ARTICLE

Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and DYS392

Almut Nebel[1,2], Dvora Filon[1], Carsten Hohoff[3], Marina Faerman[2], Bernd Brinkmann[3] and Ariella Oppenheim[1]

[1]Department of Hematology, Hebrew University, Hadassah Medical School; [2]Laboratory of Biological Anthropology and Ancient DNA, Hebrew University, Hadassah School of Dental Medicine, Jerusalem, Israel; [3]Institut für Rechtsmedizin, Westfälische Wilhelms-Universität, Münster, Germany

Deviation from the stepwise mutation model (SMM) at specific human microsatellite loci has implications for population genetic and forensic investigations. In the present study, data on six Y chromosome-specific microsatellites were pooled for 455 paternally unrelated males from six Middle Eastern populations. All chromosomes were assigned to three haplogroups defined by six binary polymorphisms. Two of the microsatellite loci tested, DYS388 and DYS392, displayed marked haplogroup-specific differences in their allele variability. A bimodal distribution of short and long alleles was observed for DYS388 in haplogroup 1 and for DYS392 in haplogroups 1 and 2. Further investigation showed that the short/long alleles segregated almost completely between genealogically distinct haplogroups defined by additional binary markers. Thus, these two loci have a discriminatory power similar to a binary polymorphism. DYS388 was characterised by an extremely low mutation rate in haplogroups 2 and 3, as was DYS392 in haplogroup 3. Sequence analysis of the repeat regions at the two loci revealed no irregularities, indicating that the triplet expansion in these loci is not controlled by sequence variation at the repeat level. A high frequency of long DYS388 alleles has, so far, been found only in populations originating in the Middle East, suggesting that this microsatellite is useful as a region-specific marker. *European Journal of Human Genetics* (2001) **9**, 22–26.

Introduction

Microsatellite loci (also termed short tandem repeats or STRs) on the human Y chromosome are increasingly being used for the study of male population history as well as for forensic analyses and paternity testing.[1,2] Both applications require a thorough understanding of the molecular mechanisms that have led to the observed mutational variability at the rapidly evolving loci. The mutation process operating at most microsatellites is thought to follow the stepwise mutation model (SMM), in which the mutations are generated by the gain or loss of a single repeat unit with equal probability.[3] The stepwise mutation model further postulates that the mutation rate is independent of the repeat number.[3] However, departure from the SMM has been demonstrated for various microsatellites in different species, including humans.[4–7] These observations indicate that the strict application of this model is an oversimplification of the complex mutational mechanisms underlying polymorphic microsatellites.

The human Y chromosome trinucleotide repeat locus DYS388 has been included in several population genetic surveys.[8–13] However, the behaviour of DYS388 appears to be inconsistent with the SMM, as was shown in two populations of Middle Eastern origin.[14,15] Additionally, another widely used microsatellite, DYS392, has recently been demonstrated to deviate from the SMM.[16] In the present study, both

microsatellite loci were analysed in a large pooled sample of six Middle Eastern populations.

Subjects and methods

Y chromosome haplotype data obtained from Israeli and Palestinian Arabs (I & P Arabs, n = 143;[17]). Moslem Kurds (n = 77) and Kurdish Jews (n = 67; manuscript in preparation) were analysed together with those already published on Ashkenazi (n = 68) and Sephardic Israelite Jews (n = 51;[14]) and Yemeni (n = 49;[15]). The DNA samples were typed for the six binary polymorphisms YAP (DYS287), 92r7, SRY4064, SRY + 465, sY81 (DYS271), Tat and the six microsatellite loci DYS19, DYS388, DYS390, DYS391, DYS392 and DYS393. The analysis followed the procedure described previously,[18] with the size of the alleles determined by automated scanning of the PCR products (ABI-310™ genetic analyser, Applied Biosystems, Foster City, CA, USA). Based on the six binary markers used, the Y chromosomes were assigned to three different haplogroups as defined earlier[14] (for haplogroup definitions, see Figure 1). Two Yemeni chromosomes, which were found to belong to a fourth haplogroup (YAP+ 92r7-C, SRY4064-A, SRY + 465-C, sY81-G, Tat-T), were excluded from the study.

