TECHNICAL COMMENT

Response to Comment on "Genetic Structure of Human Populations"

Estimates of genetic variance components depend on the type of marker used, the definitions of geographic regions, the populations sampled within these regions, the relative sample sizes from the populations, and the way in which information is combined across loci. For microsatellite markers, estimates also depend on whether the quantity whose variance is partitioned is an allele-size variable or an indicator variable for allelic presence or absence. A main purpose of our variance component estimation was to provide insight into the fine-scale population *structure* analysis in (*1*). Because the structure algorithm uses only identity and nonidentity of alleles, descriptive statistics that employ allelic indicator variables are more appropriate for understanding the dependence of *structure*-based inference on the "level of difference" among groups than are statistics that use allele size.

Excoffier and Hamilton (*2*) performed a complementary variance component analysis, demonstrating that when a subset of our data corresponding to (*3*) is studied using allele sizes, as was done in (*3*), similar estimates to (*3*) are obtained. Their smaller within-population variance component compared with that in (*1*) is consistent with the smaller estimate of (*3*) in relation to microsatellite studies that used indicator variables (*4*–*7*). However, because previous indicator-based studies of microsatellites and other markers have not all been in full agreement (*1*, *4*–*11*), a difference in the nature of the variable cannot be the sole source of differing estimates. First, the homogenizing effect of the higher mutation rates of microsatellites, in contrast with those of other markers, probably explains some of the difference of our results from nonmicrosatellite indicator-based studies (*12*). Second, consistent with past observations (*13*), the high fraction of tetranucleotide loci in our data contributes to higher within-population variance component estimates (Table 1) than are seen in dinucleotide studies (*3*, *4*, *7*). Third, the estimates vary considerably across sampling schemes within regions, and in several cases (*3*, *6*, *7*), past microsatellite samples that included multiple groups per region used populations that are among the most differentiated of the 52 groups in our data (Fig. 1). Any estimate computed with the well-separated populations that contribute to the 83.4% within-population variance component obtained by Excoffier and Hamilton (*2*) should be regarded as a lower bound.

Allele sizes are important in microsatellite analysis, and typical studies, including our use of the data from (*1*) to investigate population divergence and expansion (*14*), employ both sizes and indicator variables. However, although they are often useful, stepwise mutation models with length-independent transition probabilities, which underlie the approach used in (*2*), poorly predict microsatellite allele size distributions in the human genome compared with length-dependent models (*15*). Because of this issue and the frequent occurrence of multistep mutations (*16*), the model of Excoffier and Hamilton cannot be regarded as the "right mutation model," and the "minimum number of mutations separating the alleles" need not actually be minimal.

Finally, the main finding from studies of genetic variance components, supported by diverse analyses whose exact estimates have differed, is that the within-population variance component is much larger than the other components. The relative importance of various influences on the estimates could potentially be evaluated by further statistical analysis of the variation in the variance component estimates themselves.

Table 1. Analysis of molecular variance for 45 di-, 58 tri- and 274 tetranucleotide loci from (*1*). The samples and estimation procedure (*17*) used are the same as in (*1*).

Sample	No. of regions	No. of populations	Repeat size	Variance components and 95% confidence intervals (%)		
				Within populations	Among populations within regions	Among regions
World	1	52	2	92.2 (91.5, 92.8)	7.8 (7.2, 8.5)	
			3	92.6 (91.8, 93.2)	7.4 (6.8, 8.2)	
			4	95.4 (95.1, 95.6)	4.6 (4.4, 4.9)	
World	5	52	2	90.1 (89.2, 90.9)	2.9 (2.7, 3.2)	7.0 (6.1, 7.8)
			3	90.5 (89.5, 91.3)	2.7 (2.4, 3.0)	6.8 (6.0, 7.8)
			4	94.3 (93.9, 94.6)	2.3 (2.2, 2.4)	3.4 (3.1, 3.7)
World	7	52	2	91.4 (90.7, 92.1)	2.8 (2.5, 3.1)	5.8 (5.1, 6.6)
			3	91.8 (90.9, 92.5)	2.5 (2.3, 2.8)	5.7 (5.0, 6.5)
			4	95.0 (94.7, 95.2)	2.3 (2.2, 2.4)	2.8 (2.5, 3.0)
World-B97	5	14	2	85.9 (84.7, 86.8)	5.7 (4.9, 6.7)	8.4 (7.2, 9.7)
			3	86.2 (84.8, 87.5)	4.9 (4.2, 5.7)	8.9 (7.7, 10.3)
			4	91.2 (90.7, 91.7)	4.9 (4.6, 5.2)	3.9 (3.4, 4.4)

