

Inferring Admixture Proportions from Molecular Data: Extension to Any Number of Parental Populations

Isabelle Dupanloup and Giorgio Bertorelle

Dipartimento di Biologia, Università di Ferrara, Ferrara, Italy

The relative contribution of two parental populations to a hybrid group (the admixture proportions) can be estimated using not only the frequencies of different alleles, but also the degree of molecular divergence between them. In this paper, we extend this possibility to the case of any number of parental populations. The newly derived multiparental estimator is tested by Monte Carlo simulations and by generating artificial hybrid groups by pooling mtDNA samples from human populations. The general properties (including the variance) of the two-parental estimator seem to be retained by the multiparental estimator. When mixed human populations are considered and hypervariable single-locus data are analyzed (mtDNA control region), errors in the estimated contributions appear reasonably low only when highly differentiated parental populations are involved. Finally, the method applied to the hybrid Canary Island population points to a much lower female contribution from Spain than has previously been estimated.

Introduction

The coming together of populations that have long been isolated by geographical, ecological, or cultural barriers may give rise to a hybrid population (HP). Compared with the parental populations (PPs) they descend from, HPs have specific genetic features. In particular, HPs tend to show allele frequencies which are linear combinations of the PPs allele frequencies. This simple consideration has been used to develop several methods for estimating the relative contribution of each PP to the HP (Cavalli-Sforza and Bodmer 1971; Chakraborty 1986; Long 1991; Parra et al. 1998; Estoup et al. 1999).

In populations that have a long history of separation (and especially if mutation rates are high), different mutations may be present, and there is therefore useful information not only in allele frequency differences between populations, but also in the amount of molecular differentiation between alleles. Under these conditions, taking into account also the molecular differences between alleles when admixture proportions are estimated seems desirable.

The first attempt along these lines (Pinto et al. 1996) was followed by the derivation of an estimator (called mY) that can be applied to any type of molecular data, as long as their amount of molecular diversity can be simply related to coalescence times (Bertorelle and Excoffier 1998).

The application of mY (e.g., Hammer et al. 2000) has been limited to cases of admixture in which only two PPs contribute alleles to the gene pool of the HP. Here, we derive a system of linear equations that allows a simple extension of this model to the case in which the HP received a genetic contribution from any arbitrary number d of PPs. The behavior of our d -parental estimators of admixture proportions, and, in particular, the effect of increasing the number of estimated parameters on their errors, is evaluated by simulation. Finally,

Key words: admixture coefficients, least-squares method, coalescent, Monte Carlo simulations, human populations, mtDNA sequences.

Address for correspondence and reprints: Isabelle Dupanloup de Ceuninck, Dipartimento di Biologia, Università di Ferrara, via L. Borsari 46, 44100 Ferrara, Italy. E-mail: dpi@dns.unife.it.

Mol. Biol. Evol. 18(4):672–675. 2001

© 2001 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

we apply this method to the study of a human HP with three putative PPs.

Derivation of the Multiparental Estimators

In the admixture model considered by Bertorelle and Excoffier (1998), an ancestral population splits into two PPs, which evolve independently for τ generations. At that point, an HP is instantaneously created by combining two fractions, μ and $(1 - \mu)$, of genes taken at random from each PP. From that moment on, i.e., for t_A generations, the three populations evolve independently. Under this model, a least-squares estimator of μ , mY , was derived and applied to DNA sequences and microsatellites. Here, we use the same approach to analyze the relative contributions to the HP when d PPs are involved. We assume that all PPs contribute at once to the HP.

The mean coalescence time between a gene drawn from the HP and a gene drawn from the i th PP, $\bar{t}_{h,i}$, is given by

$$\bar{t}_{h,i} = \mu_i(\bar{t}_{i,i} + t_A) + \sum_{j \neq i}^d \mu_j \bar{t}_{i,j}, \quad (1)$$

where $\bar{t}_{i,i}$ is the mean coalescence time between two genes sampled in the same PP i , $\bar{t}_{i,j}$ is the mean coalescence time between two genes sampled in two different PPs, i and j , and μ_i (or μ_j) is the relative contribution of the i th (j th) PP to HP.

