doi: 10.1111/j.1469-1809.2006.00287.x

The Genetic Structure of Human Populations Studied Through Short Insertion-Deletion Polymorphisms

Luciana Bastos-Rodrigues[1], Juliana R. Pimenta[2] and Sergio D. J. Pena[1,2,*]

[1]*Departamento de Bioquímica e Imunologia, Universidade Federal de Minas Gerais, 31270-910 Belo Horizonte, Brazil*
[2]*GENE – Núcleo de Genética Médica, 30130-909 Belo Horizonte, MG*

Summary

In a landmark study Rosenberg *et al.* (2002) analyzed human genome diversity with 377 microsatellites in the HGDP-CEPH Genome Diversity Panel and reported that the populations were structured into five geographical regions: America, Sub-Saharan Africa, East Asia, Oceania and a cluster composed of Europe, the Middle East and Central Asia. They also observed that the within-population component accounted for 93–95%, and that the among-regions portion was only 3.6%, of the total genetic variance. We have also studied the HGDP-CEPH Diversity Panel (1064 individuals from 52 populations) with a set of 40 biallelic slow-evolving short insertion-deletion polymorphisms (indels). We confirmed the partition of worldwide diversity into five genetic clusters that correspond to major geographic regions. Using the indels we have also disclosed an among-regions component of genetic variance considerably larger (12.1%) than had been estimated using microsatellites. Our study demonstrates that a set of 40 well-chosen biallelic markers is sufficient for the characterization of human population structure at the global level.

Keywords: genetic structure, insertion-deletion polymorphisms, indels, human diversity, DNA

Introduction

Now that we have the complete DNA sequence of the euchromatic human genome (International Human Genome Sequencing Consortium, 2004) there is growing interest in characterizing human genomic variation. The conventional approach for this goal has been first to divide humanity into populations which can then be studied. However, populations often have an ambiguous meaning, being irregularly defined on the basis of "race", geography, culture, religion, physical appearance or other criteria. Rosenberg *et al.* (2002) tried to avoid this problem by studying the structure of human genome diversity without prior population assignment. In a landmark study they used the *Structure* computer program (Pritchard *et al.* 2000) that uses a cluster algorithm for inferring population structure on the basis of genotype data. A set of 377 autosomal microsatellites and the 52 populations of the HGDP-CEPH Diversity Panel (Cann *et al.* 2002) were used to study worldwide human genome variation. Without *a priori* information about the origin of individuals *Structure* was able to identify five main clusters that corresponded to major geographic regions of the globe. Rosenberg *et al.* (2002) also observed that the within-population differences among the individuals accounted for 93–95%, and that the among-regions variation was only 3.6%, of the total genetic variance.

The first estimation of the levels of human genetic variation at individual, population and regional levels was published in 1972 by Lewontin, who used blood groups, protein variants and isoenzymes to calculate values of 85.4% for the within-population, 8.3% for the among-populations-within-continent, and 6.3% for among-continents components of genetic variance. Using DNA markers other authors also obtained similar results (Barbujani *et al.* 1997; reviewed in Barbujani & Di Benedetto, 2001; Excoffier & Hamilton, 2003), leading

*Corresponding author: Sérgio Danilo Junho Pena, GENE – Núcleo de Genética Médica, Av. Afonso Pena 3111/9, 3013-909 Belo Horizonte, MG Brazil. Tel: +55-31-32848000; Fax: +55-31-32273792. E-mail: spena@gene.com.br

to the important corollary that the human species has a low level of geographical structuring that is not compatible with the existence of human races (Templeton, 1999). Observing that no previous study had estimated a within-population component of human genetic variance as high as 93–95%, Excoffier & Hamilton (2003) suggested that the microsatellite mutation model that had been used by Rosenberg *et al.* (2002) had caused an artifactual underestimation of the among-regions component. In response, Rosenberg *et al.* (2003) made a spirited defense of their mutation model and blamed the sampling scheme. We decided to approach this problem by studying exactly the same HGDP-CEPH Diversity Panel used by Rosenberg *et al.* (2002), but using a different type of genetic markers, slow-evolving diallelic short insertion-deletion polymorphisms (indels).

