European Journal of Human Genetics (1998) 6, 38–49
© 1998 Stockton Press All rights reserved 1018–4813/98 $12.00

ORIGINAL PAPER

Short tandem repeat polymorphism evolution in humans

F Calafell, A Shuster, WC Speed, JR Kidd and KK Kidd

Department of Genetics, Yale University School of Medicine, USA

Forty-five dinucleotide short tandem repeat polymorphisms were typed in ten large samples of a globally distributed set of populations. Although these markers had been selected for high heterozygosity in European populations, we found them to be sufficiently informative for linkage analysis in non-Europeans. Heterozygosity, mean number of alleles, and mean number of private alleles followed a common trend: they were highest in the African samples, were somewhat lower in Europeans and East Asians, and were lowest in Amerindians. Genetic distances also reflected this pattern, and distances modelled after the stepwise mutation model yielded trees that were less in agreement with other genetic and archaeological evidence than distances based on differentiation by drift (F_{ST}). Genetic variation in non-Africans seems to be a subset of that in Africans, supporting the replacement hypothesis for the origin of modern humans.

Keywords: Short tandem repeat polymorphisms; microsatellites; human evolution; genetic distances; replacement hypothesis; multiregional hypothesis

Introduction

Short tandem repeat polymorphisms (STRPs, also known as microsatellites) present several properties that make them particularly useful in human population genetics: the huge number of STRP loci available[1] and the relative expediency with which they can be typed allow the gathering of genetic data for a large number of loci in population samples. STRPs present high heterozygosities and genetic diversities (as measured by F_{ST}, for instance) comparable to those revealed by blood group and protein polymorphisms (the so-called *classical* polymorphisms)[2,3] or slightly lower.[4] Due to their high heterozygosity, STRPs are expected to escape the allele frequency distortions observed at loci that were ascertained on the basis of polymorphism in Europeans.[5] Mutation in STRPs appears to follow a stepwise mutation model (SMM), in which mutation events involve the gain or loss of one to a few repeat units. The product of mutation is often an already existing allele; thus, mutation in STRPs violates the basic tenet of the infinite allele model (IAM), in which every mutation generates a new allele. A strict SMM, in which the gain or loss is limited to one repeat unit per mutation event, produced allele frequency distributions consistent with those observed,[6] but did not account for the observed variability in 10 STRPs in a Sardinian sample.[7] The average of the mutation rates for STRPs is several orders of magnitude higher than for single base-pair changes in single copy DNA, but there is apparently great variation in rate among loci.[8] Chakraborty *et al*[9] analyzed a large number of loci and inferred from

Correspondence: Kenneth K Kidd, Department of Genetics, Yale University School of Medicine, 333 Cedar St, New Haven, CT 06520-8005, USA

the repeat-size variance that mutation rate of dinucleotide repeats was higher than that of trinucleotide repeats which, in turn, seemed to mutate faster than tetranucleotide repeats.

Several studies have undertaken the analysis of STRP variation in global sets of populations. Bowcock *et al*[10] typed 14 populations for 30 STRPs, most of which were dinucleotide repeats that mapped on chromosomes 13 or 15. Eight of the loci were located less than 1 cM away from their nearest neighbour, and linkage disequilibrium was not assessed. Only small samples (20 chromosomes) of each population were analyzed, and their results showed that individuals tended to cluster by population in a tree based on the proportion of shared alleles. Deka *et al*[2] tested eight dinucleotide repeats in larger samples (100 to 222 chromosomes) of eight populations. Their loci mapped on chromosome 13q; they tested for linkage disequilibrium and did not find it. Reduced levels of heterozygosity were found in Pehuenche (Chile) and Dogrib (Canada) Native Americans, as well as in New Guineans. Deka *et al*[2] also observed a significant departure from the allele frequency distributions predicted by a strict SMM. Jorde *et al*[4] typed 15 populations for 30 unlinked tetranucleotide STRPs; except for 140 chromosomes of mixed Northern European origin, most populations were represented by small numbers of chromosomes (10 to 44). The STRP results were compared to hypervariable segment II mtDNA sequences and to 30 restriction site polymorphisms. They found that, although STRP data separated African populations from the rest, the African branches were shallow and, in their interpretation, not supportive of a recent African origin of modern humans. Later, Jorde *et al*[11] added 27 additional tetranucleotide, two trinucleotide, and one dinucleotide STRPs, and focused on excess African heterozygosity, observed by pooling samples within Africa. Finally, Pérez-Lezaun *et al*[3] analyzed 20 unlinked tetranucleotide STRPs in small samples (20 to 30 chromosomes each, except for some of their European samples) of 16 populations. They compared different genetic distances and found that those not based on the SMM produced trees that were in agreement with other genetic and archaeological evidence, whereas SMM-based distances generated tree topologies that were difficult to reconcile with such evidence.

In order to estimate a number of parameters of interest in population genetics, a relatively large number of loci should be typed in samples of reasonable size (eg).[12] The combined data of a large number of loci may reveal overall patterns that could be masked by the random effects of drift or by specific selection pressures in one or a few loci. Increasing sample size leads to greater statistical power in hypothesis testing, and decreases sampling error. We present data on 45 dinucleotide STRPs, mapping on chromosomes 9, 10, and 11, at an average distance of 11.3 cM (range: 5–25 cM). These markers are part of the ABI PRISM™ Linkage Mapping Sets. Their level of polymorphism and allele frequencies have previously been determined only in a sample consisting of individuals of mostly Northern European origin, with a few individuals of Native American and African ancestry.[13] We have typed samples from ten specific populations from Africa, Europe, Asia, Australo-Melanesia, and the Americas. Sample sizes, with only one exception, ranged between 80 and 128 chromosomes. The extensive analysis of intra- and interpopulation variability in a large number of loci in large samples allows us to characterize polymorphism in this particular set of markers and to evaluate their usefulness for linkage studies in populations other than Europeans. The data also illuminated the evolution of human STRPs, in terms of the generation of polymorphism and the characterization of allele distribution. They also support one of the two hypotheses on the origin of anatomically modern *Homo sapiens*, that is, the replacement model (also known as 'Out of Africa')[14] in which modern humans have a recent, African origin, and replaced archaic *Homo*, and argue against the multiregional hypothesis,[15] according to which archaic *Homo* evolved into modern humans in parallel throughout the Old World.

