*J. Genet.*, Vol. 76, Number 3, December 1997, pp. 181–188. © Indian Academy of Sciences

# How useful are microsatellite loci in recovering short-term evolutionary history?

B. KAMESWARA RAO, SUNIL B. SIL and PARTHA P. MAJUMDER*
Anthropometry and Human Genetics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road, Calcutta 700 035, India

MS received 4 August 1997; revised received 18 December 1997

**Abstract.** Because microsatellite loci are abundant in the human genome and are highly polymorphic in most global populations, such loci have become very popular in studies on reconstructing evolutionary relationships among contemporary human populations. We have made an assessment of the efficiency of recovery of true evolutionary relationships using simulated data of microsatellite loci and a variety of distance measures. We find that allele frequency data on about 30 microsatellite loci and the use of $D_A$ (Nei *et al.* 1983) or $D_C$ (Cavalli-Sforza and Edwards 1967) distance measures with UPGMA clustering algorithm can recover true short-term evolutionary relationships with a high degree of accuracy, unless the effective sizes of the populations or mutation rates or both are very small.

**Keywords.** Simulation; evolution; tandem repeats; genetic distance; clustering.

## 1. Introduction

Allele frequency data of microsatellite loci are increasingly being used in reconstruction of evolutionary relationships of contemporary human populations (Bowcock *et al.* 1994; Deka *et al.* 1995; Jorde *et al.* 1995). Compared to other types of loci (e.g. RFLP loci), microsatellite loci have the advantage that they are highly heterozygous and have large numbers of alleles in most global populations (Jorde *et al.* 1995). Moreover, the nature of alleles at these loci suggests that their mutational patterns conform to the stepwise mutation model (SMM) proposed by Ohta and Kimura (1973). In fact, the one-step version of the SMM has been found to provide a reasonable approximation to observed allele frequency distributions at microsatellite loci (Shriver *et al.* 1993; Valdes *et al.* 1993). However, there are indications that the rate of mutation at microsatellite loci may be several orders of magnitude (one to four) higher than at traditional loci (Edwards *et al.* 1992; Weber and Wong 1993). This may be a potential disadvantage in reconstruction of evolutionary relationships because, when data of loci with high mutation rates are used, recovered relationships may be a distorted form of the true relationships.

The purpose of this study is to examine how well evolutionary relationships among contemporary populations can be reconstructed with allele frequency data of microsatellite loci under a variety of scenarios and through the use of various measures of genetic distance, some of which have been specifically developed for use with data of such loci.

*E-mail: ppm@isical.ernet.in

## 2. Methods

In the present study we used Monte-Carlo simulation methodology in which subpopulations were generated from a single ancestral population through a series of bifurcations over evolutionary time, so that the true evolutionary relationships among the populations are known. The model tree depicting true evolutionary relationships is presented in figure 1. Starting with an inital set of allele frequencies at several unlinked microsatellite loci, mutations and random genetic drift were allowed to take place independently at each locus. Eventually, allele frequencies in contemporary populations, whose true evolutionary relationships are known, were generated.

Then, starting with the allele frequencies so generated at the various microsatellite loci in contemporary populations, genetic distances (using a variety of genetic distance measures as described below) between all pairs of these populations were computed. A method of phylogenetic reconstruction was then applied on a particular distance matrix and evolutionary relationships reconstructed. Agreement of the reconstructed phylogenetic tree topology with the true topology was then checked. This procedure, when repeated a large number of times, yielded an estimate of the efficiency of recovery of the true phylogenetic tree.

Before describing details of the simulation method and distance measures used we wish to point out that, although there are various methods of phylogenetic reconstruction (Nei 1987), we have used only one method—unweighted pair group method with arithmetic mean (UPGMA), alternatively known as the average distance method (Sneath and Sokal 1973; Nei 1987). This is because (i) we have used a constant mutation rate across loci and over time in generating simulated allele frequency data; (ii) the true tree (figure 1) is a rooted tree; and (iii) it is known (Nei 1987) that under conditions (i) and (ii) UPGMA is very efficient. Further, although there are two main aspects to phylogenetic reconstruction—correctness of inferred topology and accuracy of estimated branch lengths—we have examined only one aspect, that of correctness of the inferred topology. In short-term evolutionary studies, it is this aspect that is of greater importance.

