

A Generic Estimation of Population Subdivision Using Distances Between Alleles With Special Reference for Microsatellite Loci

Yannis Michalakis* and Laurent Excoffier†

*Laboratoire d'Ecologie, Université P. & M. Curie CNRS URA 258, 75252 Paris, France and †Genetics and Biometry Laboratory, Department of Anthropology and Ecology, University of Geneva, 1211 Geneva 24, Switzerland

Manuscript received June 13, 1995
Accepted for publication November 24, 1995

ABSTRACT

Several estimators of population differentiation have been proposed in the recent past to deal with various types of genetic markers (*i.e.*, allozymes, nucleotide sequences, restriction fragment length polymorphisms, or microsatellites). We discuss the relationships among these estimators and show how a single analysis of variance framework can accommodate these qualitatively different data types.

THE analysis of molecular variance (AMOVA) (EXCOFFIER *et al.* 1992) was initially introduced as an extension of the analysis of gene frequencies (COCKERHAM 1973; LONG 1986; WEIR and COCKERHAM 1984) for molecular haplotypes in an essentially haploid system. The typical input for AMOVA consisted of a matrix of pairwise Euclidean distances between all multisite haplotypes and files containing the frequency of those haplotypes within each population. We show here the equivalence of this approach to a weighted average of single-locus treatments as defined by WEIR and COCKERHAM (1984). Under a particular definition of the distance between haplotypes, we show that the AMOVA can be applied to microsatellite data to obtain an analogue of the R_{ST} statistic recently defined by SLATKIN (1995). The AMOVA treatment thus provides a general framework for the analysis of population genetic structure, as the assumptions on the evolution of a given polymorphism can be embedded within the definition of a Euclidean distance without affecting the essential structure of the AMOVA analysis.

AMOVA as a weighted average of single-locus treatments: The AMOVA approach was initially developed to estimate population genetic structure from molecular haplotype frequencies in haploid organisms, using an analysis of variance framework (Table 1). The same framework can be used for diploid organisms. For simplicity, we will consider the case of a genetic polymorphism at several loci in a diploid organism assuming that the gametic phase is known. We will later show that this simplifying assumption is unnecessary for the kind of treatment we consider. Under this approach, each molecular haplotype i is treated as a vector of single-locus allelic states \mathbf{a}_i , of dimension equal to the number of loci considered, say m , as $\mathbf{a}_i = [a_{i1}, a_{i2}, a_{i3} \dots a_{im}]'$.

Corresponding author: Yannis Michalakis, Laboratoire d'Ecologie, Université P. & M. Curie CNRS URA 258, 7 quai St. Bernard, 75252 Paris cedex 05, France.
E-mail: ioannis.michalakis@hall.snv.jussieu.fr

The F_{ST} analogue Φ_{ST} is conventionally obtained as the ratio of the estimated variance component due to differences among P populations ($\hat{\sigma}_a^2$) over the estimated total variance ($\hat{\sigma}^2 = \hat{\sigma}_a^2 + \hat{\sigma}_w^2$) as $\Phi_{ST} = \hat{\sigma}_a^2 / \hat{\sigma}^2$. Expanding Φ_{ST} , in terms of sums of squared deviations from the mean ($SSDs$), leads to

$$\hat{\Phi}_{ST} = \frac{(2N - P) SSD(T) - (2N - 1) SSD(WP)}{(2N - P) SSD(T) - (2N - 1 - b) SSD(WP)}, \quad (1)$$

where b is equal to $n'(P - 1)$, and n' is defined in Table 1. Note that the $SSDs$ are functions of the haplotypic vectors as

$$SSD(T) = \frac{1}{2N} \sum_{i=1}^{2N} \sum_{j=1}^i \delta_{ij}^2 \quad \text{and} \quad SSD(WP) = \sum_{r=1}^P \frac{1}{2N_r} \sum_{i=1}^{2N_r} \sum_{j=1}^i \delta_{ij}^2, \quad (2)$$

where N_r is the size of r th sample, and δ_{ij}^2 is a squared Euclidean distance between haplotypes i and j defined as

$$\delta_{ij}^2 = (\mathbf{a}_i - \mathbf{a}_j)' \mathbf{W} (\mathbf{a}_i - \mathbf{a}_j), \quad (3)$$

