Letters to the Editor

Am. J. Hum. Genet. 69:904, 2001

Patterns of Y-Chromosome Variation in South Amerindians

To the Editor:

Tarazona-Santos et al. (2001) compute estimates of within- and among-group genetic variability for South Amerindian Y-chromosome samples that are thought to represent tribal populations living in various major geoecological regions of South America: the Andean highlands, the Brazilian plateau, the Chaco region, the Argentinian pampa, and the Chilean rain forest.

The samples are agglomerated into two groups, one representing populations from the Andean highlands and the other representing populations from Amazonia, the Brazilian plateau, and the Chaco. Variability estimates are computed for both subdivisions and are consequently compared, with the Andean group exhibiting higher values. For apparently unjustified reasons, a sample from the tropical forest of Ecuador that has an Amazonian origin and exhibits the highest within-group variability is excluded from the analysis, unfortunately casting doubt on the reliability of the results.

Various among-group variability estimates and their association with distances among map locations of places where samples were presumably collected are computed next. We are aware of the difficulties in obtaining Amerindian samples, but the extremely small size of some samples used in this study (the central Brazilian plateau is represented by five individuals) precludes the possibility that among-group variability statistics are unbiased estimators of *population* relationships. The lack of association between genetic and geographic distances may be a reflection of this shortcoming.

On the basis of their results, Tarazona-Santos et al. (2001) conclude that two Y-chromosome microevolutionary models that involve differential patterns of genetic drift and gene flow characterize South Amerindians. Andean populations exhibit low rates of genetic drift and high rates of gene flow, whereas populations from Amazonia, the Brazilian plateau, and the Chaco exhibit high rates of drift and low rates of gene flow. It seems to us that this is a rash generalization, if it is based on the variability estimates presented in this study. Furthermore, it presupposes that non-Andean South Amerindian tribes living far apart, in markedly different geoecological areas, can be agglomerated and treated as one homogeneous group sharing the same population structure. We are not convinced that this is a realistic assumption.

FRANCISCO ROTHHAMMER AND MAURICIO MORAGA
Programa de Genética Humana
Instituto de Ciencias Biomédicas
Facultad de Medicina
Universidad de Chile
Santiago, Chile

Reference

Tarazona-Santos E, Carvalho-Silva DR, Pettener D, Luiselli D, De Stefano GF, Martinez Labarga C, Richards O, Tyler-Smith C, Pena SDJ, Santos F (2001) Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am J Hum Genet 68:1485–1496

Address for correspondence and reprints: Dr. Francisco Rothhammer, Programa de Genética Humana, Instituto de Ciencias Biomédicas, Faculdad de Medicina, Universidad de Chile, Av. Independencia 1027, Casilla 700061, Santiago 7, Chile. E-mail: frothham@machi.med.uchile.cl

0002-9297/2001/6904-0026$02.00

Am. J. Hum. Genet. 69:904–906, 2001

Reply to Rothhammer and Moraga

To the Editor:

Rothhammer and Moraga raise objections to the conclusions in our article describing global patterns of Y-chromosome diversity among South Amerindian populations (Tarazona-Santos et al. 2001). We do not think that their criticisms are valid, for the following reasons.

First, Rothhammer and Moraga argue that our conclusions are not well grounded, since they were only based on the Y-chromosome data presented in our article. This is not correct. In the article, we present and discuss the good concordance between our Y-chromosome data and the analyses of classical marker variability

previously performed by our group (Luiselli et al. 2000; Simoni et al. 2000*b* and references therein).

Second, they criticize the size of our samples. We certainly agree that large samples are better than small ones. For small samples, large standard errors are expected, and such errors can conceal geographical patterns where they exist but cannot generate statistically significant patterns where none exist. That we observed significant differences in within-population variability means that our sample sizes were not too small—or, at least, were large enough to support our conclusions. This is further confirmed by the fact that a significant correlogram was identified using the Spatial Autocorrelation Analysis (AIDA), which means that association between genetic and geographic distances exists. Rothhammer and Moraga apparently have missed this subtle point.

Third, they criticize the aggregation of the differentiated Eastern populations to compare within-population variability among eastern and Andean populations. This, of course, has to be done carefully and, indeed, we mention in our paper (the last 23 lines of p. 1488) that this agglomeration might produce an artificial Whalund effect (i.e., it might inflate the gene diversity). However, this would create a bias acting against our conclusions and therefore has the effect of rendering our results more robust. Again, Rothhammer and Moraga have missed the point.

