European Journal of Human Genetics (1998) 6, 583–588

http://www.stockton-press.co.uk/ejhg

ORIGINAL PAPER

Polymorphism at the tetranucleotide repeat locus *DYS389* in 10 populations reveals strong geographic clustering

Burkhard Rolf[1], Eckhard Meyer[2], Bernd Brinkmann[1] and Peter de Knijff[3]

[1]*Institut für Rechtsmedizin, Westfälische-Wilhelms Universität, Münster, Germany*
[2]*Institut für Rechtsmedizin, Friedrich-Schiller-Universität, Jena, Germany*
[3]*Leiden University Medical Center, Leiden, The Netherlands*

Several short tandem repeat polymorphism loci at the non-recombining part of the Y chromosome have been described recently and are now widely used for the investigation of the history and the diversity of man.[1–5] The tetranucleotide repeat polymorphism at the *DYS389* locus consists of two repetitive stretches with different numbers of $(TCTG)_n$ $(TCTA)_m$ repeat units. To study the overall variability of this locus, 768 alleles from males from 10 human populations (two sub-Saharan African, four Caucasoid and four Asian/Amerind populations) were investigated. The alleles found in the populations of different geographic origin exhibited remarkable differences in the number and arrangement of repeats in the two repetitive stretches and up to nine different sequence variants for a single fragment length have been detected. So far 53 different alleles, ie haplotypes, have been observed. Analysis of molecular variance (AMOVA) indicates that at least 24.5% of the total genetic variance was found between the populations and that these differences were significant in most pairwise comparisons. We propose a model, in which both founder effects and genetic drift together with single step replication slippage mutations explain the picture of haplotype diversity observed with this single locus.

Keywords: tetranucleotide; DYS389; males; network; y chromosome

Introduction

Autosomal short tandem repeat (STR) systems are used in human identification/parentage,[6] linkage analysis,[7] and for the investigation of human evolution.[8] Recently, microsatellites at the male-specific part of the Y chromosome have been described and are now increasingly used for the study of male-specific lineage evolution.[4,5,9] In contrast to autosomal microsatellites, variations observed in several loci can be combined and studied via haplotype analysis, because no recombination occurs (as long as one only considers variation outside the pseudo autosomal region). As there is only one Y chromosome to every four autosomes, genetic variations on the Y chromosome are much more prone to effects of drift. This is also true of Y haplotypes in relation to mtDNA haplotypes, as the reproduction success of males can differ strongly, whereas the

Correspondence: Prof. Dr. B. Brinkmann, Institut für Rechtsmedizin, Westfälische-Wilhelms Universität, von Esmarch Straße 62, D-48129 Münster, Germany. Tel.: 49 251 8355160; Fax: 49 251 83 55158
Received 10 December 1997; revised 23 March 1998; accepted 15 April 1998

number of offspring of females is probably less variable. Paternal inheritance allows investigation of male lineages over many generations.[10]

The tetranucleotide repeat polymorphism *DYS389*[11] consists of two stretches with different numbers of $(TCTG)_n$ $(TCTA)_m$ repeat units, separated by 48 base pairs. To study the variability of this locus, 768 alleles from males from 10 populations were investigated either by sequencing or by a nested PCR approach.

Materials and Methods

Population Samples

The population samples were Ovambos (Bantu) from Namibia, Pygmies from Central Africa, Caucasoid Germans from northwest Germany, Dutch males from all over Holland, Moroccans living in Brussels, Belgium, Turks from the Adana area, Chinese (Han race) from the Shenyang area, Japanese from the Shiga area, Inuit from southwest Greenland, and Surinamese from either Surinam or living in Holland.