Weber *et al.* (2002) characterized 2,000 human diallelic short indels in the human genome. We accessed their data base (http://research.marshfieldclinic.org/) and identified 40 polymorphisms that fulfilled the following criteria: widespread chromosomal location, increasing amplicon sizes that allow multiplex analysis, and allele frequency close to 0.5 in the European population (Supplementary Table 1). Here we report our results from the application of these 40 indel markers to the study of all the samples in the HGDP-CEPG Diversity Panel. Using the *Structure* program we could reproduce the identification of five main clusters that corresponded to major geographic regions of the globe. However, our analysis of genetic variance showed considerably more structuring, with 85.7% within-population, 2.3% among-populations within-regions, and 12.1% among-regions components of genetic variance.

Materials and Methods

Populations Studied

DNA samples from 1,064 individuals were obtained from the HGPD-CEPH Human Genome Diversity Cell Line Panel (http://www.cephb.fr/HGDP-CEPH-Panel/; Cann *et al.* 2002). The individuals were sampled across all five continents and assigned to 52 different populations from seven regional groups (Africa, Europe, Middle East, Central/South Asia, East Asia, Oceania and America).

DNA Analysis

DNA from each individual was independently typed for 40 biallelic short insertion/deletion polymorphisms (indels) selected from those described by Weber *et al.* (2002) and available at http://research.marshfieldclinic.org/genetics/indels/default.asp (Supplementary Table 1). The PCR amplifications used four multiplex reaction systems, each consisting of a mix of 10–12 primer pairs (Supplementary Table 2). To each forward PCR primer a tail of the M13-40 17-mer oligonucleotide was added. The multiplex PCR assay was performed in a 10-μl final volume of the following: 1 X PCR buffer (10 mM Tris-HCl pH 8.3 or pH 9.2, 75 mM KCl, 3.5 mM $MgCl_2$), 200 μM dNTPs, 1.0 U of Platinum *Taq* DNA polymerase (Invitrogen), 20 ng of genomic DNA, 1.5 μM of M13-40 forward primer labelled with the FAM dye, 1.5 μM of each unlabelled reverse primer, and 0.1 μM of each unlabelled forward primer.

Two microlitres of each labelled PCR products was denatured in formamide solution plus MegaBACE^{TM} ET550-R Size Standard at 95°C for 5 min, and subjected to fragment analysis using a MegaBACE 1000 DNA sequencer (GE Healthcare) according to the manufacturer's instructions. Analyses of allele sizes were scored using the Genetic Profiler (version 2.2) and Fragment Profiler (version 1.2) programs (GE Healthcare).

Population Structure Analysis

We utilised the *Structure* program version 2.1 (Pritchard *et al.* 2000), available at http://pritch.bsd.uchicago.edu/software.html. This software uses multilocal genotypes to infer the structure of each population and allocate individuals to different populations. The software defines "K" clusters (where K has to be provided by the user), each of them being characterized by a set of allelic frequencies for each locus. The individuals are grouped (probabilistically) on the basis of their genotypes. We ran ten independent replicates for each value of K, which varied from 1 to 10. Every run consisted of 100,000 burn-in steps, followed by 2×10^6 Markov Chain Monte Carlo iterations, without any prior information on the population origin of each sampled individual. We used the "no admixture" model where each individual is assumed to have originated in a single population,

and as an additional parameter assumed the allele frequencies of different populations to be correlated. The graphical output of Structure was modified by the use of the *Distruct* software (Rosenberg, 2003) available at http://www.cmb.usc.edu/~noahr/district.html.

Statistical Analyses

The genetic structure of the populations and basic parameters of molecular diversity, including analyses of molecular variance (AMOVA) (Excoffier *et al.* 1992), matrix of co-ancestry coefficients (Reynolds *et al.* 1983), and Hardy-Weinberg proportions by the exact test (Guo & Thomson, 1992), were calculated using the package *Arlequin* 2.0 (Schneider *et al.* 2000) with 100,000 steps in the Markov chain. The statistical significance of F_{st} values was estimated by permutation analysis using 100,000 permutations. Multidimensional Scaling (Kruskal & Wish, 1978) was performed with the program *Statistica* for Windows, release 4.0. The Mantel test for matrix correlation was performed with a program written by Dr. Jeffrey Long of the University of Michigan Medical School and made available to us by Dr. Keith Hunley from the University of New Mexico.