### 2.1 *Computer simulation*

Consider one microsatellite locus in a population. To create subpopulations and to generate allele frequency data at this locus in each subpopulation, the following methodology was used. We started with a population of effective size $N$ which is monomorphic for a particular allele with repeat number $R$. In unit time (say a generation), this allele can mutate to $(R-1)$ or $(R+1)$ with probabilities $v_1$ and $v_2$; the total mutation rate per locus per generation being $v$. Estimates of $v$ range from $10^{-2}$ to $10^{-5}$ (Jeffreys *et al.* 1988; Edwards *et al.* 1992; Weber and Wong 1993). Usually $v_1$ and $v_2$ are taken to be equal. However, for microsatellite loci, there are indications, especially in respect of trinucleotide repeat loci which control many human diseases, that expansions may be more probable than contractions. We have therefore considered both scenarios: $v_1 = v_2$ and $v_1 < v_2$; specifically we have assumed $v_1 = v_2/2$. $v_1$ and $v_2$ were appropriately adjusted for variation in observed numbers of alleles at a locus so as to keep $v$ constant per locus per generation. Thus, using random numbers, we created mutations, with specified probabilities at each locus, in individuals in the population.

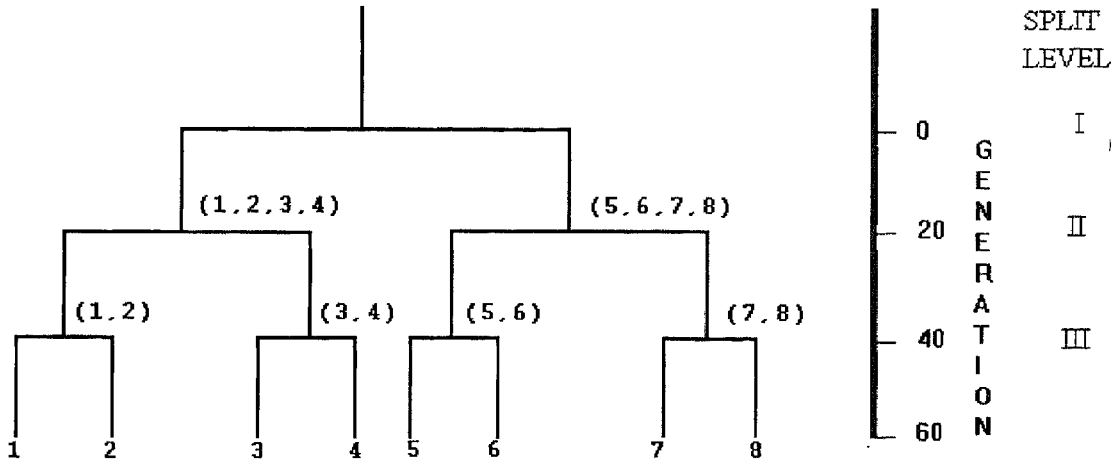


**Figure 1.** Model evolutionary tree used in computer simulation.

The next generation was formed by random sampling of $2N$ gametes from a pool containing gametes proportional to allele frequencies after mutations had taken place. The gametes were paired at random to form individuals of the next generation. This process simulated the effect of random genetic drift. We note that we have assumed, for simplicity, generations to be of constant effective size ($N$) and non-overlapping. These steps were repeated for $1/v$ generations to create an ancestral population that is in mutation–drift equilibrium at this locus. For the stepwise mutation model, the expected heterozygosity ($H$) at any locus in a population in mutation–drift equilibrium is: $H = 1 - 1/\sqrt{1 + 8Nv}$ (Ewens 1979).

