

Proceedings from
The Fourth International Symposium on Human Identification
1993

Copyright 1993 Promega Corporation
ISBN 1-882274-53-9

Global Levels of DNA Variation

Judith R. Kidd, Andrew J. Pakstis, and Kenneth K. Kidd

Department of Genetics, Yale University School of Medicine, New Haven, CT



INTRODUCTION

Other than studies of VNTR loci conducted by the forensic community (e.g., FBI Lab Division, 1993), there have been very few studies of DNA polymorphisms on a global scale. The “100 markers” study in which we participated (Bowcock *et al.*, 1987, 1991a, 1991b) was the first to look at a large number of DNA polymorphisms on a set of populations drawn from Europe, Africa, Eastern Asia, and Oceania. We learned much from that study and it has served as the core for several expansions. One recent expansion has involved a subset of 80 markers on several new populations giving more thorough “coverage” of Asia and Oceania (Mountain *et al.*, 1992). We present here some of the first results of a different study including the original five populations, plus seven others (Table 1), including for the first time five populations in the New World. We have used a set of 35 markers, mostly two-allele systems, that partly overlaps with the original 100 markers and have typed them on these twelve populations. Some of the data on three of the Amerindian populations in this study have previously been published (Kidd *et al.*, 1991). That study has received considerable attention in forensic circles and we shall return to it at the end of this paper.

CHOICE OF MARKERS

A population genetics study with objectives focused on population histories and demography cannot rely solely on the VNTR loci used in the forensic community. The reason is basically quite simple: those VNTR loci have very high mutation rates. High mutation rates obscure questions of identity by descent, even when PCR techniques allow precise identification of alleles by repeat number. Of course, since typing by Southern blot analysis does not usually allow precise resolution of VNTR alleles with nearly identical numbers of repeat units, data generated in that way are not appropriate for the types of studies in which we are interested. Harding (1992) has reviewed the use of VNTR polymorphisms in anthropological studies and has pointed out some potentially useful applications but has also noted the problems for studies of population relationships posed by the large number of alleles generated by the high mutation rate. Interestingly, it is the very large number of alleles generated by the high mutation rates that make VNTR loci so appropriate for questions of individual identification, even if allelic resolution is technically difficult.

The markers we are studying are mostly restriction site polymorphisms with two alleles. Considerable evidence is accumulating that such polymorphisms are not subject to any meaningful level of recurrent mutation within a species and that frequency differences among populations are purely a function of

random genetic drift. This allows us to use population genetics theory of a reasonably simple sort to study population similarities and the history of human expansion around the globe. In this paper we present some of the initial analyses of this data set that we believe are of relevance to the forensic community.

GLOBAL VARIATION IN ALLELE FREQUENCIES

One of the analyses of the “100 markers” data (Bowcock *et al.*, 1991a) showed considerable difference among the markers in the amount of variation among populations. In other words, when variation among populations was measured as F_{st} , some markers showed very high levels of variation and other markers showed virtually no variation. Theoretically, F_{st} is a measure of the genetic structure of the populations being studied. Knowing the structure of the populations sampled, one can predict the expected F_{st} (the variation in allele frequencies one would expect, on average) and the variation in F_{st} estimates among loci. However, when the 100 markers were compared to the expected variation of F_{st} among loci, based on certain assumptions (see Bowcock *et al.*, 1991a), some of them showed more variation than expected (i.e. high F_{st} , while others showed less variation than expected. Future studies may resolve whether this finding indicates selection operating at some loci or just an underestimate of the expected variation among loci. In either case, the amount of variation shown by a locus can be considered a “characteristic” of the locus, and deviation from the “expected” value is a measure of how well that locus estimates the structure of the populations. Conversely, the genetic data can be used to compare the average F_{st} value for populations in different combinations to illuminate hierarchical structures – this is the basis of many genetic distance measures.

That initial “100 markers” study (Bowcock *et al.*, 1991a) had only five populations and no New World populations in it. When we now repeat F_{st} calculations across the twelve populations in the current study, we continue to see that different markers show different levels of frequency variation around the globe. This is of particular interest to us because fine-scale studies of population similarities and histories require markers that show a wide range of variation. Interestingly, this is almost the opposite of what the forensic community would like for statistical evaluation of individual identity questions since polymorphic markers that have the same distribution of allele frequencies in all populations would allow calculations that were independent of geography and ethnic ancestry of the relevant individuals.