Results and Discussion

Biallelic Short Insertion-Deletion Polymorphisms (Indels)

The biallelic short indels chosen for this study, together with some of their properties, are shown in Supplementary Table 1. We tested the Hardy-Weinberg equilibrium in all 52 populations for all 40 loci by the exact method of Guo & Thompson (1992). Among the 2080 values thus obtained 94 (0.045) were significant at the 0.05 level. Therefore, the number of significant departures was less than expected on the basis of chance alone.

Analysis of Molecular Variance

We typed the 40 indels in the full HGDP-CEPH Diversity Panel, composed of 1,064 individuals from 52 different populations distributed in seven geographical regions: Europe, the Middle East, Central Asia, East Asia, Oceania, the Americas and Sub-Saharan Africa. The genotypes were then submitted to an analysis of molecular variance (AMOVA) using the *Arlequin* program (Schneider *et al.* 2000). The results of the analysis are shown in Table 1, in comparison with those of Rosenberg *et al.* (2002). If we focus on each region the results in the two studies are almost identical, the within-population component being responsible for more than 93% of the genetic variance, except for Amerindians who exhibited 11.6% of among-population-within-region variance and a corresponding lower within-population constituent. This result is not unexpected, since it is well known that the demography of Amerindians, especially in South America, has occasioned very high degrees of genetic drift that produce elevated levels of between-population gene frequency variation (Cavalli-Sforza *et al.* 1994). On the other hand, when we examined the data for the seven geographical regions our indel analysis showed a much larger among-regions component of variation (12.1%) compared with the 3.6% observed in the microsatellite study of Rosenberg *et al.* (2002), who had attributed their low value to the sampling scheme of the HGDP-CEPH Diversity Panel. Excoffier & Hamilton (2003) have already observed that the level of among-regions variance observed by Rosenberg *et al.* (2002) was smaller than other worldwide studies, and attributed this to the fact that the authors had not used a stepwise mutation model, the most appropriate model for microsatellite studies. Indeed, not taking homoplasy into account can depress the among-regions variance component (Flint *et al.* 1999; Romualdi *et al.* 2002). If one associates the relatively high mutation rate of microsatellites (Leopoldino & Pena, 2003) with the possibility of size constraints for their growth, different populations would tend to approach a common allelic distribution for these markers (Romualdi *et al.* 2002). The short biallelic indel markers that we employed are expected to have a much lower evolution rate, and thus their distribution is expected to reflect deeper events in the demographic history of populations than would of microsatellites (Romualdi *et al.* 2002).

Our results are compatible with those obtained by two worldwide studies more directly comparable with ours: one by Romualdi *et al.* (2002) with 21 *Alu* insertion-deletion polymorphisms in 32 populations

Table 1 Analysis of molecular variance (AMOVA) of the typing results with 40 short insertion-deletion polymorphisms in comparison with the results of Rosenberg *et al.* (2002). Data were analyzed with the Arlequin ver. 2.000 software (Schneider *et al.* 2000)

			40 Indels (this study)			377 Microsatellites (Rosenberg *et al.* 2002)		
			Variance components (%)			Variance components (%)		
Sample	Number of regions	Number of populations	Within populations	Among populations within regions	Among regions	Within populations	Among populations within regions	Among regions
World	1	52	87.2	12.8		94.6	5.4	
World	7	52	85.7	2.3	12.1	94.1	2.4	3.6
Africa	1	7	95.3	4.7		96.9	3.1	
Eurasia	1	21	97.7	2.3		98.5	1.5	
Eurasia	3	21	97.4	1.4	1.2	98.3	1.2	0.5
Europe	1	8	99.0	1.0		99.3	0.7	
Middle East	1	4	98.5	1.5		98.7	1.3	
Central/South Asia	1	9	98.3	1.7		98.6	1.4	
East Asia	1	18	98.7	1.4		98.7	1.3	
Oceania	1	2	93.9	6.1		93.6	6.4	
America	1	5	88.5	11.5		88.4	11.6	

