BRIEF COMMUNICATIONS

Evolution, 55(12), 2001, pp. 2601–2605

ESTIMATING THE UNBIASED ESTIMATOR θ FOR POPULATION GENETIC SURVEY DATA

JACQUELINE J. WEICKER,[1,2] ROBB T. BRUMFIELD,[3] AND KEVIN WINKER[1]
[1]*University of Alaska Museum, 907 Yukon Drive, Fairbanks, Alaska 99775-6960*
[2]*E-mail: weicker@post.harvard.edu*
[3]*Department of Zoology, Box 351800, University of Washington, Seattle, Washington 98195*

Abstract.—We consider a method of approximating Weir and Cockerham's θ, an unbiased estimator of genetic population structure, using values readily available from published studies using biased estimators (Wright's F_{ST} or Nei's G_{ST}). The estimation algorithm is shown to be useful for both model populations and real-world avian populations. However, the correlation between Wright's F_{ST} and Weir and Cockerham's θ is strong when compared among 39 empirical avian datasets. Thus, the advantage of approximating an unbiased estimator is unclear considering the small actual effect of θ's bias-removing power on empirical datasets.

Key words.—Population structure, Weir and Cockerham's θ, Wright's F_{ST}.

Received February 13, 2001. Accepted August 20, 2001.

Understanding and quantifying the genetic structure of natural populations has been a long-standing objective in evolutionary biology. Determining how genetic variation is distributed within versus among populations can provide insight into genetic population structure, levels of gene flow, historic population parameters, and the early stages of speciation. The prevailing method of describing hierarchical genetic structure has been the use of *F*-statistics, introduced by Wright (1943, 1951, 1965). F_{ST} measures the amount of genetic variation in the total sample that is due to differences among populations comprising that sample; this proportion can range from zero to one. F_{ST} is functionally equivalent to Nei's (1973) G_{ST}, which was derived using expected panmictic heterozygosity rather than the variance in allele frequencies among subpopulations (Nei 1973; Cockerham and Weir 1993). Both of these measures have been used extensively to describe population genetic structure in natural populations, particularly in allozyme studies (e.g., Nevo 1978; Evans 1987).

Although elegant in its simplicity, the estimation of F_{ST} does not account for sampling error. Weir and Cockerham (1984) developed the estimator θ to correct for the error associated with differences in allele frequency distributions between the population samples and the total sample of populations. Simulations confirmed that θ is independent of the number of groups sampled and the number of individuals sampled in each group (Weir and Cockerham 1984; Cockerham and Weir 1993). Although these authors concurred that the use of F_{ST} or G_{ST} is valid when examining diversity among populations for which there has been a complete census, they advocated the use of an unbiased estimator such as θ for most empirical studies, where sample sizes tend to be small.

An ultimate goal in quantifying population genetic structure is to understand variation among species and to determine whether there are patterns among different groups of organisms or within life zones. Making these comparisons among different allozyme studies can be problematic because of differences in loci used and in abilities to distinguish hidden alleles (for caveats, see Barrowclough 1983; Evans 1987). These limitations are somewhat mitigated because many of the same loci are used in vertebrate studies (Nevo 1978). Another possible reason not to compare allozyme studies is the differences in their sampling designs. This problem has not been adequately addressed, and its resolution may not be simple because the vast majority of allozyme studies report population structure in terms of the biased estimator F_{ST}.

It would seem that among-study comparisons using the unbiased θ would be more robust. However, calculating θ from published allele frequency tables is tedious or impossible, with hindrances often arising from missing data and typographic errors. In many cases, original data are simply unavailable; one study in which data were requested from authors of 30 publications with incomplete molecular datasets resulted in only one researcher providing the full dataset (Leberg and Neigel 1999). While this one study involved mitochondrial data, retrieving allozyme data for reanalysis would be even more difficult considering that allozyme studies tend to be older than those using mtDNA.

In this paper we consider a method of approximating θ using values readily available from published studies: F_{ST} (or G_{ST}), the number of populations sampled, and the average number of individuals sampled per population. This approximation is derived from the relationship between G_{ST} and the intraclass correlation β put forth by Cockerham and Weir (1993, p. 858). We then show how well this approximation corresponds with the actual, calculated θ and investigate the utility of the transformation in eliminating sampling bias from estimates of genetic population structure in empirical data.