Determination of Haplotypes

DNA was isolated from dried blood on cotton according to standard procedures[12] and quantified using the slot blot technique. [13] Fragment length determination was carried out on an ABI-373-DNA-Sequencer (primer R1 FAM and R4N HEX labelled) or an ALF-system (Pharmacia, Uppsala, Sweden, primer F2 FITC-labelled). The PCR-primer sequences were:

F1: CCAACTCTCATCTGTATTATCTAT

F2: TCTGTATTATCTATGTGTG

R1: TCTTATCTCCACCCACCAGA

R2: CCAGACATTGCCAAGTGTTACTTG

R3: ATACAGAAGTAGGTATAA

R4N: ATAAATAATATAAAATATAAATAAATAAA TATAGATAGACAG

Amplification was carried out with 2 ng of template DNA, 1 unit Taq-DNA-polymerase, 200 μM of each dNTP, 2 μl 10 × PCR reaction buffer containing 500 mM KCI, 100 mM Tris/HCI, 15 mM $MgC1_2$, 1% Triton X-100, pH 8.8 at 25°C, 250 μg BSA and 1 μM of each primer in a final volume of 25 μl. Amplification conditions were 94°C for 1 min followed by 30 cycles of 94°C for 45 s, 55°C for 20 s and 72°C for 30 s.

Alleles from males were sub-typed either by sequencing or by a nested PCR approach. Sequencing was carried out in both directions with the product of the PCR3 reaction on an ABI-373-DNA-Sequencer using the Taq Dye-Deoxy-Terminator Cycle Sequencing Kit (Perkin Elmer/Applied Biosystems, Foster City, CA, USA). Every sequence was determined in both directions with primer F2 and R2, respectively. For the semi-nested PCR approach, 1 μL of a 1 to 100 dilution of the product of PCR1 (Figure 1) was used as template in a semi-nested reaction PCR2 (Figure 1) using the same PCR conditions. The fragment length of the resulting four fragments was determined on the ABI-system. The haplotypes were obtained by the subtraction of the fragment lengths of the four different products. Alternatively, 1 μL of a 1 to 100 dilution of the product of PCR3 (Figure 1) was used as a template in the semi-nested reaction PCR4 (Figure 1). The fragment length of the resulting fragments was determined on the ALF-system. Together with the information from PCR5, the haplotypes could again be obtained by subtraction of the fragment lengths of the four different products. Both methods were calibrated with 30 sequenced alleles.

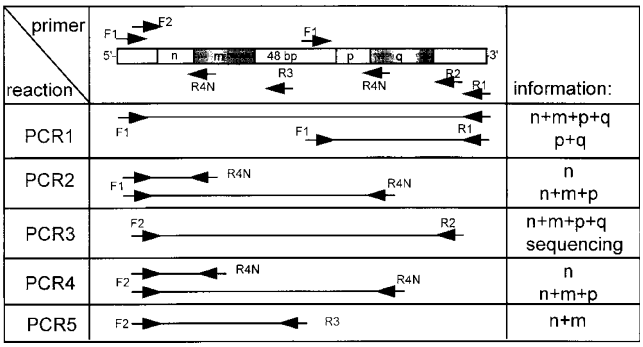

Figure 1 *Sequence structure, primer binding sites and PCR products generated from locus* DYS389. *For primer sequences and PCR conditions see Materials and Methods.*

Results

The initially developed set of primers for this locus [Genome Database, ID: G00-366-108] yielded two PCR fragments because one primer had two annealing sites.[1] To allow for direct sequencing and further sub-typing of the PCR-products, we redesigned the primers in order to obtain only one product. Fragment length determination of the resulting single PCR product of males from two sub-Saharan African populations, four Caucasoid populations and four Asian/Amerind populations gave 9 length variants from 288 to 320 bp.

The structure of all alleles is $(TCTG)_{4–7}$ $(TCTA)_{10–14}(48\,bp)(TCTG)_3(TCTA)_{8–13}$ and is possibly the result of duplication of one initial repeat motif; 53 different alleles, ie haplotypes, were observed (Table 1). The differences arise from the number and arrangement of repeats in the two repetitive stretches (Table 1). Up to nine different variants for single fragment length have been detected. The allele described by Cooper *et al*[1] (1996, EMBL accession No. 97312) consisting of $(TCTG)_5(TCTA)_{12}(48\,bp)$ $(TCTG)_3(TCTA)_{10}$ was found 58 times in seven of our 10 population samples (Table 1), haplotype #25); additionally we found eight other haplotypes of the same length but with different arrangement of the repeats. The most frequent haplotype in our study was $(TCTG)_5(TCTA)_{11}(48\,bp)(TCTG)_3(TCTA)_{10}$. It was

Table 1 Repeat structure and distribution in the populations of the observed 53 haplotyes