and on other by Watkins *et al.* (2003) with 100 *Alu* polymorphisms in 31 populations. The former obtained the following components of genetic variance: 82.9% within-populations, 8.2% among-populations-within-region and 8.9% among-regions. The latter observed that 88.6% of the genetic variance occurred within-populations, 1.9% among-populations-within-regions, and 9.6% among-regions (Watkins *et al.* 2003). Our results are also very similar to those of Bowcock *et al.* (1991) for 100 diallelic DNA polymorphisms (SNPs) tested in five populations from four continents.

As pointed out by Bowcock *et al.* (1991) the disparity of the among-populations component (F_{ST}) from one polymorphism to another may help to establish whether natural selection is playing a role or whether variation is selectively neutral. In the latter case the only force at play is drift, which we expect to be equal for all genes, since it depends only on demographic properties of the populations and not on the particular genes being studied. To investigate we plotted the observed F_{ST} values for the 100 polymorphisms against the corresponding mean gene frequency (data not shown). We then compared the number of F_{ST} values expected in the various percentile classes, calculated according to Bowcock *et al.* (1991), with the number observed. Thirty-eight out of 40 plotted points (95%) were located between the 5th percentile and the 95th percentile of the simulated F_{ST} distributions for different initial gene frequencies. In other words, at this level of resolution there was no evidence of deviation from neutrality.

Multidimensional Scaling

We tested the discrimination power of our 40-indel set by obtaining a distance matrix of the 52 populations using the Reynolds genetic measure, which is based on the F_{ST} linearized for short divergence times (Reynolds *et al.* 1983). From the matrix we undertook a Multidimensional Scaling analysis (MDS; Kruskal & Wish, 1978) using the program *Statistica*. With only two dimensions we obtained a very adequate graphical representation of the distance matrix (stress = 0.108), as shown in Figure 1. It is immediately apparent that the points corresponding to the 52 populations aggregate into five widely separated clusters that correspond to Africa, Oceania, East Asia, America and a central Europe-Middle East-Central Asia group (E-ME-CA). It is interesting to observe that the two most distant clusters are Africa and America, the exact two anchor clusters produced by the Structure program when K = 2 (see below). The E-ME-CA cluster can be separated into three population groups using prior geographical information, and then producing the separation into seven major geographical regions. Among the 52 populations, the only one misclassified according to geographical region was the Kalash population of Central Asia (open arrow).

