Maternal and Paternal Lineages of the Samaritan Isolate: Mutation Rates and Time to Most Recent Common Male Ancestor

B. Bonné-Tamir[1,*], M. Korostishevsky[1], A. J. Redd[2], Y. Pel-Or[1], M. E. Kaplan[2] and M. F. Hammer[2]

[1]*Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Ramat-Aviv, Israel*

[2]*Division of Biotechnology, University of Arizona, Tucson, AZ, USA*

Summary

The Samaritan community is a small, isolated, and highly endogamous group numbering some 650 members who have maintained extensive genealogical records for the past 13–15 generations. We performed mutation detection experiments on mitochondrial DNAs and Y chromosomes from confirmed maternal and paternal lineages to estimate mutation rates in these two haploid compartments of the genome. One hundred and twenty four DNA samples from different pedigrees (representing 200 generation links) were analyzed for the mtDNA hypervariable I and II regions, and 74 male samples (comprising 139 links) were typed for 12 Y-STRs mapping to the non-recombining portion of the Y chromosome (NRY). Excluding two somatic heteroplasmic substitutions and several length variants in the homopolymeric C run in the HVII region, no mutations were found in the Samaritans' maternal lineages. Based on mutations found in Samaritan paternal lineages, an estimate of a mutation rate of 0.42% (95% confidence interval of 0.22%–0.71%) across 12 Y-STRs was obtained. This estimate is slightly higher than those obtained in previous pedigree studies in other populations. The haplotypes identified in Samaritan paternal lineages that belong to the same haplogroup were used to estimate the number of generations elapsed since their most recent common ancestor (MRCA). The estimate of 80 generations corresponds with accepted traditions of the origin of this sect.

Introduction

Detailed and accurate genetic characterization of maternal and paternal lineages can be achieved by the analysis of the haploid compartments of the genome: mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY). As a result of their uniparental mode of inheritance and lack of recombination these segments of the genome have been developed into highly informative systems with applications in evolutionary studies, forensics, medical genetics, and genealogical reconstruction (Hammer & Zegura, 1996; Jobling *et al*. 1997; Kittles *et al*. 1998; Macaulay *et al*. 1999; Helgason *et al*. 2000; Ingman *et al*. 2000).

*Correspondence: Batsheva Bonné-Tamir, Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Ramat-Aviv 69978, Israel. Telephone: 972-3-6409318. Fax: 972-3-6409900. E-mail: bonne@post.tau.ac.il

Maternally inherited mtDNA was originally applied to human evolutionary studies, in part because of its 10- to 100-fold higher mutation rate compared with nuclear genes (Cann *et al*. 1987). Most mtDNA sequencing studies have focused on the D-loop, which is comprised of hypervariable segments HVI and HVII between nucleotide positions 16020–16400 and 48–408, respectively. While HVI has been used as a source of evolutionary information in most phylogenetic analyses of mtDNA, there is a growing database of HVII sequences from worldwide human populations in the "HVR database" at the Max-Planck-Institute for Evolutionary Anthropology in Leipzig (http://www.hvrbase.de).

A few studies of mutation rates of the mitochondrial hypervariable regions have been published (Howell *et al*. 1996, Parsons *et al*. 1997; Jazin *et al*. 1998, Sigurðardóttir *et al*. 2000; Heyer *et al*. 2001). Two studies have attempted to quantify control region mutation rates in

Table 1 Point mutations in mtDNA

	No. of Families	No. of Individuals Tested	Generation Links	Mutations	Mutation Rate
Soodyal *et al.* (1997)	5	75	108	0	<1/36
Howell *et al.* (1996)	1	45	81	2	1/25–40
Parsons *et al.* (1997)	134	>268	327	10	1/33
Sigurðardóttir *et al.* (2000)	26	272	705	3	1/232
This study	10	124	200	0	<1/61

specific kindreds: Howell *et al.* examined mutations in a large Australian kindred and Soodyall *et al.* (1997) obtained mtDNA sequences from the population of the isolated island Tristan da Cunha (Table 1). Howell *et al.* (1996) found 2 intra-lineage heteroplasmic mutations in the HVII region and detected no sequence changes in the poly C region. Soodyall *et al.* did not encounter any intra-lineage mutations of any kind in their study. Parsons *et al.* (1997) observed 10 nucleotide substitutions and only a single case of intra-lineage sequence length mutation in the poly C region. They noted that most of the individuals were heteroplasmic at low levels in the HVII poly C region.

The NRY contains the largest non-recombining block in the human genome, and by virtue of its many polymorphisms is becoming one of the most informative nuclear haplotyping systems (YCC, 2002). Recent years have witnessed an explosion in data from the NRY in human populations. This explosion has been driven by the many polymorphisms discovered on the NRY, including both binary polymorphisms such as single nucleotide polymorphism (SNPs) (Underhill *et al.* 2000), multiallelic variation within minisatellites (Jobling *et al.* 1997) and short tandem repeats (Y-STRs) (Kayser *et al.* 2001; Redd *et al.* 2002). Y-STRs have been shown to be as polymorphic as their autosomal counterparts (Roewer *et al.* 1992) and are becoming useful in studies of human identity and forensics (Jobling *et al.* 1997). To date, only three studies have attempted to estimate Y-STR mutation rates in human pedigrees. Heyer *et al.* (1997) screened 42 Canadian males, descendants from 12 "founding fathers," for 9 Y-STRs, using deep rooting pedigrees (*i.e.*, multigenerational pedigrees). They found an average mutation rate of 0.21% per generation. Bianchi *et al.* (1998) found no mutations on the Y-chromosomes of 249 father-son transmissions in CEPH (Caucasian) families, and revised downwards the average mutation rate from the literature to 0.12%. Kayser *et al.* (2000) examined 15 Y-STRs and identified 14 mutational events in 4999 meioses observed in father-son pairs. They concluded that the mutation rates (0.28%) and modes of Y-STRs and autosomal STRs are similar, and suggested that the mutational mechanism for STRs, in general, is independent from recombination.

We are not aware of any studies performed on pedigrees with multiple descendants from different generations in which both NRY and mtDNA mutation rates were estimated. The availability of exact genealogies of the Samaritan isolate enables the recognition and detection of mutations occurring between generations within lineages. The purpose of this study is to: (1) evaluate both NRY and mtDNA mutation rates in the same kindreds of a single human population; (2) increase the number of deep-rooting pedigrees used to estimate the Y-STR mutation rates; (3) compare mutation rates in populations of different origins; and (4) estimate the time to the most recent common ancestor (MRCA) of Samaritan Y chromosomes.

