Human Biology

Volume 83 | Issue 3 Article 7

2011

# Allele Frequency Estimation from Ambiguous Data: Using Resampling Schema in Validating Frequency Estimates and in Selective Neutrality Testing

Jose Manuel Nunes
*University of Geneva*, Jose.DeAbreuNunes@unige.ch

Maria Eugenia Riccio
*University of Geneva*

Jean-Marie Tiercy
*Geneva University Hospital*

Alicia Sanchez-Mazas
*University of Geneva*

Recommended Citation
Nunes, Jose Manuel; Riccio, Maria Eugenia; Tiercy, Jean-Marie; and Sanchez-Mazas, Alicia (2011) "Allele Frequency Estimation from Ambiguous Data: Using Resampling Schema in Validating Frequency Estimates and in Selective Neutrality Testing," *Human Biology*: Vol. 83: Iss. 3, Article 7.
Available at: http://digitalcommons.wayne.edu/humbiol/vol83/iss3/7

# Allele Frequency Estimation from Ambiguous Data: Using Resampling Schema in Validating Frequency Estimates and in Selective Neutrality Testing

**Abstract**

The development of molecular typing techniques applied to the study of population genetic diversity originates data with increasing precision but at the cost of some ambiguities. As distinct techniques may produce distinct kinds of ambiguities, a crucial issue is to assess the differences between frequency distributions estimated from data produced by alternative techniques for the same sample. To that aim, we developed a resampling scheme that allows evaluating, by statistical means, the significance of the difference between two frequency distributions. The same approach is then shown to be applicable to test selective neutrality when only sample frequencies are known. The use of these original methods is presented here through an application to the genetic study of a Munda human population sample, where three different HLA loci were typed using two different molecular methods (reverse PCR-SSO typing on microbeads arrays based on Luminex technology and PCR-SSP typing), as described in details in the companion article by Riccio et al. [The Austroasiatic Munda population from India and its enigmatic origin: An HLA diversity study. Hum. Biol. 38:405–435 (2011)]. The differences between the frequency estimates of the two typing techniques were found to be smaller than those resulting from sampling. Overall, we show that using a resampling scheme in validating frequency estimates is effective when alternative frequency estimates are available. Moreover, resampling appears to be the unique way to test selective neutrality when only frequency data are available to describe the genetic structure of populations.

**Keywords**

EXPECTATION-MAXIMIZATION (EM) FREQUENCY ESTIMATION, AMBIGUOUS DATA, EWENS-WATTERSON NEUTRALITY TESTING, RESAMPLING, HLA.

**Cover Page Footnote**

This work was supported by the Swiss National Science Foundation (Bern, Switzerland) Grants 3100A0-112.651, and 31003A-127.465 to A.S.M. and by the Europe Science Foundation (Strasbourg, France) Cooperation in Science and Technology Grant of Action BM0803.

# *Allele Frequency Estimation from Ambiguous Data: Using Resampling Schema in Validating Frequency Estimates and in Selective Neutrality Testing*

JOSÉ MANUEL NUNES,[1]* MARIA EUGENIA RICCIO,[1] JEAN-MARIE TIERCY,[2]
AND ALICIA SANCHEZ-MAZAS[1]

*Abstract* The development of molecular typing techniques applied to the study of population genetic diversity originates data with increasing precision but at the cost of some ambiguities. As distinct techniques may produce distinct kinds of ambiguities, a crucial issue is to assess the differences between frequency distributions estimated from data produced by alternative techniques for the same sample. To that aim, we developed a resampling scheme that allows evaluating, by statistical means, the significance of the difference between two frequency distributions. The same approach is then shown to be applicable to test selective neutrality when only sample frequencies are known. The use of these original methods is presented here through an application to the genetic study of a Munda human population sample, where three different HLA loci were typed using two different molecular methods (reverse PCR-SSO typing on microbeads arrays based on Luminex technology and PCR-SSP typing), as described in details in the companion article by Riccio et al. [The Austroasiatic Munda population from India and its enigmatic origin: An HLA diversity study. *Hum. Biol.* 38:405–435 (2011)]. The differences between the frequency estimates of the two typing techniques were found to be smaller than those resulting from sampling. Overall, we show that using a resampling scheme in validating frequency estimates is effective when alternative frequency estimates are available. Moreover, resampling appears to be the unique way to test selective neutrality when only frequency data are available to describe the genetic structure of populations.