After thus creating an ancestral equilibrium population, this population was split into two subpopulations each of size $N$. Each subpopulation's allele frequencies were determined by sampling alleles with probabilities proportional to their frequencies in the ancestral population for each locus independently. These two subpopulations were allowed to evolve independently for 20 generations before further splits, as shown in figure 1. Thus, eight contemporary populations, with known evolutionary histories, were generated through a series of three successive splits (bifurcations).

Subsequent to the final splitting of populations (split level III of figure 1), the eight populations thus generated were allowed to evolve for 20 generations. Based on the allele frequencies at $L$ independent loci in the eight contemporary populations, a matrix of pairwise distances between populations was estimated and the UPGMA method was used on this distance matrix to construct a phylogenetic tree. The topology of this estimated tree was then compared with that of the true phylogenetic tree depicted in figure 1.

### 2.2 *Distance measures used*

As mentioned earlier, it is essential to obtain estimates of genetic distances between pairs of populations for reconstructing the phylogenetic relationships among the eight contemporary populations. Let $x_{il}$ and $y_{il}$ denote the allele frequencies of the $i$th allele at the $l$th locus in two populations X and Y, respectively, where $i = 1, 2, \ldots, m_l$ = number

**Table 1.** Measures used in computation of genetic distance between populations.

| Distance measure | Reference |
|---|---|
| $D_A = 1 - \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m_l}\sqrt{x_{il}y_{il}}$ | Nei *et al.* (1983) |
| $D_C = \frac{2}{\pi L}\sum_{l=1}^{L}\sqrt{2\left(1-\sum_{i=1}^{m_l}\sqrt{x_{il}y_{il}}\right)}$ | Cavalli-Sforza and Edwards (1967) |
| $\mathrm{ASD}^1 = \frac{1}{L}\sum_{l=1}^{L}\sum_{i=1}^{m_l}\sum_{j=1}^{m_l}(i-j)^2 x_{il}y_{il}$ | Goldstein *et al.* (1995a); Slatkin (1995) |
| $D_{SW}^{1,2} = W_{XY} - \frac{1}{2}(W_X + W_Y)$ | Shriver *et al* (1995) |
| $\Delta^{1,3} = \frac{1}{L}\sum_{l=1}^{L}(\mu_{X_l} - \mu_{Y_l})^2$ | Goldstein *et al.* (1995b) |

1. $i$ denotes the $i$th allele, that is the allele with $i$ repeats.
2. $W_X = \frac{1}{L}\sum_{l=1}^{L}\sum_{i\neq j}|i-j|x_{il}x_{jl}$;

   $W_Y = \frac{1}{L}\sum_{l=1}^{L}\sum_{i\neq j}|i-j|y_{il}y_{jl}$;

   $W_{XY} = \frac{1}{L}\sum_{l=1}^{L}\sum_{i\neq j}|i-j|x_{il}y_{jl}$.
3. $\mu_{X_l} = \sum_{i=1}^{m_l} i x_{il}$; $\mu_{Y_l} = \sum_{i=1}^{m_l} i y_{il}$.

of alleles at the $l$th locus, and $l = 1, 2, \ldots, L$ = number of loci. In the present study, two values of $L$, 10 and 30, were used. Various distance measures were used; these are given in table 1. We note that the distance measures ASD, $D_{SW}$ and $\Delta$ have been proposed specifically for use with data of microsatellite loci.

## 3. Results

The precentages of recovered phylogenetic trees which agreed with the topology of the true tree are presented in table 2 for various sets of parameter values. For each set of parameter values, these estimates are based on 200 runs. (For some sets of parameter values, we performed the simulation 1000 times in batches of 100 and found that stability of this estimate was achieved in 200 runs.) Further, for each set of parameter values, agreement of the topology was examined at two split levels, levels III and II of figure 1. Agreement at split level III implied a perfect match of the entire tree topology. Agreement at split level II implied that the populations belonging to the two clusters at

**Table 2.** Percentages of correct recovery of true evolutionary relationships at two split levels (see figure 1) using five different genetic distance measures ($D_A$, $D_C$, ASD, $D_{SW}$ and $\Delta$) for various sets of parameter values.