To obtain a representative sample number for each haplogroup and to achieve greater statistical power, the chromosomes from all six populations (a total of 455 paternally unrelated individuals) were pooled. Within the three haplogroups, allele distribution and variability at all six microsatellite loci were examined. As a measure of variability two parameters were used:[19] locus diversity D (calculated as gene diversity) and repeat number variance σ^2. D and its sampling variance were estimated as described.[20]

For sequence analysis, amplification of DYS388 (GDB accession no. G00-364-862) and DYS392 (GDB accession no. G00-455-698) alleles was performed in singleplex PCRs, using the conditions reported.[18] Following separation on a 1.5% NuSieve (FMC Bioproducts, Rockland, ME, USA) agarose gel, the PCR products were cut from the gel, melted at 68°C and 2-μl aliquots were directly subjected to a ^{33}P terminator cycle sequencing (Amersham Pharmacia Biotech, Piscataway, NJ, USA) with the DYS388 L- and the DYS392 R-PCR primer, respectively.

Results

The allele distribution at DYS388 was plotted separately for each of the three haplogroups 1, 2 and 3 defined by the allelic states at six binary polymorphisms (Figure 1a). DYS388 was extremely invariable in haplogroups 2 and 3, displaying almost exclusively alleles with 12 repeats. In haplogroup 1, the alleles ranged from 10 to 19, exhibiting an evident non-Gaussian frequency distribution with two peaks, at 12 and 16 repeats. Also DYS392 seems to deviate from the SMM (Figure 1b). In haplogroups 1 and 2, allele 12 is either under-represented or completely missing, resulting in a two-peak allele distribution with mainly 11 and 13 repeats. This observation is in agreement with a report on a global survey of 48 populations (more than 3000 chromosomes) which has shown that the DYS392 allele 12 is either absent or less frequent than the adjacent alleles.[21] In haplogroup 3, the prevalent DYS392 allele is 11. In contrast, the other four microsatellites studied (the tetranucleotide loci DYS19, DYS390, DYS391 and DYS393) revealed, in each of the haplogroups, a pattern compatible with the SMM: a uni-modal allele distribution with one dominant and less frequent neighbouring alleles. The patterns seen in the pooled sample were also observed for each of the six populations separately (data not shown).

To obtain a quantitative estimate of the microsatellite variability, we computed locus diversity (D) and repeat number variance (σ^2) at DYS388 and DYS392 and compared the results with those of the other four microsatellites (Table 1). The results showed that both D and σ^2 displayed marked haplogroup-specific differences for DYS388 and DYS392, but not for the other four loci. The very low diversity and variance values for DYS388 in haplogroups 2 and 3 and for DYS392 in haplogroup 3 indicate greatly reduced mutation rates that are also seen in the histograms (Figures 1a and b). The high diversity values for DYS388 in haplogroup 1 and for DYS392 in haplogroup 2 suggest an elevated mutation rate. The corresponding extremely high variance values of 2.4 and 3.9, respectively, appear to reflect the bimodal allele distribution observed for both loci. The mean D and variance values calculated for the four loci DYS19, DYS390, DYS391 and DYS393 did not differ across the three haplogroups, probably reflecting similar haplogroup ages.

Allele-specific mutation rates at a single microsatellite locus have been linked to differences in the number, arrangement or sequence of the repeats.[5] Investigations of trinucleotide repeat expansions in genes associated with inherited human diseases (SCA1 and FMR 1) revealed that the interruption of homogenous repeat stretches by irregular units reduces the mutation rate.[22,23] A similar mechanism has been suggested to operate at microsatellites in general.[24,25] To examine whether the skewed allele distributions at DYS388 and DYS392 are caused by repeats differing in their sequence, the triplet repeat regions of both loci were analysed by direct sequencing.