Table 2 summarizes our analyses for 35 individual marker systems defined by probe and enzyme. The table includes the

average allele frequency, the range of allele frequencies, the average heterozygosity, and the F_{st} value for each polymorphic system (most of which are two-allele systems). For heterozygosity we have calculated for each population the expected proportion of heterozygotes assuming Hardy-Weinberg proportions and using the allele frequencies for that population; the average heterozygosity is the simple mean of the heterozygosities for the twelve populations. Figure 1 is a scatter plot in which each of the 35 polymorphic systems is plotted according to two of the values in Table 2: average heterozygosity and F_{st} . For population studies, we are primarily interested in those markers that fall within the high F_{st} range irrespective of their average heterozygosity. Forensic scientists, on the other hand, would be primarily interested in those markers that fall in the low F_{st} , high heterozygosity quadrant since these would give the highest discrimination among individuals with the lowest dependence on geography and ethnic origins. Obviously, the trend in forensics is for markers with greater discrimination at each locus, in part for reasons, such as mixed samples, unrelated to population genetics. However, a large number of two-allele loci of the sort just mentioned would give substantial information on a "pure" forensic sample. Exactly that concept is embodied in the AmpliType Polymarker PCR Amplification and Typing Kit marketed by Perkin-Elmer-Cetus.

ALLELES NOT SEEN IN SOME POPULATIONS

Across the 35 marker systems we have identified a total of 79 alleles. Nearly all of the alleles have been observed in all populations. All alleles at 17 of the 35 probe/enzyme systems are present in all populations. At 18 probe/enzyme combinations an allele was not observed in one or more populations. We have arbitrarily divided the alleles that are not ubiquitous into two categories: those seen in several populations and those reaching at least a 10% frequency in one population (Table 3) and those seen at frequencies less than 10% in only one or two populations (Table 4).

Table 3 shows the qualitative distribution of the few alleles that are not seen in some populations. The larger allele (14kb fragment) in the ApoB locus has not been observed in the Karitiana, Surui, Ticuna, nor Melanesians. The 9.2kb allele in the D2S91 locus (the only 5 allele locus in this set) has not been observed in the Karitiana, Surui, Ticuna, Maya, Melanesians, nor Europeans; the 9.0kb allele in the same locus has not been observed in the Surui nor Cambodians; the 7.5kb allele in that locus has only been observed in the Cambodians and the two African Pygmy groups. The 5.8kb allele in the D4S35 locus has not been observed in the Ticuna nor Chinese. The absence of the FNRBL insertion (cf. Giuffra *et al.*, 1990, for a description of this locus) has been observed in all populations except the Cambodians. The 6.2kb allele in the REN locus has not been observed in the Karitiana, Surui, Ticuna, nor Melanesians. The 6.4kb allele in the SST locus has not been observed in the Karitiana. The 15kb allele of the second Hind III system in the D4S10 locus has not been observed in the Karitiana, Surui, Ticuna, Japanese, Cambodians, nor Melanesians. The 4.5kb allele in the HLA-F locus has not been observed in the

Melanesians. The 2.4kb allele in the HLA-E locus has not been observed in the Surui, nor Melanesians. The 4.3kb allele of the *Bgl* II system in the RBP3 locus has not been observed in the Karitiana, Surui, Ticuna, Maya, nor Melanesians the 2.4kb allele of the *Msp* I system in the same locus is observed only in the Central African Republic Pygmies. The 2.4kb allele of the OS2/*Taq* I system in the D10S20 locus is observed only in the Maya and Europeans. The 7.6kb allele in the p9D11/*Taq* I system in the D13S2 locus is not observed in the Chinese, Cambodians, Melanesians, nor in either African Pygmy group. And, finally, the 1.1kb allele of the BS3/*Sac* I system in the HOXB locus was not seen in the Chinese.