No.	*n+m+q*	*n*	*m*	*q*	*Dut*	*Sur*	*Pyg*	*Inu*	*Mar*	*Tur*	*Ger*	*Jap*	*Chi*	*Ova*	*All*
1	23	4	10	9	0	1	0	0	0	0	0	0	0	0	1
2	24	4	11	9	0	0	0	0	0	0	0	0	4	0	4
3	25	4	11	10	0	0	0	0	0	1	0	1	3	0	5
4	26	4	11	11	0	0	0	0	0	2	0	3	0	0	5
5	24	4	12	8	0	0	0	0	0	0	0	2	1	0	3
6	25	4	12	9	0	1	0	0	0	0	0	2	6	0	9
7	26	4	12	10	0	1	0	0	0	0	1	3	8	0	13
8	27	4	12	11	0	2	0	0	0	0	4	5	5	0	16
9	28	4	12	12	0	0	0	0	0	0	0	1	0	0	1
10	26	4	13	9	0	2	0	0	0	1	0	5	9	0	17
11	27	4	13	10	0	3	0	0	0	3	0	2	3	0	11
12	28	4	13	11	0	1	0	0	0	2	0	1	1	0	5
13	27	4	14	9	1	0	0	0	0	0	0	2	5	0	8
14	28	4	14	10	0	0	0	0	0	0	0	0	1	0	1
15	24	5	9	10	1	0	0	0	0	0	1	0	0	0	2
16	24	5	10	9	0	1	0	0	0	1	2	0	1	0	5
17	25	5	10	10	9	1	0	0	0	4	5	0	0	0	19
18	26	5	10	11	1	1	0	0	0	1	2	1	0	0	6
19	24	5	11	8	0	0	1	0	0	1	1	0	0	0	3
20	25	5	11	9	18	0	6	4	1	9	26	1	0	0	65
21	26	5	11	10	31	8	3	8	4	23	81	2	10	0	170
22	27	5	11	11	8	3	0	20	2	12	10	5	1	0	61
23	25	5	12	8	0	0	3	0	0	0	0	0	0	0	3
24	26	5	12	9	3	1	0	1	5	7	10	0	0	1	28
25	27	5	12	10	9	2	0	1	0	11	31	2	2	0	58
26	28	5	12	11	2	4	0	14	0	5	5	8	1	0	39
27	29	5	12	12	0	1	0	9	0	0	1	2	0	0	13
28	30	5	12	13	0	0	0	1	0	0	0	0	0	0	1
29	27	5	13	9	2	0	1	0	1	1	2	0	1	0	8
30	28	5	13	10	1	0	0	1	1	11	13	1	1	0	29
31	29	5	13	11	0	0	0	3	0	2	1	2	0	0	8
32	29	5	14	10	0	1	0	0	0	1	1	0	0	0	3
33	30	5	14	11	0	0	0	0	0	1	1	0	0	0	2
34	31	5	15	11	0	0	0	0	0	1	0	0	0	0	1
35	24	6	9	9	0	0	0	0	0	0	0	0	0	1	1
36	25	6	9	10	0	0	0	0	0	0	1	0	0	0	1
37	25	6	10	9	0	1	2	0	0	0	0	0	0	2	5
38	26	6	10	10	1	0	3	0	6	0	0	0	0	6	16
39	27	6	10	11	0	0	0	0	16	0	1	0	0	0	17
40	28	6	10	12	0	0	0	0	2	0	0	0	0	0	2
41	29	6	10	13	0	0	0	0	1	0	0	0	0	0	1
42	26	6	11	9	0	1	1	0	0	2	2	0	0	1	7
43	27	6	11	10	2	9	4	0	2	4	2	0	0	20	43
44	28	6	11	11	0	2	0	0	1	3	1	0	0	1	8
45	27	6	12	9	0	2	0	0	0	2	2	0	0	3	9
46	28	6	12	10	0	4	2	0	1	1	0	0	0	8	16
47	29	6	12	11	0	2	0	0	0	0	2	0	0	2	6
48	28	6	13	9	0	0	0	0	1	0	1	0	0	0	2
49	29	6	13	10	0	0	1	0	0	1	0	0	0	2	4
50	29	7	11	11	0	0	1	0	0	0	0	0	0	1	2
51	30	7	11	12	0	0	2	0	0	0	0	0	0	0	2
52	28	7	12	9	0	0	0	0	0	1	0	0	0	0	1
53	29	7	12	10	0	0	0	0	0	1	0	0	0	1	2
SUM					89	55	30	62	44	115	210	51	63	49	768

n, m and q are the number of repeats in the three variable repetitive stretches of the locus: $(TCTG)_n$ $(TCTA)_m$ (48 bp) $(TCTG)_p$ $(TCTA)_q$.