Different alternative molecular techniques are commonly used to type human genetic polymorphisms and may produce data with distinct kinds of ambiguities (e.g., in HLA typing described by Marsh et al. 2010). For example, HLA typings

[1] Laboratory of Anthropology, Genetics and Peopling History (AGP), Anthropology Unit, Department of Genetics and Evolution, University of Geneva, Switzerland.

[2] National Reference Laboratory for Histocompatibility, Transplantation Immunology Unit, Geneva University Hospital, Switzerland.

*Correspondence to: José Manuel Nunes, Laboratory of Anthropology, Genetics and Peopling History (AGP), Anthropology Unit, Department of Genetics and Evolution, 12 rue Gustave-Revilliod, 1227 University of Geneva, Switzerland. E-mail: Jose.DeAbreuNunes@unige.ch.

*Human Biology*, June 2011, v. 83, no. 3, pp. 437–447.

KEY WORDS: EXPECTATION-MAXIMIZATION (EM) FREQUENCY ESTIMATION, AMBIGUOUS DATA, EWENS-WATTERSON NEUTRALITY TESTING, RESAMPLING, HLA.

obtained by reverse PCR–sequence-specific oligonucleotide (SSO) hybridization technique on microbeads array (usually known as SSO-Luminex) and PCR–sequence-specific primer (SSP) techniques generate highly complex genotypic distributions, as described in the companion article by Riccio et al. (this issue). In a clinical setting, whenever required, such ambiguities are resolved by extra laboratory work such as sequencing. However, the cost of such additional testing generally prevents its general use to systematically resolve all the typing ambiguities observed in an entire population sample. This represents a main problem when data of large samples of individuals are needed to estimate allele or haplotype frequencies within the scope of epidemiological or anthropological studies, because the results remain ambiguous and are hardly comparable to data provided by other laboratories. An approach that is commonly used consists in "treating" and "cleaning" the data, but this leads to nonrandom elimination in putative alleles and provides biased frequency estimates with moderate to severe deviations (Buhler 2007). An alternative approach that avoids this problem is to apply a statistical methodology accommodating ambiguous data, a very effective way of getting accurate gene frequencies and other statistics for the population under study. One must be sure, however, that this second approach provides acceptable results. In particular, how to assess, when data include ambiguities, whether two different frequency distributions estimated for the same population sample differ?

We address this question in the present study. To this aim, we use the results of the analysis of the HLA-A, -B, and -DRB1 molecular polymorphisms of the Munda population sample described in the companion article (Riccio et al. this issue). These data were produced by two different molecular methods, SSO-Luminex and PCR-SSP. The two methods produce different genotyping ambiguities at each locus, and PCR-SSP reveals fewer ambiguities than SSO-Luminex. Allele frequencies were estimated for both methods and each of the three loci by using the GENE[RATE] implementation of the Expectation-Maximization (EM) algorithm (Dempster et al. 1977; Nunes 2005), an extension capable of accommodating all kinds of ambiguous data (Nunes 2005, 2007; Nunes et al. 2010). To test the equivalence of the two estimations, we developed an original resampling procedure to compare the values of the frequency estimates. This approach also provides the basis for an original test of selective neutrality.

## Materials and Methods

**Testing the Accuracy of the Allele Frequency Estimation through a Resampling Scheme.** To compare the frequency distributions estimated for the SSO-Luminex typings with those estimated for the PCR-SSP typings, we have used a resampling approach. The frequency distribution estimated with GENE[RATE] from the ambiguous SSO-Luminex data is assumed to be the population distribution. From this distribution, a given number of random samples of the same size as the observed sample and whose genotypes do not include ambiguities are drawn. The frequency distributions of these random samples are also estimated with Gene[rate]. We then chose the usual sum of squared differences between allelic frequencies as a test statistic to compare these distributions, i.e., we computed the sum of squared differences between the *assumed* population frequencies and the frequencies of each random sample. This gives us the empirical distribution of the statistic, which is an empirical estimate of the sampling variance. Assessing the sampling variance in this way avoids making hypotheses about the number of alleles actually present in samples and avoids concerns about the accuracy of multinomial sampling.