Materials and Methods

Population Samples

Ten populations, from diverse locations around the world, were typed. These were the Mbuti Pygmies (Zaïre: 2N = 78) and the Biaka Pygmies (Central African Republic: 2N = 138) from Africa; the Danes (2N = 124) and the Druze (Northern Israel: 2N = 122) from Europe and the Middle East; Chinese Han (2N = 98), Japanese (2N = 102), and Yakut (Siberia: 2N = 102) from East Asia; the Nasioi from Bougainville, Melanesia (2N = 46); and Maya (Yucatán, Mexico; 2N = 104) and Rondônia Surui (SW Amazonia; 2N = 94) from the Americas. A more extensive description of these samples can be found in,[16,17] and on the Web page ⟨http://info.med.yale.edu/genetics/kkidd/pops.html⟩. All of these samples exist as Epstein-Barr virus transformed lymphoblastoid cell lines, which were established under approved human subjects protocols.

Marker Typing

Forty-five dinucleotide STRPs out of the 51 loci from ABI PRISM™ Linkage Mapping Sets, Panels 13 through 16, were typed; for a list of those loci, see the World Wide Web page ⟨http://info.med.yale.edu/genetics/kkidd⟩. These markers map to chromosomes 9, 10 and 11. All loci were amplified with fluorescently-labeled primer pairs and run in an ABI 373™ DNA Sequencer following manufacturer's protocols with the following modifications: the gel concentration was increased to 8% acrylamide, the electrophoresis was decreased to 15W, and the collection time was increased to 7 hours. These changes in gel concentration and wattage increased the consistency of allele-calling between and within populations, by minimizing gel–to–gel variation. Raw data were collected with the 672 GeneScan™ software and analyzed with Genotyper™ software. The decimal base pair sizes output from the Genotyper program were translated into integer values assigned as allele names.

Statistical Analysis

Allele frequencies were estimated by direct allele counting. Expected heterozygosity was estimated as one minus the sum of squared allele frequencies.[18] Linkage disequilibrium was tested in contiguous pairs of loci by estimating pairwise haplotype frequencies with the EM-HAPLO program,[19] from which independence was tested through a log-likelihood ratio χ^2 statistic. Heterozygosity, total number of alleles, and number of private alleles were found not to follow normal distributions, and, accordingly, were compared across populations with the appropriate non-parametric tests.

Genetic distances among populations were estimated through F_{ST},[20,21] and three other genetic distances that incorporate the stepwise mutation model: D_{SW},[22] $(\delta\mu)^2$,[23] and R_{ST}.[24] Neighbor-joining trees[25] were built from those distances, and the statistical robustness of their nodes were tested through a bootstrap approach:[26,27] loci were resampled with replacement 1000 times. The LSSearch program[28] was used to discern whether any tree in the topological vicinity of the neighbor-joining tree was better by least-squares or minimum length criteria. Some of these procedures were performed with the PHYLIP 3.5c package.[29]

Results

Allele Frequency Distributions and Hardy-Weinberg Equilibrium

The allele frequencies of 45 STR loci in 10 populations can be retrieved from the World Wide Web page ⟨http://info.med.yale.edu/genetics/kkidd⟩. For each locus, we included in the Web page the genotype of one individual from the CEPH families, in order to facilitate meaningful comparisons with datasets produced in the future by other authors in different populations. Globally, 620 different alleles were detected, with a range of 6 to 29 alleles per locus (mean: 13.8). The observed numbers of heterozygous individuals for every population and locus were tested against Hardy-Weinberg expectations. Out of 450 tests, 21 (4.67%) revealed statistically significant discrepancies at a 0.05 significance level, essentially the number expected; after Bonferroni correction for multiple tests, only four cases remained significant at a nominal $\alpha = 0.05$ significance level. Three of these involved locus *D11S934* in the Danes, Druze, and Nasioi. The large deficiency of heterozygotes could be attributed to a null allele (eg a base-pair variant in the primer complementary sequences) found at varying frequencies in different populations; apparent non-parental transmission consistent with segregation of a null allele was detected in family studies of Europeans (data not shown). The fourth significant heterozygote deficiency involved *D10S591* in the Danes. Null alleles in STRs have been detected in humans[30] and in bears.[31] Failure to amplify a STRP allele could be due to differential amplification if the two allelles have extremely different sizes, or to a mutation in the primer-complementary sequences. Callen *et al*[30] demonstrated an 8-bp deletion in a primer-complementary sequence, and found null alleles in 7 out of 22 dinucleotide loci in chromosome 16. The actual incidence of null alleles is not known; some instances of apparent lack of parental transmission could be, in fact, mutation events.

Even though a mean of 962 chromosomes per locus was typed, the pooled allele distributions of 28 (62.2%) markers were not continuous, with gaps of up to four intermediate allele positions. Within every population, a mean of 31.2 loci (69.3%) presented discontinuous allele frequency distributions, with a range from 26 loci in the Japanese to 36 loci in the Surui. At six loci (namely, *D9S161*, *D9S273*, *D9S290*, *D9S164*, *D10S192*, and *D11S934*) we found alleles that differ by one base pair from a perfect two-bp ladder. At four of six loci, the off-ladder alleles were specific to African populations. One of those (*D9S164*) presented a series of eight African-specific alleles interspersed within the two-bp ladder. Complex polymorphism, beyond variation in number of repeats, appears to be a common phenomenon even in dinucleotide STRPs (see[2,32] for other examples).

Linkage Disequilibrium

In order to ascertain independence among the loci analyzed, we tested for linkage disequilibrium by estimating two-locus haplotype frequencies for all pairs of adjacent loci, separately in all ten populations. We did not test linkage disequilibrium beyond adjacent loci because the high recombination fractions between loci (eg an average of 20.9 cM for next-to-adjacent pairs of loci) make linkage disequilibrium extremely unlikely.

From the haplotype frequencies, contingency tables were analyzed through a log-likelihood ratio χ^2 statistic. The number of possible haplotypes with two STRP loci is relatively high, and many will have low expected frequencies; under these conditions, an ordinary χ^2 test would perform poorly, and even a log-likelihood ratio is likely to err on the side of rejecting valid null hypotheses. Forty-three pairs of loci were tested in each of the ten populations; after Bonferroni correction, 22 tests remained significant at a nominal $\alpha = 0.05$ significance level. Sixteen of the significant tests involved the Surui. Family relations, especially paternal relations, not recorded in ethnographic field notes, could explain this observation. However, in all other populations, alleles at each of the loci studied appear to be independent of the alleles present in neighboring loci. This is not an unexpected finding, given the recombination fractions involved (eg an average of 11.3 cM between adjacent loci) and the presumable absence of extensive selective hitchhiking effects.

