6-27-2013

Genetics and the history of the Samaritans: Y-chromosomal microsatellites and genetic affinity between Samaritans and Cohanim

Peter J. Oefner
University of Regensburg, Germany

Georg Hõlzl
University of Regensburg, Germany

Peidong Shen
Stanford Genome Technology

Isaac Shpirer
Assaf Harofeh Medical Center, Israel

Dov Gefel
Barzilai Medical Center, Israel

See next page for additional authors

Authors
Peter J. Oefner, Georg Hölzl, Peidong Shen, Isaac Shpirer, Dov Gefel, Tal Lavi, Eilon Wolf, Jonathan Cohen, Cengiz Cinnioglu, Peter A. Underhill, Noah A. Rosenberg, Jochen Hochrein, Julie M. Granka, Jossi Hillel, and Marcus W. Feldman

June 27, 2013

Genetics and the history of the Samaritans: Y-chromosomal microsatellites and genetic affinity between Samaritans and Cohanim

Peter J. Oefner,[1,2] Georg Hölzl,[1] Peidong Shen,[3] Isaac Shpirer,[4] Dov Gefel,[5] Tal Lavi[6], Eilon Wolf[6], Jonathan Cohen[6], Cengiz Cinnioglu,[7] Peter A. Underhill,[7] Noah A. Rosenberg,[8] Jochen Hochrein,[1] Julie M. Granka,[8] Jossi Hillel, [6] and Marcus W. Feldman[8]

[1]Institute of Functional Genomics, University of Regensburg, Regensburg, Germany.
[2]Present address: Centre of Systems Biology, Harvard Medical School, Boston, MA, USA.
[3]Stanford Genome Technology Center, Palo Alto, CA, USA
[4]Pulmonary Institute, Assaf Harofeh Medical Center, Zerifin, Israel.
[5]Department of Medicine–C, Barzilai Medical Center, Ashkelon, Israel.
[6]Department of Genetics, Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel.
[7]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA.
[8]Department of Biology, Stanford University, Stanford, CA, USA.

Correspondence to Marcus W. Feldman, Department of Biology, Stanford University, Stanford, CA 94305-5020. Email: mfeldman@stanford.edu

Suggested running head: "Genetic origins of Samaritans"

ABSTRACT

The Samaritans are a group of some 750 indigenous Middle Eastern people, about half of whom live in Holon, a suburb of Tel Aviv, and the other half near Nablus. The Samaritan population is believed to have numbered more than a million in late Roman times, but less than 150 in 1917. The ancestry of the Samaritans has been subject to controversy from late Biblical times to the present. In this study, liquid chromatography-electrospray ionization quadrupole ion trap mass spectrometry was used to allelotype 13 Y-chromosomal and 15 autosomal microsatellites in a sample of 12 Samaritans chosen to have as low a level of relationship as possible, and 461 Jews and non-Jews. Estimation of genetic distances between the Samaritans and seven Jewish and three non-Jewish populations from Israel, as well as populations from Africa, Pakistan, Turkey, and Europe, revealed that the Samaritans were closely related to Cohanim. This result supports the position of the Samaritans that they are descendants from the tribes of Israel dating to before the Assyrian exile in 722–720 BCE. In concordance with previously published single-nucleotide polymorphism haplotypes, each Samaritan family, with the exception of the Samaritan Cohen lineage, was observed to carry a distinctive Y-chromosome short tandem repeat haplotype that was not more than one mutation removed from the six-marker Cohen modal haplotype.

Introduction

The origin of the Samaritans, a distinct religious and cultural minority in the Middle East, has generated controversy among historians, biblical scholars, and orthodox Jewish sects (Talmon 2002). According to their tradition, they are descendants of Ephraim and Manasseh, sons of Joseph, and Levitical priests, from Shechem (traditionally associated with the contemporary city of Nablus). Much of the controversy concerning their origin revolves around the conquest of the northern biblical kingdom of Israel by the Assyrians, under Sargon II, that is understood to have taken place in 722–721 BCE.

It was the custom of the Assyrians to replace the people of a conquered area by people from elsewhere. In the Nimrud Prisms (inscribed clay documents generally attributed to Sargon), the victory over Samaria (the northern kingdom) is recorded (Fuchs 1994):

> *"The inhabitants of Samaria/Samerina, who agreed [and plotted] with a king [hostile to] me not to do service and not to bring tribute [to Ashshur] and who did battle, I fought against them with the power of the great gods, my lords. I counted as spoil 27,280 people, together with their chariots, and gods, in which they trusted. I formed a unit with 200 of [their] chariots for my royal force. I settled the rest of them in the midst of Assyria. I repopulated Samaria/Samerina more than before. I brought into it people from countries conquered by my hands. I appointed my eunuch as governor over them. And I counted them as Assyrians."*
>
> Nimrud Prisms, COS 2.118D, pp. 295–296.

The biblical book of Kings describes the result of the Assyrian victory in similar terms:

> *"And the king of Assyria brought men from Babylon and from Cuthah and from Ara and from Hamath and from Sepharaim and placed them in the cities of Samaria instead of the children of Israel and they possessed Samaria…"*
>
> II Kings 17: 24

However, recalling that Hezekiah ruled the southern kingdom of Judea from 715 BCE, after the Assyrian victory, the following passage from the book of Chronicles seems to contradict the above statement from II Kings:

> *"And Hezekiah sent to all Israel and Judah and wrote letters also to Ephraim and Manasseh that they should come to the home of the Lord at Jerusalem to keep the Passover."*
>
> II Chronicles 30: 1

After the emperor Cyrus allowed the Judean exiles to return from Babylon in 538 BCE, the reconstruction of the Temple began in 520 BCE. The historian Talmon (2002) refers to disputes between the Samaritans and the leaders of the returned exiles over where to build the Temple, the Samaritans wanting it at their sacred Mount Gerezim rather than in Jerusalem. Talmon regards the claims of the book of Kings to have been an attempt by the leaders of the returning exiles to ostracize the Samaritans, who were subsequently regarded at best as second-class citizens.

During Roman times (fourth and fifth centuries CE) the Samaritan population is believed to have reached more than a million, but persecution, forced conversion, and forced migration by subsequent rulers and invaders decimated the population to the extent that they numbered 146 in the year 1917 (Ben Zvi 1957).

Samaritan writing, which resembles ancient Hebrew, is used in their Holy Scriptures. They observe the tenets of the Hebrew Bible, the Torah, but not the other parts of the Jewish scriptures. In addition, membership in the Samaritan group is transmitted along the male line, as opposed to the post-biblical rule of Jewish transmission, which is maternal. Children of Samaritan males who marry non-Samaritan females are included as Samaritans, but females who marry outside the Samaritan community are expelled.

Marriage among Samaritans is mostly endogamous, and the group is highly inbred with 84 percent of marriages between either first or second cousins. The mean inbreeding coefficient of 0.0618 is the highest recorded among human populations (Bonné-Tamir et al. 1980). Important genetic and demographic studies by Bonné and colleagues (1963, 1965, 1966) revealed differences in many traits from other Middle Eastern populations. For example, blood group O and color blindness are more frequent in Samaritans, while G6PD deficiency is less frequent. Their endogamous marriage customs and patrilineality have exacerbated the historical exclusion of the Samaritans by Orthodox Judaism, which is strictly matrilineal.