This procedure is done twice, once using the SSO-Luminex estimates, as mentioned above, and once using the PCR-SSP estimates as the population frequencies. To see how extreme is the difference observed between the estimations obtained for the two typing techniques, we compared the *observed* statistic to the empirical distributions. This observed statistic is the same in both cases, i.e., the sum of the squared differences between the PCR-SSP and the SSO-Luminex frequencies. The conclusion of the test depends on the location of the observed statistic relative to the empirical distribution. If the test statistic falls within the central 95% interval (but other significance levels can be used), it means that the difference is comparable to the sampling variation. If the test statistic falls to the right of that interval, it means that the estimates differ significantly more than expected from sampling. If, on the other hand, the test statistic falls to the left of the interval, it can be said that the observed difference is significantly smaller than expected from sampling. In other words, in this latter case, taking another sample from the same population would yield, on average, frequency differences larger than those observed when different typing techniques are used.

According to common practice in resampling (e.g., Davison and Hinkley 2006; Efron and Tibshirani 1993), the number of generated samples has been set to 1000 so that the empirical distribution provides a good approximation of the population distribution. The program implementing this, however, accepts the number of random samples as input along with the two sets of frequency estimates. The outputs of the program are the two empirical distributions, one for each PCR-SSP and SSO-Luminex estimation, and the value of the test statistic for the original samples.

**Testing Selective Neutrality through a Resampling Scheme.** Ewens-Watterson's (EW) test is a test for selective neutrality under the infinite allele model (Ewens 1972; Watterson 1978). The method used here is an improvement of the procedure implemented in Gene[rate] (Nunes 2004, 2007; Nunes et al. 2010), which is an adaptation of the classical EW test to ambiguous data. It consists in estimating the *p* values of the usual EW test for nonambiguous random samples and in using the distribution of these *p* values to assess selective neutrality. The resampling scheme consists in generating a certain number of random samples in which no individual has ambiguous genotypes. Then, the

Slatkin's version of the EW test (Slatkin 1994, 1996) is applied to each replica sample, providing a set of *p* values. The same considerations, made above, lead us to set a default value of 1,000 samples and 10,000 replicas for each EW test, but these values can be adjusted, if necessary, by indicating them explicitly to the program.

The obtained set of *p* values can be seen as an empirical distribution from which it is possible to calculate the probability of observing a *p* value smaller than a given value. Actually, however, this is not the best approach. A faster approach to test selective neutrality is to consider the set of *p* values as a multiple testing situation. As the null hypothesis is selective neutrality, an overall result requires the use of Bonferroni's (Bonferroni 1936) adjusted significance levels to detect extreme *p* values. The program that we have developed provides a count of the *p* values that are significantly lower or significantly higher than expected under the neutral hypothesis at a given significance level (default value is 5%, but the level can be set as an input to the program). As usual in a Bonferroni's multiple testing framework, a count of one significant value is sufficient to reject neutrality. If all significant counts are lower (or higher) than the critical levels, then it is reasonable to assume a deficiency (respectively, an excess) of homozygotes. If both higher and lower counts appear, neutrality is also rejected, but it is not possible, then, to suggest any selective model (and it might also be an indication of nonconformity to Hardy-Weinberg equilibrium).

If the multiple comparison approach described above is not directly applicable, for instance, because of the existence of zeros among the *p* values, the empirical distribution that can be obtained as an output of the program helps to make significance decisions that would require huge calculations otherwise (both in the number of samples and of replicas). If no *p* values are obtained between zero and Bonferroni's corrected significance level, we may suspect the zeros as being artifacts due to insufficient numbers of EW replicas or generated samples. Therefore, the neutral hypothesis cannot be rejected.

**Application.** The methods described above were applied to the data analyzed in our companion article by Riccio et al. (2011), i.e., HLA-A, -B and -DRB1 data obtained on a sample of 40 individuals of the Munda population living in the Ranchi district in Northeast India.

As described by Riccio et al. (2011), the allele frequencies were estimated with an EM algorithm, taking into account typing ambiguities—a method implemented in the GENE[RATE] program package (Nunes 2004, 2007; Nunes et al. 2010) available online (http://geneva.unige.ch) from two genetic data series: one obtained by SSO-Luminex typings, leading to a high number of genotyping ambiguities, and the other one obtained by PCR-SSP, leading to a lower number of such ambiguities.