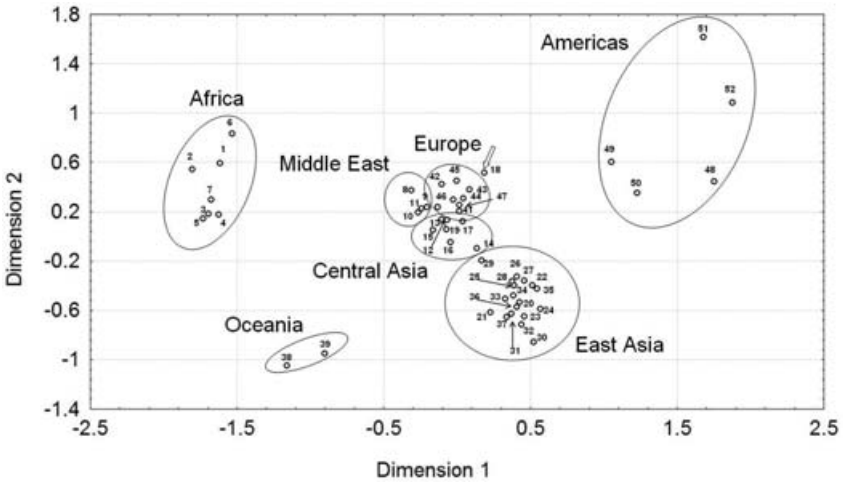

Figure 1 Multidimensional scaling plot obtained with the program *Statistica*, ver. 4.0. Each point represents one population, numbered as follows: (1) Biaka_Pygmies, (2) Mbuti_Pygmies, (3) Mandenka, (4) Yoruba, (5) Bantu_NE, (6) San, (7) Bantu_SE/SW, (8) Mozabite, (9) Bedouin, (10) Druze, (11) Palestian, (12) Brahui, (13) Balochi, (14) Hazara, (15) Makrani, (16) Sindhi, (17) Pathan, (18) Kalash, (19) Burusho, (20) Han, (21) Tujia, (22) Yizu, (23) Miaozu, (24) Oroqen, (25) Daur, (26) Mongola, (27) Hezhen, (28) Xibo, (29) Uygur, (30) Dai, (31) Lahu, (32) She, (33) Naxi, (34) Tu, (35) Yakut, (36) Japanese, (37) Cambodian, (38) Papuan, (39) NAN_Malesian, (40) French, (41) French_Basque, (42) Sardinian, (43) North_Italian, (44) Tuscan, (45) Orcadian, (46) Adygei, (47) Russian, (48) Pima, (49) Maya, (50) Colombian, (51) Karitiana, (52) Surui. The double arrow indicates the Kalash population which belongs to Central Asia and is apparently misclassified (see text). Details about the populations can be obtained at http://www.cephb.fr/HGDP-CEPH-Panel/.

Zhivotovsky *et al*. (2003) used the data from Rosenberg *et al*. (2002) to produce an MDS plot with a topography similar to ours, including the "anomalous" position of the Kalash. However, in accordance with the AMOVA results the major geographical regions appeared to be more separated in our MDS plot.

Cluster Analysis with the Structure Programs

Finally, we analyzed the indel data with the *Structure* program (Pritchard *et al.* 2000). We made ten runs each, with K varying from 1 to 10, with a burn in of 100,000 and run length of 2,000,000. All runs produced the same clusters except for those at $K \geq 7$. The results with K varying from 2 to 6 are shown in Fig. 2. At $K = 2$ the data were, as seen by Rosenberg *et al.* (2002), anchored by Africa and America, separated by a relatively large genetic distance, with East Asia very close to America and the Europe-Middle East-Central Asia block close to Africa. Excoffier (2003) pointed out that this division observed by Rosenberg *et al.* (2002) was at odds with previous results, in which a first split had often been often observed between sub-Saharan Africans and Non-Africans. Now our data with indels confirm the same result. At $K = 3$ we observe clusters of Africa, a Europe-Middle East-Central Asia-Oceania block and East Asia-America. At $K = 4$ East Asia and America separate, and at $K = 5$ we get groups that correspond to five major geographical regions (with Europe, Middle East and Central Asia clustered). With $K = 6$ the situation continues more or less unchanged with no more significant splits. The value of $K = 5$ has the highest posterior probability (0.9999).

Turakulov & Easteal (2003) computed that 65 random biallelic polymorphisms (SNPs) would be necessary for identifying distinct geographically separated populations, while Bamshad *et al.* (2003) calculated that 60 *Alu* indels would be sufficient to obtain assignment to the continent of origin with an accuracy of 90%. Our study demonstrates that a set of 40 well-chosen biallelic markers is sufficient to characterize human population structure at the global level.