DYS388 was sequenced in a total of 14 Y chromosomes obtained from Arab individuals representing all three haplogroups. The size of the sequenced repeats ranged from 10 to 18, including 8 chromosomes of haplogroups 2 and 3 with repeat number 12. DYS392 was analysed in 12 DNA samples of Moslem Kurds with alleles ranging from 10 to 16 repeats, including also three chromosomes of haplogroup 3 with 11 repeats. The results showed that apart from differences in the repeat number, the 14 DYS388 and the 12 DYS392 sequences were identical throughout. Importantly, no irregularity in the repeats was observed. This
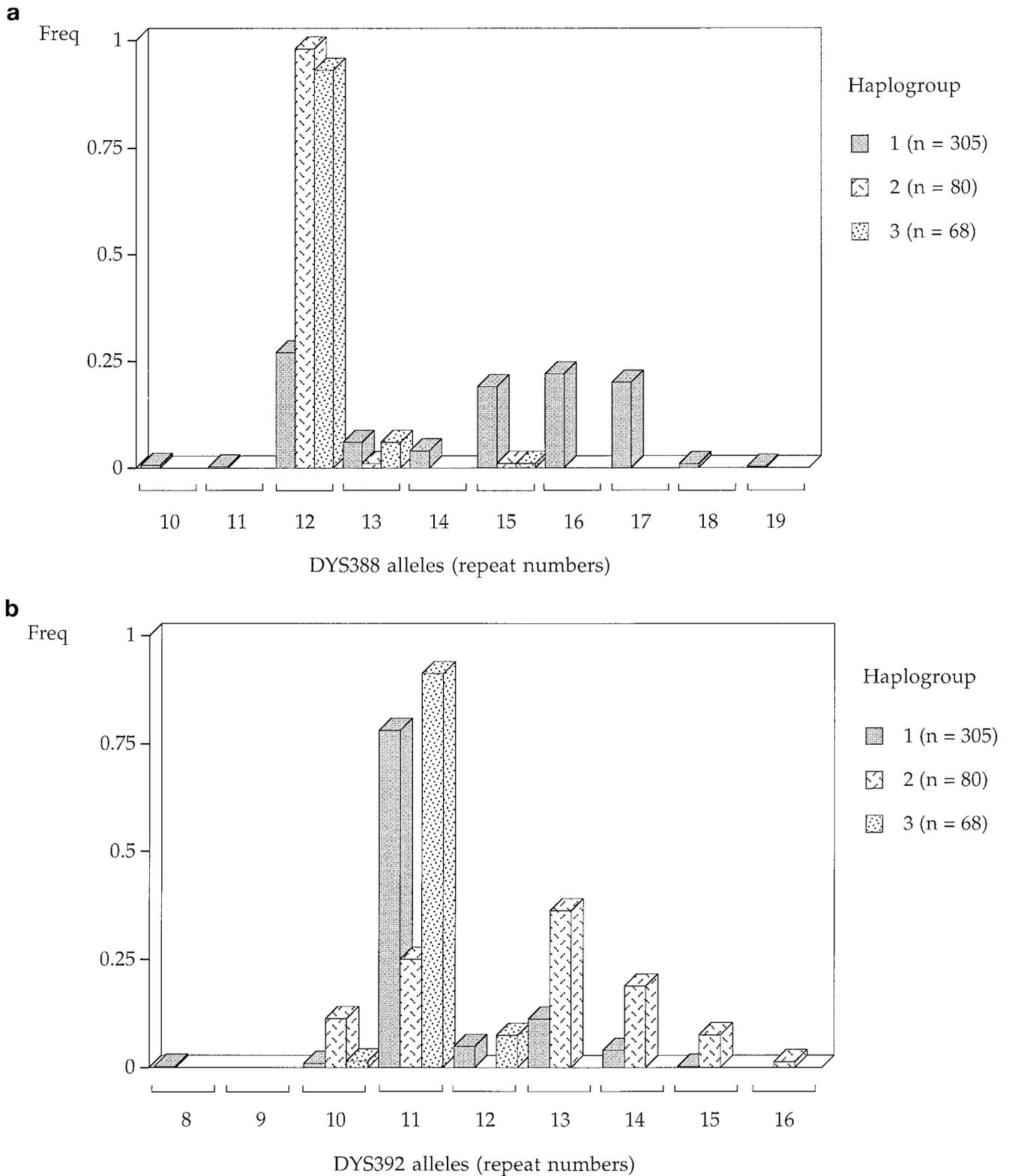
a