**Comparison between Allele Frequencies Estimated from SSO-Luminex and PCR-SSP Typings.** Our first aim was to test the null hypothesis of no significant differences between the frequencies of the two data series, at each locus. Figures 1 and 2 shows the results of the resampling procedure developed to compare these two allele frequency distributions. In Figure 1, the SSO-Luminex estimates are taken as the population frequencies that are used to draw nonambiguous random samples. The frequencies of these samples are then estimated and compared with the assumed population frequencies. The distributions shown in Figure 1 represent the test statistic, i.e., the sum of squared allele frequency differences, over the simulated data, at each locus. In all cases, the observed value (calculated from the difference between the observed SSO-Luminex and PCR-SSP frequencies, indicated by an arrow) falls at the left tail of the empirical distribution, meaning that frequency differences due to the use of the two distinct methods are less important than expected by chance as a consequence of sampling. The same conclusion is drawn when PCR-SSP frequency estimates are taken as the population frequencies, as shown in Figure 2.

These results indicate that the frequency estimations obtained with GENE[RATE] on ambiguous data can be used with confidence to represent the population under study and can be applied in population genetics analyses such as genetic distance comparisons and multivariate and correlation analyses. Although the present demonstration is empirical, the same conclusion has been reached independently for three loci (with unequal levels of polymorphism), which gives it stronger support.

**Test for Selective Neutrality.** We also tested the hypothesis of selective neutrality on the Munda sample, at each locus and for the two data series (SSO-Luminex and PCR-SSP). The results are summarized in Table 2 of the companion article by Riccio et al. (this issue). No clear rejection toward an excess or a deficit of heterozygotes is found when analyzing the SSO-Luminex frequencies. The tests applied to the PCR-SSP frequencies give the same results for loci HLA-A and -B, but not for HLA-DRB1. As the frequency estimates obtained with the two typing methods were not found to be significantly different, we further analyzed the divergence observed for neutrality at locus HLA-DRB1. According to the techniques described in Materials and Methods, we increased both the number of samples and the number of replicas, and we analyzed the distributions of the *p* values. These distributions showed that almost all values exceeded the Bonferroni's significance level, to the exception of *p* values of zero, consistently for all numbers of samples and replicas used. In order to make a decision, we set the significance level to 1% and the number of generated samples successively to 1,000, 10,000, and 100,000. The distributions of *p* values showed that the only *p* values that were smaller than the corresponding significances (1e-5, 1e-6, and 1e-7) were zeros. To avoid such situations and obtain *p* values with enough significant digits, one would need to increase the number of replicates for the EW Slatkin's test to more than 1,000,000. However, as explained in Materials and Methods, the absence of intermediate *p* values for all the bootstrapped sample sizes used lead us to