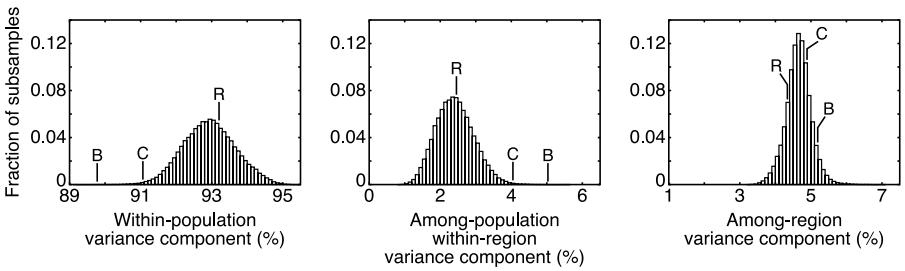

Fig. 1. Effect of sampling scheme on variance component estimates. Using the five-region design, variance components were estimated (*17*) as in (*1*) for each of 100,000 subsets of populations, sampled randomly from among the $\sim 3 \times 10^{15}$ subsets of the 52 populations in (*1*) for which all five regions were represented. Variance component estimates for a 14-population subsample corresponding to Barbujani *et al.* (*3*), a 9-population subsample corresponding to Calafell *et al.* (*7*), and the full 52-population data of Rosenberg *et al.* (*1*) are marked (B), (C), and (R), respectively. Subsample (B) is the same as subsample (B97) in (*1*). Subsample (C) includes Biaka, Druze, Han, Japanese, Maya, Mbuti, Melanesian, Surui, and Yakut. For the within-population and among-population within-region components, (B) had more extreme values than all but four of the subsets. Similar results were obtained for random subsets that included at least two populations per region.

Noah A. Rosenberg
Molecular and Computational Biology
University of Southern California
1042 West 36th Place DRB 289
Los Angeles, CA 90089, USA
E-mail: noahr@usc.edu

Jonathan K. Pritchard
Department of Human Genetics
University of Chicago
920 East 58th Street, CLSC 507
Chicago, IL 60637, USA

James L. Weber
Center for Medical Genetics
Marshfield Medical Research Foundation
Marshfield, WI 54449, USA

Howard M. Cann
Foundation Jean Dausset-CEPH
27 rue Juliette Dodu
75010 Paris, France

Kenneth K. Kidd
Department of Genetics
Yale University School of Medicine
333 Cedar Street
New Haven, CT 06520, USA

Lev A. Zhivotovsky
Vavilov Institute of General Genetics
Russian Academy of Sciences
3 Gubkin Street
Moscow 117809, Russia

Marcus W. Feldman
Department of Biological Sciences
Stanford University
Stanford, CA 94305, USA

References and Notes

1. N. A. Rosenberg *et al*., *Science* **298**, 2381 (2002).
2. L. Excoffier, G. Hamilton, *Science* **300**, 1877 (2003); www.sciencemag.org/cgi/content/full/300/5627/1877b.
3. G. Barbujani, A. Magagni, E. Minch, L. L. Cavalli-Sforza, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4516 (1997).
4. R. Deka, *Am. J. Hum. Genet*. **56**, 471 (1995).
5. L. B. Jorde *et al*., *Am. J. Hum. Genet*. **57**, 523 (1995)
6. A. Pérez-Lezaun *et al*., *Hum. Genet*. **99**, 1 (1997).
7. F. Calafell *et al.*, *Eur. J. Hum. Genet*. **6**, 38 (1998).
8. R. C. Lewontin, *Evol. Biol*. **6**, 381 (1972).
9. B. D. H. Latter, *Am. Nat*. **116**, 220 (1980).
10. M. Dean *et al*., *Am. J. Hum. Genet*. **55**, 788 (1994).
11. C. Romualdi *et al.*, *Genome Res*. **12**, 602 (2002).
12. L. Jin, R. Chakraborty, *Heredity* **74**, 274 (1995).
13. A. Ruiz Linares in *Microsatellites: Evolution and Applications*, D. B. Goldstein, C. Schlötterer, Eds. (Oxford University Press, Oxford, 1999), pp. 183–197.
14. L. A. Zhivotovsky, N. A. Rosenberg, M. W. Feldman, *Am. J. Hum. Genet*. **72**, 1171 (2003).
15. P. Calabrese, R. Durrett, *Mol. Biol. Evol*. **20**, 715 (2003).
16. X. Xu, M. Peng, Z. Fang, X. Xu, *Nature Genet*. **24**, 396 (2000).
17. Variance components were estimated assuming independence of alleles within individuals, using the framework in chapter 5 of (*18*).
18. B. S. Weir, *Genetic Data Analysis II* (Sinauer, Sunderland, MA, 1996).
19. We thank E. Minch for clarifying ambiguities in (*3*).

18 March 2003; accepted 14 April 2003