Cazes and Bonné-Tamir (1984) detailed pedigrees among the Samaritans. There are four lineages: the Tsedaka, who claim descent from the tribe of Manasseh; the Joshua-Marhiv and Danfi lineages, who claim descent from the tribe of Ephraim; and the priestly Cohen lineages from the tribe of Levi (Ben Zvi 1957; Schur 2002).

The present study aims at resolving the controversy over the origin of the Samaritans by analysis of 13 Y-chromosomal short tandem repeat (STR) markers in various Jewish and non-Jewish populations from Israel, Africa, Southwest Asia, and Europe, as well as 15 autosomal STRs in the Samaritan and Israeli samples only. Allelotyping was

accomplished by liquid chromatography-electrospray ionization-quadrupole ion trap mass spectrometry (Oberacher et al. 2001a, 2001b, 2003), which allowed not only the accurate determination of allele size but also the simultaneous detection of single-nucleotide polymorphisms (SNPs), several of which proved informative and enabled the generation of so-called SNPSTRs (Mountain et al. 2002). The study finds statistical evidence that the male lineages represented by the Y-chromosomes present in today's Samaritans are very similar to those of Cohanim, supporting the view that Samaritans have ancient roots in the Israelite population.

Materials and Methods

Subjects. Blood samples were taken from 47 Samaritans living in Holon, a city just south of Tel Aviv, after they had given their written consent according to the regulations of the 'Helsinki Committee'. Blood samples were kept at –80˚C until phenol/chloroform extraction of DNA from white blood cells. We originally sampled 27 males, but upon examination of their pedigrees, only one of any pair of individuals more closely related than great-grandfather/great-grandson was retained. The final sample comprised twelve individuals for analysis of Y-chromosomal polymorphism: two each from the Cohen and Danfi lineages, and four each from the Joshua-Marhiv and Tsedaka lineages.

In addition to the 12 Samaritan individuals, we included in the study 20 Ashkenazi Jews, 20 Iraqi Jews, 20 Libyan Jews, 20 Moroccan Jews, 20 Yemenite Jews, 17 Ethiopian Jews, and 25 Israeli Cohanim. All but the Cohanim, as well as 18 Druze and 20 Palestinians, were obtained from the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University (NLGIP). The 25 Cohanim and 19 additional Palestinians had been ascertained in Tel Aviv, after written consent had been obtained according to the regulations of the 'Helsinki Committee'. Thus, the Israeli sample included 12 Samaritans, 142 Jews, and 57 non-Jews. From the Human Genome Diversity Panel (HGDP) maintained at CEPH in Paris, 28 Bedouins, 23 individuals from Russia, including 16 Russians and seven Adygei from the Russian Caucasus, 29 Italians (including 14 Sardinians), 20 Burusho, 24 Brahui, 23 Balochi, 20 Pathan, and 20 Kalash were included in the study. Twenty-four African DNA samples were obtained from the Y-Chromosome Consortium collection, and 50 Turkish samples were selected randomly from a total of 523 samples distributed amongst 91 cities in Turkey (Cinnioglu et al. 2004). In total, 472 Y-chromosome DNA

samples from Africa, Southwest Asia, and Europe were genotyped in this study. Among the Israeli groups, one Cohen was removed from autosomal genotyping. For all analyses except Table 7, we used 24 of the Cohen Y chromosomes because autosomal genotyping was performed only on 24 Cohanim. Table 7 uses only the Y-chromosome genotypes, and all 25 Cohen Y chromosomes were used for this analysis.

Polymerase Chain Reaction (PCR). STRs were amplified by PCR, separated by liquid chromatography (LC) from unincorporated deoxynucleotides and primers, and then subjected to on-line electrospray ionization quadrupole ion trap mass spectrometry (MS) to determine the number of repeats and any deviation in base composition from that reported to GenBank.

The PCR protocol comprised an initial denaturation at 95°C for 3 min, 14 cycles of denaturation at 94°C for 20 s, primer annealing at 63-56°C with 0.5°C decrements, and extension at 72°C for 45 s, followed by 20 cycles at 94°C for 30 s, 56°C for 45 s, and 72°C for 45 s, and a final five-minute extension at 72°C. Each 20-μL PCR contained one unit of Optimase™ (Transgenomic, Omaha, NE) in 1x Optimase PCR buffer, 2.0 mM $MgCl_2$, 0.1 mM each of the four dNTPs, 0.2 μM each of forward and reverse primers (see Supplemental Table 1), and 20 ng of genomic DNA. In addition, DYS398 was amplified using AmpliTaq® Gold (Invitrogen, Carlsbad, CA) in 10 mM Tris-HCl (pH 8.3), 50 mM KCl, and 2.0 mM $MgCl_2$ (other conditions as for Optimase™). For comparison of the effect of different polymerases on quality of mass spectra, we also employed Discoverase™ dHPLC DNApolymerase (Invitrogen) in 60 mM Tris-SO_4 (8.9), 18 mM $(NH_4)_2SO_4$, and 2 mM $MgSO_4$ (other conditions as for Optimase™).

Two dinucleotide repeat marker loci (YCAIIa+b), three trinucleotide repeat loci (DYS388, DYS392, and DYS426), seven tetranucleotide repeat loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS393, DYS439), and one pentanucleotide repeat marker (DYS438) were typed in 472 Y-chromosomes. One autosomal dinucleotide (5SR1*), one autosomal trinucleotide (D4S2361), and 13 autosomal tetranucleotide repeat markers (F13B*, TPOX, D2S1400, D3S1358, D5S1456, D7S2846*, D8S1179, D10S1426, GATA48, D13S317*, FES, D16S539*, D17S1298) were also genotyped in the 238 Samaritan, Palestinian, Bedouin, Druze, and Jewish samples. For the five autosomal loci marked with *, a linked single-nucleotide polymorphism was also genotyped, producing five SNPSTRs (Mountain et al. 2002). All autosomal STR calculations involved only the STR parts of these five plus the other ten STRs.

Denaturing High-Performance Liquid Chromatography and Electrospray Ionization Quadrupole Ion Trap Mass Spectrometry. An UltiMate™ chromatograph (Dionex, Sunnyvale, CA) consisting of a solvent organizer and a micro pump was used to generate a primary eluant flow of 200 µL, which was then reduced to a constant secondary flow of 2.5 µL/min by means of a 375-µm o.d. fused silica restriction capillary of varying length with an internal diameter of 50 µm (Polymicro Technologies, Phoenix, AZ). The latter was connected to the eluant line with a 1/16″, 0.25 mm bore, stainless steel micro-cross (VICI, Houston, TX). A MicroPulse™ Pulse Damper (Restek, Bellefonte, PA), the outlet of which had been plugged, was also connected to the same cross to minimize pulsation and, consequently, background noise in the spectra. Chromatographic separation was performed in 50 x 0.2 mm i.d. monolithic, poly-(styrene/divinyl-benzene)

capillary columns (Huber et al. 2001) that had been obtained from Dionex (P/N 161409; Sunnyvale, CA, USA). Column temperature was held at 60°C in a custom-made oven made of heat-resistant Robalon S (Leripa Papertec LLC, Kimberly, WI) and measuring 13 x 5 x 6 cm (l x w x h). Temperature control was implemented by using an Omega CN3390 temperature control module (only one of the 10 channels was used in this study) and a reading type T thermocouple attached to the column. The temperature control unit was operated in on-off mode with a dead-band of 0.07°C. A nano-injection valve (model C4-1004, Valco Instruments) mounted into the oven was used to inject 500-nL volumes of polymerase chain reactions onto the column.