Noting that for d PPs there are d mean coalescence times $\bar{t}_{h,i}$ and $(d - 1)$ contributions to estimate ($\sum \mu_i = 1$), a least-squares estimator for μ_i can be computed minimizing the sum of the squares of the differences between the left- and the right-hand sides of equation (1):

$$\begin{aligned} & \left[\bar{t}_{h,1} - \mu_1(\bar{t}_1) - \sum_{j \neq 1}^{d-1} \mu_j \bar{t}_{1,j} - \left(1 - \sum_{j=1}^{d-1} \mu_j \right) \bar{t}_{1,d} \right]^2 \\ & + \left[\bar{t}_{h,2} - \mu_2(\bar{t}_2) - \sum_{j \neq 2}^{d-1} \mu_j \bar{t}_{2,j} - \left(1 - \sum_{j=1}^{d-1} \mu_j \right) \bar{t}_{2,d} \right]^2 \\ & + \dots + \left[\bar{t}_{h,d} - \left(1 - \sum_{j=1}^{d-1} \mu_j \right) (\bar{t}_d) - \sum_{j \neq d}^{d-1} \mu_j \bar{t}_{d,j} \right]^2, \quad (2) \end{aligned}$$

where \bar{t}_i is now the sum of $\bar{t}_{i,i}$ and t_A ; the right-hand side

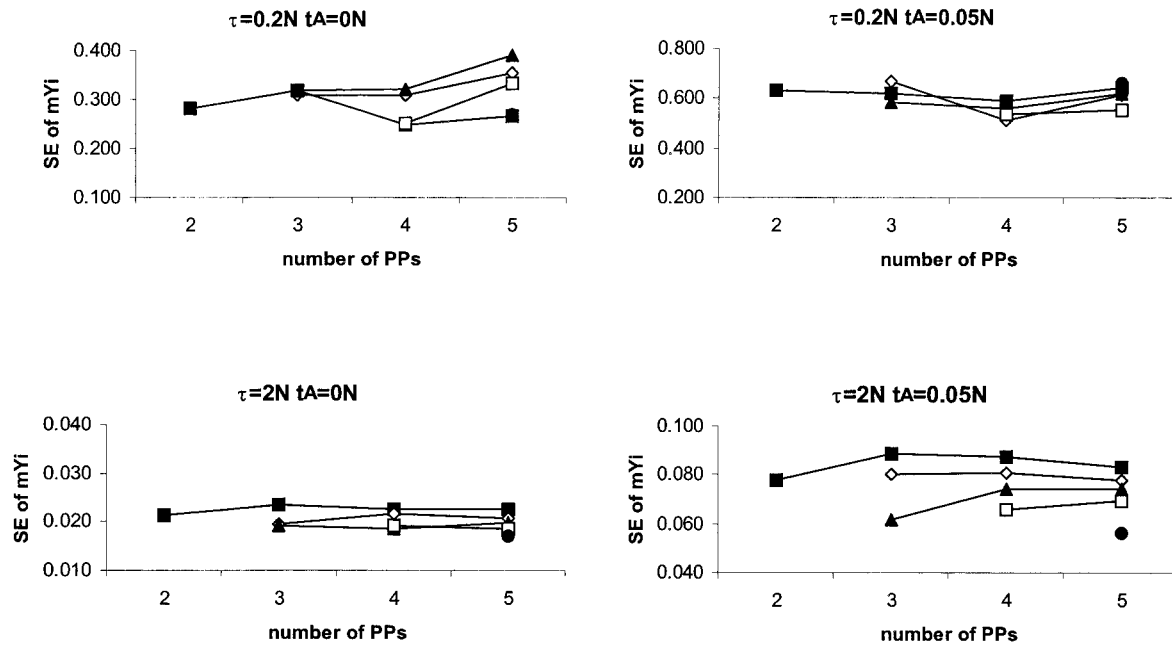


FIG. 1.—Results of the Monte Carlo simulations: standard errors of the parental estimators mY_i for different numbers of parental populations (2–5) (solid square = mY_1 ; open diamond = mY_2 ; solid triangle = mY_3 ; open square = mY_4 ; solid circle = mY_5). The true values of the relative contributions are chosen in a decreasing fashion from PP₁ to PP₅ (for $d = 2$, $\mu_1 = 0.8$ and $\mu_2 = 0.2$; for $d = 3$, $\mu_1 = 0.6$, $\mu_2 = 0.3$, and $\mu_3 = 0.1$; for $d = 4$, $\mu_1 = 0.4$, $\mu_2 = 0.3$, $\mu_3 = 0.2$, and $\mu_4 = 0.1$; for $d = 5$, $\mu_1 = 0.35$, $\mu_2 = 0.25$, $\mu_3 = 0.2$, $\mu_4 = 0.15$, and $\mu_5 = 0.05$).

of equation (1) is slightly modified because $\mu_d = 1 - \sum_{j=1}^{d-1} \mu_j$.