Within-population Allele Diversity

The mean number of different alleles per locus found in African populations was 9.67 ± 0.33 (10.31 ± 0.48 in the Biaka and 9.02 ± 0.44 in the Mbuti; (Figure 1)), and the number of different alleles per locus was successively lower in Europeans (8.90 ± 0.33), East Asians (7.83 ± 0.25), and Native Americans (6.12 ± 0.22). Except for the Americas, these values are higher than those reported by Bowcock *et al*,[10] which were based on much smaller numbers of individuals per population; many of the same individuals were included in the larger samples of those populations in this study. The difference in number of alleles was not statistically significant between Africans and Europeans (Mann Whitney's U test, $p = 0.060$), but it was between Europeans and East Asians ($p = 0.010$), and between East Asians and Native Americans ($p < 0.001$). An average of 75.98 ± 1.96% of the alleles found worldwide are found in the Biaka Pygmies. In contrast, only 38.65 ± 1.87% of the alleles ever seen occur in the Amazonian Surui. At three loci, all the different alleles ever found in the global sample were present in the Biaka.

A measure of the genetic distinctiveness of a population sample can be obtained from the number of private alleles, ie, those alleles found exclusively in one particular population. Obviously, these are private alleles only with respect to the other nine specific populations in this study. The Biaka had a total of 32 private alleles across 45 loci, the Mbuti had 18, and no other population contained more than 9 private alleles (Figure 2). The difference in number of private alleles between the Biaka and the Mbuti was not statistically significant (Mann-Whitney's U, $p = 0.119$), while it was significant at a 0.01 level between the Biaka and any non-African population, and, at a 0.05 level, between the Mbuti and any non-African population (except for the Chinese and the Japanese). Conversely, the Surui had significantly fewer private alleles than all other populations, except for the smaller Nasioi sample. Next, we pooled the samples into two groups: Africans and non-Africans. Africans presented a mean of 1.73 ± 0.30 African-specific alleles per locus, whereas non-Africans

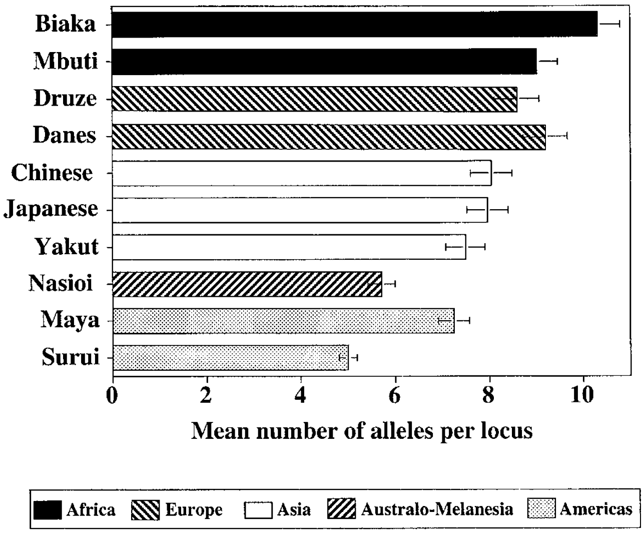

Figure 1 *Mean number of alleles per locus, estimated from 45 STRP markers. Bars represent standard errors of the mean*

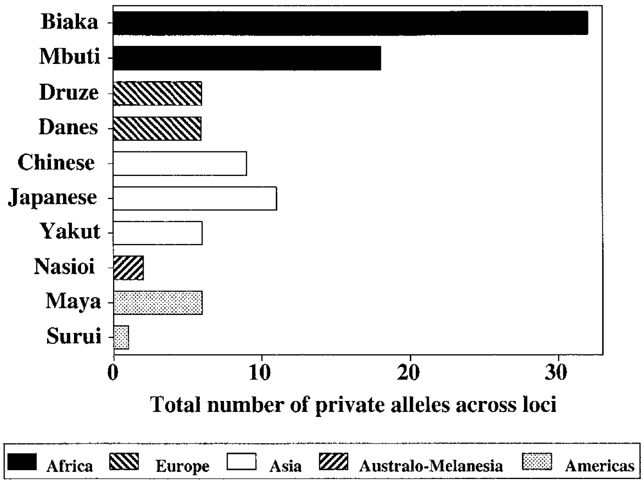

Figure 2 *Total number of alleles found exclusively in each population, across 45 loci*

Table 1 Mean, standard deviation, range and distribution of expected heterozygosities in 10 populations

	Mean	*s.d.*	*range*	*0–0.25*[a]	*0.25–0.5*	*0.5–0.75*	*0.75–1*
Biaka	0.8066	0.0087	0.5330–0.8976	–	–	9 (20%)	36 (80%)
Mbuti	0.7764	0.1059	0.4175–0.9	–	2 (4.4%)	13 (28.9%)	30 (66.7%)
Druze	0.7684	0.0077	0.5656–0.9185	–	–	16 (35.6%)	29 (64.4%)
Danes	0.7785	0.0072	0.5430–0.9007	–	–	15 (33.3%)	30 (66.7%)
Chinese	0.6745	0.1429	0.2605–0.8589	–	6 (13.3%)	22 (48.9%)	17 (37.8%)
Japanese	0.6678	0.1617	0.1488–0.8534	2 (4.4%)	4 (8.9%)	19 (42.2%)	20 (44.4%)
Yakut	0.6897	0.1327	0.2044–0.8597	1 (2.2%)	3 (6.7%)	24 (53.3%)	17 (37.8%)
Nasioi	0.6592	0.1324	0.2580–0.8367	–	4 (8.9%)	30 (66.7%)	11 (24.4%)
Maya	0.6617	0.1339	0.1665–0.8745	1 (2.2%)	1 (2.2%)	29 (64.4%)	14 (31.1%)
Surui	0.5488	0.1905	0.0503–0.8103	5 (11.1%)	8 (17.8%)	28 (62.2%)	4 (8.9%)

[a] Number and percentage of loci with heterozygosities falling in the 0–0.25, 0.25–0.5, 0.5–0.75, and 0.75–1 ranges.

as a whole presented 2.16 ± 0.27 alleles specific to non-Africans. However, the average sample size for non-Africans was more than three times larger than the average African sample size (750.9 vs. 204.7 chromosomes). We corrected for the difference in sample size by taking, for every locus, a random sub-sample with replacement from the non-African allele distribution with the sample size of the Africans, counting how many private alleles were found, and averaging over 10 000 sub-samples. Thus, when adjusted for sample size, non-Africans presented a mean of 1.22 ± 0.17 'private' alleles per locus, that is, 29.5% less than Africans. When the same procedure was applied to Europeans, East Asians, and Native Americans, the number of private alleles in those continents was always significantly *smaller* ($p < 0.001$, Wilcoxon signed rank test) than the corrected number of alleles found respectively, in non-Europeans, non-Asians, and non-Native Americans. The combined average allele frequency of private alleles was 0.060 ± 0.011 in Africans and 0.044 ± 0.015 in non-Africans ($p = 0.010$, Wilcoxon signed rank test).