The mobile phase was 25 mM butyldimethylammonium bicarbonate (BDMAB), which was prepared by passing research grade carbon dioxide gas (Praxair, Danbury, CT) through a 0.5 M aqueous solution of analytical reagent grade butyldimethylamine (Fluka, Buchs, Switzerland) until a pH value of 8.4 was reached. Single-stranded DNA fragments were eluted with a linear LC-MS grade acetonitrile (Riedel-de Haën, Sigma-Aldrich Laborchemikalien GmbH, Seelze, Germany) gradient of typically 12-24% (v/v) in 2.5 min, followed by a 2-min wash with 70% acetonitrile in 25 mM BDMAB, before re-equilibration of the column at starting conditions for 4 min. Eluting nucleic acids were detected and mass analyzed by electrospray ionization mass spectrometry (ESI-MS) using either a three-dimensional quadrupole (LCQ Advantage) or for PCR products longer than about 200 base pairs an LTQ linear ion trap mass spectrometer (both from Thermo Finnigan, San Jose, CA). The electrospray capillary (90 µm o.d., 20 µm i.d.) was positioned orthogonally to the ion source. Electrospray voltage was set at 2.5 kV and a sheath gas flow of 20 arbitrary units of nitrogen was employed. The temperature of the heated

capillary was set to 200°C. Total ion chromatograms and mass spectra were recorded on a personal computer with the Xcalibur software version 1.3 (Thermo Finnigan). Mass calibration and tuning were performed in negative ion mode with a 0.5 μM solution of an HPLC purified 60-mer heterooligonucleotide in 25 mM BDMAB, 15% acetonitrile (v/v). Raw mass spectra were recorded over a mass-to-charge (*m/z*) range of 500-2000.

Performance characteristics of LC-MS and the impact of the choice of DNA polymerase on mass spectrometric detection sensitivity and ability to detect SNPs are given in the Appendix.

DNA Sequencing. Amplicons that showed deviations from the biomolecular mass computed from the reference sequence deposited in GenBank (Supplemental Table 2) were treated with exonuclease I and shrimp alkaline phosphatase (USB Corporation, Cleveland, OH) for 30 min at 37°C and 15 min at 80°C to remove excess deoxynucleotide triphosphates and amplimers. Bidirectional dideoxy sequencing was performed with the Applied Biosystems (Foster City, CA) Dye Terminator Cycle Sequencing Kit. Sequencing reactions were purified by solid-phase extraction using either Sephadex G-50 (Amersham Pharmacia Biotech, Piscataway, NJ) or CentriSep (Princeton Separations, Adelphia, NJ) spin columns and then run on an Applied Biosystems 3730 DNA sequencer. Sequence traces were aligned and analyzed with SeqScape v.2.5 (Applied Biosystems).

Genotyping of Y-Chromosome Single-Nucleotide Polymorphisms. A total of 84 Y-chromosomal SNPs were genotyped by DHPLC (Xiao and Oefner 2001) for the assignment of Y-chromosomes to one of a total of 67 haplogroups (Underhill et al. 2001). One

Bedouin and one Cohen Y-chromosome could not be assigned to any haplogroup because of insufficient DNA for genotyping. There were no missing data for the Y-chromosomes. For the autosomes, there were missing data. The following lists the populations and numbers of loci with less than 9% missing data: Samaritans, 12 loci Libyan Jews, 15 loci; Moroccan Jews, 15 loci; Druze, 14 loci; Bedouin, 13 loci; Iraqi Jews, 14 loci; Cohanim, 14 loci; Ethiopian Jews, 15 loci; Ashkenazi Jews, 10 loci; Palestinian, 15 loci; Yemeni Jews, 13 loci.

Statistical Analysis. For both Y-chromosomes and autosomes, expected heterozygosity is first calculated per locus and then averaged over loci. The values are obtained using Arlequin 3.5 (Excoffier and Lischer, 2011). Per locus Y-chromosomal heterozygosities are corrected to be comparable to autosomal values using the formula $H_{corr} = 4H_{uncorr}/(3H_{uncorr} + 1)$ (Perez-Lezaun et al. 1997). Averages and standard deviations were computed over the per locus values. Gene diversity, which is calculated for Y-chromosome haplotypes and corrected for sample size, was also reported by Arlequin 3.5.

F_{ST} genetic distance was computed using *Arlequin* 3.5. We corrected the Y F_{ST} values for comparison to autosomal values using the formula $F_{ST\text{corr}} = F_{ST\text{uncorr}}/(4 - 3F_{ST\text{uncorr}})$ (Perez-Lezaun et al. 1997). We also calculated Nei's (1972) genetic (standard) distance *D* using the formula $D = -\ln[(1-P_{XY})/((1-P_X)(1-P_Y))^{1/2}]$, where P_{XY} is the number of pairwise differences between populations (per locus and averaged over loci), and P_X and P_Y are the number of pairwise differences within populations (per locus and averaged over loci). Correction for sample size (Nei 1978) is $-\ln[(1-P_{XY})/(G_XG_Y)^{1/2}]$, where $G_X = [2n_X(1-P_X)-1]/(2n_X-1)$ for autosomes, and $G_X = [n_X(1-P_X)-1]/(n_X-1)$ for the Y chromosome, and n_X is

the number of individuals in the sample from population x. Locus-by-locus F_{ST} calculations were also obtained from *Arlequin* 3.5. Statistical comparisons were made using non-parametric statistics, either Mann-Whitney or Wilcoxon signed-ranks tests, which test whether two samples are drawn from the same population when the two sample variances may differ.

Genetic divergence (Goldstein et al. 1995), assuming a stepwise mutation model (Ohta and Kimura 1973, Goldstein and Schlötterer 1999), was estimated as

$$(\delta\mu)^2 = (\hat{\mu}_A - \hat{\mu}_B)^2,$$

where $\hat{\mu}_A$ and $\hat{\mu}_B$ are the number of repeats in samples from populations *A* and *B*, respectively. The expected value of $(\delta\mu)^2$ after *T* generations of separation between populations *A* and *B* is $2\omega T$, where ω is the effective mutation rate and is given by the actual mutation rate times the variance in mutational jump size (Zhivotovsky and Feldman 1995). $(\delta\mu)^2$ averaged over loci was reported from *Arlequin* 3.5.

For affinity propagation (AP) based clustering of allelotypes we used the *R* package *APCluster* (Bodenhofer et al. 2011). This approach incorporates the clustering algorithm AP (Frey and Dueck 2007) for finding clusters in a given dataset and allelotypes that are the most representative for each cluster. These are called *exemplars*. Members of a cluster are determined by passing real-valued "messages" between the points of a dataset. The messages describe the affinity that one data point has for selecting another as its cluster center. In AP the desired number of clusters can be adjusted via a parameter called *input preference*. The input preference can be regarded as the intention of a given sample to be representative of its respective cluster. In the work presented here, we tuned the input preference in an iterative approach to reach the desired number of partitions. The starting

value for the optimization process was always set to the median of the input similarities, as proposed by Frey and Dueck (2007). Dendrograms were created by exemplar-based agglomerative clustering, which produced a hierarchy of clusters using the results of an AP run. For computation of clusters, the microsatellite data were imported into *R* and subjected to analysis via AP without further data normalization.

For the autosomal dataset, the *R* function *daisy*, which is provided in the R package *cluster* (Maechler et al. 2013), was used. This function allows the handling of missing values and combines numeric values, *i.e.* the number of repeats, with associated non-numeric SNP alleles into a single non-numeric variable for the calculation of distance measures as input for AP.