We would like to stress that the purpose of our study was not to infer the origins of the Samaritans as a population. The Samaritan isolate represents a very small group that underwent an immense reduction in numbers, and currently consists of a few closely related kindreds that claim to be descended from a small number of founders.

The Samaritans

The Samaritan community in the Middle East survives as a distinct religious and cultural sect and constitutes one of the oldest and smallest ethnic minorities in the world. Some 650 individuals comprising the total group of present day Samaritans trace their ancestry over a

period of more than 2,000 years to the Biblical Israelite tribes of Ephraim and Menashe. As a religious sect, the Samaritans broke away from the main stream of Judaism around the fifth century B.C. They attained their greatest numbers and significance as an independent nation during the Roman period, but from the 6th century onwards, misfortunes and oppressions caused a gradual decline in their numbers. From a nation of at least several thousand individuals they became a small sect, reaching their lowest number of only 122 members in 1853 (Petermann, 1860). Despite prediction of their imminent extinction at the beginning of the 20th century (Gini, 1933), they have gradually increased in numbers and are presently located in two localities some 20 miles apart—Nablus and Holon—in the same geographical region which they have never left.

Throughout their history, the Samaritans adhered to an endogamous marriage system that was practiced not only within the limits of the community but also often within the limits of the family. Extensive demographic and genetic investigations of the Samaritan community have been carried out since the 1960's (Bonné, 1963; Bonné, 1966a,b; Roberts & Bonné, 1973).

The Samaritans exhibit a unique genetic profile with extreme allele frequencies at almost all loci that have been examined; these frequencies differ significantly from those found in both Jewish and non-Jewish populations of the region (Bonné-Tamir, 1980). The community is highly inbred with 84% marriages contracted between either first or second cousins; their mean inbreeding coefficient of 0.0618 is the highest recorded for any human population (Bonné-Tamir, 1980). The genetic constitution of the present day population derives from only 45 founders. Detailed pedigrees documenting the last 13–15 generations (ca 400 years) enable counting the total number of descendants from each founder (Cazes & Bonné-Tamir, 1984). Since the Samaritans maintain extensive and detailed genealogical records, it is possible to construct accurate pedigrees and specific maternal and paternal lineages.

Subjects and Methods

Ten maternal lineages were identified in the Samaritan isolate (Table 2 and Figures 1a–1b). Three of them descend from non-Samaritan women who married into the community in recent times. Altogether 124 DNA samples (57 males and 67 females) were screened for the mtDNA hypervariable I (HVI) and hypervariable II (HVII) control regions by single strand conformation polymorphism (SSCP) analysis. Direct sequencing was further performed on 46 samples for the HVI and 54 for the HVII control regions, in order to confirm the SSCP

Table 2 mtDNA sequence types found among the Samaritans

	HVI												HVII						N*
	1	1	1	1	1	1	1	1	1	1	1	1							
	6	6	6	6	6	6	6	6	6	6	6	6							
	1	1	2	2	2	2	2	2	2	3	3	3			1	1	1	2	
	2	7	1	5	7	8	9	9	9	0	1	6	6	7	5	5	9	6	
	6	2	9	6	8	8	4	6	8	9	8	2	4	3	0	2	5	3	
Reference**	T	T	A	C	C	T	C	C	T	A	A	T	C	A	C	T	T	A	
Lineages 1, 2, 3, 9 (type A)	C	.	.	.	.	C	T	T	.	.	.	.	.	G	.	.	C	G	58
Lineages 4, 5 (type B)	.	.	.	T	.	.	.	.	.	G	C	.	.	G	.	C	.	G	47
Lineage 8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	11
Lineage 6	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	C	.	G	3
Lineage 7	C	.	.	.	.	.	.	.	.	.	.	C	T	.	.	.	.	G	2
Lineage 10	.	C	G	.	T	.	.	.	.	.	.	.	.	G	T	C	.	G	3

*N = number of individuals typed in each lineage type.

**Anderson *et al*. 1981.

The first two lineage types shown (A, B) and lineage 8 are 'true' Samaritan lineages, while the lower three (lineages 6, 7 and 10) are known to descend from non-Samaritan females (Syria, Egypt and Russia).

In HVII, lineage 7 has an additional C residue at position 58 and all the lineages have insertions in the polycytosine 'C stretch' region, i.e., up to two additional C residues at position 309 and an additional C residue at position 315.

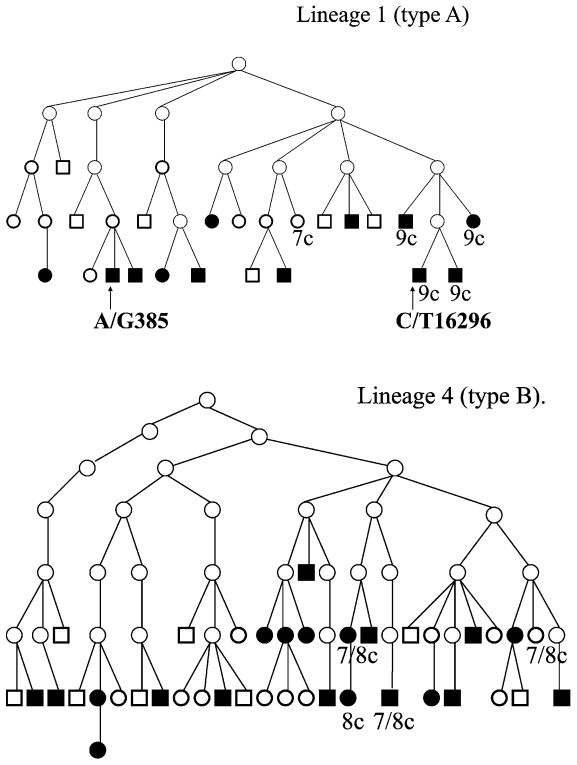

Figure 1a–b Poly C sequence and heteroplasmic states in Samaritan maternal lineages. Individuals represented with solid circles or solid squares were checked by SSCP and sequencing; those represented in bold were checked only by SSCP; 7, 8 and 9 refer to different numbers of C residues (those unmarked in lineage 1 carry 8 Cs and those unmarked in lineage 4 carry 7 Cs); arrows point to heteroplasmic states.

results. To examine Y chromosome mutation rates, we typed 12 Y-STRs in a total of 74 DNA samples from males whose family and direct ancestor-descendant relationships are known for the last seven generations. These males descend from only four paternal ancestors whose lineages are designated C, D, M and Z (See Figures 2a–2d). The Samaritan study was approved by the Tel-Aviv University institutional review board.