In summary, the amount of genetic variability exclusive to Africans is larger than that of other continents (although a meaningful comparison with the smaller Australo-Melanesian sample was not possible), and it is even larger than that of all other continents combined.

Mean and standard deviation of expected heterozygosities are shown in Table 1; most expected heterozygosities fell above 0.5 (Table 1). Depending on the populations, zero to six loci (0–13.3%) presented heterozyosities below 0.5, the only exception being the Surui, with 13 (28.9%) loci below 0.5. Thus, even though they were selected for polymorphism in Europeans, most of the markers in this particular set of STRPs are informative enough to be used in genetic linkage studies in non-European populations. The mean expected heterozygosity (Figure 3) was highest in Africa (0.792 ± 0.010), it was slightly lower in Europe (0.773 ± 0.008), and it declines in East Asia (0.677 ± 0.013), and especially in the Americas (0.605 ± 0.018). These values are comparable to those found by Jorde *et al*,[11] and by Bowcock *et al*;[10] sample size does not appear to have a strong effect on the estimation of heterozygosity. All pairwise differences in heterozygosity between the continental groupings of populations were statistically significant at a 0.05 level (Mann-Whitney's U test).

Allele Size Variance

Variance can be used to roughly characterize the distribution of the number of repeats in a STRP. At equilibrium, and in the absence of bounds to allele length,

$$v = 2N_e\mu\sigma_m^2 \tag{1}$$

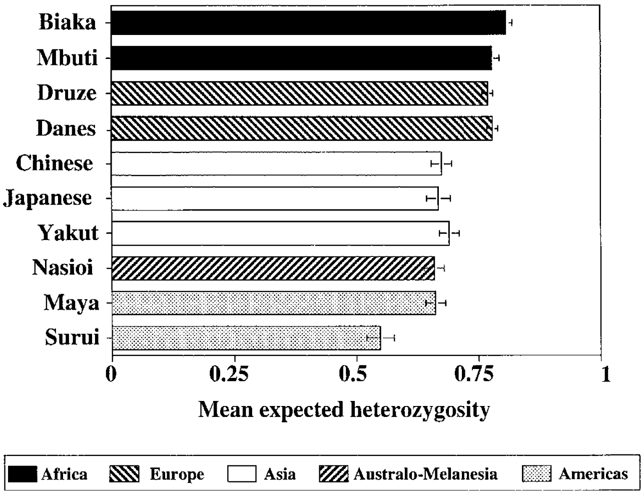

Figure 3 *Mean expected heterozygosity in 45 STRP loci. Bars are standard errors of the mean*

where v is the variance in allele size, N_e is effective population size, μ is the mutation rate, and σ^2_m is the variance of the number of repeats gained or lost in every mutation event.[24,33–35] The average variance in allele size (an indirect measure of repeat number, Figure 4) was 8.11 ± 0.86 in African populations, 6.48 ± 0.67 in Europeans, 6.20 ± 0.69 in East Asians, and 5.65 ± 0.82 in Native Americans. The distribution of variances across loci was highly skewed to the right, with one locus (*D11S922*) presenting a mean variance of 41.61, more than six times the overall mean variance. African populations showed a mode around 5–7; modal values are shifted to lower values in Europeans and East Asians, resulting in lower medians (Table 2), and the distributions of variances across loci in Native Americans present peaks at lower values (< 2), lower medians, and long right tails (Table 2).

Taking logarithms in Equation (1), a linear model is obtained.[9] We tested differences in the logarithm of the variance by means of one way analysis of variance (ANOVA), which revealed statistically significant differences across loci ($p < 0.001$). These can be interpreted as differences in the pattern of mutation (the $\mu\sigma^2_m$ term), as population size is constant across loci. A separate ANOVA showed significant differences by population ($p < 0.001$); thus, different populations could have significantly different present or historical effective population sizes. In a survey of different published datasets, Chakraborty *et al*[9] found significant differences for loci within and between motif length classes, but did not find statistically significant differences across populations, probably because Native American populations were not included in those datasets. Neither the variance nor its logarithm fit a normal distribution, in violation of the requirements for ANOVA. Thus, we performed a non-parametric test (Kruskal-Wallis' H) and found that variances were statistically significantly different ($p < 0.001$) among loci and among populations. However, a non-parametric test does not allow evaluation of fit to a linear model.

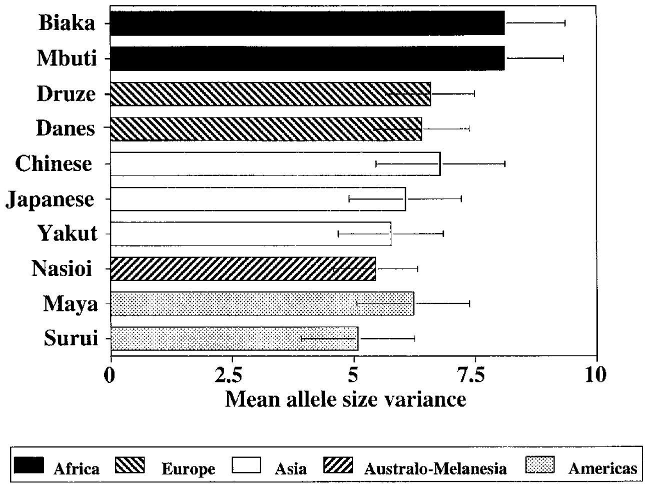

Figure 4 *Allele size variance, averaged across 45 STRPs. Allele size is expressed in number of repeats; bars represent standard errors of the mean*

Table 2 Mean, median, and range of the variance in repeat number in 45 STRP loci in 10 populations

	Mean	*Median*	*Minimum*	*Maximum*
Biaka	8.11	6.25	0.86	53.67
Mbuti	8.11	6.05	0.77	47.98
Druze	6.39	4.59	0.53	43.05
Danes	6.57	5.00	1.04	37.16
Chinese	6.78	4.15	0.26	54.97
Japanese	6.06	3.98	0.25	47.50
Yakut	5.76	4.83	0.33	46.15
Nasioi	5.44	4.39	0.21	32.90
Maya	6.22	3.40	0.70	36.73
Surui	5.08	1.95	0.22	40.90

In summary, African populations present allele distributions with more alleles, more private alleles, and a higher variance than other populations, whereas American populations have fewer alleles, fewer private alleles, and repeat number distributions that were less variable on average, but more variable across loci. The implications of these findings will be discussed in detail below.