| Effective population size ($N$) | $v$ | $v_1/v_2$ | $E(H)$ | No. of loci ($L$) | Split level | $D_A$ | $D_C$ | ASD | $D_{SW}$ | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 0·01 | 1·0 | 0·380 | 10 | III | 18·5 | 16·0 | 5·5 | 15·0 | 13·0 |
| | | | | | II | 44·5 | 44·0 | 27·5 | 36·5 | 31·0 |
| | | | | 30 | III | 70·5 | 68·5 | 29·0 | 62·5 | 47·5 |
| | | | | | II | 79·5 | 75·5 | 50·5 | 76·0 | 58·0 |
| | | 0·5 | | 10 | III | 14·5 | 8·0 | 5·0 | 12·0 | 5·5 |
| | | | | | II | 37·0 | 32·5 | 19·0 | 36·0 | 26·5 |
| | | | | 30 | III | 62·0 | 56·5 | 28·0 | 58·0 | 39·5 |
| | | | | | II | 69·0 | 64·0 | 50·0 | 70·0 | 62·0 |
| 20 | 0·001 | 1·0 | 0·072 | 10 | III | 5·0 | 5·5 | 0·5 | 2·0 | 2·0 |
| | | | | | II | 22·5 | 21·5 | 17·5 | 20·5 | 20·0 |
| | | | | 30 | III | 15·0 | 12·5 | 6·5 | 13·5 | 17·0 |
| | | | | | II | 45·5 | 36·5 | 34·5 | 43·0 | 42·0 |
| | | 0·5 | | 10 | III | 3·5 | 3·5 | 1·5 | 1·5 | 1·5 |
| | | | | | II | 24·0 | 20·5 | 16·5 | 22·0 | 23·0 |
| | | | | 30 | III | 24·5 | 14·5 | 3·0 | 17·0 | 11·0 |
| | | | | | II | 46·5 | 40·5 | 32·0 | 39·0 | 43·0 |
| 20 | 0·0005 | 1·0 | 0·038 | 10 | III | 0·0 | 0·0 | 0·0 | 0·0 | 0·0 |
| | | | | | II | 16·0 | 15·0 | 15·0 | 17·0 | 15·0 |
| | | | | 30 | III | 11·5 | 9·0 | 4·5 | 9·5 | 7·5 |
| | | | | | II | 38·0 | 29·0 | 28·5 | 35·5 | 37·0 |
| | | 0·5 | | 10 | III | 0·0 | 0·0 | 0·0 | 0·0 | 0·0 |
| | | | | | II | 11·5 | 8·5 | 7·5 | 11·5 | 10·0 |
| | | | | 30 | III | 7·0 | 5·5 | 2·0 | 4·5 | 4·5 |
| | | | | | II | 26·0 | 26·0 | 23·0 | 28·5 | 27·5 |
| 100 | 0·01 | 1·0 | 0·667 | 10 | III | 32·0 | 32·0 | 2·0 | 14·0 | 5·0 |
| | | | | | II | 60·0 | 50·0 | 7·0 | 44·0 | 27·0 |
| | | | | 30 | III | 91·5 | 89·5 | 2·0 | 79·5 | 51·0 |
| | | | | | II | 92·5 | 91·5 | 27·5 | 84·0 | 66·0 |
| | | 0·5 | | 10 | III | 28·5 | 23·5 | 0·0 | 15·0 | 6·5 |
| | | | | | II | 50·5 | 44·0 | 9·5 | 33·5 | 22·5 |
| | | | | 30 | III | 83·5 | 79·5 | 3·0 | 69·5 | 40·0 |
| | | | | | II | 83·5 | 81·5 | 20·0 | 75·5 | 55·5 |
| 100 | 0·001 | 1·0 | 0·255 | 10 | III | 11·5 | 15·0 | 0·0 | 5·5 | 2·0 |
| | | | | | II | 39·0 | 44·5 | 9·5 | 37·5 | 20·0 |
| | | | | 30 | III | 66·0 | 57·0 | 1·5 | 42·5 | 25·0 |
| | | | | | II | 73·0 | 75·0 | 15·0 | 58·5 | 52·0 |
| | | 0·5 | | 10 | III | 12·0 | 12·0 | 0·0 | 12·5 | 6·5 |
| | | | | | II | 39·5 | 34·0 | 13·0 | 33·0 | 22·5 |
| | | | | 30 | III | 55·0 | 47·5 | 1·5 | 49·5 | 28·5 |
| | | | | | II | 68·0 | 66·5 | 18·5 | 64·0 | 47·0 |
| 100 | 0·0005 | 1·0 | 0·155 | 10 | III | 6·5 | 7·0 | 0·0 | 1·5 | 0·0 |
| | | | | | II | 28·0 | 27·0 | 2·0 | 17·5 | 15·0 |
| | | | | 30 | III | 48·5 | 44·0 | 0·0 | 39·5 | 23·5 |
| | | | | | II | 60·5 | 56·0 | 15·5 | 55·0 | 51·5 |
| | | 0·5 | | 10 | III | 7·5 | 8·5 | 0·0 | 4·5 | 8·5 |
| | | | | | II | 36·0 | 28·0 | 3·0 | 25·0 | 17·0 |
| | | | | 30 | III | 47·5 | 34·5 | 0·0 | 34·5 | 23·5 |
| | | | | | II | 66·5 | 56·0 | 15·5 | 51·5 | 52·5 |