Principal component analysis was performed using *XLSTAT 2013* (Addinsoft, Paris, France).

RESULTS

Gene Diversity of Samaritans and other Israeli Populations. Genotypes were obtained by means of liquid chromatography-electrospray ionization-quadrupole ion trap mass spectrometry that produces more detailed information than standard genotyping of fluorescently labeled microsatellites by means of capillary electrophoresis (see Appendix). Table 1 shows the six distinct Samaritan Y chromosome STR haplotypes. The haplotypes are identical within the Joshua-Marhiv and Tsedaka lineages. There is a single repeat difference at DYS 391 in the Samaritan Cohen lineage, and a single repeat difference at DYS 390 in the Danfi lineage. The former had been already observed by Bonné-Tamir et

al. (2003), who had typed twelve Y-chromosomal STRs in 74 Samaritan males. Two of the markers they had used, DYS385a and DYS385b, were not included in our sample of 13 markers, and they typed nine members of the Cohen lineage including five individuals who were first-degree relatives. Note that each of the four Samaritan Y-chromosomal lineages is associated with a different SNP haplogroup, shown in the last column of Table 1 and reported earlier by Shen et al. (2004). Haplotype distances, computed as the total number of repeat differences summed over loci, between pairs of Samaritan individuals are shown in Table 2, where it is clear that the Cohen and Joshua-Marhiv lineages are further from the Danfi and Tsedaka lineages than the latter two are from each other.

In Table 3, the variability in these Y-chromosomal markers in Samaritans is compared to that in our non-Samaritan sample. Both average gene diversity across loci and average number of alleles per STR marker are lower in the Samaritans; this is largely due to the three monomorhpic markers in Samaritans (Table 1: DYS19, DYS392, and YCAIIb). Apart from the Samaritans, the Bedouin, Druze, and Palestinian samples show lower mean gene diversity, which may be due to a higher frequency of cousin marriages in those groups than in the Jewish populations. For autosomal markers, Table 4 records the average gene diversity, average allele number, and average expected heterozygosity in Samaritans, which are also lower than in the other ten Israeli populations.

Both Y-chromosomal and autosomal genetic distances were calculated between the Samaritans and individual Israeli populations, and they are reported in Table 5. The Y-chromosomal distances are based on the thirteen STR markers listed in “Methods”, and the autosomal distances are computed for the fifteen STRs and also for the five SNP-STRs. *Arlequin* 3.5 gives the F_{ST} values based on allelotypes and on haplotypes: both are

reported in Table 5 for the Y chromsomes. Table 5 also lists Nei's standard genetic distance (Nei 1972) and his distance adjusted for sample size.

It is clear from Table 5 that the Samaritan Y chromosomes are closest to those of the Cohanim, by a considerable margin for most distance estimates. It is also interesting that apart from the Cohanim, the closest Y chromosomal distances to the Samaritans are those of the Yemeni Jews and the Bedouins (and, for $(\delta\mu)^2$, the Libyan Jews). Importantly, the autosomal distances of the Samaritans to the other populations do not show the special closeness to the Cohanim or to any other Jewish population.

A locus-by-locus comparison, using single-locus analyses from *Arlequin* 3.5, was made for the Y-chromosomal and autosomal STRs between the Samaritans and the combined Jewish populations (excluding the Ethiopian Jews) and between the Samaritans and the non-Jewish populations. The corresponding F_{ST} values are recorded in Table 6. The Y chromosomal comparison between the Samaritan-Jewish distances and the Samaritan-non-Jewish distances shows that the former are significantly lower than the latter (two-sided Wilcoxon signed ranks test $p = 0.003$). This is not true, however, of the autosomal distances ($p = 0.103$).

Table 7 reports three overall pairwise comparisons for both Y and autosomal data: Samaritans vs. Jewish, Samaritans vs. non-Jewish, and Jewish vs. non-Jewish population. Again the Samaritans are closer to the Jewish populations than they are to the non-Jewish for the Y-chromosomal STRs, but not for the autosomal STRs. The Jewish and non-Jewish groups are the closest of the three pairs, which is not surprising given the small number of markers and the earlier finding by Rosenberg et al. (2001), based on twenty

STRs, that these two groups were difficult to distinguish in data of comparable size to the current study.

Affinity Propagation Clustering of Y-STR Haplotypes. Affinity Propagation uses the max-product algorithm to search and score configurations of random variables in a factor graph (Frey and Dueck 2007). It does so by simultaneously considering all data points as potential prototypes or exemplars and passing around soft information until a subset of exemplars emerges around which clusters of similar data points are formed. In contrast to hierarchical agglomerative clustering it avoids hard decisions, thereby reducing the chance of making erroneous choices when forming clusters. Clusters can then be joined by exemplar-based agglomerative clustering on the basis of a matrix of pairwise similarities to obtain a cluster hierarchy or a dendrogram, wherein the heights of the vertical lines measure the similarity of two clusters, i.e., similarity decreases with increasing heights (Bodenhofer et al. 2011).

The corresponding dendrogram of Y-chromosome haplotype clusters (Fig. 1) and the frequencies of the respective clusters in the 20 populations studied as well as the relationship of haplotype clusters to Y-chromosome haplogroups are depicted in Table 8. Cluster 1 on the far left of the dendrogram stands alone among the 26 haplotype clusters. It represents 10 of 29 Italians that belong to haplogroup I-M26 and are distinguished by the unique YCAIIa,b motif 11,21. Next to it is a clade of five clusters that includes the remaining 19 Italian, as well as all Russian and Burusho Y-chromosomes haplotyped. The clear distinction of the Burusho Y-chromosomes from the other four Pakistani populations studied and their apparent affinity to European populations appears to support a

recent linguistic study that found Burushaski personal and demonstrative pronouns in their entirety to be closely related to the Indo-European pronominal system in addition to extensive grammatical correspondences in the nominal and verbal systems (Çasule 2012). It is now believed that the Burushaski language descended most probably from Phrygian, an ancient Indo-European language and population believed to have originated on the Balkan Peninsula in today's Macedonia before migrating to Asia Minor, where the Phrygians dominated most of western and central Anatolia between 1200-700 BC. A previous study of 113 autosomal microsatellites in extant Pakistani and Greek populations also concluded that there was evidence for a southeastern European contribution to the gene pool of the Burusho and the Pathan that probably predated the invasion of the Indian subcontinent in 327-323 BC by Alexander the Great (Mansoor et al. 2004).

Another distinct clade of four clusters comprises 19 of 24 Yoruba Y-chromosomes included in this study, with another three Yoruba Y-chromosomes belonging to the closely related cluster 11. The only non-Yoruba individuals assigned to clusters 7-11 were a Brahui G-P15 and a Kalash R-M207 Y-chromosome, respectively. The remaining 15 clusters capture all Israeli populations including the Ethiopian Jews, the Turkish population, and with the exception of all Burusho and a single Y-chromosome each from the Brahui and Kalash, the remaining Pakistani chromosomes. Generally, the closely related STR haplotypes captured by each of these clusters tend to belong to the same haplogroup. This is particularly obvious for closely related clusters 17, 18, and 19, which capture 123 (93.2%) of the 132 J-haplogroup individuals included in the study. The only non-J individuals included in this group of clusters were an I-M170 and an I-M253 Y-chromosome, respectively, both of which originated in Turkey and belong to cluster 19.