It may seem as though there are many alleles that have been "lost" in the populations of the New World. (Obviously, "lost" is used here as a shorthand for "not seen in our sample; a low frequency cannot be excluded as a possibility.") However, inspection shows that of the 72 alleles observed in most populations, none is "lost" in all populations of the New World. There are, however, three cases of an allele being very rare in the New World sample that deserve comment. One allele at the five-allele 9.2kb locus – the 9.2kb allele – has been observed at only a 1% frequency in only one population of the New World, the Jemez Pueblo. If the presence of this allele is an indication of past gene flow into this population, it would be difficult to specify the source population (on the basis of those reported in this study) as that allele has not been observed in Europeans and has estimated frequencies of only 4-5 % in the two African Pygmy groups; its frequency is estimated at 22% in Japanese, 16% in Cambodians, and 8% in Chinese. Conversely, there is an allele, the 2.4kb allele, at the OS2/*Taq* I system at the D10S20 locus observed in only two populations of the twelve studied: 1% in the Maya and 10% in the Europeans. The single observation of this allele in Maya is, in this case, a possible indication of some European contribution to the ancestry of this sample of Maya. Likewise, there is a single observation of the 6.2kb allele at the OS3/*Taq* I system of the same locus in both Maya and Jemez Pueblo. This allele is observed elsewhere only in Europeans and at a frequency of 12%. This could also indicate some gene flow from Europeans.

Six alleles have not been observed in any of the three populations of South America. Three of these, at D2S91 and D10S20 (two alleles), were discussed above with respect to admixture. In addition, three others were not seen: the 14kb allele of ApoB, the 15kb allele of the D4S10/*Hind* IIIb system, and the 6.2kb allele of REN. The 14kb ApoB allele is uncommon to rare in all other populations studied, not reaching an estimated frequency higher than the 16% seen in the Chinese. The 15kb allele of D4S10/*Hind* IIIb reaches estimated frequencies of 28% (Central African Republic Pygmies) and 37% (Zaire Pygmies) in Africa, 17% in Europeans, but was present at only 1% in Chinese and not seen at all in Japanese, Cambodians, nor Melanesians – it is also infrequent in North America (2% in Jemez Pueblo and 7% in Maya). The 6.2kb allele at the REN locus – not observed in South America – ranges in estimated frequency in the Old World from 32% to 55%, and in North America from 16% in the Maya to 28% in the Jemez Pueblo. A seventh allele, the RBP3/*Bgl* II 4.3kb allele, is observed rarely in the New World (less than 1% in the Karitiana and 5% in the

Jemez Pueblo), and ranges in frequency in the Old World from 4% in the Zaire Pygmies to 26% in the Japanese (it was not observed among the Melanesians).

PRIVATE POLYMORPHISMS

Of the alleles observed in this study, very few are *private* in the sense that they have been observed in only a few people worldwide. Table 4 summarizes the findings. The 7.5kb allele of the OS5/*Msp* I system at D2S91 was observed only in low frequencies in the Cambodians and Central African Republic Pygmies. The 2.2kb allele of the pKO82/*Pst* I system in the D4S10 locus was observed only in the Ticuna. The 17.0kb allele at the D13S2/*Msp* I system is observed only in the Central African Republic Pygmies. The 3.0kb allele of the H20-1-3/*Msp* I system of the HOX2F locus was observed only in the two African Pygmy groups. Finally, the 2.4kb allele of the H4/*Msp* I system in the RBP3 locus was observed only in the Central African Republic Pygmies.

Given the low frequencies of these alleles and their occurrence in only one or two populations, they contribute no significant information to the relationships among these particular populations. We also doubt they would have any forensic value. They may be useful in studies of other populations closely related to those in which they are seen since quite probably they represent “new” mutations that have arisen since *H. sapiens* spread around the world. One possible exception to that is the 7.5kb allele in the D2S91 locus which is seen in Cambodia and Africa. If this is the identical allele (and not two similar alleles indistinguishable by Southern blotting), then there may have been two independent mutation events or this was a rare allele in early *H. sapiens* and has been lost from (or is very rare and not detected in) most populations and only remains (or has been detected) in these two widely separated populations. Of course, more recent gene flow in either direction between Cambodia and Africa cannot be ruled out but seems unlikely given the “equal” and very low allele frequencies in both populations. Only considerable additional study could clarify this uncertainty.

POPULATION-SPECIFIC ALLELES

The forensic community would benefit greatly if it were possible to identify the ethnic origins of a human sample. Our data (summarized in Tables 3 and 4) suggest that individual markers that identify specific ethnic groups will be rare and difficult to apply in forensic cases. Consider, for example, the “best case” in our existing data: the 2.4kb allele defined by OS2/*Taq* I at the D10S20 locus. Globally we have seen this allele only in Europeans (with an allele frequency of 10%) and one Mayan individual. That Mayan individual may have a remote ancestor who was European but an independent mutational event cannot be ruled out. (Unpublished haplotype data at this locus support the gene flow explanation.) Assuming this allele was found in the DNA from skeletal remains, it would be likely that the individual had a European ancestor (or one from a closely related non-European population such as those in the Middle East). In a country such as Japan this logic might be