METHODS

Cockerham and Weir (1993) presented the following formulation of G_{ST} (their eq. 4), the same formulation used by Slatkin and Barton (1989):

$$G_{ST} = \frac{(r-1)\beta + \dfrac{r-1}{2M-1}}{r - \beta + \dfrac{r-1}{2M-1}}, \qquad (1)$$

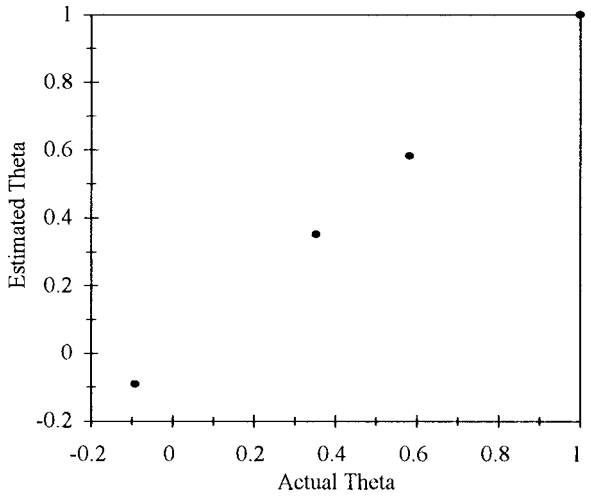

FIG. 1. Actual θ versus estimated θ, using simulated populations. This shows the relationship between the full calculation of Weir and Cockerham's θ (1984), called actual θ, and θ as estimated through equation (2). Adjusted $r^2 = 1.00$, $P < 0.01$, $n = 4$.

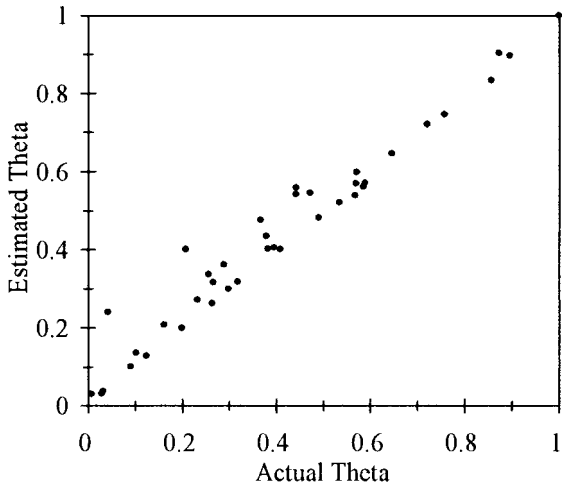

FIG. 2. Actual θ versus estimated θ, using empirical data from 39 studies. This shows the relationship between the full calculation of Weir and Cockerham's θ (1984), called actual θ, and θ as estimated through equation (2). Adjusted $r^2 = 0.96$, $P < 0.01$, $n = 39$.

where r is the number of populations and M is the number of individuals per population. β is a ratio of functions involving the two measures F_0 and F_1, and has the same statistical calculations as θ. For this approximation, they may be regarded as equivalent. Solving for θ (or β), the equation becomes

$$\theta = \frac{G_{ST}\left(r + \frac{r-1}{2M-1}\right) - \frac{r-1}{2M-1}}{G_{ST} + r - 1}. \quad (2)$$

This approximation of θ can be calculated if the values of G_{ST}, r, and M are known.

To demonstrate the efficacy of this approximation, four different simulated datasets were constructed. Each dataset contained 18 individuals divided into three (r) populations of equal size ($M = 6$). For each simulated dataset, individuals were assigned one of three alleles at each of 32 diploid, autosomal loci. These numbers of populations, sample sizes, loci, and alleles all approach the average of several representative allozyme studies on birds (Capparella 1987, 1988; Peterson 1990, 1992; Bates 1993; Brumfield 1993; Brumfield and Capparella 1996; Winker et al. 2000). Alleles were assigned to each individual such that the following four models were created. The first model consisted of panmictic populations sharing all alleles at identical frequencies. A second model consisted of populations with only private alleles, suggesting fixation of those alleles. The other two models had varied frequencies of shared alleles, representing populations with intermediate amounts of gene flow.