(Continued)

**Table 2.** (Continued)

| Effective population size ($N$) | $v$ | $v_1/v_2$ | $E(H)$ | No. of loci ($L$) | Split level | $D_A$ | $D_C$ | ASD | $D_{SW}$ | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|---|
| 300 | 0·01 | 1·0 | 0·800 | 10 | III | 59·0 | 54·0 | 0·0 | 39·0 | 12·0 |
| | | | | | II | 62·0 | 61·0 | 3·0 | 50·0 | 30·0 |
| | | | | 30 | III | 97·0 | 90·0 | 0·0 | 82·0 | 43·0 |
| | | | | | II | 97·0 | 93·0 | 6·0 | 88·0 | 60·0 |
| | | 0·5 | | 10 | III | 62·0 | 54·0 | 0·0 | 33·0 | 13·0 |
| | | | | | II | 68·0 | 63·0 | 1·0 | 53·0 | 35·0 |
| | | | | 30 | III | 99·0 | 94·0 | 0·0 | 83·0 | 35·0 |
| | | | | | II | 99·0 | 95·0 | 4·0 | 85·0 | 56·0 |
| 300 | 0·001 | 1·0 | 0·458 | 10 | III | 34·0 | 32·0 | 0·0 | 11·0 | 2·0 |
| | | | | | II | 63·0 | 65·0 | 7·0 | 34·0 | 26·0 |
| | | | | 30 | III | 91·0 | 88·0 | 0·0 | 72·0 | 33·0 |
| | | | | | II | 91·0 | 89·0 | 7·0 | 83·0 | 59·0 |
| | | 0·5 | | 10 | III | 34·0 | 31·0 | 0·0 | 15·0 | 5·0 |
| | | | | | II | 54·0 | 50·0 | 1·0 | 44·0 | 31·0 |
| | | | | 30 | III | 92·0 | 84·0 | 2·0 | 74·0 | 34·0 |
| | | | | | II | 95·0 | 91·0 | 8·0 | 86·0 | 60·0 |
| 300 | 0·0005 | 1·0 | 0·326 | 10 | III | 26·0 | 24·0 | 0·0 | 17·0 | 3·0 |
| | | | | | II | 61·0 | 54·0 | 1·0 | 44·0 | 27·0 |
| | | | | 30 | III | 83·0 | 75·0 | 0·0 | 66·0 | 33·0 |
| | | | | | II | 85·0 | 79·0 | 13·0 | 70·0 | 52·0 |
| | | 0·5 | | 10 | III | 30·0 | 30·0 | 0·0 | 16·0 | 12·0 |
| | | | | | II | 58·0 | 56·0 | 1·0 | 52·0 | 28·0 |
| | | | | 30 | III | 78·0 | 73·0 | 0·0 | 71·0 | 34·0 |
| | | | | | II | 83·0 | 78·0 | 9·0 | 80·0 | 55·0 |