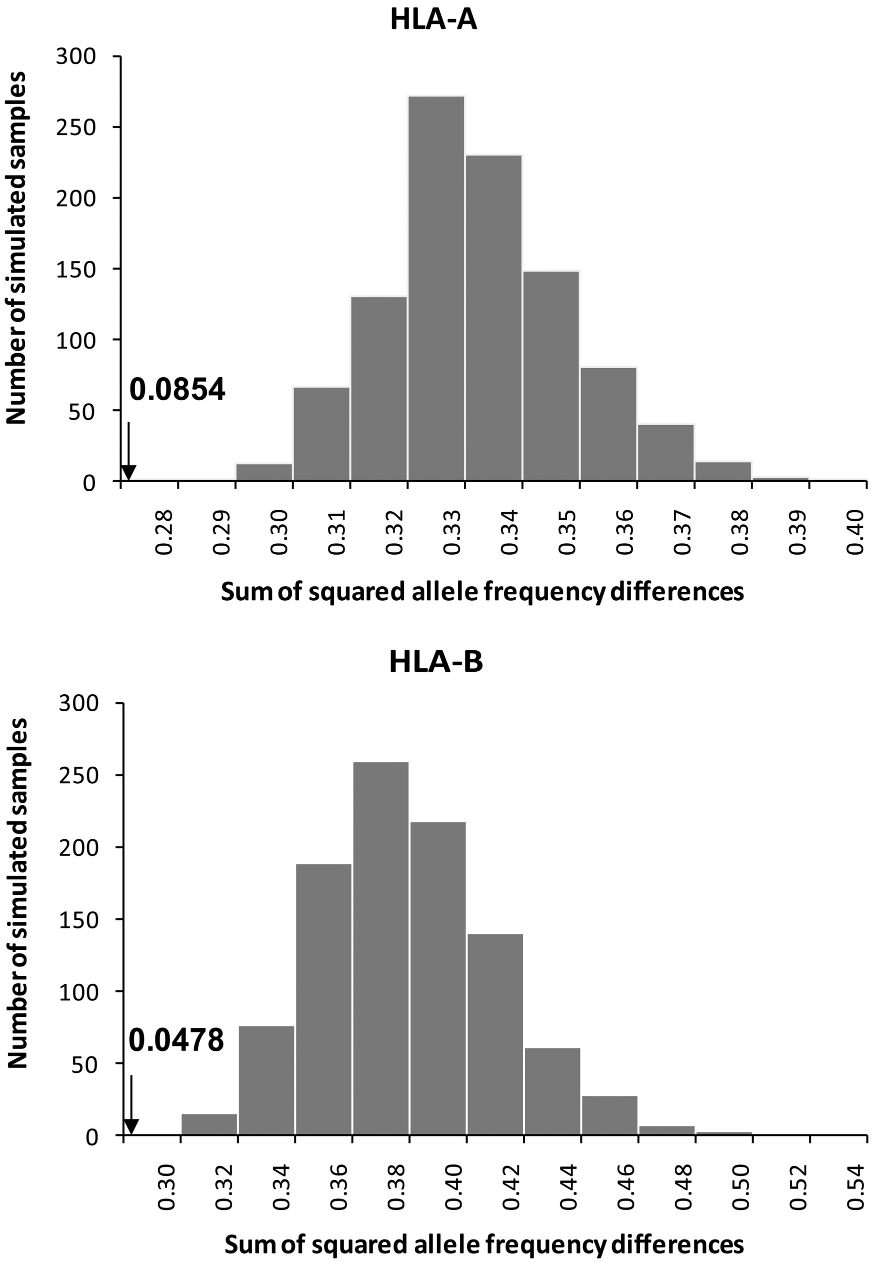

**Figure 1.** Results of the resampling procedure developed to compare allelic frequency distributions using SSO-Luminex estimates. In all cases the observed value (calculated from the difference between the observed SSO-Luminex and PCR-SSP frequencies, indicated by an arrow) falls at the left tail of the empirical distribution.

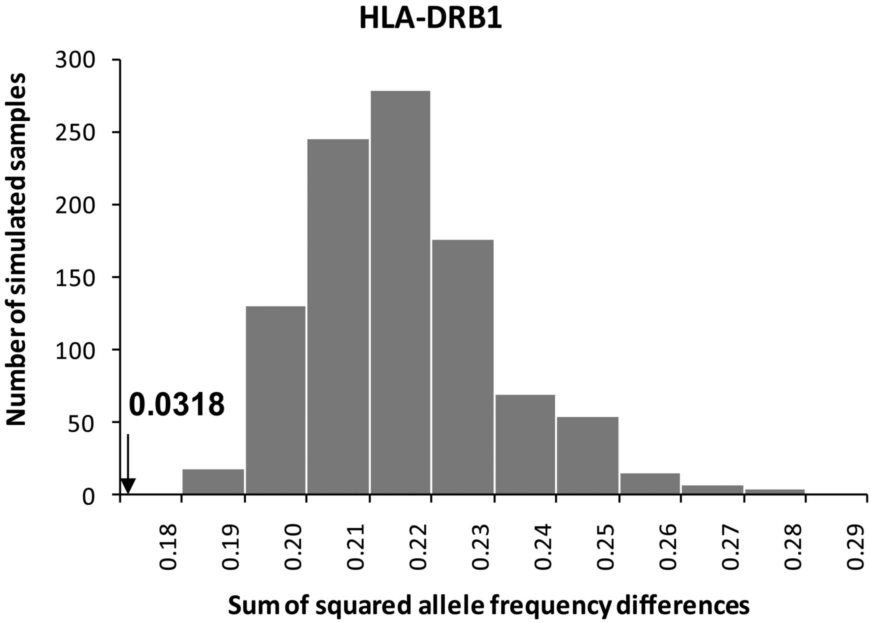

**Figure 1.** (*continued*)

consider these zeros as numerical artifacts. Therefore, the test of selective neutrality cannot be rejected at the HLA-DRB1 locus typed by PCR-SSP; we thus conclude that all loci are compatible with the hypothesis of selective neutrality for both typing techniques.