For each simulated dataset, Wright's F_{ST} was calculated using BIOSYS-1 (Swofford and Selander 1981). This program was selected because it has been the predominant tool for analyzing allozyme data and was used in the empirical studies cited. In addition, Weir and Cockerham's θ (1984) was calculated for each dataset using the program GDA 1.0 (Lewis and Zaykin 1999). This value shall be referred to as ''actual θ.'' Bootstrapping over loci obtained estimates of upper and lower limits for actual θ after 5000 repetitions.

Finally, the conversion formula above (eq. 2) was used to calculate the approximation of θ, or estimated θ, based on F_{ST}- (from BIOSYS-1), r-, and M-values for each simulated dataset. Actual θ- and estimated θ-values for each dataset were compared using linear regression.

Additionally, empirical data were gathered from allozyme studies of New World landbirds (Capparella 1987, 1988; Peterson 1990, 1992; Bates 1993; Brumfield 1993; Brumfield and Capparella 1996; Winker et al. 2000). From these studies 39 species were selected for which the allozyme frequency data were available or could be derived, and these data were reanalyzed. In some cases where species/subspecies distinctions were unclear, data from different species were combined to form single species groups following the lead of the author. For the purpose of this study, as long as populations are closely related enough to make allozyme comparisons meaningful, it does not matter whether the populations are considered different subspecies or species.

As with the simulated datasets, F_{ST}-values were calculated using BIOSYS-1 (Swofford and Selander 1981). To eliminate any discrepancies due to typographical or data-entry errors, reanalyzed values were checked against those reported in the literature. When the F_{ST}-value was not published but Nei's genetic distance was, this measure was calculated in reanalysis to provide confirmation. GDA 1.0 (Lewis and Zaykin 1999) was used to calculate actual θ for each set of populations, and equation (2) was used to estimate θ, based on directly calculated values for F_{ST}, r, and M. Actual θ and estimated θ values for each empirical dataset were compared using linear regression.

RESULTS

Using simulated datasets, equation (2) produced estimations of θ that correlated nearly perfectly with actual θ (Fig. 1). When the estimations were not the same as the actual

TABLE 1. Actual and estimated θ calculated for each model set of populations. Actual values for θ (Weir and Cockerham 1984) and their associated 95% confidence intervals were determined using GDA (Lewis and Zaykin 1999). Estimated θ was determined using the following variables: Wright's F_{ST} (from BIOSYS-1, Swofford and Selander 1981), *r* (the number of populations sampled), and *M* (the average number of individuals sampled per population).

Model	Actual θ	Confidence interval[1]	Estimated θ	F_{ST}	*r*	*M*
1	−0.09091	−0.09091 to −0.09091	−0.09091	0.000	3	6
2	1.00000	1.00000 to 1.00000	1.00000	1.000	3	6
3	0.58298	0.41329 to 0.73269	0.58338	0.519	3	6
4	0.35341	0.18270 to 0.52788	0.35319	0.314	3	6

[1] Estimated using the bootstrap, 5000 repetitions. Because the bootstrapping is done over individual loci, population samples that are invariant within subpopulations have confidence intervals valued at zero.

values of θ, they were well within the 95% confidence intervals determined by bootstrapping (Table 1).

Using the empirical datasets, equation (2) performed almost as well in estimating θ (Fig. 2), with an adjusted r^2 of 0.96. Only one of 39 taxa had an estimated θ that fell outside of the 95% confidence intervals around actual θ; the estimated θ of *Amazona farinosa* was higher than its actual θ (Table 2).

One might assume that those datasets that produced the largest confidence intervals around θ when bootstrapped

TABLE 2. Weir and Cockerham's (1984) θ (actual and estimated) calculated for empirical data sets. See Methods for details.