where \mathbf{W} is a square $m \times m$ weighting matrix that allows us to deal with possible interactions among loci and unequal locus weighting schemes (see EXCOFFIER *et al.* 1992). If loci are assumed independent and are given equal weight, \mathbf{W} is the identity matrix \mathbf{I} and

$$\delta_{ij}^2 = \sum_{k=1}^m (a_{ik} - a_{jk})^2. \quad (4)$$

Under this assumption of interlocus independence, the sums of squared deviations can be partitioned into m single-locus components as

$$\hat{\Phi}_{ST} = \frac{(2N - P) \sum_{i=1}^m SSD(T)_i - (2N - 1) \sum_{i=1}^m SSD(WP)_i}{(2N - P) \sum_{i=1}^m SSD(T)_i - (2N - 1 - b) \sum_{i=1}^m SSD(WP)_i}, \quad (5)$$

TABLE 1
Analysis of molecular variance framework

Source of variation	d.f.	SSD	MSD	E(MSD)
Among populations	$P - 1$	$SSD(AP)$	$SSD(AP)/(P - 1)$	$\sigma_w^2 + n' \sigma_a^2$
Among genes within populations	$2N - P$	$SSD(WP)$	$SSD(WP)/(2N - P)$	σ_w^2
Total	$2N - 1$	$SSD(T)$		

$$n' = \frac{1}{P - 1} \left(2N - \sum_{i=1}^P \frac{(2N_i)^2}{2N} \right)$$

where P is the number of sampled populations, N_i is the sample size for the i th population, and $N = \sum_i N_i$, the total number of sampled individuals.

$$= \sum_{k=1}^m \hat{\sigma}_{ai}^2 / \sum_{k=1}^m \hat{\sigma}_i^2, \quad (5b)$$

and $\hat{\Phi}_{ST}$ is therefore equal to the weighted average of single-locus ratio estimator ($\hat{\theta}_w$) defined by WEIR and COCKERHAM (1984), shown to be essentially unbiased. This AMOVA treatment, although initially considering a chromosome to be the segregating unit within a population, is thus formally equivalent to a treatment where each locus would be considered independently in turn, provided that $\mathbf{W} = \mathbf{I}$.

The AMOVA treatment has already been applied to multilocus nuclear data in diploids by PEACALL *et al.* (1995). These authors have also described the use of AMOVA to estimate intra-individual variance components and measures of inbreeding such as F_{IS} and F_{IT} . The AMOVA can also be compared to other estimators of population differentiation based on molecular diversity. HUDSON *et al.* (1992) defined an estimator ($\langle F_{ST} \rangle$) of population genetic structure from DNA sequence data as

$$\langle F_{ST} \rangle = 1 - \frac{H_w}{H_b}, \quad (\text{HUDSON } et al. 1992) \quad (6)$$

where H_w is the mean number of nucleotide differences between DNA sequences sampled from the same population, and H_b is the mean number of differences between sequences sampled from different populations (HUDSON *et al.* 1992). $\langle F_{ST} \rangle$ is therefore similar to $\hat{\theta}_w$ calculated by treating each nucleotide site as a single locus, and then averaging over sites. LYNCH and CREASE (1990) defined another estimator of population differentiation inferred from DNA sequences, N_{ST} , which differs from $\langle F_{ST} \rangle$ only in not including the Jukes-Cantor correction for multiple substitutions per site (HUDSON *et al.* 1992).

It is worth mentioning that all the estimators discussed so far have been designed to estimate the same population parameter F_{ST} . Another class of estimators (see *e.g.*, TAKAHATA and PALUMBI 1985) has been designed to estimate another population parameter, G_{ST} , defined by NEI (1973). They mainly differ from the estimators of F_{ST} by considering a weighted average of differences among alleles drawn at random from the

whole collection of populations, instead of a weighted average of differences among alleles drawn from different populations. The differences between the two parameters and their estimators are reviewed in CHAKRABORTY and DENKER-HOPFE (1991) and in COCKERHAM and WEIR (1993).