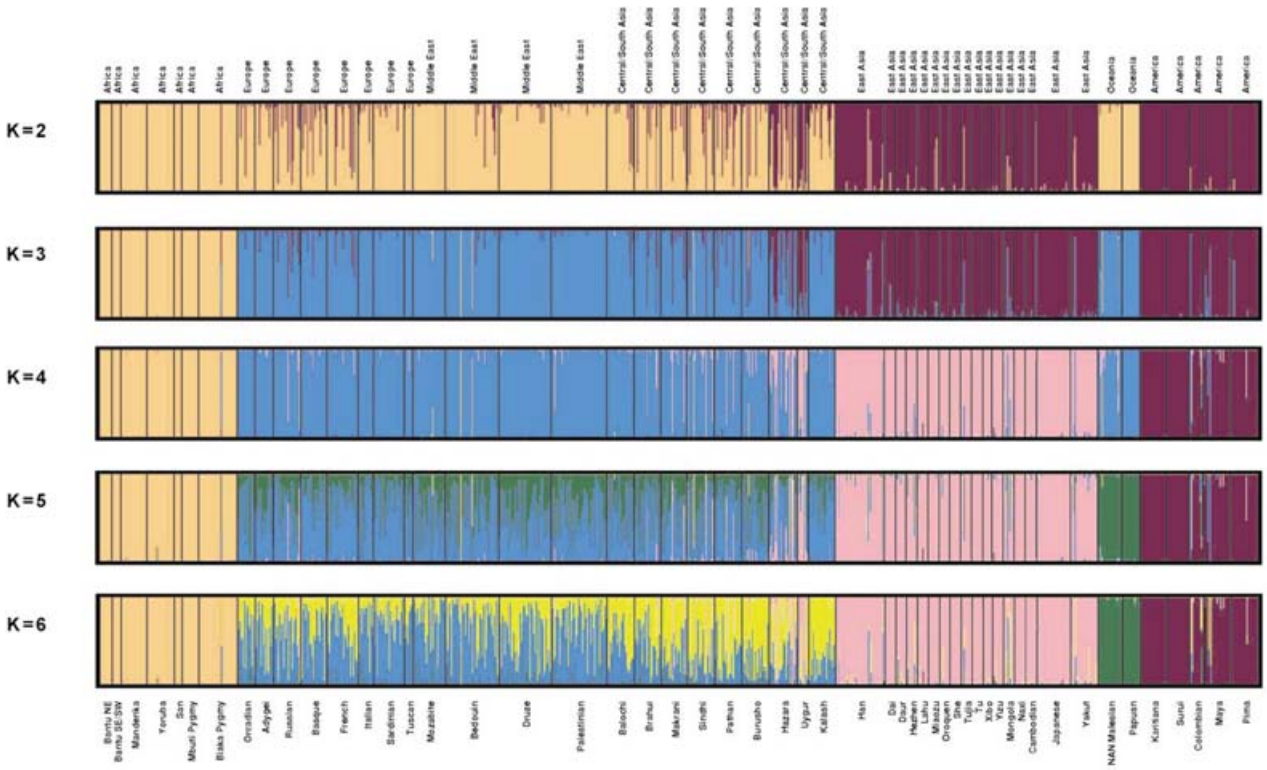

Figure 2 Estimated population structure of 52 human populations studied with 40 diallelic short insertion-deletion polymorphisms. Each of the five horizontal bars is composed of thin vertical lines representing all 1064 individuals. The lines are coloured dependent on the individual's estimated membership fractions and divided into K clusters; the value of K is stated on the left. Vertical black lines separate the individuals into 52 different populations, identified by the labels on the bottom. Ten *Structure* runs were performed for each K with a burn-in of 100,000 runs and run length of 2,000,000. The graph was prepared with the *Distruct* software. Details about the populations can be obtained at http://www.cephb.fr/HGDP-CEPH-Panel/.

Verification of Possible Biases

As explained above, one of the criteria for the choice of these specific indels was allele frequency close to 0.5 in the European population. For the chosen loci the average frequency of the long allele in Europeans, as determined by Weber *et al.* (2002), was 0.51 (0.42 in Amerindians, 0.50 in Japanese and 0.61 in Africans). The fact that the indels had been chosen for their high variability in Europeans will lead to biases in the comparison of allele frequencies and gene diversity among the various regions. However, this should not inevitably affect the partition of genetic variance. Indeed, our within-population and among-population-within-region components of the total genetic variance are practically identical to those of Rosenberg *et al.* (2002). We checked this further using three approaches. First, we studied the distribution of worldwide Fst values. Kidd *et al.* (2004), studying 369 biallelic markers, observed that the distribution of worldwide Fst values had a mean of 0.138 and was skewed to the right (skewness = 1.082). They attributed the skewness to the fact that the markers had been ascertained by being shown to be polymorphic with moderate to high heterozygosities in a non-African population. The Fst distribution for the 40 indels had a mean of 0.141, which is very similar to the value obtained by Kidd *et al.* (2004). Moreover, it was not significantly different from normality and had a skewness of only 0.669, not significantly different from zero. Thus we could not reveal evidence for a bias. We then compared our distance matrix of the 52 populations using the Reynolds genetic measure (Reynolds *et al.* 1983) with a distance matrix calculated from the data of Rosenberg *et al.* (2002), which can be considered essentially free of ascertainment bias. The test revealed a highly significant correlation of 0.48 ($p < 0.0001$). The fact that the correlation was not perfect may be related to the fact that, as already discussed, the indels are expected