meaningful but that is hardly helpful in the U.S. where the African American, Hispanic, Amerindian, and even Asian populations all have significant levels of European ancestry. Of course, failure to find the 2.4kb allele in a sample carries essentially no information since most people around the world (including most Europeans) lack this allele. This particular marker will be of great interest for, and possibly valuable in, studies of relationships among European populations and those in adjoining parts of the world. While this population-specific allele is not in itself forensically valuable, it might become forensically useful if it could be combined with several others with similar or complementary patterns. As more different markers are studied, more examples of this rare type of marker may be found.

AVERAGE HETEROZYGOSITY OF POPULATIONS

Europeans have the highest mean heterozygosity of any of the twelve populations. We believe this is, in part, a bias introduced because most of these markers were first identified as polymorphisms in samples of Europeans and those with higher heterozygosities were the ones that were published. Unfortunately, it is impossible to identify the magnitude of this bias. Table 5 shows mean heterozygosity data both for each population and combined into regional averages to facilitate regional comparisons. Average heterozygosity of Europeans is, in any case, not much greater than that seen in North America, Asia, and Africa. Only two regions show much decrease in mean heterozygosity: South America and Melanesia. Even for those populations, all small linguistically distinct populations, the maximum decrease from the mean heterozygosity in Europe is 30%. Alternatively, as a way of circumventing the European bias, the maximum decrease can be calculated relative to 0.326 which is the average heterozygosity of natives of North America, Asia, and Africa (i.e. average heterozygosity of 7 population samples including Maya, J. Pueblo, Chinese, Japanese, Cambodians, C.A.R. Pygmies, and Zaire Pygmies). Calculated in this way, the maximum reduction in mean heterozygosity is 22%, shown by the Karitiana. However relative heterozygosity is measured, even extremely isolated human populations retain most of the heterozygosity found in the largest populations.

COMMENTS ON AMERINDIANS

Analyses of gene frequency data of classical markers have long been known to show large amounts of variation among New World populations (Salzano and Callegari-Jacques, 1988) while there is much less variation among populations in Europe and among populations in Eastern Asia. A very comprehensive analysis of a large compendium of such data has recently been completed by Cavalli-Sforza *et al.* (1993). Their analysis shows that for a given geographic distance between two populations, New World populations show a higher F_{st} than populations of any other continent.

In the late 1980's we had the opportunity to obtain samples from two Amazon basin populations living in Rondonia province of Brazil. One of these was particularly interesting to us because it is one of the smallest, most inbred, linguistically isolated groups known: the Karitiana. The Karitiana are a very small, probably expanding, tribal group of people living near the Madeira River of southwestern Brazil's Rondonia Province. The entire population lives in a single village and numbers less than 150 people (F.L. Black, personal communication; Professor Black, Dept. of Epidemiology and Public Health at Yale University visited this group in 1987 as part of a larger study of native Brazilians). They may be the only speakers of Ariken, an offshoot of the Tupi linguistic stock. Endogamy prevails among the Karitiana and most marriages are consanguineous in several ways. Because of their linguistic distinctiveness, the general isolation of Amerindian groups and Amazon groups in particular, and the documented family structure that shows this group is really just one family, the Karitiana should represent an extreme example of random genetic drift in humans. In 1991 we published a study that included several of the VNTR loci used by forensic laboratories as a minor part of the data (Kidd *et al.*, 1991). Mention of those data in court testimony by KKK prior to publication, the requirement by the court that the data be given to the Defense attorneys, and KKK's technically erroneous statement ("no two Karitiana have the same genotype") (Chakraborty and Kidd, 1991), have propelled the Karitiana into the forefront of the U.S. legal debate on DNA typing. Unfortunately, this prominence is the result of what we consider to be misrepresentation and misuse of the data by defense witnesses. Of course there are allele frequency differences between the Karitiana and other Amazon populations; that could have been predicted from hundreds of studies of classical markers and the known high levels of variation among Amerindian groups, especially Amazonian populations, as discussed above. Such allele frequency variation is irrelevant to questions of substructure in large urban areas populated largely by descendants of Europeans, i.e., in most of the U.S.A. Moreover, the data on the Karitiana do not constitute a forensic database; they were collected for a different purpose using procedures appropriate for that purpose but not for forensics.