Relationship between Slatkin's R_{ST} and Φ_{ST} : SLATKIN (1995) has recently defined R_{ST} , a G_{ST} analogue for microsatellite data at a single locus, taking into account the difference between microsatellite allelic sizes, very much like others have incorporated the differences in electrophoretic mobility among allozymes (*e.g.*, RICHARDSON and SMOUSE 1976; RICHARDSON *et al.* 1977) or the differences between allelic phenotypic effects for quantitative characters (*e.g.*, CHAKRABORTY and NEI 1982) to evaluate population differences. R_{ST} is defined as

$$R_{ST} = \frac{\bar{S} - S_w}{\bar{S}}, \quad (\text{SLATKIN, 1995}) \quad (7)$$

where S_w and \bar{S} are the average squared difference in allele size between pairs of genes within populations and between pairs of genes taken from a collection of P populations, respectively (SLATKIN 1995). The size of the i th allele (a_i) is here simply equal to the number of repeats it carries. If one defines the pairwise allelic distance equivalent to (4) for a single locus as the square of the difference in the number of repeats between two alleles, one can show that under SLATKIN's notations and assumptions,

$$\hat{\Phi}_{ST} = \frac{S_B - S_w}{S_B}, \quad (8)$$

where S_B is the average squared difference in allele size between pairs of genes from different populations. It follows that the relationship between SLATKIN's \hat{R}_{ST} and at a single locus is

$$\hat{\Phi}_{ST} = \frac{(1 - c)\hat{R}_{ST}}{1 - c\hat{R}_{ST}}, \quad (9)$$

where $c = (2N - P)/(2NP - P)$, and the difference between \hat{R}_{ST} and $\hat{\Phi}_{ST}$ is analogous to that between \hat{G}_{ST} and \hat{F}_{ST} . The equilibrium value of the parameter

(ρ_{ST}) estimated by Φ_{ST} for microsatellite data has been determined by ROUSSET (personal communication), who also first derived the relationship between \hat{R}_{ST} and $\hat{\Phi}_{ST}$.

For the treatment of multilocus data, SLATKIN (1995) suggested to use a weighted average across loci similar to that defined in (5b). It implies that microsatellite multilocus data can be analyzed with a single AMOVA if one sums the squared differences in allele size over loci, as in (4).

Computing Φ_{ST} in diploid populations: Typically, if one has molecular data on physically linked loci, such as DNA sequence data, restriction fragment length polymorphisms, or a combination of microsatellite data, diploid individuals may be heterozygous at more than one locus and the gametic phase may be ambiguous. However, as no linkage information is necessary to compute Φ_{ST} , diploid multilocus heterozygous genotypes need not be resolved to estimate the amount of population genetic structure for codominant markers. Instead of first trying to resolve the haplotypes of each individual and to estimate their sample frequencies, which may be a complex procedure (EXCOFFIER and SLATKIN 1995; LONG *et al.* 1995), one could proceed as follows: (1) define the gametic phase at random and thus two dummy haplotypes for each individual of each population, (2) calculate the dummy haplotype frequencies by simply counting them and (3) define a matrix of the squared Euclidean distances among all dummy haplotypes. The definition of this matrix depends on whether one wants to calculate statistics equal to θ [the single locus F_{ST} analogue defined by COCKERHAM (1973), where distances between alleles are equal to one if alleles are different, or zero if they are identical], $\bar{\theta}_w$ (the weighted average defined by WEIR and COCKERHAM (1984), where distances are the sum of allelic differences over all loci), or R_{ST} -like (defined for microsatellite data, where the distances are the sum of squared differences in allele size over all loci). One would next carry out an AMOVA analysis as if the dummy haplotypes and their frequencies were the real ones. The locus-by-locus structure of Equation 5 guarantees that the sums of square deviations both within or among populations are insensitive to the choice of the haplotypic phase in the diploid individuals.