Furthermore, we have now made the following calculations from our published data. (1) The 95% confidence interval (CI) of average gene diversity in the eastern populations, when the Cayapa sample is included, is 0.398–0.459, which does not overlap with the 95% CI of average gene diversity in Andean populations (0.463–0.524). (2) When *Rst* values for the eastern part of the continent are recalculated excluding one small sample each time ($n < 9$), they are always >23% ($P < .01$). Therefore, (1) our conclusions are still valid when the Cayapa sample, from Ecuadorian Amazonia, is considered an eastern population, and (2) the higher level of between-population differentiation is not an artifact of some small sample. We still think the Cayapa should be analyzed separately, and our reason for including them in the article was to illustrate that, in the future, our model can incorporate new elements, allowing for the inclusion of tribes with peculiar population histories, such as the amalgamation of Amazonian and Andean tribes.

By definition, models are working simplifications of reality. They should be continuously tested for goodness-of-fit as new data arise and, as a consequence of this, may be reinforced, modified, or rejected. Anyhow, model building is essential in science. The model proposed by us is very simple. South American genetic-variability data are scanty when compared, for instance, with data about Europe. For this reason, our model did not incorporate detailed migratory routes or estimates of the times when these migrations occurred. Future data may allow such refinements to be built in. Nevertheless, we think even a simple model should be based on accurate comparisons, the statistical significance of which must be assessed—which means that, one way or another, "probabilities or likelihoods should be estimated" (Simoni et al. 2000*a*).

Rothhammer and Moraga consider our results insufficient for any conclusions. However, Rothhammer and Silva (1989, 1992) proposed a much more complicated model, claiming genetic evidence of demic expansion accompanying the diffusion of manioc cultivations from central Amazonia to the Andean area, on even scantier data. Although we recognize that Rothhammer and Silva's proposal may be more appealing than our simple model, their fascinating tale about the migration of manioc farmers is not supported by any statistical test but, rather, is based on a cline inferred from synthetic genetic maps in an area where data are scanty or absent altogether (see figs. 1 and 2 of Rothhammer and Silva [1992]). Sokal et al. (1999) showed that, when samples are few and distant in space, synthetic maps obtained by interpolation often suggest a geographic trend, even when the data are spatially random.

We suspect that, in the case of our model, a simple unconvincing statement, even if authoritative, is not sufficient to discredit it. We are ready to accept that further data, or even an accurate reanalysis of our data, could challenge our model, but this seems not to be the case with Rothhammer and Moraga's criticisms. We think that, at the moment, the data about genetic variability of South Amerindians (at least for classical markers and molecular Y-chromosome variability) support our model rather than any of its alternatives.

Acknowledgments

We especially thank Guido Barbujani for suggestions and criticisms.

Eduardo Tarazona-Santos,[1,3,*] Denise R. Carvalho-Silva,[1,4] Davide Pettener,[3] Donata Luiselli,[3] Gian Franco De Stefano,[5] Cristina Martinez-Labarga,[5] Olga Rickards,[5] and Chris Tyler-Smith,[6] Sérgio D. J. Pena,[1] and Fabrício R. Santos[2]

Departamentos de [1]Bioquímica e Imunologia and [2]Biologia Geral, Universidade Federal de Minas Gerais, Minas Gerais, Brazil; [3]Area di Antropologia, Dipartimento di Biologia e. s., Università di Bologna, Bologna, Italy; [4]The Research School of Biological Sciences, Australian National University, Canberra; [5]Dipartimento di Biologia, Università di Roma "Tor Vergata," Roma; and [6]Department of Biochemistry, University of Oxford, Oxford

References

Luiselli D, Simoni L, Tarazona-Santos E, Pastor S, Pettener D (2000) Genetic structure of Quechua-speakers of Central Andes and geographic patterns of gene frequencies in South Amerindian populations. Am J Phys Anthropol 113:5–17

Rothhammer F, Moraga M (2001) Patterns of Y-chromosome variation in South Amerindians. Am J Hum Genet 69:904 (in this issue)

Rothhammer F, Silva C (1989) Peopling of Andean South America. Am J Phys Anthropol 78:403–410