What about the Karitiana samples with identical typings? Yes, they exist if a wide enough matching window is used. Though we suspect that small size differences really do exist for alleles at some loci, the experiments to confirm this are not important to our research and have not been done. If we apply an appropriate match window to the Karitiana dataset on VNTR loci that was released on discovery in *United States v. Yee* (1991), we find that there are THREE pairs of individuals with "identical" typings at six loci. A composite of the Karitiana kindred has been assembled by Prof. F. L. Black from Fundacao Nacional do Indio family lists (1985) and a genealogy collected by Mr. Alan Vogel and was made available to us by Prof. F. L. Black. This is a socially defined pedigree and cannot be assumed to be a completely correct representation of the true biological relationships. Figure 2 shows a small segment of that pedigree and indicates our best prior estimate of which individuals correspond to the six relevant samples. One of those pairs ("C" in Figure 2) is an

obvious collection or laboratory error in which one sample was duplicated. This was discovered when we examined the typings of over two dozen other markers and found them identical for the two DNA samples. Moreover, the duplicated DNA sample that was labeled as the two individuals indicated as "C" in Figure 2 does not perfectly fit in either place in the pedigree. In one case there is an incompatibility with a sample identified as a child; in the other there are incompatibilities with one "parent" and two "sibs" (who are compatible among themselves).

The other two "matches" do represent DNA samples from different individuals with VNTR phenotypes that are a forensic "match" across six loci. How likely is that to happen by chance and is it relevant to use of DNA markers in the U.S.? We have not calculated the exact probability because it has not seemed scientifically important given the following information. First, relatively few alleles exist in the Karitiana for each of these loci. Thus, no single-locus genotype has an exceedingly small expected frequency in this population – in sharp contrast to populations in the U.S. However, the main reason the "matches" are irrelevant to the probability of an unrelated individual having the same genotype is that both pairs probably involve closely related individuals. As can be seen from the indications on the pedigree (Figure 2), one of the pairs ("A") is believed to represent two siblings who are offspring of a man and his half-brother's daughter (uncle and half-niece). The second pair ("B") is believed to be the child of one of that previous pair and another sibling of that first pair. This second pair is related, all at the same time, as uncle-niece, as half cousins twice removed, and as half second cousins – hardly unrelated! For two siblings of a non-consanguineous marriage the probability of genotypic identity at several infinitely polymorphic systems is $.25^n$ where n is the number of such unlinked loci. Where the loci have somewhat restricted variation and where the parents are consanguineous, this probability is altered greatly. Thus, in a population such as the Karitiana, which is not a large population analogous to anything in the U.S. but rather a single kindred in which parents of the current generation are usually closely related in multiple ways, it is not surprising to find siblings who are identical at several loci. Indeed, in a population as inbred and isolated as the Karitiana, it is surprising that there are not more individuals who are "identical" at several VNTR loci.

Is there any relevance of the Karitiana? The data seem relevant to us only in the way they were originally intended to be used *United States v. Yee* (1991) – no single multi-locus VNTR genotype is common in any population, even in such extreme examples of genetic drift and inbreeding as the Karitiana. Although the nature of the population was described briefly by us (Kidd *et al.*, 1991) and mentioned in court testimony by KKK, no individual who has used the Karitiana data in court testimony to attempt to discredit DNA evidence has ever seen the pedigree or asked about the specific relationships of the individuals who "match". It is a disservice to the courts when individuals with such partial knowledge of the relevant facts testify as scientific experts.

ACKNOWLEDGMENTS

This work was supported in part by Dissertation Improvement Grant (NSF DBS9214097) (to J.R.K.) from the National Science Foundation, by NSF grant (NSF DBS 9208917) (to K.K.K.), and by NIH grant MH39239 (to K.K.K.). We thank the investigators cited in Table 2 for making available the probes used in this study and those investigators cited in Table 1 who collected samples, established cell lines, and made them available to us.