PCR, SSCP, and sequencing mitochondrial DNA

HVI was amplified using the primers L15926 and H16498 (Comas *et al*. 1995). HVII was amplified using the primers forward: 5′ GTTCCCCTTAAGACATC 3′ (16543 –16563) and reverse: 5′ TTCCCCGCCGTGTGGCTGGCTGGTT 3′(457–437). The HVI and

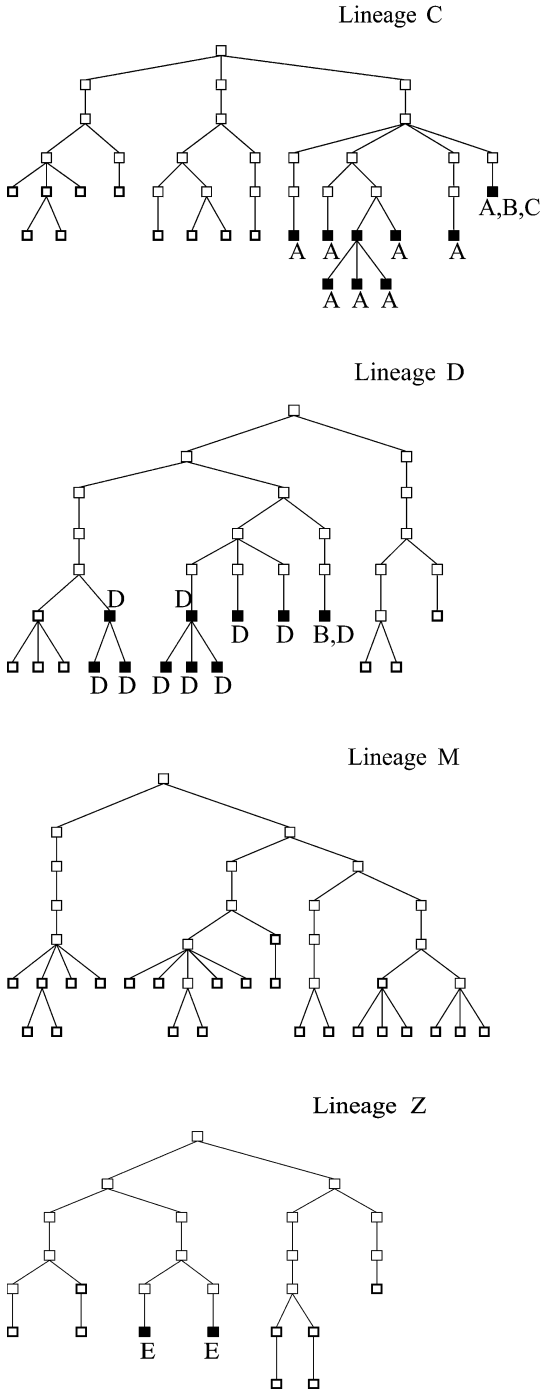

Figure 2a–d Y-chromosomal mutations in Samaritan paternal lineages. Individuals in whom Y-chromosomal DNA was analyzed are shown as bold squares; those shown solid were carriers of repeat changes: (A) *DYS391* 11 → 10, (B) *DYS19* 14 → 15, (C) *DYS388* 12 → 13, (D) *DYS389I* 14 → 15, (E) *DYS385a* 14 → 17. Note: three scenarios for mutations at *DYS389I* in lineage D are considered (see Results section).

HVII fragments were too large for direct SSCP analysis, so they were cut into usable fragments with restriction enzymes. HVI DNA was digested with *Bfa*I and *Ssp*I and 10 ng of the combined fragments from each individual were dephosphorylated, 5′ labeled with ^{32}P, and run on non-denaturing MDE acrylamide gels. The gel and buffer conditions were 10% glycerol at 800 volts for 16–20 hours. HVII DNA was digested with *Mn*II and treated as for HVI. Individuals exhibiting gel shifts were then sequenced using Sequitherm Cycle Sequencing kits (Epicentre Technologies).

Typing STRs on the NRY

All individuals were genotyped for 12 Y-STRs using two multiplex PCR reactions (Redd *et al*. 2002). The 12 Y-STRs included *DYS19*, *DYS385a*, *DYS385b*, *DYS388*, *DYS389I*, *DYS389II-I*, *DYS390*, *DYS391*, *DYS392*, *DYS393*, *DYS426*, and *DYS439*. In addition, male samples were typed with a set of binary markers (Y-SNPs) to determine which of 18 major Y chromosome binary haplogroups they belonged to. See Karafet *et al*. (2002) for a list of markers, methods, and typing strategies. Samples that had the 12f2a-8-kb allele were also typed with M172 (YCC, 2002).

Statistical methods

A method to estimate the mutation rate for mtDNA lineages in which no mutations were found by sequencing has been detailed by Soodyall *et al*. (1997). In cases where the data include results of DNA sequencing and SSCP, the possibility that some mutations are missed in the SSCP analysis should be taken into account. The probability of observing no mutations in the sample (P_0) is the probability of finding no mutations in the generation links confirmed by sequencing, multiplied by the probability of finding no mutations in the generation links screened only by SSCP. If the generation links screened only by SSCP are the end-links in the pedigree, the P_0 value can be evaluated by a simple formula:

$$P_0 = (1 - \mu)^n (1 - \alpha\mu)^k$$

where n is the number of generation links tested by sequencing, k is the number of generation links tested by SSCP only, μ is mutation rate and α is the probability of finding a mutation using SSCP. The SSCP analysis detects at least 80% of all point mutations (Sheffield *et al*. 1993), therefore α was set at 0.8. To estimate the upper limit for the mutation rate μ, P was set at 0.05 (Soodyall *et al*. 1997). The 95% confidential interval for the mutation rate μ for NRY markers was evaluated according to Sigurðardóttir *et al*. (2000).