## Discussion and Conclusion

There are two difficulties concerning the estimation of gene frequencies: one relates to the estimation with ambiguities and the other to the comparison of samples for which precise genotypes are not known. The questions raised by the existence of ambiguous typings are well documented in the literature (Adams et al. 2004; Leffell 2002; Lind et al. 2010; Scott et al. 1998; Swelsen et al. 2004; Voorter et al. 2007), but the approach of improving the resolution is simply not feasible in general for population-based studies. As the vast majority, if not almost all, of the computer programs are not able to handle frequency estimation for arbitrarily ambiguous data [e.g., Arlequin (Excoffier et al. 2005), PyPOP (Lancaster et al. 2003), Phase (Stephens and Donnelly 2003)]), we have developed an extension of the EM algorithm (Nunes 2005) capable of dealing with them. Actually, this corresponds to the ability to use an EM-like algorithm for noncodominant data. Although this is useful by itself, the problem of comparing ambiguous estimates for noncodominant data cannot be handled by the usual homogeneity test that requires data with nonambiguous genotypes (Sokal and Rohlf 1995).

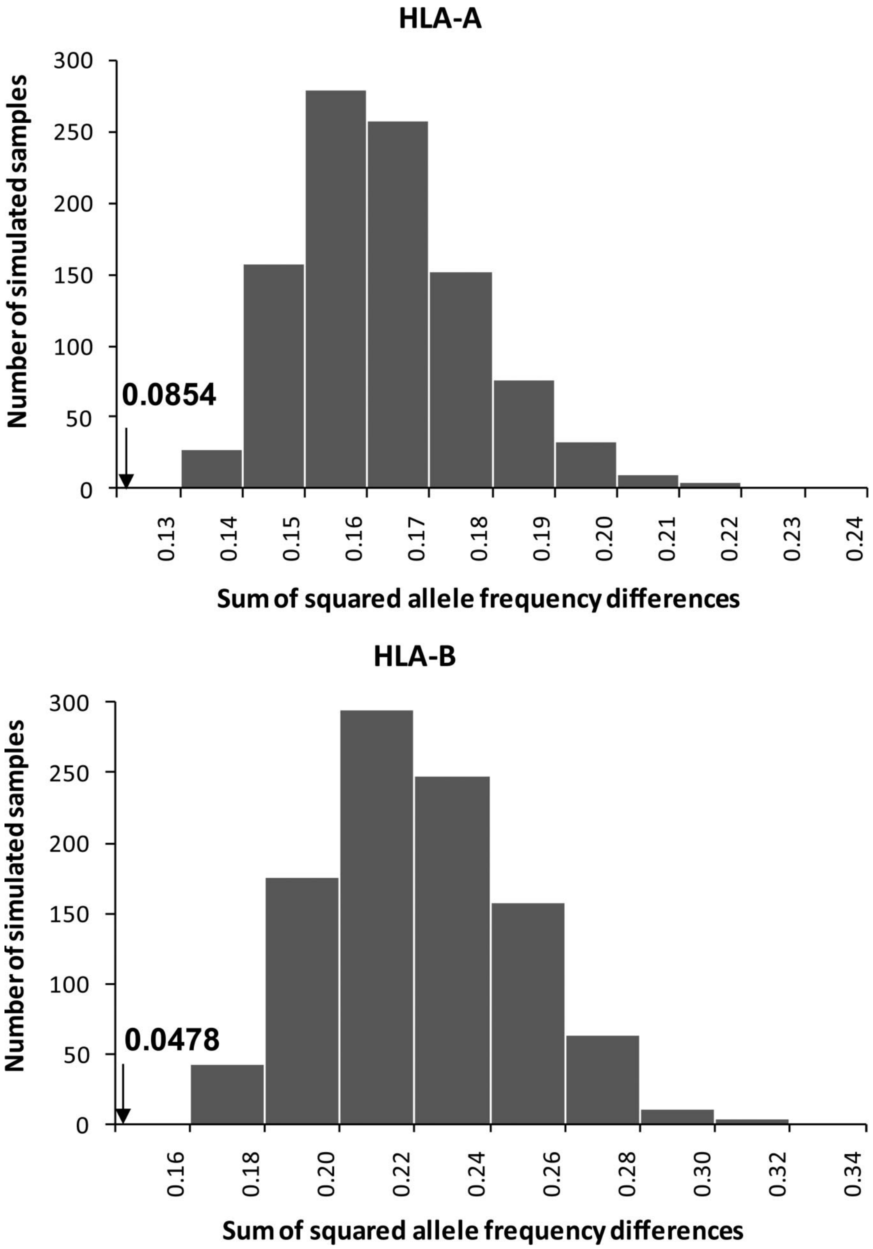

**Figure 2.** Results of the resampling procedure developed to compare allelic frequency distributions using PCR-SSP estimates. In all cases the observed value (calculated from the difference between the observed SSO-Luminex and PCR-SSP frequencies, indicated by an arrow) falls at the left tail of the empirical distribution.