The estimators are thus computed solving the ($d - 1$) linear equations obtained deriving equation (2) for the ($d - 1$) unknowns. This system turns out to be simply described by the general k th equation:

$$\alpha_k \mu_k + \sum_{j \neq k}^{d-1} \beta_j \mu_j = \gamma_k$$

with

$$\alpha_k = \sum_i^d (\bar{t}_{i,k} - \bar{t}_{i,d})^2$$

$$\beta_j = \sum_i^d (\bar{t}_{i,d} - \bar{t}_{i,j})(\bar{t}_{i,d} - \bar{t}_{i,k})$$

$$\gamma_k = \sum_i^d (\bar{t}_{i,d} - \bar{t}_{h,i})(\bar{t}_{i,d} - \bar{t}_{i,k}).$$

Testing the Estimators

The statistical behavior of the multiparental estimators of μ_i , mY_i , was analyzed in three different ways. First, we simulated the genealogy of genes sampled in the PPs and the HP assuming different values of μ_i , τ , t_A , and d . The “true” (i.e., imposed in the simulation) values of the relative contribution of each PP were then compared with their estimates obtained in each simulated sample by our proposed method. Second, we used real samples of human sequences at the mtDNA control region to create artificially admixed populations, and again we evaluated the performance of the method to

estimate the known proportions of parental genes in the artificial HP. Finally, we estimated the female contribution of North African Berbers, Sub-Saharan Guineans, and Spanish mainlanders into the Canarian human population, and we compared our results with those obtained with other methods, considering also the historical evidence of the admixture process in this islands.

Monte Carlo Simulation of HP and PPs

The genealogies of samples of 60 DNA sequences from the HP and each PP were reconstructed following a coalescent approach (Hudson 1990). Mutations were then introduced assuming an infinite-sites model with $\theta = 2Nu = 10$ (N = haploid effective population size; u = mutation rate per locus per generation). For various combinations of the parameters μ , τ , t_A , and d , 1,000 genealogies were generated, thus allowing an empirical evaluation of the bias and the standard error of the estimators. Mean coalescence times were estimated from the average number of nucleotide differences.

In general, the d -parental estimators seemed to retain the properties of the two-parental estimator of admixture proportions (Bertorelle and Excoffier 1998). The estimator bias, unless very short divergence times among parental populations were assumed, was almost negligible (results not reported). On the other hand, the standard error became reasonably low only when PPs had diverged for a number of generations τ in the range of the population size or higher (see fig. 1). As observed for the two-parental estimator, the results reported in figure 1 also suggest that the age of the admixture event (t_A) affects the precision of the estimates. For example, for PPs with an effective population size of 500, reliable

estimates of admixture proportions are expected if the PPs diverged at least 500 generations ago. Even in this case, however, if the admixture event occurred 25 generations ago, the errors in the estimates can increase by a factor 3 or 4. As noted earlier (Bertorelle and Excoffier 1998), these results suggest some conditions ($\tau > 1N$, $t_A < 0.05N$) for the applicability of the estimator mY_i to single-locus data. If these conditions are not fulfilled, several loci with similar mutation rates should be simultaneously analyzed.

Interestingly, we did not observe any systematic increase in the errors of the estimates when the number of parental populations contributing to the HP gene pool increased from 2 to 5. On the other hand, at least when the PPs were more genetically differentiated ($\tau = 2N$), the standard errors tended to be slightly higher (but the coefficient of variation tended to be lower) for the PPs with larger contributions to the HP.