found 170 times and was present in nine of our 10 population samples. Other haplotypes were unique in our study and seen only once. To investigate relations between the haplotypes observed, a network was constructed.[1,2,14] Assuming a single step mutation model,[15,16] haplotypes differing by only one repeat are called adjacent and connected by a line. Due to the three variable parameters, the resulting network is a three-dimensional figure. Haplotypes sharing the number of repeats in one motif type, eg four in the most 5'-repetitive region, are placed in a plane. All 53 haplotypes observed could be connected in this way. Seven haplotypes are connected with six others, the maximum number of connections that could be achieved in this model. Four haplotypes are only connected once (Figure 2).

Figure 2 highlights the individual haplotypes found in the respective populations. Almost all population-specific haplotypes can be connected via the single step lines; however these population-specific nets do not overlap to a great extent. The Asian/Inuit haplotypes all start with four or five TCTG repeats in the most 5'-repeat cases. The haplotypes of Caucasoid males (Dutch, Germans, Turks and Moroccans) all start with

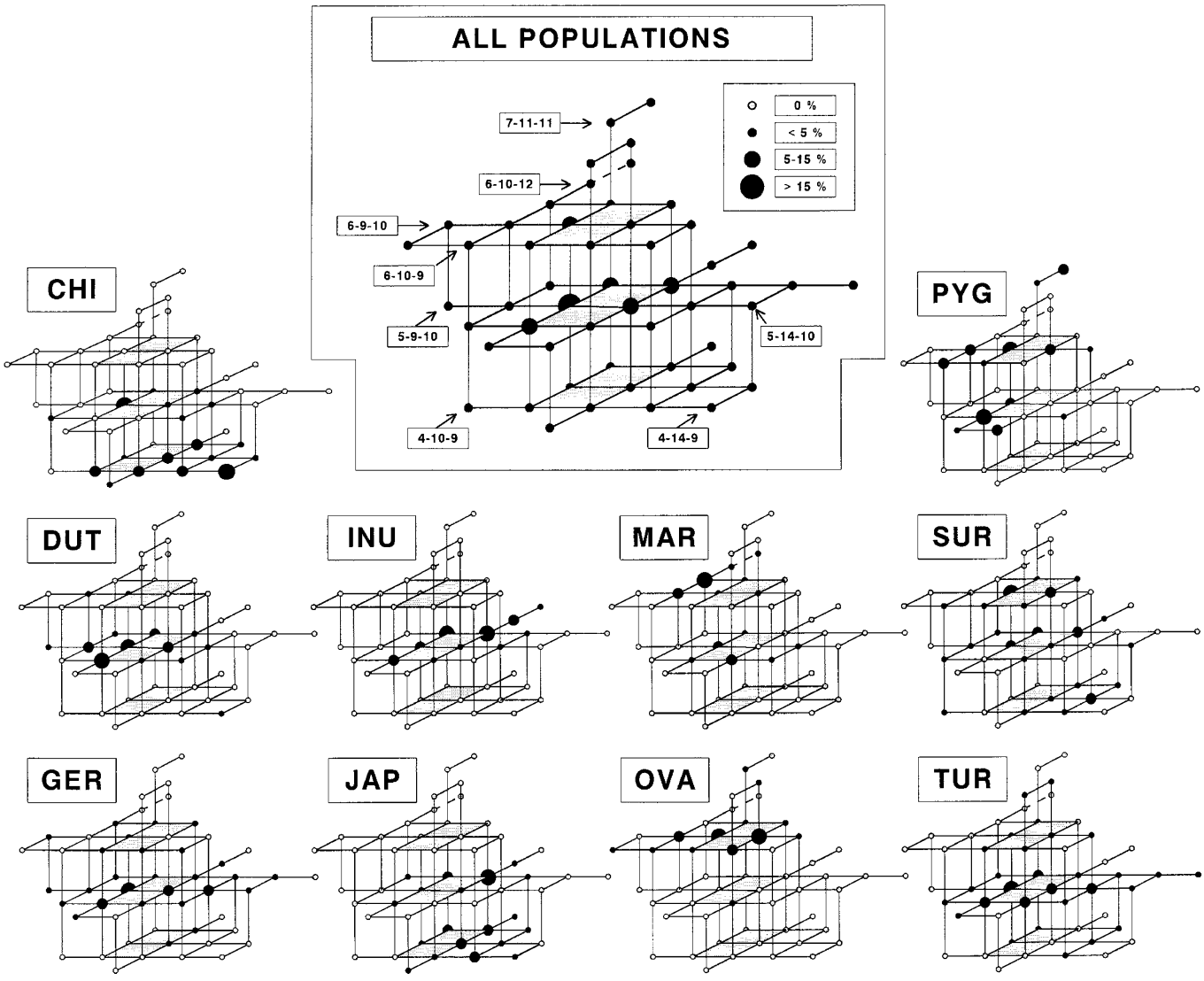

Figure 2 *Minimum spanning network of haplotypes of* DYS389 *locus. All 53 haplotypes observed in this study are shown. Haplotypes differing by only one repeat are called adjacent and connected by a line. The resulting networks indicate possible transitions caused by single step mutations. Every dimension in the figure indicates length changes in one of the three regions. Frequency of the haplotypes is indicated by the size of the nodes.*

Table 2 PhiST-distances between pairs of populations (below diagonal). Above diagonal is the probability for the statistical significance of the observed distance, derived from 10 000 iterations

1	*Dut*	*Sur*	*Pyg*	*Inu*	*Mor*	*Tur*	*Ger*	*Jap*	*Chi*	*Ova*
Dut		0.0005	0.0002	0.0000	0.0000	0.0000	0.0077	0.0000	0.0000	0.0000
Sur	0.0869		0.0000	0.0000	0.0001	0.5293	0.0345	0.0000	0.0000	0.0000
Pyg	0.1385	0.1327		0.0000	0.0018	0.0000	0.0000	0.0000	0.0000	0.0188