b

Figure 1 Allele distribution **a)** at DYS388 and **b)** at DYS392 in three haplogroups. Y chromosomes were pooled from six Middle Eastern populations. The haplogroups are defined by the allelic states at six binary polymorphisms as follows: haplogroup 1: YAP⁻, 92r7-C, SRY4064-G, SRY + 465-C, sY81-A, Tat-T; haplogroup 2: YAP⁻, 92r7-T, SRY4064-G, SRY + 465-C, sY81-A, Tat-T; haplogroup 3: YAP⁺, 92r7-C, SRY4064-A, SRY + 465-C, sY81-A, Tat-T.

Table 1 Locus diversity (D) and repeat number variance (σ^2)

Locus	Haplogroup 1		Haplogroup 2		Haplogroup 3	
	D	σ^2	D	σ^2	D	σ^2
DYS19	0.516±0.027	0.423	0.661±0.047	1.189	0.598±0.049	0.873
DYS388	0.802±0.008	3.946	0.050±0.034	0.124	0.140±0.056	0.183
DYS390	0.673±0.020	0.955	0.650±0.044	0.828	0.708±0.030	0.941
DYS391	0.450±0.027	0.304	0.504±0.012	0.252	0.424±0.063	0.255
DYS392	0.373±0.034	0.914	0.762±0.024	2.429	0.166±0.058	0.086
DYS393	0.531±0.026	0.588	0.613±0.029	0.466	0.430±0.057	0.246
Mean[a]	0.543±0.094	0.568±0.283	0.607±0.072	0.684±0.412	0.540±0.138	0.579±0.380

[a]Mean D and σ^2 values are based on the four loci DYS19, DYS390, DYS391 and DYS393 within each haplogroup.

finding indicates that the triplet expansion in these loci is not controlled by sequence variation at the repeat level. Furthermore, the sequencing data confirmed the number of repeats determined by the GeneScan analysis. The obtained sequences were the same as those in the 'Database of the Forensic Laboratory for DNA Research' at the University of Leiden.[26] To date, no uniform microsatellite nomenclature has been agreed upon. According to the recommendations of the International Society of Forensic Haemogenetics (ISFH, now International Society of Forensic Genetics, ISFG),[27] which were also adopted in this study, the DYS388 repeat is $(ATT)_n$ and not $(ATA)_n$, as was previously reported, see Kayser *et al*[2] (Appendix), de Knijff *et al*[21] (Appendix). It should be noted that when the DYS388 trinucleotide repeat is read as 'ATT', a PCR product of, for example, 128 bp corresponds to 12 repeats, and not to 11 when read as 'ATA'. Similarly, following the ISFG guidelines, the DYS392 repeat is $(TAT)_n$ and not $(ATT)_n$. Either reading results in the same repeat number.