Artificial Hybrid Populations

We also simulated artificial HPs by pooling individuals extracted from real samples of human populations. On the basis of a multidimensional analysis of 61 samples of human populations typed for hypervariable region I of mtDNA (Excoffier and Schneider 1999), we chose one group of genetically rather homogeneous samples in Europe (Bavarians, Cornish, English, Germans, Welsh) and one group of genetically differentiated samples (!Kung, Australians, Japanese, Nootka, Saami). Using the sample allele frequencies as probabilities, and separately for each group of samples, we generated artificial PP samples of 100 sequences and a sample of 100 sequences (the artificial HP) extracted with fixed relative proportion from the PPs. This procedure was repeated 1,000 times, and the mean and standard errors of the estimators were then analyzed. Compared with the previous Monte Carlo simulations, the coalescent structure did not change among replicates; the standard errors we compute in this section therefore do not include the stochastic factors associated to the gene genealogy.

Figure 2 shows the results obtained for the two groups of populations when the relative contributions were fixed at 0.6 for one PP and 0.1 for four others. The bias was virtually absent in both cases, but the error of our estimators seemed reasonably low only when very different populations were used as artificial PPs. In other words, single-locus analysis of admixture processes does not seem feasible for human populations if they are only slightly differentiated, as is the case for most European groups. Again, we expect that only the simultaneous analysis of several loci could provide more reliable estimates, as is also the case, for example, for trees summarizing the evolutionary relationships of populations (Mountain and Cavalli-Sforza 1997). Finally, the results of this analysis seem to indicate a positive relationship between the error of the estimated contribution of a PP and its level of genetic variability. A set of simulations (whose results are not reported in detail) in which parental populations had different effective siz-

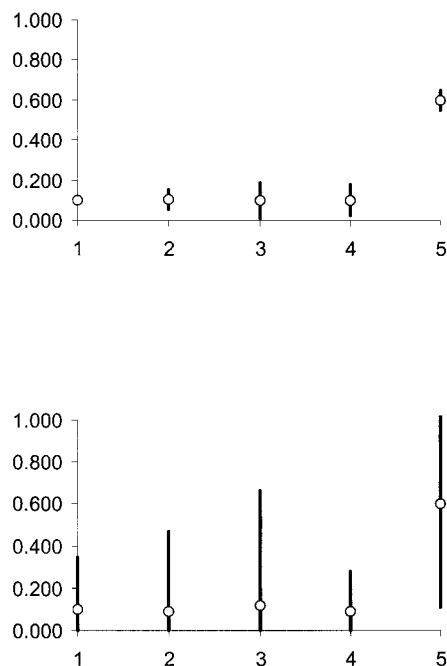


FIG. 2.—Results of the simulations for which an artificial HP was created by pooling real samples of human mtDNA HV1 sequences. Average $mY_i \pm 1$ SE within the range [0; 1] computed from 1,000 simulated admixture events are reported. *Above*, The parental populations are !Kung, Australians, Japanese, Nootka, and Saami. *Below*, The PPs are Bavarians, Cornish, English, Germans, and Welsh. In both cases, the fixed contribution of the last PP in the lists above was set to 0.6, whereas the other PPs all contributed with a fixed proportion of 0.1.

es supported the conclusion that the genetic variability of a PP and the error of its estimated contribution to the HP are positively correlated.

Application to a Real Case of Admixture

Pinto et al. (1996) estimated that the Spanish contribution to the Canarian mtDNA pool was 36%, about half the contribution previously estimated from autosomal loci (Roberts et al. 1966; Pinto et al. 1994). This difference was explained by an asymmetry of the female and male contributions, which was also supported by historical evidence. Berber and Guinean contributions to Canarian mtDNA were estimated to be 43% and 21%, respectively. When the estimator presented here was applied to the data of Pinto et al. (1996), we obtained the following bootstrap average coefficients (\pm SE): Spaniards, 0.17 (\pm 0.35); Berbers, 0.69 (\pm 0.34); and Guineans, 0.14 (\pm 0.08). Bootstrap coefficients and SEs were obtained using the procedure described in Bertorelle and Excoffier (1998).

Our results suggest, therefore, that a large majority of Canarian mtDNAs have a North African Berber origin and that the Spanish contribution was limited. These results seem more consistent than previous ones with the known history of the Spanish occupation and the presumed relationship between the pre-occupation people and the North African Berbers. We nevertheless note the large standard error associated with the estimates.