Estimation of the time to MRCA was performed based on the stepwise model of the mutation process. A good fit for the stepwise model for Y STR was shown by a number of researchers (Edwards *et al*.1992; Brinkmann *et al*. 1998; Kayser *et al*. 2001; Walsh, 2001). The time t to MRCA was estimated as the value for which the expected (according to the model) difference in the number of repeats was equal to the observed difference d between two lineages (i.e., as the average across the markers typed). A 95% confidence interval of the time to MRCA was determined as the range of t values for which the expected standard 95% interval included the observed difference d. The formulae for the mean difference and the standard deviation in the number of repeats between two t linked Y chromosomes, as well as the formula for the range interval boundaries, are presented in the Appendix. Another method of estimating the time to MRCA between a pair of haplotypes based on a Bayesian analysis was recently published (Walsh, 2001). To estimate the time to MRCA for a haplotype set (i.e., more than two haplotypes) a topology of the origin tree should be inferred (Korostishevsky *et al*. 2001). Based on such a tree topology the number of generations to each node of the tree can be estimated. This estimate corresponds to the average difference in the number of repeats between pairs of lineages for which the same node of the tree indicates their MRCA.

Results

Mitochondrial DNA

Table 2 summarizes the mtDNA hypervariable sequences observed among 10 contemporary Samaritan maternal lineages. Each one of three maternal lineages deriving from recently introduced Jewish women (lineage 6, 7, 10) had a separate haplotype. Direct sequencing of the HVI and HVII regions in remaining Samaritan lineages identified three different sequence types. Lineages 1, 2, 3, and 9, involving 191 individuals, share one

sequence type (type A, Table 2) and may be considered a single extended lineage. A number of similar types have been found in Europe and the Middle East, but none of them includes the T to C transition at position 16,288 (Table 2). Lineages 4 and 5, involving 141 individuals, share another sequence type (type B) which includes a unique transversion (A to C at position 16,318). This type appears to occur only in the Samaritans, with no related types reported in Europe or in the Middle East (Pel-Or *et al*. 1997). Lineage 8 (comprising 22 individuals all of whom are descendants of one female who lived 6 generations ago) conforms to the Anderson reference sequence in HVI, but differs from it in the HVII control region by the addition of two C residues in the polycytosine "C-stretch" region. Such additions are apparently common (Hauswirth & Clayton, 1985).

Two intra-lineage nucleotide changes (one in HVI and one in HVII) were identified in lineage 1 (Figure 1a). Both were found in a heteroplasmic state in two males. A change from A to G at position 385 was detected in one individual but not in his sister, brother, or mother. The origin of a change from T to C at position 16296 is somewhat less clear, as the mother of the carrier was not available for testing. This mutation was not detected, however, in the carrier's brother or in his two aunts. Both heteroplasmic states thus appear to be somatic, non-hereditary mutations.

Poly C sequences are found in both hypervariable regions. HVI contains the sequence 5C-T-4C between co-ordinates 16184 and 16193, according to Anderson's sequence. In HVII, the poly C region consists of 7 Cs followed by a T, and then 5 additional Cs between nucleotides 303 and 315, according to Anderson's sequence. This repetitive poly C tract in HVII is known to have a high mutation rate that changes the length of the sequence (Hauswirth & Clayton, 1985; Andrews *et al*. 1999; Howell & Smejkal, 2000). All Samaritan lineages contained 6 Cs in the 3′ part and from 7 to 10 Cs in the 5′ region of the HVII poly C tract. At least 6 sequence length mutations in the HVII region were observed among the Samaritans. In lineage 1 (Figure 1a) several related members carried 9 Cs in place of the 8 Cs that are typical for this lineage. In lineage 4 (Figure 1b), which is characterized by 7C residues, several poly C sequence length changes were observed.

In our study, 127 of the 200 generation links were confirmed by sequencing. When these figures were used for computing the upper limits for mutation rates, μ yielded a value of 0.016 (see method section). This means that, according to our study, the mutation rate at the hypervariable regions is no more than one new mutation every 61 generations, with a confidence of 95%. In other words, if the mutation rate was higher than this, there would have been a 95% chance of finding at least one non-somatic mutation in our study.

The NRY

Four major NRY STR haplotypes were identified in the Samaritans, with a different haplotype in each lineage: C, D, M and Z (Table 3). All four lineages had the same allele at one STR (*DYS392*), while each lineage carried a different allele at two STRs (*DYS385b* and *DYS439*). All other STRs exhibited two or three alleles in the four Samaritan lineages. The average pairwise difference in repeat number among the four Samaritan lineages was 0.94 per marker (95% CI =: 0.69–1.20).

Altogether the following mutation events were observed (Table 3): (1) a gain of three repeats (from 14 to 17) at *DYS385a* in two cousins in lineage Z (see Figure 2d); (2 and 3) two mutation events at *DYS389I* which occurred independently in two separate branches in lineage D (see Figure 2b); (4) a loss of one repeat at *DYS391* (from 11 to 10) in lineage C (see Figure 2a), (5) a gain from 12 to 13 at *DYS388* in lineage C; (6 and 7) a gain of one repeat from 14 to 15 at *DYS19*

Table 3 The Y STR haplotypes in four Samaritan paternal lineages

Lineage	*DYS19*	*DYS385a*	*DYS385b*	*DYS388*	*DYS389I*	*DYS389II-I*	*DYS390*	*DYS391*	*DYS392*	*DYS393*	*DYS426*	*DYS439*
C	14/15*	17	19	12/13*	13	18	24	11/10*	11	13	11	16
D	14/15*	14	16	15	14/15*	16	24	10	11	12	11	15
M	14	13	18	16	13	17	23	11	11	12	11	14
Z	14	14/17*	17	15	13	16	23	10	11	12	12	17

*Mutations identified among members in the same lineage.