to have a smaller mutation rate than microsatellites and that they may more sensitive to deeper events in the human genealogical tree. As a final test, following Urbanek *et al.* (1996), we chose the 11 loci among our indels with highest gene diversity in Europe (set E) and the 11 loci with highest diversity in Africa (set A). We then performed separate worldwide AMOVA analyses with the two sets, obtaining virtually identical results, as follows: for set E and set A, respectively, the within-population components of variance were 86.35% and 88.38% respectively, the among-groups components were 10.84% and 9.05%, and the among-populations-within-groups components were 2.82% and 2.56%. We conclude that in our study the ascertainment bias apparently did not significantly affect the partition of variance.

Conclusions

In summary, we have studied the same worldwide population sample as Rosenberg *et al.* (2002) with a set of 40 biallelic slow-evolving short insertion-deletion polymorphisms. We found that the genetic structure of the populations included in the HGDP-CEPH Diversity Panel is best portrayed by a picture of the world divided into genetic clusters that tightly correspond to five geographic regions: America, Sub-Saharan Africa, East Asia, Oceania and a group composed of Europe, the Middle East and Central Asia. We have also shown that with our set of indels we disclose an among-regions component of genetic variance considerably larger than was estimated by Rosenberg *et al.* (2002) using microsatellites.

Population studies have suggested that genetic variation is essentially continuous throughout space among humans (Romualdi *et al.* 2002). This knowledge is apparently at odds with the regional discontinuity observed by Rosenberg *et al.* (2002) and by us. Serre & Paabo (2004) have proposed that such discontinuity might be an artifact imposed by the sample structure of the HGDP-CEPH Diversity Panel. To address this issue more directly, it will probably be necessary in the future to switch the emphasis of worldwide panels from populations to individuals (Cavalli-Sforza, 2005).

Acknowledgements

This work was partly supported by a grant from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We are grateful to Dr. Howard Cann of the Fondation Jean Dausset who provided the HGDP-CEPH Diversity Panel, and to Dr. Jeffrey Long of the University of Michigan Medical School and Dr. Keith Hunley from the University of New Mexico for making available to us software for the analysis of matrix correlation. We thank Rodrigo Richard Gomes for software development and Tales Silva for his help in running *Structure*. Neuza A. Rodrigues and Kátia Barroso provided expert technical assistance.

References

Bamshad, M. J., Wooding, S., Watkins, W. S. *et al.* (2003) Human population genetic structure and inference of group membership. *Am J Hum Genet* **72**, 578–589.

Barbujani, G. & Di Benedetto, G. (2001) Genetic variances within and between human groups. In: *Genes, Fossils and Behaviour* (eds. P. Donnelly & R. A. Foley), pp. 63–77. IOS press, Amsterdam.

Barbujani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L. L. (1997) An apportionment of human DNA diversity. *Proc Natl Acad Sci* **94**, 4516–4519.

Bowcock, A. M., Kidd, J. R., Mountain, J. L. *et al.* (1991) Drift, Admixture, and Selection in Human Evolution: A Study with DNA Polymorphisms. *Proc Natl Acad Sci* **85**, 839–843.

Cann, H. M., de Toma, C., Cazes, L. *et al.* (2002) A human genome diversity cell line panel. *Science* **296**, 261–262.

Cavalli-Sforza, L. L. (2005) The Human Genome Diversity Project: past, present and future. *Nat Rev Genet* **6**, 333–340.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1994) *The history and geography of human genes*, pp. 1–551. Princeton University Press, Princeton, NJ.

