopulation, w_i is as previously defined (Eq. 1 a) and,

$$\overline{p^2_{kx}} = \sum_i w_i p^2_{ikx} . \qquad (9b)$$

Weir and Cockerham (1984) calculate $F_{IS}(kx)$ by using heterozygote frequencies:

$$F_{IS}(kx) = B_{kx}/(B_{kx} + C_{kx}) \quad \text{(Table 7)} , \qquad (10)$$

Table 6. F-statistics analysis at the *ADH* locus by Nei's (1977) method

Locus (K): *ADH*

F(IS(IKX)) values

	Subpopulation no. (I)				
Allele (X)	1.	2.	3.	4.	5.
1	–	-0.010	–	-0.010	-0.010
2	0.376	0.299	0.200	0.307	0.554
3	0.137	0.256	-0.101	0.220	0.256
4	0.285	0.381	0.047	-0.075	-0.063
5	–	-0.010	-0.010	–	–
F (IS (IK))	0.282	0.293	0.088	0.207	0.294
HS (IK)	0.418	0.481	0.448	0.404	0.510
HO (IK)	0.300	0.340	0.408	0.320	0.360

F-statistics for individual alleles: Nei

Allele (X)	F(IS(KX))	F(ST(KX))	F(IT(KX))
1	-0.010	0.004	-0.006
2	0.352	0.005	0.355
3	0.176	0.006	0.181
4	0.114	0.015	0.127
5	-0.010	0.006	-0.004
	0.236	0.008	0.242
	F(IS(K))	F(ST(K))	F(IT(K))

HO(K) =0.3454
HS(K) =0.4520
HT(K) =0.4556

Table 7. F-statistics analysis at the ADH locus by Wier and Cockerham's (1984) method

F-statistics for individual alleles: W & C

Allele (X)	F(IS(KX))	F(ST(KX))	F(IT(KX))
1	0.000	-0.008	-0.008
2	0.361	-0.010	0.354
3	0.186	-0.008	0.179
4	0.124	-0.002	0.122
5	0.000	-0.006	-0.006
	0.245	-0.003	0.243
	F(IS(K))	F(ST(K))	F(IT(K))

where

$$B_{kx} = [r(\bar{n}-1)]^{-1}\left[\sum_i n_i H_S(ikx) - (2\bar{n})^{-1}(2\bar{n}-1)\sum_i n_i H_0(ikx)\right], \tag{10a}$$

$$C_{kx} = (r\bar{n})^{-1}\sum_i n_i H_0(ikx), \tag{10b}$$

and

$$\bar{n} = \sum_i n_i/r. \tag{10c}$$

F_{IS} is next calculated as the average correlation among all alleles at a locus. With Nei's method,

$$F_{IS}(k) = 1 - [H_0(k)/H_S(k)] \quad \text{(Table 6)}, \tag{11}$$

where $H_0(k)$ and $H_S(k)$ are the weighted averages of $H_0(ik)$ and $H_S(ik)$ over all subpopulations. Using Weir and Cockerham's formulae,

$$F_{IS}(k) = B_k/(B_k+C_k) \quad \text{(Table 7)}, \tag{12}$$

where

$$B_k = [r(\bar{n}-1)]^{-1}\left[\sum_i n_i H_S(ik) - (2\bar{n})^{-1}(2\bar{n}-1)\sum_i n_i H_0(ik)\right], \tag{12a}$$

$$C_k = (r\bar{n})^{-1}\sum_i n_i H_0(ik), \tag{12b}$$

and $\bar{n}$ is as previously defined (Eq. (10c)).

F_{IS} is computed as the average correlation among all alleles at loci. By Nei's method:

$$F_{IS} = \sum_k F_{IS}(k)\cdot H_S(k)/\sum_k H_S(k) \quad \text{(Table 8)}. \tag{13}$$

With Weir and Cockerham's method:

$$F_{IS} = B/(B+C) \quad \text{(Table 9)}, \tag{14}$$

where

$$B = \sum_k B_k \tag{14a}$$

and

$$C = \sum_k C_k. \tag{14b}$$

Nonrandom mating arising from reproductive isolation among subpopulations

Wright (1965) stated that F_{ST} is "the ratio of the actual variance in allele frequencies among subpopulations to the maximum possible variance under complete isolation of subpopulations". This is,

$$F_{ST}(kx) = s_{kx}^2/\bar{p}_{kx}(1-\bar{p}_{kx}). \tag{15}$$

Nei's (1977) formulations of F_{ST} are all based on Eq. (15) and so are all positive. But Cockerham (1969) and Weir and Cockerham (1984) observed that F_{ST} can be negative when there is more variation in allele frequencies within subpopulations than between subpopulations. This agrees with Wright's original (1951) definition of F_{ST} as correlation between random gametes within a subpopulation.