Table 4 Number of mutations identified in each of the 12 STRs*

Lineage	Links	*DYS19*	*DYS385a*	*DYS385b*	*DYS388*	*DYS389I*	*DYS389II-I*	*DYS390*	*DYS391*	*DYS392*	*DYS393*	*DYS426*	*DYS439*
C	41**	1	0	0	1	0	0	0	1	0	0	0	0
D	34	1	0	0	0	2	0	0	0	0	0	0	0
M	40	0	0	0	0	0	0	0	0	0	0	0	0
Z	24	0	1	0	0	0	0	0	0	0	0	0	0
TOTAL	139	2	1	0	1	2	0	0	1	0	0	0	0

*Each mutation was confirmed by retyping and/or sequence analysis.
**In 5 loci, *DYS19*, *DYS385a*, *DYS385b*, *DYS390* and *DYS391*, only 40 links were used.

which occurred independently in lineages C and D. It should be noted that mutation 1 (three-step mutation) may in fact be a case of two mutations (one-step and two-step mutations). Three scenarios for mutations 2 and 3 are possible: two gains from 14 to 15, two losses from 15 to 14, one gain and one loss. Two of these three scenarios imply that allele 14 (closest to allele 13, common for the other three lineages) is the ancestor allele in lineage D. In summary, twelve STRs with a minimum of seven mutational events were observed in 1663 links (Table 4). This yields an average mutation rate of 0.42% per marker, with a 95% confidence interval of 0.22%–0.71%.

The SNP typing revealed that lineage C was YAP^+ and a member of haplogroup E-P2. The other three Samaritan lineages (M, Z, and D) carried the 12f2a-8 kb allele and were members of haplogroup J. The Z and D lineages had the derived allele at M172 and, hence, were members of haplogroup J-M172. Lineage M was ancestral at M172 and was part of paragroup J^*.

Discussion

The known and detailed genealogical structure of the Samaritan community provides the opportunity to construct haplotypes for the characterization of maternal and paternal lineages, and to identify mutations occurring between generations within each kindred. It should be stated that in the many studies carried out in the past on the distribution of markers and polymorphisms in the Samaritan community (Bonné, 1966a; Bonné-Tamir, 1980) no paternity incompatibilities were ever encountered. Therefore, the pedigree construction is regarded as very accurate and any differences observed within families are likely to result from mutation rather than mistaken paternity.

Although heteroplasmy in the control region of the human mitochondrial genome was not described until 1995 (Comas *et al*. 1995), more recent studies indicate that control region heteroplasmy is rather frequent (Tully *et al*. 2000). However, the long-term fate of mutants found in heteroplasmic states is not clear, and it is difficult to derive an estimate of the mutation rate based on the presence of heteroplasmy (Sigurðadottir *et al*. 2000). Furthermore, in our study both of the heteroplasmic nucleotide substitutions observed are apparently somatic, and thus have no effect on the long-term mitochondrial mutation rate. Length changes in the poly C regions are apparently quite common, and not representative of the general mtDNA mutation rate. We followed Parson *et al*.'s (1997) example and did not include mutations observed in the poly C tract. Thus, in our study, the three heteroplasmic cases involving the HVII poly C sequence, as well as the six length variants in the poly C sequence, were not included in our estimate of the Samaritan control region mutation rate. In contrast to the situation for paternal lineages, no new mutations were detected among the ten different Samaritan control region lineages. The only new variations seen in the Samaritan isolate results from marriages to ethnically diverse women who recently entered into the community, rather than to mutations occurring throughout the generations. Compared with results obtained by others on mtDNA mutation rates (Table 1), our upper limit estimate of the mutation rate of 1/61 mutations per generation is in close agreement with those previously published. When we estimate the overall mutation rate taking into account all data cited in Table 1, we obtain a figure of 0.011 (95%CI: 0.007–0.014). Note that if one wishes to include the two heteroplasmic substitutions detected in lineage 1 the mutation rate would be 2/200, a value that does not contradict the aforementioned mutation rate.

Examination of father-son pairs for Y chromosome mutations as performed by Kayser *et al.* (2000) provides knowledge on mutations occurring only in two living generations. In contrast, the Samaritans' lineage structure, with connecting relationships of living descendants to their paternal ancestors, allowed us to determine both the most likely individual or individuals in whom the mutation occurred, and the time elapsed since the mutation occurred. For example, the loss of a single *DYS391* repeat in lineage C must have occurred 4 or 5 generations ago (Figure 2a). The *DYS389I* mutation in lineage D must have occurred two generations ago, while the second must have occurred 4–6 generations ago (Figure 2b). These distinctions lead to a more accurate description of the mutation process. Examination of deep rooting pedigrees for Y chromosome mutations was performed by Heyer *et al.* (1997). However, in their pedigrees it was not possible to determine whether the mutations reflected a loss or a gain, because they appeared in two living individuals who were connected to a common ancestor via six or seven previous generations. When compared with the two previous studies reporting on mutation rates (Heyer *et al.* 1997; Kayser *et al.* 2000), the Samaritans' rate of 0.42% appears to be somewhat faster. It should be noted that the standard error range overlaps in the two previous studies and our study. When we estimate the overall mutation rate for all three studies, a figure of 0.34% (95%CI: 0.24–0.48) was obtained.

The differences encountered among the four lineages in their major NRY haplotypes (Table 5) are, on the whole, not very large. The larger differences are between the C lineage and each of the other three lineages, reflecting divergence between haplogroups E and J. The difference between C and the other lineages in number of repeats is 14.67 (95% CI = 14.01–15.32), while there is an average of 8.0 repeat differences (95% CI = 6.04–9.96) among the D, M and Z lineages. This indicates a close relationship between paragroup J* and its descendant haplogroup, J-M172. The reconstructed tree for Samaritan lineages based on 12 STRs is presented on Figure 3. This reconstruction is in a good accordance with a tree inferred for NRY haplogroups (YCC, 2002).

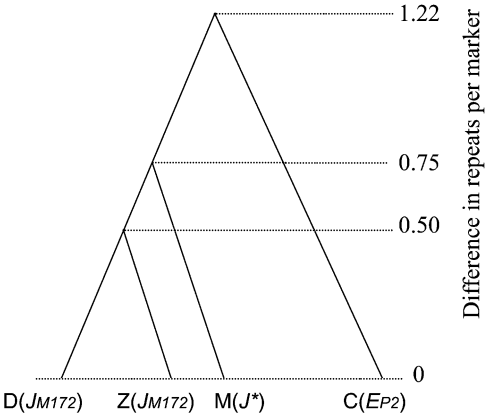

Figure 3 The origin tree of Samaritan male lineages based on 12 STRs. The tree was constructed using the UPGMA method (PHYLIP, 1995). A distance matrix comprised of the repeat difference per marker among Samaritan lineages was used.