The use of dummy haplotypes will yield correct estimates of population differentiation, even if the loci are statistically linked, as long as all loci are given equal weight (WEIR and COCKERHAM 1984). Special care is however required for testing the significance of these estimates. AMOVA currently tests the significance of population differentiation estimates by randomly permuting whole haplotypes, thus constructing an empirical null distribution of the estimator under the hypothesis of complete linkage among loci. If the individual loci composing the haplotypes are totally linked and the gametic phase is known, such as for animal mito-

chondrial markers or DNA nucleotide sequences, then permuting the haplotypes is the correct testing procedure. If the loci composing the haplotypes are statistically independent, then permuting dummy haplotypes across populations is a conservative procedure, in the sense that significance levels will be overestimated because the empirical null distribution will be more platykurtic than the true null distribution. If, however, some of the loci are statistically linked and the gametic phase is not known, then permuting the dummy haplotypes will lead to erroneous significance levels. In that case, an empirical testing procedure appears difficult to build without an exact knowledge of the pattern of disequilibrium among loci. It would thus appear necessary to estimate the disequilibrium patterns and incorporate them in the analysis of the population genetic structure, a procedure opened to investigation.

A personal computer-based computer program (WINAMOVA) is available to perform the AMOVA treatment up to two hierarchical levels (individuals in populations and populations into groups) allowing for unequal sample sizes, and to test the significance of variance components and Φ -statistics using the permutation approach described above. It can be retrieved by anonymous ftp on acasun1.unige.ch, directory pub/comp/win/amova or by connecting to <http://acasun1.unige.ch/LGB/Software/WinDoze/amova>.

We thank MONTY SLATKIN and PETER SMOUSE for their useful comments on the manuscript, and FRANÇOIS ROUSSET for sharing unpublished material and pointing out the difference between ρ_{ST} and R_{ST} . L.E. was supported by a Swiss National Foundation grant No. 32-37821.93.

LITERATURE CITED

- CHAKRABORTY, R., and H. DANKER-HOPPE, 1991 Analysis of population structure: a comparative study of different estimators of Wright's fixation indices, pp. 203–254 in *Handbook of Statistics*, Vol. 8, edited by C. R. RAO and R. CHAKRABORTY. Elsevier Science Publishers, Amsterdam.
- CHAKRABORTY, R., and M. NEI, 1982 Genetic differentiation of quantitative characters between populations or species I. Mutation and random genetic drift. *Genet. Res.* **39**: 303–314.
- COCKERHAM, C. C., 1973 Analysis of gene frequencies. *Genetics* **74**: 679–700.
- COCKERHAM, C. C., and B. WEIR, 1993 Estimation of gene flow from F-statistics. *Evolution* **47**: 855–863.
- EXCOFFIER, L., P. SMOUSE and J. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**: 479–491.
- EXCOFFIER, L., and M. SLATKIN, 1995 Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol. Biol. Evol.* **12**: 921–927.
- HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. *Genetics* **132**: 583–589.
- LONG, J. C. 1986 The allelic correlation structure of Gainj and Kalam speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics* **112**: 629–647.
- LONG, J. C., R. C. WILLIAMS and M. URBANEK, 1995 An E-M algorithm and testing strategy for multiple-locus haplotypes. *Am. J. Hum. Genet.* **56**: 799–810.
- LYNCH, M., and T. J. CREASE, 1990 The analysis of population survey data on DNA sequence variation. *Mol. Biol. Evol.* **7**: 377–394.

- NEI, M., 1973 Analysis of gene diversity in subdivided populations. Proc. Natl. Acad. Sci. USA **70**: 3321–3323.
- PEAKALL, R., P. E. SMOUSE and D. R. HUFF, 1995 Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss *Buchloë dactyloides*. Mol. Ecol. **4**: 135–147.
- RICHARDSON, R. H., and P. E. SMOUSE, 1976 Patterns of molecular variation. I. Interspecific comparisons of electromorphs in the *Drosophila mulleri* complex. Biochem. Genet. **14**: 447–466.
- RICHARDSON, R. H., P. E. SMOUSE and M. E. RICHARDSON, 1977 Patterns of molecular variation. II. Associations of electromorphic mobility and larval substrate within species of the *Drosophila mulleri* complex. Genetics **85**: 141–154.
- SLATKIN, M. 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**: 457–462.
- TAKAHATA, N., and S. R. PALUMBI, 1985 Extranuclear differentiation and gene flow in the finite island model. Genetics **109**: 441–457.
- WEIR, B. S., and C. C. COCKERHAM, 1984 Estimating F-statistics for the analysis of population structure. Evolution **38**: 1358–1370.

Communicating editor: W. F. EWENS