Genetic Diversity Among Populations

Mean F_{ST} among all populations was 0.0682 ± 0.0034, with a range from 0.0279 to 0.1189. Other global F_{ST} values reported for STRPs are 0.0340,[4] 0.0859,[3] and 0.1072.[2] Since F_{ST} measures variances among populations, discrepancies in F_{ST} values among populations can be due to the number and choice of populations typed. By continent, the highest interpopulation diversity was found within the Americas (F_{ST} = 0.0352 ± 0.0028) and Africa (F_{ST} = 0.0330 ± 0.0026), while the greater interpopulation similarities were found among the East Asian (F_{ST} = 0.0183 ± 0.0016) and European (F_{ST} = 0.0146 ± 0.0011) samples.

Several genetic distances were computed among the population samples: F_{ST}[20,21] D_{SW},[22] $(\delta\mu)^2$,[23] and R_{ST}.[24] All distance matrices were positively correlated (Table 3) and all correlation coefficients were statistically significantly different from zero ($p < 0.001$) by the Mantel test.[36] However, most correlation coefficients ranged between 0.6 and 0.75, which means that only 36% to 56% of the variance of one genetic distance can be explained by another genetic distance. Therefore, we kept all four distance measures when performing tree analysis.

Goldstein *et al*[23] showed that, under mutation-drift equilibrium, $(\delta\mu)^2$ increases linearly with time, with a slope equal to twice the mutation rate, and, with the dataset published by Bowcock *et al*,[10] estimated the age of the African *vs* non-African split at 156 000 years ago, with a 95% confidence interval of 75 000–287 000 years. The largest $(\delta\mu)^2$ value we found was 4.45 between the Mbuti and the Maya. Applying the same parameters suggested by Goldstein *et al*,[23] that is, a mutation rate of 5.6×10^{-4} and a generation time of 27 years, we obtain a total date (ie, the time needed to accumulate the genetic distance between the Mbuti and the Maya) for the tree of 107 000 years, with a 95% confidence interval (estimated by 1000 bootstrap iterations as described by Goldstein *et al*[23]) of 43 500–220 000 years, which overlaps with the estimate by Goldstein *et al*.[23] The estimate of divergence time depends linearly on the mutation rate. However, using a more accurate average mutation rate estimate (7.8×10^{-4},[13] which is based on typing over 2000 loci (including most of those used in the present study) in large families, we obtained a similar time estimate: 77 000 years with a 95% confidence interval of 31 000 to 158 000 years ago.

Neighbor-joining trees were built for all four genetic distances (Figures 5a–d), and the robustness of their branches was tested by bootstrap. The LSSearch program[28] showed no better (by least squares or minimum length) tree topologies in the vicinity of the neighbor-joining trees. The F_{ST} and D_{SW} trees (Figures 5a and 5b) showed the same topology: African samples clustered on one end, followed by Europeans, Nasioi Melanesians, East Asians, and Native Americans. The same topology was obtained by Cavalli-Sforza *et al*[21] with 42 populations and 40 blood groups and protein polymorphisms, totaling 120 independent alleles. For comparison, we can estimate conservatively the range of independent alleles in the 45 STRPs as 94–230 (in the Surui and Biaka, respectively), obtained by counting the number of alleles with frequencies larger than 0.05 and subtracting one at every locus. Bootstrap supports were high, with only one value slightly below 50%, and similar between the F_{ST} and D_{SW}. In particular, branches grouping respectively Africans, Europeans, and Native Americans presented bootstrap supports above 95%. The $(\delta\mu)^2$ tree (Figure 5c) displayed a topology that is harder to reconcile with other genetic and archeological evidence: for instance, the Japanese appear at the base of the branch leading to the Africans. Moreover, bootstrap supports were lower than for the F_{ST} and D_{SW} trees. Finally, R_{ST} (Figure 5d) shared the same tree topology with F_{ST} and D_{SW}, except for the position of the Melanesian Nasioi, which clustered with the Brazilian Surui.

Table 3 Mantel correlation coefficients[36] between genetic distances

	D_{SW}	$(\delta\mu)^2$	R_{ST}
F_{ST}	0.8731[a]	0.6840[a]	0.6851[a]
D_{SW}		0.9231[a]	0.7200[a]
$(\delta\mu)^2$			0.6311[a]

[a] $p < 0.001$; significance was estimated from 100 000 matrix reshufflings.

In terms of branch robustness, previous genetic evidence, and expectations based on current hypotheses on human origins. F_{ST} and D_{SW} performed similarly to each other and better than both R_{ST} and $(\delta\mu)^2$. These results agree with those found by Pérez-Lezaun *et al*,[3] although their trees were even more discrepant.

Discussion

A total sample of 962 chromosomes from a global set of 10 populations was typed for 45 dinucleotide STR loci. These loci were originally selected for polymorphism in a sample comprising mostly families of mixed European ancestry[13] in order to obtain markers suitable for genetic linkage analysis. However, linkage studies are not restricted to European populations (see References 37–39 among many possible examples). Many genetic disorders are confined to specific populations or areas; and isolated populations, which are especially suitable for mapping mutations responsible for genetic disorders, are often of non-European ancestry. In spite of the European bias in their selection, we have shown that almost all of these markers are informative enough for linkage studies in non-Europeans. Even in populations such as the Surui that have reduced genetic diversity because of genetic drift, three quarters of these dinucleotide STRPs have heteozygosities above 0.5, that is, they are more informative than any diallelic marker.

Polymorphism in STRPs is not restricted to perfect multiples of the repeat unit. In our dataset, six out of 45 loci had alleles interspersed within, but not belonging to, a two-bp ladder. Deka *et al*[2] found such alleles in three out of eight dinucleotide STRPs, and Pérez-Lezaun *et al*[40] found similar off-ladder alleles at two out of twenty tetranucleotide STRPs. Sequence analysis is needed to determine whether the source of these 'irregular' alleles lies in the flanking regions,[41] or within the repeat region. It has been shown that STR stretches containing imperfect repeats have a lower mutation rate than perfect repeats,[32,42] and thus off-ladder alleles could add an additional layer of complexity to STRP mutation patterns. Since they exist at more than 10% of the loci surveyed, the nature(s) of the mutations that generate them must be a significant factor in STRP evolution.

Off-ladder and null alleles are two examples of complex polymorphism at STR loci; base substitutions could also be present in the repeat sequences,[43] but the methods we used did not allow for their detection. Such complex polymorphism defies attempts to model STRP evolution as the result of an SMM. However, in humans, such a distinction might not be relevant if, as discussed below, drift contributed significantly to shaping STRP allele frequencies.