Table 5 Differences between lineages in number of marker repeats

lineages compared	total difference*	difference per marker	SD value
D/Z	6	0.50	0.20
M/D	9	0.75	0.19
M/Z	9	0.75	0.26
C/D	15	1.25	0.37
C/M	14	1.17	0.44
C/Z	15	1.25	0.32

*Total difference is a sum of the differences in number of repeats for the 12 markers.

Who is the C lineage? According to the Samaritan tradition, males from this lineage are descendants of the biblical tribe of Levi. Following the Jewish custom, these males are assigned religious duties and serve as priests (Cohanim; Cohen is the Hebrew word for priest). In the investigation of binary polymorphisms (SNPs) typed in the Samaritans, the C lineage was found to be haplogroup E-P2. It is well known that while STRs are fast evolving microsatellites, binary markers have a low mutation rate and represent unique mutation events (de Knijff, 2000.) Hence, in our attempt to use STR haplotypes for assessing the number of generations that separate the lineages from their most recent common ancestor, we included only the lineages that belong to the same haplogroup/paragroup.

Using our inferred mutation rate of 0.42% per generation, and assuming a stepwise mutation model (see

Statistical methods), we obtained a figure of 80 (95%CI: 38–192) generations that separated the two Samaritan haplogroup J-M172 lineages from their most recent common ancestor. If we assume 25–30 years per generation, we arrive at an estimate of 2,000–2,400 years ago. This estimate is in concordance with the Samaritans' own belief that they are descendants of the original Israelite tribes that populated the central region of Samaria and then split from the Jews over religious issues, such as the location of the temple. It was also of interest to compare the Samaritans' results with those obtained from typing the same 12 Y-STRs in a sample of Jewish priests—Cohanim (Behar, Skorecki & Hammer, unpublished results). We computed the TMRCA of the M lineage and the Cohen Modal Haplotype, which are both members of paragroup J*. A similar time estimate of 98 (95%CI: 47–232) generations was obtained. It is interesting to note that Thomas *et al*. (1998), in their study of the origin of the Cohen modal haplotype as defined by six Y-STRs, arrived at a coalescence time of 106 (95% CI: 84–130) generations (based on the method of average square difference among pairs of haplotypes). If we estimate TMRCA based on STR markers, for three Samaritan lineages (D, Z and M) whose haplogroups evolved one from the other (Underhill *et al*. 2000), we arrive at 118 (95%CI: 58–218) generations. However, it should be noted that all of the above time estimates do not take into account uncertainty in mutation rates for STR markers, and possible deviation from the stepwise model.

In light of the history of the Samaritan sect, together with recent results on Y-STR haplotypes, we compared our data with those published for several Middle Eastern populations (Nebel *et al*. 2000; Frenkel-Arons *et al*. 2000, and personal communication). The comparison relates to six STRs (*DYS19*, *DYS388*, *DYS390*, *DYS391*, *DYS392*, *DYS393*) common to the different studies. Table 6 shows that 3 Ashkenazi Jews out of a total of 143 have the same haplotype as the Z lineage, while two individuals share haplotypes with the M lineage. Two Palestinians (out of 79) also share their haplotype with the Z lineage. Three of 34 Libyan haplotypes, as well as 2/35 Moroccan haplotypes, were also shared with the Samaritans. Thus, all four Samaritan lineage haplotypes are found in Jewish and non-Jewish Middle Easterners, suggesting common ancestry for these

Table 6 Populations* in which Samaritan haplotypes were observed

Lineage	Mr	Lb	Tu	Bu	AJ	Pa
C	1	1	0	1	0	0
D	0	1	0	0	0	0
M	0	0	0	0	2	0
Z	1	1	0	0	3	2
Total typed	35	34	35	35	143	79

*Mr = Moroccan Jews, Lb = Libyan Jews, Tu = Turkish Jews, Bv = Bulgarian Jews, AJ = Ashkenazi Jews, Pa = Palestinians. The comparison relates to six STRs (*DYS19, DYS388, DYS390, DYS391, DYS392, DYS393*) common to the different studies. The data is from Frenkel-Arons, N., unpublished.

populations (see also Nebel *et al*. 2001). Of particular interest is the finding that out of an extensive analysis of 986 males from 20 globally dispersed human populations, only 3 individuals (0.3%) shared Samaritan haplotypes; all three are from Buenos Aires in Argentina (Kayser *et al*. 2000). It should be emphasized again, that the observed similarity includes only six STRs, and that identical haplotypes are expected to be found in different populations as a result of convergence (Heyer *et al*. 1997; Walsh, 2001). This, in combination with the fact that the Samaritans have experienced extreme genetic drift, means that one should use caution when attempting to infer the genetic affinities of the Samaritans.

Appendix

Let $P_t(n)$ be the probability of the difference (n) in number of repeats at an STR between two Y-chromosomes (t linked). Then under the stepwise model with a mutation rate μ the following equations become

$$P_{t+1}(0) = (1-\mu)P_t(0) + \frac{\mu}{2}P_t(1)$$

$$P_{t+1}(1) = (1-\mu)P_t(1) + \mu P_t(0) + \frac{\mu}{2}P_t(2)$$

$$P_{t+1}(1 < n < t) = (1-\mu)P_t(n) + \frac{\mu}{2}P_t(n-1) + \frac{\mu}{2}P_t(n+1)$$

$$P_{t+1}(t) = (1-\mu)P_t(t) + \frac{\mu}{2}P_t(t-1)$$

$$P_{t+1}(t+1) = \frac{\mu}{2}P_t(t) \qquad (1A)$$

The sum of the 1A equations, multiplied by $0, 1, \ldots, t+1$ respectively, gives the recurrent equation for the mean difference $\hat{N}_t$

$$\hat{N}_{t+1} = \hat{N}_t + \mu P_t(0) \tag{2A}$$

where $P_t(0)$ is given by

$$P_t(0) = \sum_{i=0}^{[t/2]} \frac{t!}{i!i!(t-2i)!}\left(\frac{\mu}{2}\right)^{2i}(1-\mu)^{t-2i} \tag{3A}$$

A simple consequence of the 2A equation is

$$\hat{N}_{t+1} = \mu \sum_{i=0}^{t} P_i(0) \tag{4A}$$

The sum of the 1A equations, multiplied by $0, 1^2, \ldots, (t+1)^2$ respectively, gives the recurrent equation for the mean of squared difference $\hat{N}_t{}^2$

$$\hat{N}_{t+1}^{\,2} = \hat{N}_t{}^2 + \mu \tag{5A}$$

and

$$\hat{N}_t{}^2 = t\mu \tag{6A}$$

Hence, the standard deviation σ_t is given by

$$\sigma_t = \sqrt{t\mu - (\hat{N}_t)^2} \tag{7A}$$

For k markers with the same mutation rate μ, the SD interval becomes

$$\hat{N}_t \pm \sigma_t/\sqrt{k} \tag{8A}$$

In Figure 4, according to the stepwise model ($\mu = 0.0042$), we present the expected mean difference and the standard interval for 12 markers as a function on the generation number t up to MRCA.