REFERENCES

- Bowcock A.M., C. Bucci, J.M. Hebert, J.R. Kidd, K.K. Kidd, J.S. Friedlaender, and L.L. Cavalli-Sforza, 1987. Study of 47 DNA markers in five populations from four continents. *Gene Geography* 1:47-64.
- Bowcock, A.M., J.M. Hebert, J.L. Mountain, J.R. Kidd, J.A. Rogers, K.K. Kidd, and L.L. Cavalli-Sforza, 1991b. Study of an additional 58 DNA markers in five human populations from four continents. *Gene Geography* 5: 151-173.
- Bowcock, A.M., J.R. Kidd, J.L. Mountain, J.M. Hebert, L. Carotenuto, K.K. Kidd, and L.L. Cavalli-Sforza, 1991a. Drift, admixture and selection in human evolution: A study with DNA polymorphisms. *Proc. Natl. Acad. Sci., U.S.A.* 88: 839-843.
- Cavalli-Sforza, L.L., and A. Piazza, 1993. Human genomic diversity in Europe: A summary of recent research and prospects for the future. *European Journal of Human Genetics* 1: 3-18.
- Cavalli-Sforza, L.L., A. Piazza, and P. Menozzi, 1993. *The History and Geography of Human Genes*. Princeton University Press, Princeton NJ 1993, in press.
- Chakraborty, R. and K.K. Kidd, 1991. The utility of DNA typing in forensic work. *Science* 254: 1735-1739.
- Federal Bureau of Investigation, Laboratory Division. *VNTR Population Data: A Worldwide Study*, 5 volumes, published by the Forensic Science Research and Training Center, FBI Academy, Quantico, Virginia, 1993.
- Giuffra, L.A., P. Lichter, J. Wu, J.L. Kennedy, A.J. Pakstis, J.A. Rogers, J.R. Kidd, H. Harley, T. Jenkins, D.C. Ward, and K.K. Kidd, 1990. Genetic and physical mapping and population studies of a fibronectin receptor, β -subunit-like sequence on human chromosome 19. *Genomics* 8: 340-346.
- Harding, R.M., 1992. VNTRs in Review. *Evolutionary Anthropology* 1: 62-71.
- Kidd, J.R., F.L. Black, K.M. Weiss, I. Balazs, and K.K. Kidd, 1991. Studies of three Amerindian populations using nuclear DNA polymorphisms. *Human Biology* 63: 775-794.
- Mountain, J.L., A.A. Lin, A.M. Bowcock, and L.L. Cavalli-Sforza, 1992. Evolution of modern humans: evidence from nuclear DNA polymorphisms. *Philosophical Transactions of the Royal Society of London B.* 337: 159-165.
- Salzano, F.M., and S.M. Callegan-Jacques. *South American Indians: A Case Study of Evolution*, Clarendon Press, Oxford, 1988.
- United States v. Yee*, 134 F.R.D. 161 (N.D. Ohio 1991)

Table 1. Population Samples

Population	Average Persons Typed	Coordinates	Collected By	Cell Lines Established By
Karitiana Amazon Basin	57	10S; 63W	P. Black	J.R. Kidd
Surui Amazon Basin	48	11S; 62W	P. Black	J.R. Kidd
Ticuna Amazon Basin	67	3.5N; 68.5W	D. Lawrence	D. Wallace
Maya Yucatan	53	19N; 91W	K. Weiss	J.R. Kidd
Jemez Pueblo Southwest U.S.A	44	35.5N; 106.7W	D. Goldman	D. Goldman
Japanese	56	35.5N; 140E	L. Cavalli-Sforza & J.R. Kidd	L. Cavalli-Sforza & J.R. Kidd
Chinese	62	28N; 115E	L. Cavalli-Sforza & J.R. Kidd	L. Cavalli-Sforza & J.R. Kidd
Cambodians	25	12N; 105E	K. Dumars	L. Cavalli-Sforza
Nasioi Melanesia	23	6S; 155E	J. Friedlaender	L. Cavalli-Sforza & J.R. Kidd
Europeans	49	50N; 10E	K.K. Kidd	J.R. Kidd
Biaka Pygmy Central African Republic	68	3.4N; 18E	L. Cavalli-Sforza	J.R. Kidd & L. Cavalli-Sforza
Mbuti Pygmy Zaire	39	1.2N; 29E	L. Cavalli-Sforza	J.R. Kidd & L. Cavalli-Sforza

Table 2. F_{st} and Average Heterozygosity at 35 Genetic Systems Across 12 World Populations