After the discovery of STRPs and the suggestion that they mutate by adding or subtracting repeat units, there

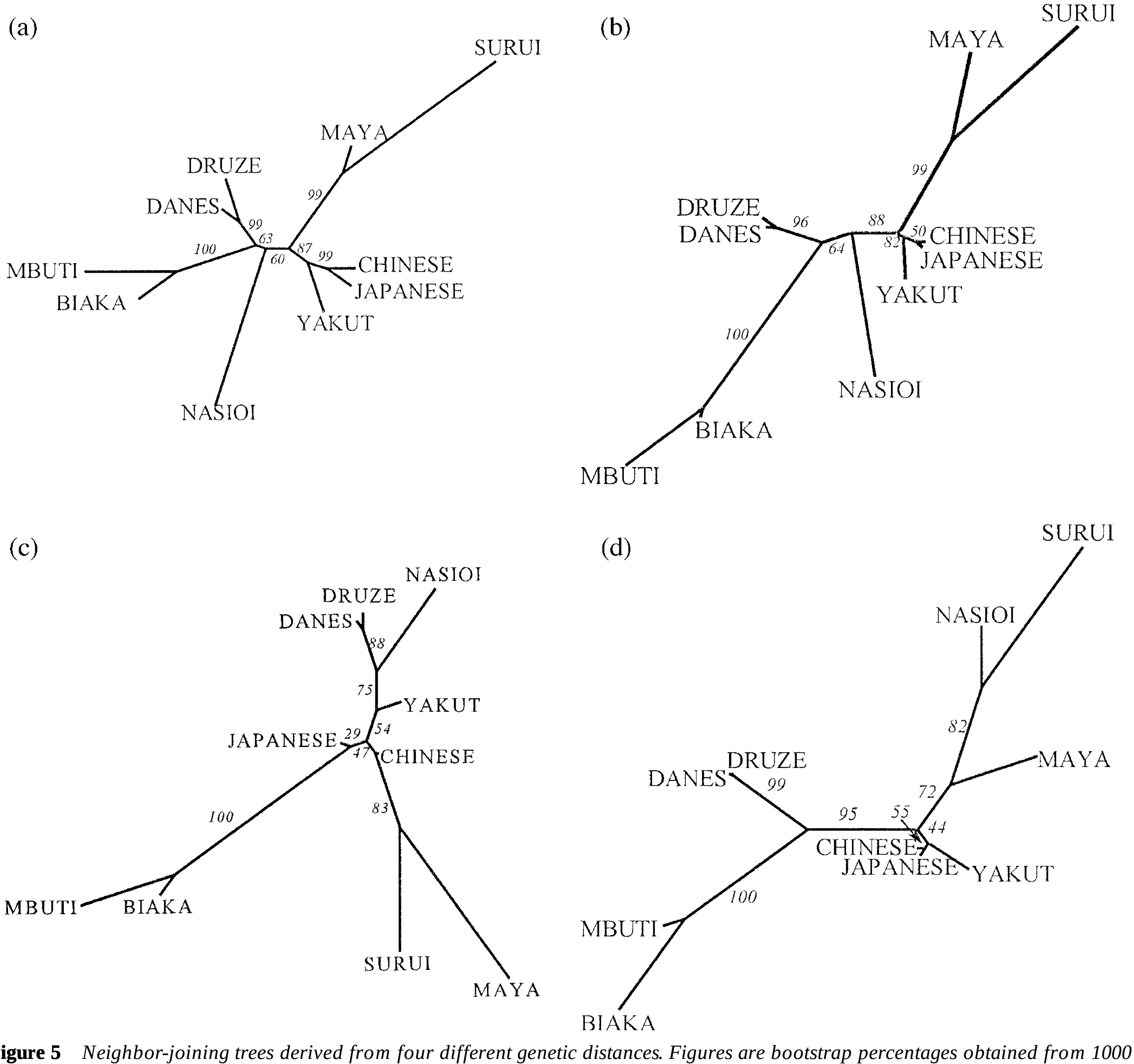

Figure 5 *Neighbor-joining trees derived from four different genetic distances. Figures are bootstrap percentages obtained from 1000 replicates: (a) F_{ST}; (b) D_{SW}; (c) $(\delta\mu)^2$; (d) R_{ST}*

has been a notable effort to adapt population genetics parameters that assumed no mutation or an infinite allele model to the stepwise mutation model, and, in particular, several genetic distance measures have been devised specifically for STRPs. We have used three such STRP-specific genetic distances, and we have found, using the criteria of reasonableness and agreement with archaeological data and other (non-STRP) genetic data, and in agreement with,[4] that F_{ST} performs as well as, or, in some instances, better than the STRP-specific distance measures. Thus, modeling genetic distances after the SMM does not seem to improve their performance in human STRP analysis. Several reasons can explain this apparent paradox, namely, sampling variance, deviation from the SMM, and genetic drift.

Sampling variance

Zhivotovski and Feldman[44] suggested that the variance of $(\delta\mu)^2$ grows with the number of loci, and the square of the time elapsed since two populations diverged; Nauta and Weissing[34] noted the high sampling error of D_0, the mean pairwise difference in number of repeats within a population, a parameter related to R_{ST} and $(\delta\mu)^2$. Takezaki and Nei[45] simulated the evolution of loci under the SMM in several populations and showed that IAM-appropriate distances, such as Cavalli-Sforza and Edwards' chord distance and Nei *et al*'s D_A generally outperformed D_{SW} and $(\delta\mu)^2$ in recovering the correct topology of a population tree. They attributed that result to the large sampling variances of D_{SW} and $(\delta\mu)^2$. This could also explain why there was greater discrepancy between the trees generated by Pérez-Lezaun *et al*,[4] with smaller sample sizes than between ours.

Deviation from the SMM

This paper and previous reports show that STRP polymorphism is by no means confined to simple variation in repeat number: off-ladder alleles, null alleles, and base substitutions have been repeatedly found. Interruptions of the repeat stretch have been shown to decrease mutation rate,[32] and longer alleles appear to have higher mutation rates.[42] In five loci, intermediate alleles were missing from all ten population samples, which could be due to a mutation involving a large jump in the number of repeats occurring in a very deep branch of the coalescent process,[6] and/or to some mechanical or selective constraint. Thus, departure from a strict SMM could be partly responsible for discontinuous allele distributions. Moreover, allele size can be constrained by upper and lower bounds:[34] selection against large repeat stretches is seen in the genetic disorders caused by large expansions in some trinucleotide STRPs, although we have not found evidence for such limits in the loci we have analyzed (the variance in repeat number was one-third of that expected under a bound model). The complexity of STRP polymorphism might be over-simplified by a strict SMM, and the distance measures suggested for STRPs might not capture the complexity of their mutation pattern. However, some attempts have been made to increase the sophistication of the models implemented in STRP-specific genetic distances. Thus, Feldman *et al*[46] devised D_L, a version of $(\delta\mu)^2$ corrected for size boundaries to allele variation. Unfortunately, D_L requires that the mutation rate and the maximum number of alleles be known and equal across loci, and we and Chakraborty *et al*[9] have shown that mutation rate is likely to vary across loci.