Let d be the observed difference in number of repeats between two Y-haplotypes (average for the markers typed). An estimate of the time to MRCA was evaluated as the t value for which the expected difference $\hat{N}_t$ equals d (see equations: 3A, 4A). A 95% range interval of the estimate was defined as the range of t values for which the expected standard 95% interval includes the observed difference d. Assuming Gaussian distribution for the expected difference d, the range interval was evaluated as:

$$\left\{t : \hat{N}_t - U_{1-p}\frac{\sigma_t}{\sqrt{k}} \le d \le \hat{N}_t + U_{1-p}\frac{\sigma_t}{\sqrt{k}}\right\} \tag{9A}$$

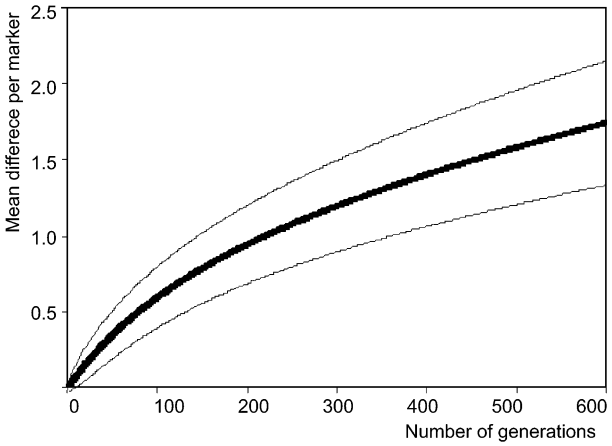

Figure 4 The mean difference in the number of repeats per marker between two Y-chromosomal haplotypes (thick line) and the SD interval (thin lines) as a function of time to MRCA according to the stepwise model of the mutation process: 12 markers and a mutation rate of 0.0042.

One may find the U_x value in a statistical table for the standard Gaussian distribution: it is the argument of the distribution function equaled x (e.g., $U_{0.95} = 1.64$, $U_{0.99} = 2.33$).

Acknowlegments

We thank the members of the Samaritan community for their long-standing cooperation in our study. This work was supported by research grants from the Applebaun Foundation (114115 to BBT), the Israel Academy of Science (0110041 to BBT), the National Institute of Justice (2000-IJ-CX-K006 to MFH), and the National Institute of General Medical Sciences (GM-53566 to MFH). We would also like to thank Asher Berry for his helpful comments on the manuscript.

References

Andrews, R. M., Kubacka, I., Chinnery, P. F., Lightowlers, R. N., Turnbull, D. M. & Howell, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**, 147.

Bianchi, N. O., Catanesi, C. I., Bailliet, G., Martinez-Maringnac, V. L., Bravi, C. M., Vidal-Rioja, L. B., Herrera, R. *et al.* (1998) Characterization of ancestral and derived Y chromosomal haplotypes of new world populations. *Am J Hum Genet* **63**, 1862–1871.

Bonné-Tamir, B. (1963) The Samaritans: a demographic study. *Hum Biol* **35**, 61–89.

Bonné-Tamir, B. (1966a) Genes and phenotypes in the Samaritan isolate. *Am J Phys Anthrop* **24**, 1–19.

Bonné-Tamir, B. (1966b) Are there Hebrews left? *Am J Phys Anthrop* **24**, 134–145.

Bonné-Tamir, B. (1980) The Samaritans: a living ancient isolate, In: *Population Structure and Genetic Disorders* (eds Eriksson A. W., Forsius H. R., Nevallina H. R., Workman P. L., Norio R. K.), pp 27–41, London: Academic press.

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

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987) Mitochondrial DNA and human evolution. *Nature* **325**, 31–36.

Cazes, M. H. & Bonné-Tamir, B. (1984) Genetic evolution of the Samaritans. *J. of Biosoc* **16**, 177–187.

Comas, D., Paabo, S. & Bertranpetit, J. (1995) Heteroplasmy in the control region of human mitochondrial DNA. *Genome Research* **5**, 89–90.

Frenkel-Arons, N., Korostishevsky, M., Pel-Or, Y., Berry, A., Woolf, R., Hillel, Y., Livshits, G., Bonné-Tamir, B. (2000) Mitochondrial DNA variation in Mediterranean populations. *Third Symposium on Human Genetics in the Post-Genomic Age*, Maagan, Israel. Abstract P50.

Gini, C. (1933) I Samaritani. *Genus* **1**, 117–146.

Hammer, M. F. & Zegura, S. L. (1996) The role of the Y chromosome in human evolutionary studies. *Evol Anthropol* **5**, 116–134.

Hammer, M. F., Redd, A. J., Wood, E. T., Bonner, M. R., Jarjanazi, H., Karafet, T. M., Santachiara-Benerecetti, S., Oppenheim, A., Jobling, M. A., Jenkins, T., Ostrer, H. & Bonné-Tamir, B. (2000) Jewish and middle eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Nat Acad Sci USA* **97**, 6769–6774.

Hammer, M. F., Karafet, T. M., Redd, A. J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H. & Zegura, S. L. (2001) Hierarchical patterns of global human Y-chromosome diversity *Mol Biol and Evol* **18**, 1189–1203.

Hauswirth, W. W. & Clayton, D. A. (1985) Length heterogeneity of a conserved displacement-loop sequence in human mitochondrial DNA. *Nucleic Acids Res* **13**, 8093–8104.

Helgason, A., Sigurðardóttir, S., Gulcher, J. R., Ward, R. & Stefansson, K. (2000) MtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* **66**, 999–1016.