Taxon	Actual θ	Confidence interval	Estimated θ	F_{ST}	*r*	*M*
Crypturellus berlepschi/*C. cinereus*[1]	0.87346	0.65812–1.00000	0.90341	0.879	2	1.50000
Leucopternis plumbea/*L. schistacea*[1]	1.00000	1.00000–1.00000	1.00000	1.000	2	1.00000
Micrastur plumbeus/*M. gilvicollis*[1]	0.89630	0.64103–1.00000	0.89655	0.856	2	2.00000
Pyrrhura melamura[2]	0.57143	0.00000–1.00000	0.59875	0.514	2	2.50000
Pionus menstruus[2]	0.20805	−0.02280–0.46253	0.40119	0.413	3	2.66667
Amazona farinosa[2]	0.04280	−0.09091–0.14286	0.24019	0.274	2	2.00000
Nyctiphrynus ocellatus/*N. rosenbergi*[1]	0.72157	0.52903–0.85135	0.72126	0.623	2	3.00000
Threnetes ruckeri/*T. leucurus*[2]	0.58932	0.33908–0.78690	0.57216	0.527	3	4.00000
Eutoxeres aquila[2]	0.37965	0.00848–0.60036	0.43506	0.440	3	2.66667
Trogon rufus[2]	0.28958	−0.08571–0.55066	0.36225	0.421	3	2.00000
Baryphthengus martii[1]	0.49099	0.02697–0.76088	0.48251	0.437	3	4.66667
Glyphorynchus spirurus[2]	0.31855	0.05759–0.61069	0.31862	0.296	3	5.00000
Glyphorynchus spirurus[3]	0.16231	0.08319–0.23825	0.20790	0.230	8	8.50000
Dendrocolaptes certhia[2]	0.36719	0.04801–0.57968	0.47674	0.463	3	3.00000
Automolus rubiginosus[2]	0.10300	−0.04234–0.23261	0.13592	0.149	2	3.50000
Sclerurus mexicanus[2]	0.85739	0.56331–1.00000	0.83368	0.816	3	2.33333
Xenops minutus[2]	0.64648	0.44286–0.81887	0.64665	0.615	3	3.00000
Myrmotherula axillaris[2]	0.53512	0.06245–0.86469	0.52153	0.476	3	4.33333
Microrhopias quixensis[2]	0.09160	−0.03704–0.16712	0.10184	0.198	3	2.66667
Myrmoborus myotherinus[4]	0.23305	0.05054–0.46120	0.27159	0.261	4	7.50000
Pithys albifrons[4]	0.03283	−0.01976–0.09538	0.03757	0.057	4	13.00000
Tityra semifasciata[2]	0.29905	0.07407–0.48123	0.29966	0.376	3	2.00000
Pipra coronata[4]	0.26419	0.03800–0.48397	0.26326	0.233	5	30.80000
Chiroxiphia pareola[4]	0.40955	0.27711–0.61941	0.40192	0.262	2	22.50000
Myiobius barbatus/*M. sulphureipygius*[2]	0.44267	0.21013–0.64049	0.55951	0.571	4	2.75000
Mionectes olivaceus[2]	0.39584	0.05843–0.65630	0.40575	0.371	3	4.66667
Henicorhina leucosticta[1]	0.56808	0.28136–0.76209	0.54006	0.563	4	2.50000
Microcerculus marginatus[2]	0.58548	0.14914–0.78779	0.56246	0.518	3	4.00000
Cyphorhinus arada/*C. phaeocephalus*[2]	0.44289	−0.07243—0.66093	0.54290	0.512	3	3.33333
Turdus albicollis/*T. assimilis*[2]	0.38304	0.09070–0.57927	0.40285	0.385	3	3.66667
Microbates cinereiventris[2]	0.57037	0.20833–0.81373	0.57083	0.513	2	2.00000
Atlapetes brunneinucha[5]	0.26645	0.11529–0.39383	0.31672	0.287	4	10.50000
Pitylus grossus[2]	0.00842	−0.03490–0.05872	0.03003	0.152	3	2.66667
Chlorospingus ophthalmicus[5]	0.25702	0.06311–0.47566	0.33748	0.297	4	14.50000
Chlorophanes spiza[2]	0.47259	0.12566–0.71169	0.54618	0.533	3	2.66667
Tersina viridis[2]	0.20000	0.00000–0.38462	0.20000	0.250	2	2.0000
Limnothlypis swainsonii[6]	0.02938	0.00167–0.07585	0.03095	0.043	5	21.80000
Aphelocoma coerulescens[7]	0.12456	0.04986–0.15973	0.12875	0.126	5	18.20000
Aphelocoma unicolor[7]	0.75771	0.29411–1.00000	0.74615	0.686	3	6.33333

[1] Brumfield (1993).
[2] Brumfield and Capparella (1996).
[3] Bates (1993).
[4] Capparella (1987).
[5] Peterson (1992).
[6] Winker et al. (2000).
[7] Peterson (1990).