Inu	0.3650	0.1517	0.4382		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Mor	0.2485	0.2006	0.1626	0.3133		0.0000	0.0000	0.0000	0.0000	0.0002
Tur	0.0947	–0.0041	0.1996	0.1852	0.2771		0.0089	0.6040	0.0000	0.0000
Ger	0.0300	0.0251	0.1742	0.2938	0.3104	–0.0231		0.0000	0.0000	0.0000
Jap	0.3318	0.1563	0.3976	0.1983	0.4430	0.1344	0.2576		0.0002	0.0000
Chi	0.3807	0.2681	0.4317	0.4419	0.5509	0.2181	0.3188	0.1020		0.0000
Ova	0.3309	0.2208	0.0669	0.4925	0.1512	0.2689	0.3194	0.5206	0.5867	

four one (one only), five and six TCTG-repeats, whereas the sub-Saharan African samples start with five, six or seven (Table 1)

Discussion

To explain the variability and population specificity observed with this locus, we assume the presence of alleles with four, five, six and seven TCTG-repeats in an ancestral population. Migration processes have spread males of this ancestral population around the world. Founder effects and genetic drift will probably have reduced the initial diversity (eg chromosome number) and thus the variation of the number of TCTG-repeats in modern populations. The haplotype diversity observed within populations today is caused mainly by the number of repeats in both $(TCTA)_n$ motifs of the locus. Therefore new mutations are more likely to occur in the longer TCTA strand than in the shorter 5'-TCTG strand. Interestingly, the autosomal tetranucleotide repeat in the intron 31 of the von Willbrandt Factor gene (VWA) has a similar repeat structure of $(TCTG)_4(TCTA)_{8–17}$. Variation in length of this STR is almost exclusively due to variations in the number of TCTA repeats, indicating the higher mutability of this repeat motif. The fact that the frequent haplotypes are observed in the centre of the network further supports the single step mutation model, as does the study by Heyer *et al*.[17] who investigated Y chromosomal STRs in deep rooted pedigrees in order to estimate mutation rates. All mutations observed in this study were single-step mutations; the mutation rate was as high as 0.21% per generation. Therefore the haplotype diversity and the geographic clustering observed is a result of both founder effects and genetic drift together with a high rate of (single-step) mutations.