Further, clusters 18 and 19 accounted for 10 of 12 (83.3%), 20 of 24 (83.3%), and 18 of 28 (64.3%) of the Samaritan, Cohen, and Bedouin Y-chromosomes studied, respectively, while their relative frequency in the other populations investigated did not exceed 30%. The close relationship of the Samaritan, Cohen, and Bedouin chromosomes is also evident from principal component analysis (PCA) of all pairwise Jewish and non-Jewish population $(\delta\mu)^2$ genetic distances computed from the 13 Y-chromosome microsatellite loci (Fig. 2A). Moreover, this PCA plot also shows that PC1, which captured 94.37% of the variance in the data, was responsible for the clear distinction of the Italian, Russian, and Burusho Y-chromosomes that was also obvious from Affinity Propagation-based clustering. Interestingly, PCA of all pairwise Jewish and non-Jewish population F_{ST} values failed to separate the Italian, Russian, and Burusho Y-chromosomes as clearly (Fig. 2B). However, the Samaritan, Cohen, and Bedouin Y-chromosomes clearly group together.

Affinity Propagation based clustering of the 238 Jewish and non-Jewish individuals collected in Israel based on 15 autosomal STRs did not set any of the 11 populations apart from the others (Supplemental Figure 3). Interestingly, irrespective of the number of clusters Affinity Propagation was instructed to generate, over a range of 5 to 16, the 12 Samaritans were always allocated to four different clusters. In contrast to the Y-chromosome, however, the four Samaritan lineages could not be assigned to separate clusters. Principal component analysis, on the other hand, clearly separated the Samaritans from the other Israeli populations, including the Cohanim and Bedouins, irrespective of the measure of genetic distance used (Supplemental Figure 4). Distinction of the Samaritans is most likely driven by the limited diversity of autosomal allelotypes found in

Samaritans compared to other populations rather than the presence of distinct allelotypes among Samaritans, since Affinity Propagation-based clusters 4-7 capture (with the exception of the Ethiopian Jewish and the Bedouin population) between 65 and 77% of the members of the remaining populations. Finally, inclusion of SNPs identified in the course of LC/MS-based allelotyping of the autosomal STRs did not improve resolution.

Discussion

The genetic study of the origin of the Samaritans may assist in the estimation of the historical value of biblical sources and their chronology. Their origin has been a contentious issue for millennia, leading to discrimination against the Samaritans and, as a consequence, to their near extinction at the hands of the various rulers of the southern Levant. Addressing it by means of scientific evidence has been impossible until recently due to the paucity of data. This study, which complements an earlier study based on simple sequence polymorphism discovered by the re-sequencing of 7,280 bp of non-recombining Y-chromosomes and 5,622 bp of coding and hypervariable segment I mitochondrial DNA sequences in Samaritans and neighboring Jewish and non-Jewish populations (Shen et al. 2004), begins to provide an informative genotypic database for the Samaritans and assesses their genetic affinity with their historical neighbors.

In recent years, several studies have applied genetic polymorphisms to compare Jews of various ethnic origins (Ostrer 2001). Hammer et al. (2000) used 18 biallelic Y chromosome markers to study the paternal gene pool of various Jewish and Middle-Eastern populations. Their results suggested that modern Jewish Y-chromosome diversity derived mainly from a common Middle Eastern source population rather than from ad-

mixture with neighboring non-Jewish populations during and after the Diaspora. Nebel et al. (2000) used six microsatellite and 11 single nucleotide polymorphism (SNP) markers on the Y-chromosome to reveal two modal haplotypes of Israeli and Palestinian Arabs (~14% and ~8% for the two haplotypes). They demonstrated that the Y-chromosome distribution in Arabs and Jews was similar but not identical and suggested a relatively recent common ancestry. Rosenberg et al. (2001) studied 20 unlinked autosomal microsatellites in six Jewish and two non-Jewish populations and found that the Libyan Jewish group retained a genetic signature distinguishable from those of the other populations. They also identified evidence of some similarity between Ethiopian and Yemenite Jewish groups, reflecting possible migration in the Red Sea region. Nebel et al. (2001) analyzed six Middle Eastern populations (three Jewish and three non-Jewish populations residing in Israel) for 13 binary polymorphisms and six microsatellite loci. Their results showed that in comparison with data available from other relevant populations in the region, Jews were found to be more closely related to groups in the north of the Fertile Crescent (Kurds, Turks, and Armenians) than to their Arab neighbors (Palestinian Arabs, and Bedouins).

Thomas et al. (2002) analyzed the maternally inherited mitochondrial DNA from each of nine geographically separated Jewish groups, eight non-Jewish host populations, and an Israeli Palestinian Arab population. Their results suggested that most Jewish communities were founded by relatively few women, that the founding process was independent in different geographical areas, and that subsequent genetic input from females in the surrounding populations was limited.

Recent studies of microsatellite (Kopelman et al. 2009; Listman et al. 2010) and single-nucleotide (Atzmon et al. 2010; Behar et al. 2010; Campbell et al. 2012) polymorphisms on autosomes have been able to statistically distinguish European, North African and Middle Eastern Jewish populations from their non-Jewish neighbors. Our data show that next to the Cohanim, the closest Y-chromosomal group to the Samaritans, using F_{ST}, is the Yemeni Jews, while using $(\delta\mu)^2$ it is the Bedouins, whose autosomes, with F_{ST}, are actually the closest to those of the Samaritans (Table 5). These relationships are interesting in light of the close connection between Yemeni Jews and Bedouins shown by the neighbor-joining tree in Ostrer and Skorecki (2013), which is based on autosomal SNPs.

Microsatellite data from markers on the Y chromosome distinguish between Samaritans and other populations in the area. The Samaritans have fewer alleles per microsatellite locus than the other populations. This can be explained by their exceptionally small population size and by the high degree of inbreeding inside the community. A related finding of allelic paucity in ß-thalassemia genes among Samaritans was reported by Filon et al. (1994). The Samaritan Y-chromosomes are significantly closer to those of the Jewish groups than to Palestinians. Exact tests for population differentiation using the Y markers also distinguish Samaritans from Palestinians but not Samaritans from Jews. The Y-chromosome distance of the Samaritans from Palestinians is significantly greater than that of the autosomes.

Among the 12 Y-chromosomes analyzed, seven haplotypes were found. Two were in the Cohen lineage, one in the Joshua-Marhiv lineage, two in the Danfi lineage, and two in the Tsedaka lineage. The relationships between these chromosomes are shown as a matrix in Table 2, where the off-diagonal elements record the total number of single-

repeat steps, summed over the whole chromosome. Each lineage has at most very minor differences among its members. Bonné et al. (2003) claim that the custom of endogamous marriages among Samaritans is practiced not only within the limits of the community but also often within each lineage. The Samaritan Cohen lineage is clearly very different from all others, and there is an indication that the Tsedaka lineage separates from the other three lineages. In other words, the nine Y-chromosome STR markers seem to resolve the four lineages. The broader tribal categories, however, are less clear. The distance of Joshua-Marhiv to Danfi, which is actually greater than that to Tsedaka, is not in accord with the proposed ancient origins, namely Menasseh for Tsedaka, Levi for Cohen, and Ephraim for Danfi and Joshua-Marhiv, respectively. The four lineages seem clearer as genetic groups than the three tribes. The separation of the Samaritan Cohen lineage from the others is reflected in the large distances in the first two rows of Table 2. Bonné et al. (2003) have also reported that the Samaritan Cohen lineage represents a different Y haplogroup from all other Samaritan lineages. Here, as shown in other studies (e.g., Jobling 2001), Y chromosome haplotypes are surrogates for surnames.