In "Genestats", F_{ST} is first calculated as the correlation between random gametes from subpopulations bearing allele x. Nei calculates $F_{ST}(kx)$ using observed and expected homozygote frequencies:

$$F_{ST}(kx) = (\overline{p_{kx}^2} - \bar{p}_{kx}^2)/(\bar{p}_{kx} - \bar{p}_{kx}^2) \quad \text{(Table 6)}. \tag{16}$$

Weir and Cockerham estimate $F_{ST}(kx)$ by using heterozygote frequencies:

$$F_{ST}(kx) = A_{kx}/(A_{kx}+B_{kx}+C_{kx}) \quad \text{(Table 7)}, \tag{17}$$

where

$$A_{kx} = n_c^{-1}\left[(r-1)^{-1}\sum_i n_i V_{ikx} - (r(\bar{n}-1))^{-1}\cdot\left(\sum_i n_i H_S(ikx) - 1/2\sum_i n_i H_0(ikx)\right)\right], \tag{17a}$$

$$V_{ikx} = (p_{ikx} - \bar{p}_{kx})^2, \tag{17b}$$

$$C^2 = r(\bar{n}^2(r-1))^{-1}\left[\left(\sum_i n_i^2/r\right) - \bar{n}^2\right], \tag{17c}$$

and

$$n_c = \bar{n}\,(1-(C^2/r))\,. \tag{17d}$$

F_{ST} is next estimated as the average correlation among alleles at a locus in random gametes. Nei's statistic is,

$$F_{ST}(k) = 1-[H_S(k)/H_T(k)] \quad \text{(Table 6)}\,, \tag{18}$$

where

$$H_T(k) = 1-\sum_x \bar{p}^2_{kx}\,. \tag{18a}$$

Weir and Cockerham calculate $F_{ST}(k)$ with the formulae:

$$F_{ST}(k) = A_k/(A_k+B_k+C_k) \quad \text{(Table 7)}\,, \tag{19}$$

where

$$A_k = n_c^{-1}\Big[(r-1)^{-1}\sum_i n_i V_{ik} - (r(\bar{n}-1))^{-1} \cdot \Big(\sum_i n_i H_S(ik) - 1/2\sum_i n_i H_0(ik)\Big)\Big]\,, \tag{19a}$$

and

$$V_{ik} = \sum_x V_{ikx}\,. \tag{19b}$$

F_{ST} is computed as the average correlation among alleles at loci in random gametes. Employing Nei's method:

$$F_{ST} = \sum_k H_T(k)\cdot F_{ST}(k)/\sum_k H_T(k) \quad \text{(Table 8)}\,. \tag{20}$$

Weir and Cockerham's formula is:

$$F_{ST} = A/(A+B+C) \quad \text{(Table 9)}\,, \tag{21}$$

where

$$A = \sum_k A_k \tag{21a}$$

and B and C are previously defined (Eqs. (14a, b)).

Table 8. Summary of F-statistics according to Nei

Mean F-statistics over all loci: Nei			
Locus (K)	F(IS(K))	F(ST(K))	F(IT(K))
AMY	0.000	0.026	0.026
ADH	0.236	0.008	0.242
PGM	−0.035	0.016	−0.019
SOD	0.045	0.004	0.049
Mean	0.109	0.015	0.123

Table 9. Summary of F-statistics and standard deviations according to Weir and Cockerham

Mean F-statistics over all loci: W & C			
Locus (K)	F(IS(K))	F(ST(K))	F(IT(K))
AMY	0.011	0.022	0.032
ADH	0.245	−0.003	0.243
PGM	−0.025	0.010	−0.015
SOD	0.055	−0.005	0.050
Mean	0.119	0.008	0.126
S.D. (X)	0.05619	0.00576	0.05073

Nonrandom mating in the total population

Equation (6) is the correlation between alleles in individuals from the entire population. This correlation is calculated in terms of alleles ($F_{IT}(k\,x)$, Tables 6 and 7). It is also estimated as an average among all alleles at a locus ($F_{IT}(k)$, Tables 6 and 7). As a summary statistic, F_{IT} is computed as an average among alleles at loci (Tables 8 and 9).

Estimation of the variance in F-statistics

Weir and Cockerham's method permits the variance of the three F-statistics to be estimated. They advocate a "jackknife" procedure in which estimates are made by successively eliminating one locus at a time. Thus, for example, the variance of F_{ST}, for y loci is,

$$\mathrm{Var}(F_{ST}) = y^{-1}(y-1)\sum_L [F_{ST}(L) - y^{-1}\sum F_{ST}(L)]^2\,, \tag{22}$$

where for locus k,

$$F_{ST}(L) = \sum_{k\neq L} A_k/\sum_{k\neq L}(A_k+B_k+C_k)\,. \tag{22a}$$

The variances of F_{IS} and F_{ST} are calculated in an analogous manner. "Genestats" prints the standard deviations of the three F-statistics (Table 9), rather than variances, to provide additional significant digits.