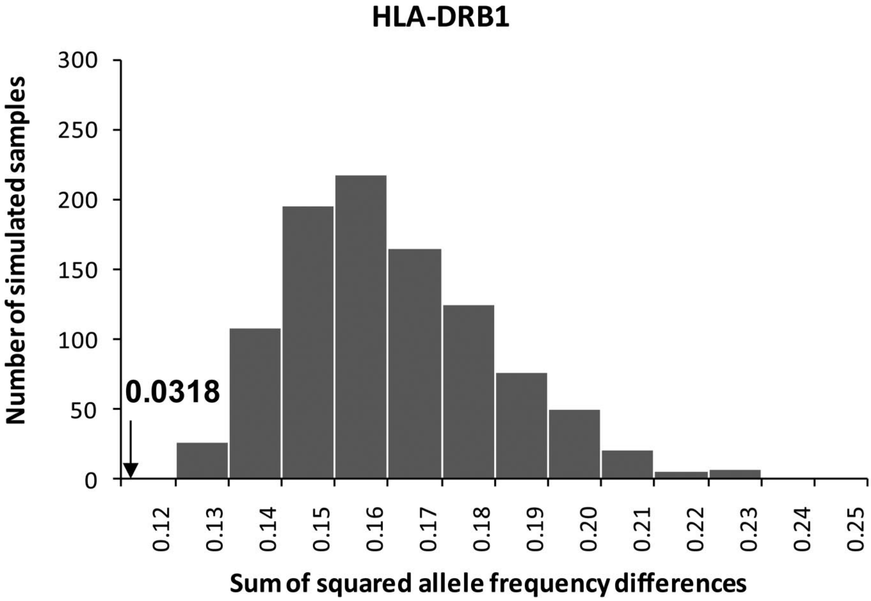

**Figure 2.** (*continued*)

To our knowledge, this is the first study evaluating the consistency of allele frequency distributions estimated from genotypic data obtained by two different molecular approaches. Here, we have applied a new approach to HLA data obtained by SSO-Luminex and PCR-SSP methods at three different loci (HLA-A, -B, and -DRB1) for the same Austroasiatic Munda population sample from India (see Riccio et al., this issue). This allowed us to assess, by an original resampling procedure, whether the frequency estimations obtained using highly ambiguous SSO-Luminex typing data could be accepted as equivalent to those obtained by PCR-SSP data.

The EM algorithm accommodating ambiguous data implemented in GENE[RATE] is a useful tool allowing to avoid arbitrary elimination of alleles from ambiguous genotypes and to prevent both a loss of information and additional typing work. In the present study, we presented a procedure allowing to assess whether HLA frequency distributions estimated on ambiguous data can be used with confidence in population genetics analyses. Of course, prudence is required while making comparisons of specific allelic frequencies estimated from ambiguous data between populations, as discussed in the companion article by Riccio et al. (this issue). In particular, allelic frequencies should not be used as indicators of the presence or absence of any given allele.

We believe that the most useful result of this study for researchers working with frequency data is the simple and easy way, developed here, of assessing selective neutrality. It may be argued about the power of such a test; however, if a rejection of neutrality still holds after the detailed inspection, considering both multiple testing and the actual $p$ value distributions, which we have illustrated, then it can be taken as very likely.

Overall, using a resampling scheme in validating frequency estimates proves to be most useful when alternative frequency estimates are available, for example, due to the application of different typing techniques. Also, using a resampling approach currently appears to be the unique way to test selective neutrality when only frequency data are available to describe the genetic structure of populations.

*Acknowledgments* This work was supported by the Swiss National Science Foundation (Bern, Switzerland) Grants 3100A0-112.651, and 31003A-127.465 to A.S.M. and by the Europe Science Foundation (Strasbourg, France) Cooperation in Science and Technology Grant of Action BM0803.