Among a number of Jewish populations of either Ashkenazi or Sephardic origin, an important component in the sharing of Y-chromosomes is the Cohen Modal Haplotype (CMH), first described by Thomas et al. (1998). The CMH is defined by alleles 14, 16, 23, 10, 11, and 12 at the STR loci DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393, respectively (Table 3). The CMH was observed 23 times in the present study: in eight of our 25 Cohanim, three Ashkenazi, two Iraqi, one Libyan and one Yemenite Jew, as well as three Brahui, two Turks, one Baluch, and one Italian, respectively. Nine of our 12 Samaritans (Table 1) were only one step removed from the CMH, as was the

case for eight Cohanim, six Bedouins, five Turks, two Palestinians, two Moroccan Jews, two Druze, one Baluch, and one Libyan, Iraqi and Yemenite Jew each. Y-chromosome similarities reflected the large number of haplotypes in the sample, which was too small to include only those haplotypes that are observed in at least one population (e.g., Thomas et al. 2000 used a threshold of ten percent) in the calculation of identity. It is interesting that in terms of the number of single repeats separating the haplotypes of the Samaritan lineages from the CMH, the distances between CMH and C_1 and C_2 were 6 and 7, respectively, while the distances between CMH and the other Samaritan haplotypes were 1, with the single exception of D_1. This suggests that, contrary to expectation on the basis of their family names, the Tsdaka, Joshua-Marhiv and Danfi lineages share a common ancestor with the paternally inherited Jewish high priesthood more recently than does the Samaritan Cohen lineage.

There are two main hypotheses for the origin of Samaritans. The first, which is argued by the orthodox Jewish authorities and a few modern scholars (Kaufman 1956), is that Samaritans are not Israelites at all but were brought to Israel by the Assyrian king when he conquered Israel (722–720 BC) and exiled its people (II Kings 17: 23–24). If this view were true, assuming that modern Jewish populations are continuous with the ancient Jewish populations, we would not expect similarity of Samaritans and modern Jewish populations. The second hypothesis, which is argued by the Samaritans themselves, is that they are descendants of Israelites who remained in Israel after the Assyrian conquest and diverged from the mainstream more than 2500 years ago. They remained isolated until the present time (although foreign elements from the surrounding Arabic people have been incorporated into their style of life). The Israeli historian S. Talmon

(2002) supports the Samaritans' claim that they are mostly descendants of the tribes of Ephraim and Menasseh that remained in Israel after the Assyrian conquest. His opinion is that the statement in the Bible (II Kings 17: 24) is tendentious and intended to ostracize the Samaritans from the rest of Israel's people (see also Cogan and Tadmor 1988). In fact, II Chronicles 30: 1 may be interpreted as confirming that a large fraction of the tribes of Ephraim and Menasseh (i.e., Samaritans) remained in Israel after the Assyrian exile. In comparing Samaritans to Jews and to Palestinians, the latter comprise a local neighboring reference population. In his book, Ben Zvi (1957) indicates that under the rule of the Moslems (end of the thirteenth century), the Samaritan population gradually declined and they were moved to Egypt, Syria, and to other Middle Eastern locations. Gene flow from these local populations to the Samaritans could then have occurred.

Taken together, our results suggest that there has been gene flow between non-Samaritan females and the Samaritan population to a significantly greater extent than for males. The male lineages of the Samaritans, on the other hand, seem to have considerable affinity with those of the five non-Ethiopian Jewish populations examined here. These results are in accordance with expectations based on the endogamous and patrilineal marriage customs of the Samaritans and provide support for an ancient genetic relationship between Samaritans and Israelites.

Acknowledgments

We thank the Samaritan community of Holon for their participation and collaboration in this study. Research supported in part by NIH grant GM28428.

References

Atzmon, G., L. Hao, I. Pe'er, C. Velez, A. Pearlman, P. F. Palamara, B. Morrow, E. Friedman, C. Oddoux, E. Burns, and H. Ostrer. 2010. Abraham's children in the genome era: major Jewish diaspora populations comprise distinct genetic clusters fwith shared Middle Eastern ancestry. *Am. J. Hum. Genet.* 86: 850-859.

Behar, D. M., B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, J. Parik, S. Rootsi, G. Chaubey, I. Kutuev, G. Yudkovsky, E. K. Khusnutdinova, O. Balanovsky, O. Semino, L. Pereira, D. Comas, D. Gurwitz, B. Bonne-Tamir, T. Parfitt, M. F. Hammer, K. Skorecki, and R. Villems. 2010. The genome-wide structure of the Jewish people. *Nature* 466: 238-242.

Ben Zvi, I. 1957. *The exiled and the redeemed*. Philadelphia, Jewish Publication Society of America.

Bodenhofer, U., A. Kothmeier, and S. Hochreiter. 2011. APCluster: an R package for affinity propagation clustering. *Bioinformatics* 27:2463-2464.

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

Bonné, B. 1966. Genes and phenotypes in the Samaritan isolate. *Am. J. Phys. Anthropol*. 24:1-19.

Bonné, B., A. Adam, I. Ashkenazi, and M. Bat-Miriam. 1965. A preliminary report on some genetical characteristics in the Samaritan population. *Am. J. Phys. Anthropol*. **23**, 397-400.

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

Bonné-Tamir, B., M. Korostishevsky, A.J. Redd, Y. Pel-Or, M.E. Kaplan, and M. Hammer. 2003. Maternal and paternal lineages of the Samaritan isolate: Mutation rates and time to most recent common male ancestor. *Ann. Hum. Genet.* 67:153-164.

Campbell, C. L., P. F. Palamara, M. Dubrovsky, L. R. Botigué, M. Fellous, G. Atzmon, C. Oddoux, A. Pearlman, L. Hao, B. M. Henn, E. Burns, C. D. Bustamante, D. Comas, E. Friedman, I. Pe'er, and H. Ostrer. 2012. North African Jewish and non-Jewish populations form distinctive, orthogonal clusters. *Proc. Natl. Acad. Sci. USA* 109: 138655-13870.

Çasule, I. 2012. Correlation of the Burushaski pronominal system with Indo-European and phonological and grammatical evidence for a genetic relationship. JIES 40:59-153.

Cazes, M.H., and B. Bonné-Tamir. 1984. Genetic evolution of the Samaritans. *J. Biosoc. Sci.* 16:177-187.

Cinnioglu, C., R. King, T. Kivisild, E. Kalfoglu, S. Atasoy, G.L. Cavalleri, A.S. Lillie, C.C. Roseman, A.A. Lin, K. Prince, P.J. Oefner, P. Shen, O. Semino, L.L. Cavalli-Sforza, and P.A. Underhill. 2004. Excavating Y-chromosome haplotype strata in Anatolia. *Hum. Genet.* 114:127-148.

Cogan, M., and H. Tadmor. 1988. *II Kings (AB)*, pp. 210–211. Anchor Bible, Doubleday, N.Y.

Excoffier, L., and H. Lisscher. 2011. *Arlequin* 3.5. An integrated software package for population genetic data analysis. URL: http://cmpg.unibe.ch/software/arlequin3.

Filon, D., V. Oron, S. Krichevski, A. Shaag, Y. Shaag, T.C. Warren, A. Goldfarb, Y. Shneor, A. Koren, M. Aker, A. Abramov, E.A. Rachmilewitz, D. Rund, H.H. Kaza-

zian, and A. Oppenheim. 1994. Diversity of beta-globin mutations in Israeli ethnic-groups reflects historic events. *Am. J. Hum. Genet.* 54:836-843.