Discussion

There is ample evidence from investigations in yeast, *Drosophila* and humans that the mode and tempo of microsatellite evolution are dependent on the repeat number at a given allele.[4,5,7,19] Short alleles with 10 triplets or fewer appear to mutate extremely seldom in human microsatellites.[5] This could be caused by physical constraints imposed by the DNA replication machinery. Thus, the reduced mutation rate at DYS388 in haplogroups 2 and 3 and at DYS392 in haplogroup 3 may be due to a low repeat number allele on the founder chromosomes. Alternatively, sequence variations and repetitive structures in the flanking region of several microsatellite loci have been suggested to influence the expansion of homogeneous stretches.[5] Such differences, though not yet detected at DYS388 or DYS392, may also underlie the haplogroup-specific mutation rates at the two loci. In addition, alleles of a certain size may be inherently unstable on a particular haplogroup background (eg DYS392 allele 12 in haplogroup 2). Regardless of the molecular mechanism, the resulting low mutation rate allows us to suggest that some of the alleles prevalent today, such as the DYS388 allele 12 in haplogroups 2 and 3 or the DYS392 allele 11 in

haplogroup 3, represent the founder alleles of these lineages.

The three haplogroups do not seem to differ widely in their time of origin, as seen in the similar mean diversity and variance values across the haplogroups. The older a lineage, the higher will be its average Y chromosome microsatellite variance, until the mutation-drift equilibrium is reached.[28] The average variances observed in our data (Table 1) are much lower than the expected value (2.52) at mutation-drift equilibrium.[28] Thus, the high variability of DYS388 and DYS392 in haplogroups 1 and 2, respectively, cannot be attributed to an older age of the lineages.

The distribution of the short and long alleles of DYS388 in haplogroup 1 and of DYS392 in haplogroups 1 and 2 suggests that they belong to genealogically distinct haplogroups. This hypothesis was confirmed in a recent study (ref. 17 and manuscript in preparation) in which the three haplogroups were further defined by five additional binary polymorphisms (M9, M13, M17, M20 and SRY10831). Our results showed that the long and short alleles of both DYS388 and DYS392 segregated almost completely between these newly defined haplogroups. We postulate that complete segregation will be achieved once a high-resolution haplogroup classification is available.

A high frequency of long DYS388 alleles has, so far, been found only in populations originating in the Middle East[15] and appears to be characteristic for the large majority of haplotypes in haplogroup 1, defined by a C at M9 (haplogroup 1A). The only other haplogroup known today with a high incidence of long DYS388 alleles is the Y chromosome lineage Hg 9.[12] This haplogroup is characterised by the p12f2-8kb allele[29] on a genetic background compatible with haplogroup 1A.[12] Hg 9 has also primarily a Mediterranean distribution, with its main concentration in the Middle East.[30] Therefore, we hypothesise that haplogroups 1A and Hg 9, the only ones known to exhibit long DYS388 alleles, are closely related or even identical. We suggest that the restriction of high repeat DYS388 alleles to genealogical lineages (haplogroup 1A/Hg 9) with a limited geographic distribution gives this microsatellite the discriminatory power similar to a binary polymorphism and renders it useful as a haplogroup- and region-specific marker.

In conclusion, our study shows that the deviation from the SMM at DYS388 and DYS392 is haplogroup-specific. This finding should be taken into consideration when these microsatellites are applied to population genetic studies and/or forensics.

References

1 Jobling M, Pandya A, Tyler-Smith C: The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med* 1997; **110**: 118–124.

2 Kayser M, Caglià A, Corach D *et al*: Evaluation of Y-chromosomal STRs: A multicenter study. *Int J Legal Med* 1997; **110**: 125–133.

3 Freimer NB, Slatkin M: Microsatellites: evolution and mutational processes. In: *Variation in the Human Genome*. Wiley: Chichester, 1996; pp 51–72.

4 Wierdl M, Dominska M, Petes T: Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* 1997; **146**: 769–779.

5 Brinkmann B, Klintschar M, Neuhuber F, Hühne J, Rolf B: Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* 1998; **62**: 1408–1415.

6 Forster P, Kayser M, Meyer E *et al*: Phylogenetic resolution of complex mutational features at Y-STR DYS390 in aboriginal Australians and Papuans. *Mol Biol Evol* 1998; **15**: 1108–1114.