In order to compare differences in haplotype distribution and the amount of genetic variation included in the Y haplotypes we followed the suggestions of Excoffier *et al*[18,19] and Michalakis and Excoffier[20] and performed an 'analysis of molecular variance', (AMOVA) as previously explained in detail by us.[3] The WINAMOVA and ARLEQUIN programs were used for these analyses. Surprisingly, 24.51% of the variance was inter-population and the remaining 76.49% intra-population difference. Bootstrapping of the PhiST distance between the populations revealed that 38 of the 45 comparisons were statistically significant (Table 2).

Interestingly, others using network approaches for the investigation of Y chromosomal haplotypes[1,2] did not observe strong geographic clusters. These authors assumed that high rates of new mutations randomised the haplotypic structures. The data presented in this work allow us to define population-specific haplotypes, because sequencing revealed that part of the locus is highly conserved. Thus, locus *DYS389* will become a useful tool for the investigation of human evolution.

References

1 Cooper G, Amos W, Hoffman D, Rubinsztein DC: Network analysis of human Y microsatellite haplotypes. *Hum Mol Genet* 1996; **5**: 1759–1766.
2 Deka R, Jin L, Shriver MD, Yu LM *et al*: Dispersion of human in chromosome haplotypes based on five microsatellites Y global populations. *Genome Res* 1996; **6**: 1177–1184.
3 Roewer L, Kayser M, Dieltjes P *et al*: Analysis of molecular variance (AMOVA) of Y chromosome specific microsatellites in two closely related human populations. *Hum Mol Genet* 1996; **5**: 1029–1033.
4 Lineares AR, Nayer K, Goldstein DB *et al*: Geographic clustering of human Y chromosome haplotypes. *Ann Hum Genet* 1996; **60**: 401–408.
5 Sajantila A, Salem A-H, Savolainen P, Bauer K, Gierig C, Pääbo S : Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc Natl Acad Sci USA* 1996; **93**: 12035–12039.

6 Gill P, Ivanov PL, Kimpton C *et al*: Identification of the remains of the Romanov family by DNA analysis.. *Nat Genet* 1994; **6**: 130–135.
7 Hearn CM, Ghosh S, Todd JA: Microsatellites for linkage analysis of genetic traits. *Trends Genet* 1992; **8**: 288–294.
8 Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL: High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 1994; **368**: 455–457.
9 Mitchell JR, Hammer MF: Human evolution and the Y chromosome. *Curr Opin Genet Dev* 1996; **6**: 737–742.
10 Jobling MA, Tyler-Smith C Fathers and sons: the Y chromosome and human evolution. *Trends Genet* 1995; **11**: 449–456.
11 Murray JC: Genome DataBase: 366108 1993
12 Wiegand P, Budowle B, Rand, S, Brinkmann B: Forensic validation of the STR systems SE33 and TC11. *Int J Legal Med* 1993; **105**: 315–320.
13 Waye JS, Presley LA, Budowle B, Shutler GS, Fourney RM: A simple and sensitive method for quantifying human genomic DNA in forensic specimen extracts. *Biotechniques* 1989; **7**: 852–855.
14 Bandelt HJ, Forster P, Sykes BC, Richards MB: Mitochondrial portraits of human populations using median networks. *Genetics* 1995; **141**: 743–753.
15 Weber JL, Wong C: Mutation of human short tandem repeats. *Hum Mol Genet*, 1993; **2**: 1123–1128.
16 Valdes AM, Slatkin M, Freimer NB: Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* 1994; **133**: 737–749.
17 Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P: Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees: *Hum Mol Genet* 1997; **6**: 799–803.
18 Excoffier L, Smouse PE: Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: Molecular Variance Parsimony. *Genetics*, 1994; **136**: 343–359.
19 Michalakis Y, Excoffier LA: Generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci: *Genetics* 1996; **142**: 1061–1064.
20 Excoffier L, Smouse PE, Quattro JM: Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 1992; **131**: 479–491.