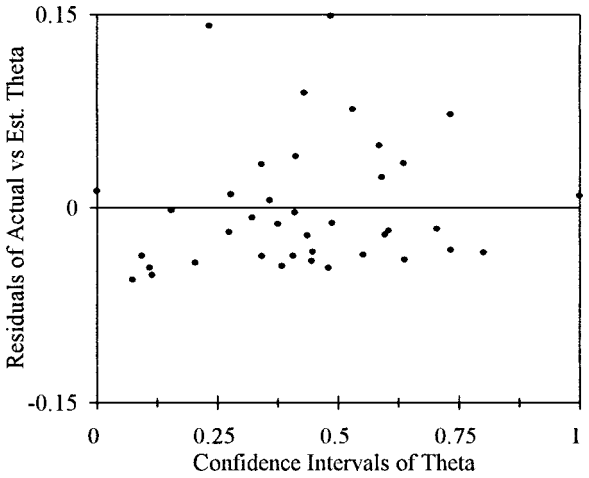

FIG. 3. The magnitude of residuals from Figure 2 do not correspond with the 95% bootstrap confidence intervals around actual θ.

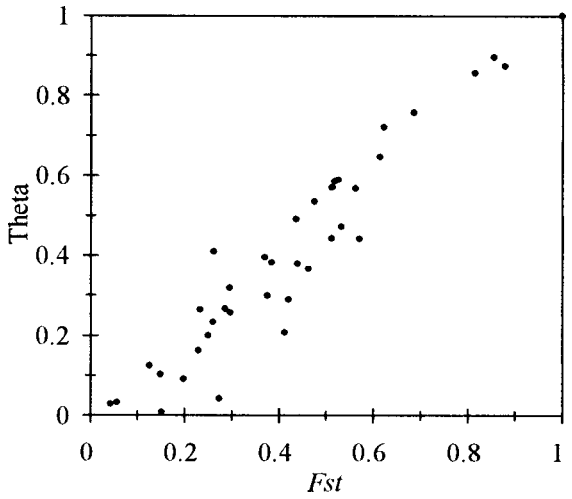

FIG. 4. Wright's (1943) F_{ST} or Nei's (1993) G_{ST} (as reported in each study and recalculated here) versus Weir and Cockerham's (1984) θ. Adjusted $r^2 = 0.91$, $P < 0.01$, $n = 39$.

would have the largest residuals in Figure 2; that is, they would have the least accurate estimations of θ. This is not the case, however; there is no correlation between the magnitude of the residuals (whether positive or negative) around estimated θ and the magnitude of the bootstrapped 95% confidence intervals associated with actual θ (Fig. 3). The bootstrapping is performed across loci and reflects inconsistencies that may arise from the effects of a particular locus. The problematic effects of specific loci are not revealed in the process of estimating θ because the differences in the effects of loci influence equation (2) only indirectly through the value of F_{ST}.

DISCUSSION

The estimation algorithm for θ (eq. 2) appears to be a sound method for approximating θ when full datasets are not available for the actual calculation of θ. In cases where only the number of populations, the number of individuals, and the value of Wright's F_{ST} are known, equation (2) is a feasible way of estimating θ. However, although the correlations between estimated and actual θ were highly significant for both simulated and real-world populations, equation (2) did not perform quite as well with the empirical data (Figs. 1, 2).

What explains the difference in how well equation (2) approximates θ for empirical datasets? One reason is that equation (1) is simplified by the use of *M*, an average of the number of individuals per population sample. Studies with equal-sized population samples are not compromised by the use of an average population size in the estimation of θ, whereas datasets that include population samples of different sizes may be affected by the use of a mean, whether arithmetic, harmonic, or otherwise.

Datasets with equal population samples had, on average, a smaller difference between actual θ and estimated θ than those with unequal population samples (comparing a mean difference of 0.00026 to 0.045; paired independent *t*-test not assuming equal variances, df = 31.007, two-tailed $P < 0.001$). The six taxa that showed the greatest differences between actual and estimated θ (*A. farinosa, Pionus menstruus, Myiobius* sp., *Dendrocolaptes certhia, Cyphorhinus* sp., and *Chlorospingus ophthalmicus*) all had unequal population sample sizes. Of the six taxa that showed the least differences between actual and estimated θ, five had equal population sample sizes (*Leucopternis* sp., *Tersina viridis, Glyphorynchus spirurus, Micrastur* sp., and *Nyctiphrynus* sp.). The one with unequal population sample sizes, *Xenops minutus*, is not far from equality, with two population samples of three individuals and one sample of two individuals.