this level in the reconstructed tree were 1, 2, 3 and 4 and 5, 6, 7 and 8, irrespective of whether these clusters bifurcated as (1,2), (3,4) and (5,6), (7,8). Thus the difference between the percentages at split levels II and III given in table 2 provides an estimate of the percentage of reconstructed trees that match at level II but are imperfect matches when the entire tree is considered.

The following features are evident from table 2: (i) there is an overall pattern of positive correlation between expected heterozygosity, $E(H)$, and the percentage of recovery of true evolutionary relationships; (ii) with data on 10 microsatellite loci, the percentage of recovery is unacceptably low even for high values of $E(H)$, irrespective of the distance measure used; (iii) even when data of 30 loci are used, for values of expected average heterozygosity, $E(H)$, less than 0·1, the percentage of recovery is extremely low ($< 50\%$); (iv) the distance measures $D_A$ and $D_C$ perform satisfactorily , with $D_A$ performing slightly better; (v) the distance measures $\Delta$ and ASD perform unsatisfactorily; (vi) the use of $D_{SW}$ cannot be recommended even though its performance is not as poor as ASD or $\Delta$; (vii) for values of $E(H) > 0{\cdot}5$, the percentage of recovery of the true topology is very high (80% or higher) when data of 30 loci are used, and at such levels of expected heterozygosity a match at split level II implies a match at split level III, that is a perfect match; (viii) for values of $E(H)$ in the range 0·2–0·5, the recovery at a higher level of evolutionary bifurcation (Split level II) may be acceptable when data of 30 loci are used, even though there is a high proportion of imperfect matches at the level of the entire phylogenetic tree (split level III); (ix) at any level of expected heterozygosity,

efficiency of recovery is higher when expansion and contraction rates at microsatellite loci are equal than when these rates are unequal.

## 4. Discussion

The present study was prompted by the increasing use of microsatellite loci in reconstructing evolutionary relationships among contemporary human populations, and by the proliferation of distances measures (such as ASD, Δ, $D_{SW}$) designed specifically for use with data on microsatellite loci. We have performed simulations using a symmetric bifurcative evolutionary model. This bifurcation model seems likely to be valid for human subpopulations which commonly form through successive fissions over time (K. C. Malhotra, personal communication), although the symmetry may not hold in reality. Our simulation study has shown that the traditional distance measures $D_A$ (Nei *et al.* 1987) and $D_C$ (Cavalli–Sforza and Edwards 1973) are far superior for the purpose of capturing true evolutionary histories to those devised for use with data of microsatellite loci. Unless the expected heterozygosity is below 0·2 (that is, a small effective population size and a very low mutation rate), the recovery rate of true relationships with either measure is quite high. While this work was in progress, Takezaki and Nei (1996) published results of a study with similar conclusions. However, there are some notable differences between our results and those published by Takezaki and Nei (1996). Before pointing these out, we note that although Takezaki and Nei (1996) had also used eight populations in their study, their true phylogenetic tree was asymmetric. We have used a symmetric tree (figure 1). Previous studies (Nei 1987) indicate that, at least from allele frequency data of biochemical markers, the correct recovery of phylogenetic relationships is easier when the true tree is symmetric than when it is asymmetric. Thus our study is complementary to that of Takezaki and Nei (1996). In fact, since the true tree is never known in practice, the results of this study and those of Takezaki and Nei (1996) may serve as upper and lower bounds respectively. The two major differences between the results of these two studies are: (i) for a symmetric true tree the percentage of recovery of correct topology is many times higher than for an asymmetric tree, and (ii) $D_{SW}$ performs much better for a symmetric tree than for an asymmetric tree, although even for a symmetric tree its performance may not always be good. The one disturbing fact emanating from our results is that the efficiency of recovery of the correct tree topology is dependent on the ratio of expansion to contraction rates at microsatellite loci, although at higher values of expected heterozygosity there is virtually no dependence on this ratio. We have not investigated the consequences of variability of mutation rates over evolutionary time or along the various branches of the evolutionary tree. We suspect that the efficiency of reconstruction will decline when mutation rates are variable. Further, we have used only one method (UPGMA) of phylogenetic reconstruction. The major reasons for not using the currently popular neighbour-joining (NJ) method are that the true evolutionary tree that we have used is a rooted tree and we have used constant mutation rates. The neighbour-joining method produces an unrooted tree, and it is known that under the constant mutation rate model UPGMA provides the most satisfactory performance for reconstruction of rooted trees (Nei 1987).