*Received 6 September 2010; revision accepted for publication 7 December 2010.*

## Literature Cited

Adams, S. D., K. C. Barracchini, D. Chen et al. 2004. Ambiguous allele combinations in HLA Class I and Class II sequence-based typing: When precise nucleotide sequencing leads to imprecise allele identification. *J. Transl. Med.* 2:30.

Bonferroni, C. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*. 8:3–62.

Buhler, S. 2007. Etude du polymorphisme moléculaire des gènes HLA de classes I et II à l'échelle mondiale: Analyse de la diversité nucléotidique dans les populations. Ph.D. diss., Geneva, Switzerland: University of Geneva, 251.

Davison, A. C., and D. V. Hinkley. 2006. *Bootstrap Methods and Their Application*. Cambridge, U.K.: Cambridge University Press, 592.

Dempster, A., N. Laird, and D. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series A (General)*. 39:1–38.

Efron, B., and R. J. Tibshirani. 1993. *An Introduction to the Bootstrap*. London, U.K.: Chapman and Hall.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* 3:87–112.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol. Bioinform. Online.* 1:47–50.

Lancaster, A., M. P. Nelson, D. Meyer et al. 2003. PyPop: A software framework for population genomics: Analyzing large-scale multi-locus genotype data. *Pac. Symp. Biocomput*. 514–525.

Leffell, M. S. 2002. MHC polymorphism: Coping with the allele explosion. *Clin. Appl. Immunol. Rev.* 3:35–46.

Lind, C., D. Ferriola, K. Mackiewicz et al. 2010. Next-generation sequencing: The solution for high-resolution, unambiguous human leukocyte antigen typing. *Hum. Immunol.* 71: 1033–1042.

Marsh, S. G., E. D. Albert, W. F. Bodmer et al. 2010. Nomenclature for factors of the HLA system, 2010. *Tissue Antigens* 75:291–455.

Nunes, J. M. 2004. Tools for efficient HLA data handlings. HLA 2004, Immunobiology of the Human MHC. In *Proceedings of the 13th International Histocompatibility Workshop and Congress*, J. Hansen and B. Dupont, eds. Seattle, WA: IHWG Press.

Nunes, J. M. 2005. Counting Genes. Ph.D. diss. Instituto de Ciências Biomédicas Abel Salazar, Porto, Portugal: University of Porto.

Nunes, J. M. 2007. Tools for analysing ambiguous HLA data. *Tissue Antigens* 69(Suppl. 1):203–205.

Nunes, J. M., M. E. Riccio, S. Buhler et al. 2010. Analysis of HLA population data (AHPD) submitted to the 15th International Histocompatibility/Immunogenetics workshop by using the GENE[RATE] computer tools accommodating ambiguous data (AHPD project report). *Tissue Antigens* 76:18–30.

Riccio, M. E., J. M. Nunes, M. Rahal et al. 2011. The Austroasiatic Munda population from India and its enigmatic origin: An HLA diversity study. *Hum. Biol.* 38:405–435.

Scott, I., J. O'Shea, M. Bunce et al. 1998. Molecular typing shows a high level of HLA class I incompatibility in serologically well matched donor/patient pairs: Implications for unrelated bone marrow donor selection. *Blood* 92:4864–4871.

Slatkin, M. 1994. An exact test for neutrality based on the Ewens sampling distribution. *Genet. Res.* 64:71–74.

Slatkin, M. 1996. A correction to the exact test based on the Ewens sampling distribution. *Genet. Res.* 68:259–260.

Sokal, R. R., and F. J. Rohlf. 1995. *Biometry.* New York, NY: Freeman. 887.

Stephens, M., and P. Donnelly. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am. J. Hum. Genet.* 73:1162–1169.

Swelsen, W. T., C. E. Voorter, and E. M. van den Berg-Loonen. 2004. Ambiguities of human leukocyte antigen-B resolved by sequence-based typing of exons 1, 4, and 5. *Tissue Antigens* 63:248–254.

Voorter, C. E., E. Mulkers, P. Liebelt et al. 2007. Reanalysis of sequence-based HLA-A, -B and -Cw typings: How ambiguous is today's SBT typing tomorrow. *Tissue Antigens* 70:383–389.

Watterson, G. A. 1978. The homozygosity test of neutrality. *Genetics* 88:405–417.