Ultimately two questions must be raised, the first of which concerns the utility of equation (2). Although removing the sampling bias inherent in Wright's F_{ST} has been deemed important for making cross-study comparisons, the bias itself is trivial compared to the error involved in the approximation of θ. It is clear that the correlation between Wright's F_{ST} and Weir and Cockerham's θ (shown using the same 39 empirical datasets) is strong (Fig. 4). The difference between the two values (which would include the adjustment for bias) is not much greater than the difference between estimated and actual θ. The statistical bias of F_{ST} introduced by sampling error is not large enough to warrant the conversion of F_{ST} to an estimation of θ using equation (2). The second question is whether the problem of statistical bias in F_{ST} has been overstated. It would appear that the bias introduced by the use of F_{ST} may be negligible, especially relative to other potentially large errors involved in comparing natural populations across studies.

ACKNOWLEDGMENTS

We thank K. E. Schwaegerle, J. A. Cook, and an anonymous reviewer for discussion and suggestions on drafts of this manuscript.

LITERATURE CITED

Barrowclough, G. F. 1983. Biochemical studies of microevolutionary processes. Pp. 223–261 *in* A. H. Brush and G. A. Clark Jr., eds. Perspectives in ornithology. Cambridge Univ. Press, Cambridge, U.K.

Bates, J. M. 1993. The genetic effects of forest fragmentation on

Amazonian forest birds. Ph.D. diss., Louisiana State University and Agricultural and Mechanical College, Baton Rouge, LA.
Brumfield, R. T. 1993. Avian biodiversity in northwestern South America. M.S. thesis, Illinois State University, Normal, IL.
Brumfield, R. T., and A. P. Capparella. 1996. Historical diversification of birds in northwestern South America: a molecular perspective on the role of vicariant events. Evolution 50:1607–1624.
Capparella, A. P. 1987. Effects of riverine barriers on genetic differentiation of Amazonian forest undergrowth birds. Ph.D. diss., Louisiana State University and Agricultural and Mechanical College, Baton Rouge, LA.
———. 1988. Genetic variation in neotropical birds: implications for the speciation process. Acta XIX Congr. Int. Ornithol. 2: 1658–1664.
Cockerham, C. C., and B. S. Weir. 1993. Estimation of gene flow from *F*-statistics. Evolution 47:855–863.
Evans, P. G. H. 1987. Electrophoretic variability of gene products. Pp. 105–162 *in* F. Cooke and P. A. Buckley, eds. Avian genetics. Academic Press, London.
Leberg, P. L., and J. E. Neigel. 1999. Enhancing the retrievability of population genetic survey data? An assessment of animal mitochondrial DNA studies. Evolution 53:1961–1965.
Lewis, P. O., and D. Zaykin. 1999. Genetic data analysis: computer program for the analysis of allelic data. Ver. 1.0 (d12). Free program distributed by the authors via http://lewis.eeb.uconn.edu/lewishome/gda/
Nei, M. 1973. Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. 70:3321–3323.
Nevo, E. 1978. Genetic variation in natural populations: patterns and theory. Theor. Popul. Biol. 13:121–177.
Peterson, A. T. 1990. Evolutionary relationships of the *Aphelocoma* jays. Ph.D. diss., University of Chicago, Chicago, IL.
———. 1992. Phylogeny and rates of molecular evolution in the *Aphelocoma* jays (Corvidae). Auk 109:133–147.
Slatkin, M., and N. H. Barton. 1989. A comparison of three methods for estimating average levels of gene flow. Evolution 43: 1349–1368.
Swofford, D. L., and R. B. Selander. 1981. BIOSYS-1:a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. J. Hered. 72:281–283.
Weir, B. S., and C. C. Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution 38:1358–1370.
Winker, K., G. R. Graves, and M. J. Braun. 2000. Genetic differentiation among populations of a migratory songbird: *Limnothlypis swainsonii*. J. Avian Biol. 31:319–328.
Wright, S. 1943. Isolation by distance. Genetics 28:114–138.
———. 1951. The genetical structure of populations. Ann. Eugen. 15:323–354.
———. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution 19: 395–420.

Corresponding Editor: S. Karl