Excoffier, L. & Hamilton, G. (2003) Comment on "Genetic structure of human populations." *Science* **300**, 1877.

Excoffier, L. (2003) Human diversity: our genes tell where we live. *Curr Biol* **13**, R134–136.

Excoffier, L., Smouse, P. E. & Quattro, J. M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Flint, J., Bond, J., Rees, D. C. *et al.* (1999) Minisatellite mutational processes reduce Fst. estimates. *Hum Genet* **6**, 567–576.

Guo, S. W. & Thompson, E. A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **2**, 361–372.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. (2004) *Nature* **431**, 931–945.

Kidd, K. K., Pakstis, A. J., Speed, W. C. *et al.* (2004) Understanding human DNA sequence variation. *J Hered* **95**, 406–20.

Kruskal, J. B. & Wish, M. (1978) *Multidimensional Scaling*. SAGE Publications, New York.

Leopoldino, A. M. & Pena, S. D. (2003) The mutational spectrum of human autosomal tetranucleotide microsatellites. *Hum Mutat* **21**, 71–79.

Lewontin, R. C. (1972) The apportionment of human diversity. *Evol Biol* **6**, 381–398.

Pritchard, J. K., Stephens, M. & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Reynolds, J., Weir, B. S. & Cockerham, C. C. (1983) Estimation of the co-ancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767–779.

Romualdi, C., Balding, D., Nasidze, I. S. *et al*. (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* **12**, 602–612.

Rosenberg, N. A. (2004) distruct: a program for the graphical display of population structure. *Mol Ecol* **4**, 137–138.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L. *et al*. (2002) Genetic structure of human populations. *Science* **298**, 2381–2385.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L. *et al*. (2003) Response to comment on "Genetic structure of human populations". *Science* **300**, 1877.

Schneider, S., Roessli, D. & Excoffier, L. (2000) *Arlequin ver 2.000: A software for population genetics data analysis*. Genetics and Biometry Laboratory, University of Geneva, Switzerland.

Serre, D. & Paabo, S. (2004) Evidence for gradients of human genetic diversity within and among continents. *Genome Res* **14**, 1679–1685.

Templeton, A. R. (1999) Human races: A genetic and evolutionary perspective. *Am Anthropol* **100**, 632–650.

Turakulov, R. & Easteal, S. (2003) Number of SNPs loci needed to detect population structure. *Hum Hered* **55**, 37–45.

Urbanek, M.-G., Goldman, D. & Long, J. C. (1996) The apportionment of dinucleotide diversity in Native Americans and Europeans: a new approach to measuring gene identity reveals asymmetric patterns of divergence. *Mol Biol Evol* **13**, 943–953.

Watkins, W. S., Rogers, A. R., Ostler, C. T. *et al*. (2003) Genetic variation among world populations: Inferences from 100 *Alu* insertion polymorphisms. *Genome Res* **13**, 1607–1618.

Weber, J. L., David, D., Heil, J. *et al*. (2002) Human diallelic insertion/deletion polymorphisms. *Am J Hum Genet* **71**, 854–862.

Zhivotovsky, L. A., Rosenberg, N. A. & Feldman, M. W. (2003) Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *Am J Hum Genet* **72**, 1171–1186.

Supplementary Material

The following supplementary material is available for this article online:

Table S1. Short diallelic insertion-deletion polymorphisms used in this study (Weber *et al*. 2002).

Table S2. Amplification primers for the diallelic insertion-deletion polymorphisms used in this study (Weber *et al*. 2002).

Web Site References

http://pritch.bsd.uchicago.edu/software.html; *Structure* 2.1 download

http://research.marshfieldclinic.org/genetics/indels/default.asp; Human Insertion/Deletion Polymorphisms

http://www.cephb.fr/HGDP-CEPH-Panel/; HGDP-CEPH Human Genome Diversity Cell Line Panel

http://www.cmb.usc.edu/∼noahr/distruct.html; Software *Distruct* download

Received: 13 October 2005
Accepted: 7 December 2005