Chromosome Location	Locus Symbol	Probe/Enzyme	Allele Symbol	Mean Freq.	Freq. Range	Avg. Het.	F_{st}	Source
1q32	REN	hREN/ <i>Hind</i> III	6.2	0.250	0.000-0.545	0.296	0.2123	Chirgwin
1q23-q25	AT3	AT3/ <i>Pst</i> I	7+5	0.482	0.148-0.986	0.352	0.2955	Orkin
2p24-p23	APOB	AB1/ <i>Eco</i> R I	14	0.055	0.000-0.164	0.098	0.0600	Scott
2p23-p15	D2S6	pXG-18/ <i>Taq</i> I	4.6	0.689	0.340-0.941	0.364	0.1486	Szabo
2p16-p15	D2S91	OS5/ <i>Msp</i> I	12	0.380	0.149-0.761	0.535	0.1424	Miki
2q34-q37	ALPI	pcD98#7pGEM/ <i>Pst</i> I	1.6	0.328	0.029-0.750	0.356	0.1934	Harris
3q28	SST	SST/ <i>Eco</i> RI	6.4	0.159	0.000-0.652	0.216	0.1941	Bell
4p16.3	D4S10	pKO82/ <i>Pst</i> I	5.6	0.848	0.634-0.977	0.233	0.1019	Gusella
4p16.3	D4S10	pKO82/ <i>Taq</i> Ia	3.0	0.801	0.663-0.935	0.300	0.0555	Gusella
4p16.3	D4S10	pKO82/ <i>Hind</i> IIIa	4.9	0.302	0.060-0.843	0.323	0.2361	Gusella
4p16.3	D4S10	pKO82/ <i>Hind</i> IIIb	17.5	0.923	0.632-1.000	0.112	0.2112	Gusella
4p11-q21	MT2P1	pHM6/ <i>Eco</i> RI	4.6	0.691	0.486-0.946	0.368	0.1371	Tsui
4p11-q11	D4S35	G9-20/ <i>Msp</i> I	5.5	0.896	0.606-1.000	0.163	0.1218	Gusella
6p21.3	HLA-E	pRS5.1.0/ <i>Pst</i> I	2.8	0.158	0.000-0.522	0.216	0.1889	Weissman
6p21.3	HLA-F	p5.4SH3/ <i>Hind</i> III	5.4	0.809	0.519-1.000	0.262	0.1535	Gruen
10q11.2	RBP3	H4/ <i>Bg</i> II	6.3A	0.928	0.735-1.000	0.120	0.0966	Bridges
10q11.2	RBP3	H4/ <i>Msp</i> I	3.0B	0.795	0.533-0.986	0.280	0.1424	Bridges
10q22-q26	D10S20	OS2/ <i>Taq</i> I	11	0.492	0.265-0.784	0.440	0.1354	Miki
10q22-q26	D10S20	OS2/ <i>Hind</i> III	7.4	0.508	0.265-0.789	0.437	0.1255	Miki
10q22-q26	D10S20	OS2/ <i>Taq</i> I	5.5	0.493	0.206-0.740	0.440	0.1482	Miki
10q11.1-q11.2	D10S97	pKW6dSc/ <i>Eco</i> RI	9	0.496	0.137-0.778	0.428	0.1438	Kidd
11p15.5	TH	pJ4.7/ <i>Taq</i> I	6.0	0.254	0.027-0.569	0.315	0.1683	Mallet
11p15.2-p15.1	CALCA	pTT42/ <i>Taq</i> I	8.5	0.400	0.083-0.726	0.409	0.1478	Nelkin
11q22-q23	DRD2	h2GD1/ <i>Taq</i> Ia	6.6	0.451	0.171-0.696	0.441	0.1104	Civelli
12q22-q24.2	PAH	pPAH247/ <i>Bg</i> II	0.08	0.444	0.158-0.781	0.419	0.1512	Woo
13q22	D13S2	p9D11/ <i>Msp</i> I	15	0.827	0.571-0.990	0.256	0.1061	White
13q22	D13S2	p9D11/ <i>Taq</i> I	7.6	0.132	0.000-0.378	0.189	0.1770	White
13q22	D13S2	p9D11/ <i>Pst</i> I	2.1	0.252	0.064-0.500	0.355	0.0593	White
17q21-q22	HOX2B	BS3/ <i>Sac</i> I	2.0	0.882	0.681-1.000	0.190	0.0861	Kidd
17q21-q22	HOX2F	H2-1-3/ <i>Msp</i> I	2.0	0.625	0.445-0.823	0.445	0.0555	Kidd
17q21-q22	HOX2G	H2-1-5/ <i>Taq</i> I	3.8	0.216	0.023-0.536	0.278	0.1802	Kidd
17q23.2-q25.3	TK1	pHTK9/ <i>Hind</i> III	4.5	0.364	0.023-0.754	0.335	0.2778	Kidd
19p	FNRBL	pGEM1-32/Many	PRES	0.678	0.256-1.000	0.377	0.1356	Ruoslahti
20p12	D20S5	pRI2.21/ <i>Pvu</i> II	2.7	0.171	0.022-0.331	0.268	0.0577	Shaw
20p12	D20S5	pRI2.21/ <i>Msp</i> I	3.8	0.222	0.010-0.543	0.299	0.1361	Shaw