Results of house fly electrophoresis

The mean F-statistics (Tables 8 and 9) demonstrate an excess of homozygotes in a house fly population sampled at 5 locations. The locus F-statistics pinpoint *ADH* as the principal source of the excess. Allele F-statistics (Tables 6 and 7) show heterozygote deficiencies for *ADH* alleles 2, 3 and 4. Subpopulation statistics (Table 6) reveal that the deficiencies were consistent among subpopulations. Table 5 demonstrates that this heterozygote deficiency was statistically significant in subpopulation 1. It was also significant in subpopulations 2 and 5 (not shown).

Heterogeneity in the frequency of *AMY* alleles contributed most to the mean F_{ST} value (Tables 8 and 9). *AMY* alleles 4 and 5 contributed most to the overall variance. Table 3 demonstrates that the heterogeneity in the frequencies of *AMY* 4 and 5 was significant. Inspection of Table 4 suggests that most of the heterogeneity arose in subpopulation 5.

Discussion

Evaluation of F-statistics

Wright's (1951) original derivation of F-statistics was through correlation analysis. Recent workers (Nei 1967; Weir and Cockerham 1984) maintain this approach, and so it is often assumed that departures from random mating alone cause F-statistics to deviate from zero. In practice F-statistics can assume large values for a variety of reasons. Selection acting on heterozygotes can affect F_{IS}. Selection in subpopulations can alter F_{ST}. In addition to natural causes, technical problems may cause significant F-statistics. Inadequate resolution of allozymes may lead to inaccurate estimates of heterozygote frequencies and thus alter F_{IS}. Null alleles can generate a false excess of homozygotes and positive F_{IS} values. Inconsistent resolution of an allozyme from different samples might increase heterogeneity and thus inflate F_{ST}. In short, a wide array of biological and technical effects can generate positive and negative F-statistics.

For the foregoing reasons F-statistics must be interpreted cautiously. The statistics describe breeding structure but do not alone identify the causes. For example, a variety of sources could have generated the *ADH* heterozygote deficiency in house flies. Had inbreeding been the sole cause, a uniform homozygote excess would have been noted at all loci. The statistics demonstrate that the excess was homogeneous among subpopulations, but offer no further clues as to the cause of the excess.

Comparison of methods for estimating F-statistics

This is the first application of Weir and Cockerham's method to genotypic data from natural populations. Their method had a notable effect on F-statistics (Tables 6–9). F_{ST} values became consistently smaller. Weir and Cockerham's method thus provided a more conservative estimate of the degree of differentiation among subpopulations. Weir and Cockerham's F_{IS}, however, consistently assumed slightly more positive values than Nei's F_{IS}, demonstrating that their method estimates greater amounts of inbreeding. In unpublished work, we applied "Genestats" to larger data sets in which at least 100 flies were sampled from 6 subpopulations. Increasing the sample size diminished the differences in estimates provided by Nei's and Weir and Cockerham's methods. Weir and Cockerham's method should therefore be routinely applied in studies where sample sizes are fewer than, say, 100 individuals.

Frequently breeding structure studies must rely on smaller and less uniform samples than those available when studying house flies. It would be interesting to reevaluate previous studies of breeding structure where samples were inconsistent and small. It will be interesting to see how Weir and Cockerham's method affects conclusions in future studies on the breeding structure of natural populations.

Acknowledgements. We thank Dr. B. S. Weir for discussion of F-statistics and for kindly providing his unpublished manuscript. We thank Dr. E. Pollak for critically reviewing the manuscript. Dr. James Hoekstra made valuable comments on the structure of the program.

References

Cockerham CC (1969) Variance of gene frequencies. Evolution 23:72–84

Nei M (1977) F-statistics and analysis of gene diversity in subdivided populations. Ann Hum Genet 41:225–233

Swofford DL, Selander RB (1981) BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. J Hered 72: 281–283

Weir BS, Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. Evolution 38:1358–1370

Workman PL, Niswander JD (1970) Population studies on southwestern Indian tribes. 2. Local genetic differentiation in the Papago. Am J Hum Genet 22:24–49

Wright S (1951) The genetical structure of populations. Ann Eugenics 15:323–354

Wright S (1965) The interpretation of population structure by F-statistics with special regards to systems of mating. Evolution 19:395–420

Wright S (1969) Evolution and the genetics of populations, vol 2. The theory of gene frequencies. University of Chicago Press, Chicago