Heyer, E., Puymirat, J., Dielties, P., Bakker, E. & de Knijff, P. (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Molecular Genet* **6**, 799–803.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J. & Labuda, D. (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* **69**, 1113–1126.

Howell, N., Kubacka, I. & Mackey, D. A. (1996) How rapidly does the human mitochondrial genome Evolve? *Am J Hum Genet* **59**, 501–509.

Howell, N. & Smejkal, C. B. (2000) Persistent heteroplasmy of a mutation mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. *Am J Hum Genet* **66**, 1589–1598.

Ingman, M., Kaessmann, H., Paabo, S. & Gyllensten, U. (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* **408**, 708–713.

Jazin, E., Soodyall, H., Jalonen, P., Lindholm, E., Stoneking, M. & Gyllensten, U. (1998) Mitochondrial mutation rate revisted: hot spots and polymorphism. *Nat Genet* **18**, 109–110.

Jobling, M. A., Pandia, A. & Tyler-Smith, C. (1997) The Y chromosome forensic analysis and paternity testing. *Int J Legal Med* **110**, 118–124.

Karafet T. M., Osipova L. P., Gubina M. A., Posukh O. L., Zegura S. L. & Hammer M. F. (2002) High levels of Y chromosome differentiation among Native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol* **74**. 761–789.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C. *et al.* (2000) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet* **66**, 1580–1588.

Kayser, M., Krawczac, M., Excoffier, L., Dieltjes, P., Corach, D., Pascali, V., Gehrig, C., Berlini, L. F., Jespersen, J., Bakker, E., Roewer, L. & de Knijff, P. (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human population. *Am J Hum Genet* **68**, 990–1018.

Kittles, A. R., Perola, M., Peltonen, L., Bergen, A., Aragon, R., Virkkunen, M., Linnoila, M., Goldman, D. & Long, L. (1998) Dual origins of Finns revealed by Y chromosome haplotype variation. *Am J Hum Genet* **62**, 1171–1179.

Korostishevsky, M., Ginzburg, E., Bonné-Tamir, B. (2001) Mutation origin reconstruction based on adjacent haplotypes. *The First Workshop on Information Technologies Application to Problem of Biodiversity and Dynamics of Ecosystem in North Eurasia*, Novosibirsk. Russia. Abstract P219.

Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonné-Tamir, B., Sykes, B. & Torroni, A. (1999) The emerging tree of west Eurasian mtDNA: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* **64**, 232–249.

Nebel A., Filon, D., Weiss, D. A., Weale, M., Weale, M., Faerman, M., Oppenheim, A. & Thomas, M. G. (2000) High resolution Y chromosome microsatellite haplotypes of

Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews. *Hum Genet* **107**, 630–641.

Nebel A., Filon, D. M., Brinkmann B., Majumder P. P., Faerman, M. & Oppenheim, A. (2001) The Y chromosome pool of Jews as a part of genetic landscape of the Middle East. *Am J Hum Genet* **69**, 1095–1112.

Parsons, T. J., Muniec, D. S., Sullivan, K., Woodyatt, N., Alliston-Greiner, R., Wilson, M. R., Berry, D. L., Holland, K. A., Weedn, V. W., Gill, P. & Holland, M. M. (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet* **15**, 363–368.

Pel-Or, Y., Korostishevsky, M., Kalinsky, H. & Bonné-Tamir, B. (1997) Sequence types and mutation rate of the hypervariable I region of mitochondrial DNA in the Samaritans. *International Conference on Molecular Biology and Evolution*, Abstract P26.

Petermann, H. Von (1860) Reisen im Orient. Leipzig: Von Veit and Co.

PHYLIP (1995) Phylogeny Inference Package, release 3.57c. Department of Genetics, University Washington, USA.

Redd, A. J., Agellon, A. B., Kearney, V. A., Contreras V. A., Karafet, T., Park, H., de Knijff, P., Butler, J. M. & Hammer, M. F. (2002) Forensic value of fourteen novel STRs on human Y chromosome. *Forensic Science International* **130**, 97–111.

Roberts, D. F. & Bonné, B. (1973) Reproduction and inbreeding among the Samaritans. *Social Biology* **20**, 64–70.

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M. & de Knijff, P. (1996) Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. *Hum Mol Genet* **5**, 1029–1033.

Sheffield, V. C., Beck, J. S., Kwitek, A. E., Sandstrom, D. W., Stone, E. M. (1993) The sensitivity of single-strand conformation polymorphism analysis for the detection of single base substitutions. *Genomics* **16**, 325–332.

Sigurðardóttir, S., Helgason, A., Gulcher, J. R., Stefansson, K. & Donneelly, P. (2000) The mutation rate in the human mtDNA control region. *Am J Hum Genet* **66**, 1599–1609.

Skorecki, K., Selig, S., Blazer, S., Bradman, R., Bradman, N., Warburton, P. J., Ismajlowicz, M. & Hammer, M. F. (1977) Y chromosomes of Jewish priests. *Nature* **385**, 32.

Soodyall, H., Jenkins, T., Mukherjee, A., Toit, E., Roberts, D. F. & Stoneking, M. (1997) The founding mitochondrial DNA lineages of Tristan da Cunha Islanders. *Am J Phys Antropol* **104**, 157–166.

Thomas, M. G., Skorecki, K., Parafitt, T., Bradman, N. & Goldstein, D. B. (1998) Origins of Old Testament priests. *Nature* **394**, 138–140.

Tully, L. A., Parsons, T. J., Steighner, R. J., Holland, M. M., Marino, M. A. & Prenger, V. L. (2000) A sensitive gradient-gel electrophoresis assay reveals a high frequency of heteroplasmy in hypervariable region 1 of the human mtDNA control region. *Am J Hum Genet* **67**, 432–443.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L. & Oefner, P. J. (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**, 358–361.

Walsh, B. (2001) Estimating the time to the most recent common ancestor for Y chromosome of mitochondrial DNA for a pair of individuals. *Genetics* **158**, 897–912.

Wilson, I. J. & Balding, D. J. (1998) Genealogical inference from microsatellite data. *Genetics* **150**, 499–510.

Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y chromosomal binary haplogroups. *Genome Res* **12**, 339–348.

Received: 27 August 2002
Accepted: 21 November 2002