Since microsatellite loci are abundant in the human genome and since these loci are highly polymorphic in most global populations, our study clearly shows that these loci

are very useful in recovering evolutionary relationships. Allele frequency data on about 30 microsatellite loci should suffice in human genome diversity studies in most regions of the world.

## Acknowledgements

This study was supported in part by a grant from the Department of Biotechnology, Government of India (to P.P.M.). We are grateful to Professor M. Nei and Dr N. Takezaki for helpful comments during the progress of this work, and also for making some of their unpublished results available to us.

## References

Bowcock A. M., Ruiz-Linares A., Tomfohrde J., Minch E., Kidd J. R. and Cavalli-Sforza L. L. 1994 High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368: 455–457

Cavalli-Sforza L. L. and Edwards A. W. F. 1967 Phylogenetic analysis: methods and estimation procedures. *Am. J. Hum. Genet.* 19: 233–257

Deka R., Jin L., Shriver M. D., Yu L. M., DeCroo S., Hundreiser J., Bunker C. H., Ferrell R. and Chakraborty R. 1995 Population genetics of dinucleotide $(dC–dA)_n$. $(dG–dT)_n$ polymorphisms in world populations. *Am. J. Hum. Genet.* 56: 461–474

Edwards A., Hammond H. A., Jin L., Caskey C. T. and Chakraborty R. 1992 Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* 12: 241–253

Ewens W. J. 1979 *Mathematical population genetics.* (Berlin: Springer)

Goldstein D. B., Ruiz-Linares A., Cavalli-Sforza L. L. and Feldman M. W. 1995a An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139: 463–471

Goldstein D. B., Ruiz-Linares A., Cavalli-Sforza L. L. and Feldman M. W. 1995b Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* 92: 6723–6727

Jeffreys A. J., Royle N. J., Wilson V. and Wong Z. 1988 Spontaneous mutation rates to new length alleles at tandem-repetitive loci in human DNA. *Nature* 332: 278–281

Jorde L. B., Bamshad M. J., Watkins W. S., Zenger R., Fraley A. E., Krakowiak P. A., Carpenter K. D., Soodyall H., Jenkins T. and Rogers A. R. 1995 Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am. J. Hum. Genet.* 57: 523–538

Nei M. 1987 *Molecular evolutionary genetics* (New York: Columbia University Press)

Nei M., Tajima F. and Tateno Y. 1983 Accuracy of estimated phylogenetic trees from molecular data. *J. Mol. Evol.* 19: 153–170

Ohta T. and Kimura M. 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in finite populations. *Genet. Res.* 22: 201–204

Shriver M. D., Jin L., Chakraborty R. and Boerwinkle E. 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* 134: 983–993

Shriver M. D., Jin L., Boerwinkle E., Deka R., Ferrell R. E. and Chakraborty R. 1995 A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.* 12: 914–920

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

Sneath P. H. A. and Sokal R. R. 1973 *Numerical taxonomy* (San Francisco: Freeman)

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

Valdes A. M., Slatkin M. and Freimer N. B. 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 133: 737–749

Weber J. L. and Wong C. 1993 Mutation of human short tandem repeats. *Hum. Mol. Genet.* 2: 1123–1128