7 Schlötterer C, Ritter R, Harr B, Brem G: High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Mol Biol Evol* 1998; **15**: 1269–1274.

8 Pérez-Lezaun A, Calafell F, Seielstad M *et al*: Population genetics of Y-chromosome short tandem repeats in humans. *J Mol Evol* 1997; **45**: 265–270.

9 Pérez-Lezaun A, Calafell F, Comas D *et al*: Sex-specific migration patterns in central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* 1999; **65**: 208–219.

10 Kittles R, Perola M, Peltonen L *et al*: Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet* 1998; **62**: 1171–1179.

11 Kittles R, Bergen A, Urbanek M *et al*: Autosomal, mitochondrial and Y chromosome DNA variation in Finland: evidence for a male-specific bottleneck. *Am J Phys Anthropol* 1999; **108**: 381–399.

12 Bosch E, Calafell F, Santos F *et al*: Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. *Am J Hum Genet* 1999; **65**: 1623–1638.

13 Sasaki M, Shioni H: The polymorphisms of DYS388 and DYS392 on the Y chromosome in Japanese and German populations. *Int J Legal Med* 1999; **112**: 132–133.

14 Thomas M, Skorecki K, Ben-Ami H, Parfitt T, Bradman N, Goldstein D: Origins of Old Testament priests. *Nature* 1998; **394**: 138–140.

15 Thomas M, Parfitt T, Weiss D *et al*: Y chromosomes travelling South: The Cohen Modal Haplotype and the origins of the Lemba – The 'Black Jews of Southern Africa'. *Am J Hum Genet* 2000; **66**: 674–686.

16 Forster P, Röhl A, Lünnemann P *et al*: A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet* 2000; **67**: 182–196.

17 Nebel A, Filon D, Weiss DA *et al*: High resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews. *Hum Genet*, 2000; **107**: 630–641.

18 Thomas M, Bradman N, Flinn H: High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y chromosome. *Hum Genet* 1999; **105**: 577–581.

19 Carvalho-Silva D, Santos F, Hurtz M, Salzano F, Pena S: Divergent human Y-chromosome microsatellite evolution rates. *J Mol Evol* 1999; **49**: 204–214.

20 Nei M: *Molecular Evolutionary Genetics*. Columbia University Press: New York, 1987.

21 de Knijff P, Kayser M, Caglià A *et al*: Chromosome Y microsatellites: Population genetic and evolutionary aspects. *Int J Legal Med* 1997; **110**: 134–140.

22 Chung M-Y, Ranum L, Duvick L, Servadio A, Zoghbi H, Orr H: Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nat Genet* 1993; **5**: 254–258.

23 Eichler E, Holden J, Popovich B *et al*: Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nat Genet* 1994; **8**: 88–94.

24 Garza J, Slatkin M, Freimer N: Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size. *Mol Biol Evol* 1995; **12**: 594–603.

25 Goldstein D, Clark A: Microsatellite variation in North American populations of *Drosophila melanogaster*. *Nucleic Acids Res* 1995; **23**: 3882–3886.

26 Forensic Laboratory for DNA Research, Leiden University Medical Center, the Netherlands: http://ruly70.medfac.leidenuniv.nl/~fldo.

27 Bär W, Brinkmann B, Budowle B *et al*: DNA recommendations. Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. *Int J Legal Med* 1997; **110**: 175–176.

28 Goldstein D, Zhivotovsky L, Nayar K, Ruiz Linares A, Cavalli-Sforza L, Feldman M: Statistical properties of the variation at linked microsatellite loci: Implications for the history of human Y chromosomes. *Mol Biol Evol* 1996; **13**: 1213–1218.

29 Casanova M, Leroy P, Boucekkine C *et al*: A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* 1985; **230**: 1403–1406.

30 Semino O, Passarino G, Brega A, Fellous M, Santachiara-Benerecetti A: A view of Neolithic demic diffusion in Europe through two Y chromosome-specific markers. *Am J Hum Genet* 1996; **59**: 964–968.