Genetic Drift

Beside the deviation factors noted, genetic drift could also have contributed extensively to the creation and maintenance of discontinuous allele distributions. F_{ST} and other genetic distances were modeled to be linear with separation time between populations and constant population size if differentiation occurred mainly or exclusively by drift. This could the case be for STRP evolution in humans, as argued by Pérez-Lezaun *et al*.[3] Drift rather than mutation certainly shaped allele frequencies in genetic polymorphisms such as blood groups and protein electromorphs,[21] RFLPs,[47] and Alu sequence insertions,[48] which have a much slower mutation rate. At the *CD4* locus, Tishkoff *et al*[17] showed that, in the CD4 pentanucleotide, mutation had not restored the allelic diversity in non-Africans to the level observed in Africans in, according to their interpretation, about 5000 generations. It has been shown[18,49] that STRP and flanking RFLP alleles at the *DRD2* locus are in strong disequilibrium in multiple populations and that variation in allele frequency at the STRP is accounted for by variation in frequency for the entire haplotype, ie mutation has not restored linkage equilibrium.

Recently,[50–52] it has been suggested that part of the genetic variation in the β-globin locus and in the Y chromosome might have been generated in Asia and introduced into Africa by an ancient 'back to Africa' migration. If that were the case for a significant fraction of the genome, we would expect to see a shorter genetic distance between Africa and Asia than between Africa and other continents. This is not the case (Figure

5a–5d]: in the present analysis of 45 loci, European rather than East Asian populations are the closest to Africans. A significant gene flow from Asia to Africa could have resulted in Africa and Asia sharing a higher number of alleles than Africa and other populations. Across 45 loci, 22 alleles are shared by African and Asian populations and are absent elsewhere; the number of alleles present in Africa and Europe and not elsewhere is considerably larger: 34. Although we cannot rule out a small gene flow from Asia to Africa, large genetic contributions from Asia to Africa seem unlikely.

Several measures of intra-population genetic variability, such as number of alleles, expected heterozygosity, and variance in allele size, showed a consistent pattern in our results. Genetic variability is highest in Africa, intermediate in Europe and Africa, and lowest in the Americas. The Biaka and the Mbuti presented intermediate haplotype diversities in the mtDNA control region,[53] and at the *DMPK* locus (Tishkoff *et al*, in preparation) when compared to other African populations. The Mbuti, but not the Biaka, seemed to have an elevated haplotype diversity at the Y chromosome.[50] Thus, there is no consistent evidence that our choice of African populations biased the estimates of intra-population diversity. Genetic trees show Africans at one end and, at successive splits, Europeans, East Asians, and Native Americans. A higher heterogeneity in Africans and/or a genetic tree in which Africans branch first have been repeatedly shown by different types of nuclear genetic markers: STRPs,[3,10,11,17,23] single nucleotide polymorphisms,[47] minisatellites,[54] haplotyes,[16,17,55] and Alu sequences.[56] This could be explained by a recent African origin of anatomically modern humans, which is consistent with the date estimates obtained by us and by Goldstein *et al*,[23] which place an upper limit for the generation of STRP variation at same 220 000 years ago. Alternatively, and as pointed out by Relethford,[57] a larger population size in Africa could account for a higher African heterozygosity and for the apparent greater time depth of African genetic diversity. However, such a scenario would probably not be able to accommodate our observation that, on average across 45 nuclear loci, African populations present a significantly higher number of private alleles, and, thus, genetic variation outside of Africa tends to be a subset of that of Africa. This has also been observed for the *CD4* locus,[17] for the *DRD2* locus,[16,49] and for the Y chromosome.[50] The relatively high levels of migration needed to synchronize evolution in a multi-regional model would also continuously spread the whole African variation across the globe, and it seems more likely that one or a few main 'Out of Africa' migration events spread a subset of the variation generated in Africa.

Acknowledgements

We appreciate the assistance of Beth Wilson and Tara Coleman in maintaining the lymphoblastoid cell lines. The Applied Biosystems Division of Perkin-Elmer provided Linkage Mapping Sets and other materials. Supported in part by USPHS grants MH39329 and MH50390 (to KKK), and NSF grant SBR-940834 (to JRK). FC is the recipient of a postdoctoral fellowship from CIRIT, Autonomous Government of Catalonia.

References

1 Dib C *et al*: A comprehensive genetic map of the human genome based on 5624 microsatellites. *Nature* 1996; **380**: 152–154.

2 Deka R *et al*: Population genetics of dinucleotide (dC-dA)n (dG-dT)n polymorphisms in World populations. *Am J Hum Genet* 1995; **56**: 461–474.

3 Pérez-Lezaun *et al*: Microsatellite variation and the differentiation of modern humans. *Hum Genet* 1997; **99**: 1–7.

4 Jorde LB *et al*: Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet* 1995; **57**: 523–538.

5 Rogers AR, Jorde LB: Ascertainment bias in estimates of average heterozygosity. *Am J Hum Genet* 1996; **58**: 1033–1041.

6 Valdes AM, Slatkin M, Freimer NB: Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 1993; **133**: 737–749.

7 Di Rienzo A *et al*: Mutational processes of simple sequence repeat loci in human populations. *Proc Natl Acad Sci USA* 1994; **91**: 3166–3170.

8 Weber J, Wong C: Mutation of human short tandem repeats. *Hum Molec Gen* 1993; **2**: 1123–1128.

9 Chakraborty R *et al*: Relative mutation rates at di-, tri- and tetranucleotide microsatellite loci. *Proc Natl Acad Sci USA* 1997; **94**: 1041–1046.

10 Bowcock AM *et al*: High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 1994; **368**: 455–457.

11 Jorde LB *et al*: Microsatellite diversity and the demographic history of modern humans. *Proc Natl Acad Sci USA* 1997; **94**: 3100–3103.

12 Astolfi P, Kidd KK, Cavalli-Sforza LL: A comparison of methods for reconstructing evolutionary trees. *Syst Zool* 1981; **30**: 156–169.