Conclusions

Using a simple model with two parental populations and one admixed population, Bertorelle and Excoffier (1998) showed that the estimation of admixture proportions using a specific method for the analysis of molecular data was feasible and efficient when the gene pool of the PPs was sufficiently differentiated. Here, we have extended their method to any possible number of PPs.

A Monte Carlo simulation study shows that the multiparental estimators behave in a way very similar to that of the two-parental estimator. The number of parameters to estimate increases with the number of PPs, but so does the information contained in the data. This is probably the reason for the constancy of the standard errors with the numbers of PPs considered.

Simulating artificial HP using human mtDNA sequences, we showed that in our species, the level of divergence between populations from different continents is probably large enough to allow reliable estimates of admixture proportions based on a single locus. This is certainly not true for closely related populations, such as those in the Canarian example, where the analysis of several loci seems necessary.

Finally, it is important to remember that the estimators of admixture proportion proposed here were derived assuming a specific population model. Indeed, suppose two parental populations contributed to the hybrid population the same amount of genes, but at different times. Larger numbers of mutations are expected to accumulate between the hybrid population and the parental population which contributed its genes earlier; this may lead to a decreased similarity between them, producing an underestimation of the more ancient contribution. Small deviations from the model probably have limited effects on the admixture proportion estimates, but this point needs to be further clarified.

Acknowledgments

We thank Guido Barbujani and Laurent Excoffier for stimulating talks. I.D. was supported by a Swiss NSF grant for prospective researchers. The computer program Admix 2.0, to compute the multiparental estimators of admixture proportions and their bootstrap standard errors, is available from I.D. on request.

LITERATURE CITED

- BERTORELLE, G., and L. EXCOFFIER. 1998. Inferring admixture proportions from molecular data. *Mol. Biol. Evol.* **15**:1298–1311.
- CAVALLI-SFORZA, L. L., and W. F. BODMER. 1971. The genetics of human populations. W. H. Freeman and Company, San Francisco.
- CHAKRABORTY, R. 1986. Gene admixture in human populations: models and predictions. *Yearb. Phys. Anthropol.* **29**: 1–43.
- ESTOUP, A., J. M. CORNUET, F. ROUSSET, and R. GUYOMARD. 1999. Juxtaposed microsatellite systems as diagnostic markers for admixture: theoretical aspects. *Mol. Biol. Evol.* **16**: 898–908.
- EXCOFFIER, L., and S. SCHNEIDER. 1999. Why hunter-gatherers do not show signs of Pleistocene demographic expansions. *Proc. Natl. Acad. Sci. USA* **96**:10597–10602.
- HAMMER, M. F., A. J. REDD, E. T. WOOD et al. (12 co-authors). 2000. Jewish and middle eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl. Acad. Sci. USA* **97**:6769–6774.
- HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44. *in* D. FUTUYMA and J. ANTONOVICS, eds. *Oxford surveys in evolutionary biology*. Oxford University Press, Oxford, England.
- LONG, J. C. 1991. The genetic structure of admixed populations. *Genetics* **127**:417–428.
- MOUNTAIN, J. L., and L. L. CAVALLI-SFORZA. 1997. Multilocus genotypes, a tree of individuals, and human evolutionary history. *Am. J. Hum. Genet.* **61**:705–718.
- PARRA, E. J., A. MARCINI, J. AKEY et al. (11 co-authors). 1998. Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**:1839–1851.
- PINTO, F., V. M. CABRERA, A. M. GONZALEZ, J. M. LARRUGA, A. NOYA, and M. HERNANDEZ. 1994. Human enzyme polymorphism in the Canary Islands. VI. Northwest African influence. *Hum. Hered.* **44**:156–161.
- PINTO, F., A. M. GONZALEZ, M. HERNANDEZ, J. M. LARRUGA, and V. M. CABRERA. 1996. Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences. *Ann. Hum. Genet.* **60**:321–330.
- ROBERTS, D. F., M. EVANS, E. W. IKIN, and A. E. MOURANT. 1966. Blood groups and the affinities of the Canary Islanders. *Man* **1**:512.

JEFFREY LONG, reviewing editor

Accepted December 11, 2000