——— (1992) Gene geography of South America: testing models of population displacement based on archeological evidence. Am J Phys Anthropol 89:441–446

Simoni L, Calafell F, Pettener D, Bertranpetit J, Barbujani G (2000*a*) Reconstruction of prehistory on the basis of genetic data. Am J Hum Genet 66:1177–1179

Simoni L, Tarazona-Santos E, Luiselli D, Pettener D (2000*b*) Genetic differentiation of South America native populations inferred from classical markers: from explorative analysis to a working hypothesis. In: Renfrew C (ed) America past, America present: genes and languages in the Americas and beyond. McDonald Institute for Archeological Research, Cambridge, pp 123–134

Sokal RR, Oden NL, Thomson BA (1999) A problem with synthetic maps. Hum Biol 71:1–13

Tarazona-Santos E, Carvalho-Silva D, Pettener D, Luiselli D, De Stefano GF, Martinez-Labarga C, Rickards O, Tyler-Smith C, Pena SDJ, Santos FR (2001) Genetic differentiation in South Amerindians is related to environmental and cultural diversity: evidence from the Y chromosome. Am J Hum Genet 68:1485–1496

Address for correspondence and reprints: Dr. Fabrício R. Santos. Departamento de Biologia Geral, ICB, UFMG, Av. Antônio Carlos 6627, CP 486, 31.270-010, Belo Horizonte, MG, Brazil. E-mail: fsantos@ icb.ufmg.br

* Present affiliation: Department of Biology, University of Maryland, College Park, MD.

0002-9297/2001/6904-0027$02.00

Am. J. Hum. Genet. 69:906–912, 2001

Comparisons of Two Methods for Haplotype Reconstruction and Haplotype Frequency Estimation from Population Data

To the Editor:

Haplotype reconstruction is an important issue, both in population genetics and in the identification of complex disease genes. Stephens et al. (2001) proposed a new statistical method (called the "PHASE method" in the following discussion, after the name of their computer program) for haplotype reconstruction based on phase-unknown marker genotype data from unrelated individuals in a population. On the basis of their simulations using coalescent models, they found that the PHASE method can reduce the error rate by >50% relative to the maximum-likelihood method, implemented via the expectation-maximization (EM) algorithm (Xie and Ott 1993; Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995). One limitation of their study is the fact that their simulations are based on coalescent models, which may not be good approximations of human population evolutionary histories. In fact, the authors acknowledge that "there simply do not exist enough real data sets, with known haplotypes for sequence or closely linked markers, to allow sensible statistical comparisons of different methods" (Stephens et al. 2001; p. 982). In this letter, we report a comparison of the two methods; our comparisons involve phase-known genotype data sets, as well as simulations using empirical population haplotype frequency data. Our results show that, in general, for most of the populations studied, there is no significant difference between the PHASE method and the EM method, both in the average error rate for haplotype reconstruction and in the discrepancy (see the report by Stephens et al. [2001] for definitions of these measures) between the estimated and true sample haplotype frequencies.

For our simulations based on empirical population haplotype frequency data, we used population haplotype frequencies for four loci (RET, COMT, HOXB and D4S10, with 3, 4, 5, and 6 polymorphisms, respectively) found in samples of four populations: European Americans, San Francisco Chinese, Biaka, and Maya. We use these four populations to represent the populations from four different continents. Descriptions of the populations and of the samples of those populations, as well as the haplotype definitions, can be found in ALFRED (Osier et al. 2001; ALFRED Web site). For each locus and each population, we randomly chose $2n$ haplotypes according to the haplotype frequencies and then randomly paired the haplotypes to form a population of n individuals with phase-known genotypes. The abilities of the two methods to reconstruct these haplotypes from the resulting data, ignoring phase information, were then evaluated. Twenty independent replicates for each population-locus combination were generated to compare the two haplotype reconstruction methods.

To estimate the haplotype frequencies, we implemented the EM algorithm in a computer program that analyzes the simulated data sets with the starting point of equal frequencies for every possible haplotype. We expect that any of the programs implementing the EM algorithm should yield similar results. Following Stephens et al. (2001), we specify the haplotype pair for an individual by choosing the most probable haplotype pair consistent with the individual's multisite genotype. The program developed by Stephens et al. (2001) was used to evaluate the performance of the PHASE method with