Frey, B.J., and D. Dueck. 2007. Clustering by passing messages between data points. *Science* 315:972-976.

Fuchs, A. 1994. *Die Inschriften Sargons II*. Aus Khorsabad, Göttingen.

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

Goldstein, D.B., C. Schlötterer. 1999. *Microsatellites.* Oxford View Press. Oxford UK.

Hammer, M.F., A.J. Redd, E.T. Wood, M.R. Bonner, H. Jarjanzi, T. Karafat, S. Santachiara-Benerecetti, A. Oppenheim, M.A. Jobling, T. Jenkins, H. Ostrer, and B. Bonné-Tamir. 2000. Jewish and Middle Eastern non-Jewish populations share a common pool of Y-Chromosome biallelic haplotypes. *Proc. Natl. Acad. Sci. USA* 97:6769-6774.

Heyer, E., J. Puymirat, P. Dieltjes, E. Bakker, and P. de Knijff, P. 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6:799-803.

Jobling, M. 2001. In the name of the father: surnames and genetics. *Trends Genet*. 17:353-357.

Kaufmann, Y. 1977. *History of the religion of Israel: volume IV*. Ktav Pub. House, New York.

Kayser, M., A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim,

S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, P.M. Schneider, R. Szibor, J. Teifel-Greding, G. Weichhold, P. de Knijff, and L. Roewer. 1997. Evaluation of Y-chromosomal STRs: a multicenter study. *Int. J. Legal Med.* 110:125–133.

Kayser, M., L. Roewer, M. Hedman, L. Henke, J. Henke, S. Brauer, C. Krüger, M. Krawczak, M. Nagy, T. Dobosz, R. Szibor, P. De Knijff, M. Stoneking, and A. Sajantila. 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.

Kopelman, N. M., L. Stone, D. Gefel, M. W. Feldman, J. Hillel, and N. A. Rosenberg. 2009. Genomic microsatellites identify shared Jewish ancestry intermediate between Mediterranean and European populations. *BMC Genet*. **10**: 80.

Listman, J. B., D. Hasin, H. R. Kranzler, R. T. Malison, A. Mutirangura, A. Sughondhabirom, E. Aharonovich, B. Spivak, and J. Gelernter. 2010. Identification of population substructure among Jews using STR markers and dependence on reference populations included. *BMC Genet.* 11: 48.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2013. cluster: Cluster Analysis Basics and Extensions. *R* package version 1.14.4.

Mansoor, A., S. Mazhar, A. Hameed, S. Rehman, S. Siddiqi, M. Papaioannou, L.L. Cavalli-Sforza, S.Q. Mehdi, Q. Ayub. 2004. Investigation of the Greek ancestry of populations from northern Pakistan. *Hum. Genet.* 114:484-490.

Mountain, J. L., A. Knight, M. Jobin, C. Gignoux, A. Miller, A. A. Lin, and P. A. Underhill. 2002. SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes

for inference of population history and mutational processes. *Genome Res.* 12: 1766-1772.

Nebel, A., D. Filon, D.A. Weiss, M. Weale, M. Faerman, A. Oppenheim, and M.G. Thomas. 2000. High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with haplotypes of Jews. *Hum. Genet.* 107:630-641.

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

Nei, M. 1972. Genetic distance between populations. *Am. Nat.* 106:283–292.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89:583–590.

Oberacher, H., P. J. Oefner, W. Parson, and C. G. Huber. 2001a. On-line liquid chromatography mass spectrometry: a useful tool for the detection of DNA sequence variation. *Angew. Chem. Int. Ed.* 40:3828-3830.

Oberacher, H., W. Parson, R. Muhlmann, and C. G. Huber. 2001b. Analysis of polymerase chain reaction products by on-line liquid chromatography-mass spectrometry for genotyping of polymorphic short tandem repeat loci. *Anal. Chem.* 73:5109-5115.

Oberacher, H., C. G. Huber, and P. J. Oefner. 2003. Mutation scanning by ion-pair reversed-phase high-performance liquid chromatography-electrospray ionization mass spectrometry (ICEMS). *Hum. Mutat.* 21:86-95.

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

Ostrer, H. 2001. A genetic profile of contemporary Jewish populations. *Nature Rev. Genet*. 2:891-898.

Ostrer, H., and K. Skorecki. 2013. The population genetics of the Jewish people. *Hum. Genet.* 132: 119-127.

Pérez-Lezaun, A. F. Calafell, M. Seielstad, E. Mateu, D. Comas, E. Bosch, and J. Bertranpetit. 1997. Population genetics of Y-chromosome short tandem repeats in humans. *J. Mol. Evol.* 45: 265-270.

Rosenberg, N.A., E. Woolf, J.K. Pritchard, T. Schaap, D. Gefel, I. Shpirer, U. Lavi, B. Bonné-Tamir, J. Hillel, and M.W. Feldman. 2001. Distinctive genetic signatures in the Libyan Jews. *Proc. Natl. Acad. Sci. USA* 98:858-863.

Schur, N. 2002. The Samaritans in the Mamluk and Ottoman periods and in the twentieth century. In: *The Samaritans* (E. Stern and H. Eshel, eds.) (In Hebrew.) Yad Ben-Zvi Press, Jerusalem.

Shen, P., T. Lavi, T. Kivisild, V. Chou, D. Sengun, D. Gefel, I. Shpirer, E. Woolf, L. Hillel, M.W. Feldman, and P.J. Oefner. 2004. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum. Mutat.* 24:248-260.

Talmon, S. (2002) Biblical traditions on Samaritan history. In: *The Samaritans* (E. Stern and H. Eshel, eds.) (In Hebrew.) Yad Ben-Zvi Press, Jerusalem.

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

Thomas, M.G., T. Parfitt, D.A. Weiss, K. Skorecki, J.F. Wilson, M. le Roux, N. Bradman, and D.B. Goldstein. 2000. Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba—the "black Jews of southern Africa". *Am. J. Hum. Genet.* 66:674–686.

Thomas, M.G., M.E. Weale, A.L. Jones, M. Richards, A. Smith, N. Redhead, A. Torroni, R. Scozzari, F. Gratrix, A. Tarekegn, J.F. Wilson, C. Capelli, N. Bradman, and D.B. Goldstein. 2002. Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am. J. Hum. Genet*. 70:1411-1420.

Underhill, P.A., G. Passarino, A.A. Lin, P. Shen, R.A. Foley, M. Lahr, P.J. Oefner, and L.L. Cavalli-Sforza. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* 65:43-62.

Zhivotovsky, L.A., and M.W. Feldman. 1995. Microsatellite variability and genetic distances. *Proc. Natl. Acad. Sci. USA* 92:11549-11552.

Zhivotovsky, L.A., L. Bennett, A.M. Bowcock, and M.W. Feldman. 2000. Human population expansion and microsatellite variation. *Mol. Biol. Evol*. 17:757-767.