Table 3. Alleles That Were Not Seen in Some Populations

Locus	Probe/ Enzyme	Allele in KB	Populations											
			Kari	Suru	Ticu	Maya	Jeme	Japa	Chin	Camb	Mela	Euro	CPyg	ZPyg
APOB	AB1/ <i>Eco</i> RI	14.0	0	0	0	1	1	1	1	1	0	1	1	1
D2S91	OS5/ <i>Msp</i> I	9.2	0	0	0	0	1	1	1	1	0	0	1	1
D2S91	OS5/ <i>Msp</i> I	9.0	1	0	1	1	1	1	1	0	1	1	1	1
D4S10	pKO82/ <i>Hind</i> IIIb	15.0	0	0	0	1	1	0	1	0	0	1	1	1
D4S35	G9-20/ <i>Msp</i> I	5.8	1	1	0	1	1	1	0	1	1	1	1	1
D10S20	OS2/ <i>Taq</i> I	2.4	0	0	0	1	0	0	0	0	0	1	0	0
D10S20	OS3/ <i>Taq</i> I	6.2	0	0	0	1	1	0	0	0	0	1	0	0
D13S2	p9D11/ <i>Taq</i> I	7.6	1	1	1	1	1	1	0	0	0	1	0	0
FNRBL	pGEM-32/Many	absent	1	1	1	1	1	1	1	0	1	1	1	1
HLA-E	pRS5.1.0/ <i>Pst</i> I	2.8	0	0	1	1	1	1	1	1	0	1	1	1
HLA-F	p5.4SH3/ <i>Hind</i> III	4.5	1	1	1	1	1	1	1	1	0	1	1	1
HOX2B	BS3/ <i>Sac</i> I	1.1	1	1	1	1	1	1	0	1	1	1	1	1
RBP3	H4/ <i>Bgl</i> II	4.3	1	0	0	0	1	1	1	1	0	1	1	1
REN	hREN/ <i>Eco</i> R I	6.2	0	0	0	1	1	1	1	1	0	1	1	1
SST	pgHS7/ <i>Eco</i> R I	6.4	0	1	1	1	1	1	1	1	1	1	1	1

Only those alleles generally polymorphic but not seen in some populations are tabulated. The presence of the allele in a population is indicated by a “1”; the absence of the allele in the sample examined is indicated by a “0”. For full names of the populations see Table 1.

Table 4. Private Polymorphisms

Locus	Probe/Enzyme	Allele	Observation
D2S91	OS5/ <i>Msp</i> I	7.5kb	2% Cambodians, 2% C.A.R. Pygmy
D4S10	pKO82/ <i>Pst</i> I	2.2kb	7% Ticuna
D13S2	p9D11/ <i>Msp</i> I	17.0kb	2% C.A.R Pygmy
HOX2F	H2-1-3/ <i>Msp</i> I	3.0kb	2% C.A.R. Pygmy 1% Z. Pygmy
RBP3	H4/ <i>Msp</i> I	2.4kb	2% C.A.R. Pygmy

There are only five alleles classified as private in the sense that they occur only rarely in very few populations. These are additional rare alleles at loci that are otherwise polymorphic in all or most populations. Those alleles are shown with their defining system and allele size along with their frequencies in the population(s) in which they were observed.

Table 5. Average Heterozygosity for Each Population Based on 35 Separate Genetic Marker Systems

Population	Average Het.	Standard Error	Regional Avg. Het.
Karitiana	0.253	0.032	
Surui	0.286	0.033	
Ticuna	0.261	0.031	0.267
Maya	0.337	0.025	
Jemez Pueblo	0.328	0.025	0.333
Chinese	0.324	0.030	
Cambodians	0.316	0.030	
Japanese	0.323	0.026	0.321
Nasioi Melanesians	0.272	0.032	0.272
Europeans	0.384	0.023	0.384
C.A.R. Pygmy	0.337	0.025	
Zaire Pygmy	0.320	0.028	0.329