13 Gyapay G *et al*: The 1993–1994 Généthon human genetic linkage map. *Nat Genet* 1994 **7**: 246–339.

14 Stringer CB, Andrews P: Genetic and fossil evidence for the origin of modern humans. *Science* 1988; **239**: 1263–1268.

15 Wolpoff MH: Multiregional evolution: the fossil alternative to Eden. In: Mellar P, Stringer C (eds.)*The Human Revolution: Behavioural and Biological Perspectives on the Origins of Modern Humans*, Princeton University Press, Princeton, NJ, 1989, pp62–108.

16 Castiglione CM *et al*: Evolution of haplotypes at the *DRD2* locus. *Am J Hum Genet* 1995; **57**: 1445–1456.

17 Tishkoff SA *et al*: Global patterns of linkage disequilibrium at the *CD4* locus and modern human origins. *Science* 1996; **271**: 1380–1387.

18 Weir BS: *Genetic data analysis II*. Sinauer Associates, Sunderland, MA, 1996.

19 Hawley ME, Kid KK: HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotyes. *J Hered* 1995; **86**: 409–411.

20 Wright S: The genetical structure of populations. *Ann Eugenet* 1951; **15**: 323–354.

21 Cavalli-Sforza LL, Menozzi P, Piazza A: *History and geography of human genes*. Princeton University Press, Princeton, NJ, 1994.

22 Shriver MD *et al*: A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol Biol Evol* 1995; **12**: 914–920.

23 Goldstein DB, Ruiz-Linares A, Cavalli-Sforza LL, Feldman MW:Genetic absolute dating based on microsatellites and the origin of modern humans. Proc Natl Acad Sci USA 1995; **92**: 6723–6727.

24 Slatkin M: A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 1995; **139**: 457–462.

25 Saitou N, Nei M: The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 1987; **4**: 406–425.

26 Efron B: The jackknife, the bootstrap, and other resampling plans. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1982.

27 Felsenstein J: Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985; **39**: 783–791.

28 Kidd KK, Sgaramella-Zonta LA: Phylogenetic analysis: concepts and methods. *Am J Hum Genet* 1971; **23**: 235–252.

29 Felsenstein J: PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* 1989; **5**: 164–166.

30 Callen DF *et al*: Incidence and origin of 'null' alleles in the (AC)n microsatellite markers. *Am J Hum Genet* 1993; **52**: 922–927.

31 Paetkau D, Strobeck C: The molecular basis and evolutionary history of a microsatellite null allele in bears. *Mol Ecol* 1995; **4**: 519–520.

32 Jin L *et al*: Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proc Natl Acad Sci USA* 1996; **93**: 15285–15288.

33 Moran PAP: Wandering distributions and the electrophoretic profile. *Theor pop biol* 1975; **8**: 318–330.

34 Nauta MJ, Weissing FJ: Constraints on allele size at microsatellite loci: implications for genetic differentiation. *Genetics* 1996; **143**: 1021–1032.

35 Kimmel M, Chakraborty R, Stivers DN, Deka R: Dynamics of repeat polymorphisms under a forward-backward mutation model: within and between variability at microsatellite loci. *Genetics* 1996; **143**: 549–555.

36 Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; **27**: 209–220.

37 Takiyama Y *et al*: The gene for Machado-Joseph disease maps to human chromosome 14q. *Nat Gen* 1993; **4**: 300–304.

38 Yu CE *et al*: Positional cloning of the Werner's syndrome gene. *Science* 1996; **272**: 258–262.

39 Norman RA *et al*: Absence of linkage of obesity and energy metabolism to markers flanking homologues or rodent obesity genes in Pima Indians. *Diabetes* 1996; **45**: 1229–1232.

40 Pérez-Lezaun A *et al*: Allele frequencies for 20 microsatellite loci in a worldwide population survey. *Hum Hered* 1997; **47**: 189–196.

41 Grimaldi MC, Crouau-Roy B: Microsatellite allelic homoplasy due to variable flanking sequences. *J Mol Evol* 1997; **44** 336–340.

42 Murray A *et al*: The role of size, sequence and haplotype in the stability of *FRAXA* and *FRAXE* alleles during transmission. *Hum Molec Gen* 1997; **6**: 173–184.

43 Pérez-Lezaun A: Identification of a base pair substitution at the tetranucleotide tandem repeat locus *DHFRP2*(*AAAC*)n using non-denaturing gel electrophoresis. *Int J Leg Med* 1996; **109**: 159–160.

44 Zhivotovsky LA, Feldman MW: Microsatellite variability and genetic distances. *Proc Natl Acad Sci USA* 1995; **92**: 11549–11552.

45 Takezaki N, Nei M: Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* 1996; **144**: 389–399.

46 Feldman MW, Bergman A, Pollock DD, Goldstein DB: Microsatellite genetic distances with range constraints: analytic description and problems of estimation. *Genetics* 1997; **145**: 207–216.

47 Bowcock AM *et al*: Drift, admixture and selection in human evolution: A study with DNA polymorphisms. *Proc Natl Acad Sci* 1991; **88**: 839–843.

48 Batzer MA *et al*: African origin of human-specific polymorphic Alu insertions. *Proc Natl Acad Sci USA* 1994; **91**: 12288–12292.

49 Kidd KK *et al*: DRD2 haplotypes containing the *TaqI A1* allele: implications for alcoholism research. *Alcohol Clin Exp Res* 1996; **20**: 697–705.

50 Hammer MF *et al*: The geographic distribution of human Y chromosome variation. *Genetics* 1997; **145**: 785–805.

51 Altheide TK, Hammer MF: Evidence for a possible Asian origin of YAP + Y chromosomes. *Am J Hum Genet* 1997; **61**: 462–466.

52 Harding RM *et al*: Archaic African *and* Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* 1997; **60**: 772–789.

53 Watson E *et al*: mtDNA sequence diversity in Africa. *Am J Hum Genet* 1996; **59**: 437–444.

54 Armour JA *et al*: Minisatellite diversity supports a recent African origin for modern humans. *Nat Gen* 1996; **13**: 154–160.

55 Kidd KK, Kidd JR: A nuclear perspective on human evolution. In: Boyce AJ, Mascie-Taylor CGN (eds.) *Molecular Biology and Human Diversity* Cambridge University Press, Cambridge, UK, 1996, pp242–264.

56 Knight A *et al*: DNA sequences of the Alu elements indicate a recent replacement of the human autosomal genetic complement. *Proc Natl Acad Sci USA* 1996; **93**: 4360–4364.

57 Relethford J: Genetics and modern human origins. *Evol Anthropol* 1995; **4**: 53–63.