APPENDIX

Liquid Chromatography-Mass Spectrometry (LC-MS) of Short Tandem Repeats. Denaturing capillary electrophoresis with fluorescence detection is still the method of choice for sizing short tandem repeats (STRs). Its precision of ±0.2 nucleotides in length generally suffices to ensure with 99.7% confidence the identity of a PCR-amplified dinucleotide repeat-containing fragment of 350 base pairs (Wenz et al. 1998); genotyping errors, however, remain a common occurrence with this approach (Ewen et al. 2000). In contrast, the high mass accuracy of approximately 0.01% (corresponding to ±0.04 nucleotides for a 350-bp fragment) of electrospray ionization ion trap mass spectrometry not only permits the detection of length but also single-base substitutions in STRs (Oberacher et al. 2008). Further, mass spectrometry has the advantage that measurements do not require fluorescent or radioactive labeling or the inclusion of size markers or allelic ladders. The on-line coupling of ion-pair reversed-phase high-performance liquid chromatography (IP-RP-HPLC) to electrospray-ionization mass spectrometry provides critical desalting and purification of amplicons from unincorporated deoxynucleotides and primers, whose preferential ionization would otherwise impede detection of the amplicons in the lower femtomole range (see also Supplemental Figure 1). Further, IP-RP-HPLC performed at elevated temperatures (>60°C) provides a simple means of denaturing double-stranded PCR products into their complementary single-stranded components, thereby doubling the operational size range and enabling two independent mass measurements for every amplicon, namely for the forward and the reverse strand, which typically differ enough in mass to be resolved (Supplemental Table 2) (Xiao & Oefner 2001; Hoelzl and Oefner 2004). Further, with this approach in addition to obtaining sizes of STR alleles, it be-

comes feasible to detect single-base substitutions and their respective linkage to the STR alleles (Oberacher et al. 2002). While transitions and transversions go undetected in double-stranded DNA fragments, as they result in either no difference or a difference of only one mass unit in case of the replacement of an AT-base pair (617.4 Da) with a GC-base pair (618.4 Da), base substitutions in single-stranded DNA can be identified unequivocally due to mass differences of at least 9 Da (A>T) up to 40 Da (G>C). The mass accuracy necessary to detect a shift in mass due to an A>T mutation in a 100-mer single-stranded sequence (molecular mass of approximately 31,500) has to be at least 0.014%, which is the standard accuracy of the ion trap mass spectrometers used in this study. As the sizes and molecular masses of all STRs but one, namely DYS426, exceeded 100 base pairs, A>T or T>A mutations might have gone undetected. Still, several single-nucleotide substitutions within and, in particular, flanking the microsatellite sequence, resulting in mass shifts of 15 (G>A) and 24 Da (A>C or C>A), respectively, were detected (Supplemental Table 3). Use of more highly priced time-of-flight mass analyzers would have afforded detection of any single base exchange in nucleic acids with sizes up to 250 nucleotides (Oberacher and Parson 2007).

The choice of thermostable DNA polymerase is of utmost importance for efficient mass spectrometric sizing of PCR-amplified STRs (Oberacher et al. 2006). DNA polymerases with intrinsic 3′>5′ exonuclease activity can proofread repeat deletion intermediates occurring due to enzyme slippage, thus lowering the frequency of deletion mutants by 2- to 10-fold (Kroutil and Kunkel 1999). Absence of 3′-adenylation activity and, thus, of mono- and diadenylated amplicons further improves detection sensitivity, as these PCR artefacts will otherwise compete with the PCR product of interest for ionization.

Consequently, spectra of STRs amplified with the proofreading ploymerases Optimase™ and Discoverase™ yield significantly improved signal-to-noise ratios for the major allele(s) in comparison to AmpliTaq® Gold-generated amplicons (Supplemental Figure S1). Aside from differences in PCR fidelity, provision of polymerases in storage buffers devoid of detergents eliminates the detrimental effect of detergents on performance of both reversed-phase high-performance liquid chromatography (Hecker 2003) and mass spectrometry (Oberacher et al. 2006).

The 3′-5′ proofreading exonuclease activity of Optimase™ not only has the known ability to remove a mismatched 3′ terminal base, but also at least the penultimate 3′ terminal base from the primer prior to extension and incorporation of the correct base matching that of the template. This feature is exemplified by DYS438, which contained a total of three base substitutions, one of which was located in the pentanucleotide repeat itself, while the other two were observed upstream of the repeat region (Supplemental Table 3). Of the latter two, the G>A transition (M391) was located at the penultimate position of the 3′ end of the forward primer. Detection of this base substitution came somewhat as a surprise because the primer sequence is typically incorporated into the newly synthesized strand, as can be seen from the sequence trace generated from a template amplified with the non-proofreading Ampli Taq Gold Polymerase from Applied Biosystems (Supplemental Figure 2). To confirm that the single nucleotide polymorphism, which mimics M17, is indeed located within the priming region, a new primer pair was designed. With the latter, the presence of M393 could be confirmed using both Optimase™ and AmpliTaq® Gold Polymerase generated templates.

Another peculiarity of Optimase™ is the lack of 5′-3′ exonuclease activity and, thus, its inability to degrade oligonucleotide probes that have annealed to the template strand during extension (Holland et al. 1991). This is exemplified in the present study for the duplications of the compound STR DYS389 that are separated by 52 bp and share duplicated priming sites for the forward primer. Consequently, whenever the forward primer hybridizes to both priming sites, Optimase™ will amplify preferentially only the shorter fragment containing DYS389I, while amplification of the longer fragment is aborted. In contrast, AmpliTaq® Gold Polymerase, due its 5′-3′ exonuclease activity, will degrade the shorter extension product and preferentially amplify the longer allele DYS389II. For that reason, it was necessary to amplify DYS389 with both Optimase™ and AmpliTaq® Gold Polymerase.

References

Ewen, K.R., M. Bahlo, S.A. Treloar, D.F. Levinson, B. Mowry, J.W. Barlow, and S.J. Foote. 2000. Identification and analysis of error types in high-throughput genotyping. *Am. J. Hum. Genet.* 67:727-736.

Hecker, K.H. 2003. Basics of automated, high-accuracy mutation screening with the WAVE® nucleic acid fragment analysis systems. In: *Genetic Variance Detection: Nuts & Bolts of DHPLC in Genomics* (ed. K.H. Hecker), pp. 15-34. Eagleville: DNA Press LLC.

Hoelzl, G., and P. Oefner. 2004. Microsatellite genotyping by LC/MS using the Finnigan LTQ linear ion trap mass spectrometer. *Application Note 339*, Thermo Electron Corporation.

Holland, P.M., R.D. Abramson, R. Watson, and D.H. Gelfand. 1991. Detection of specific polymerase chain reaction product by utilizing the 5'-3' exonuclease activity of *Thermus aquaticus* DNA polymerase. *Proc Natl Acad Sci USA* 88:7276-7280.

Huber, C.G., A. Premstaller, H. Oberacher, W. Xiao, G.K. Bonn, and P.J. Oefner. 2001. Mutation detection by capillary denaturing high-performance liquid chromatography using monolithic columns. *J. Biochem. Biophys. Meth.* 47:5-20.

Kroutil, L.C., and T.A. Kunkel. 1999 Deletion errors generated during replication of CAG repeats. *Nucleic Acids Res.* 27:3481-3486.

Oberacher, H., P.J. Oefner, G. Hölzl, A. Premstaller, K. Davis, and C.G. Huber. 2002. Re-sequencing of multiple single nucleotide polymorphisms by liquid chromatography-electrospray ionization mass spectrometry. *Nucleic Acids Res.* 30:e67.

Oberacher, H., H. Niederstätter, B. Casetta, and W. Parson. 2006. Some guidelines for the analysis of genomic DNA by PCR-LC-ESI-MS. *J. Am. Soc. Mass Spectrom.* 17:124-129.

Oberacher, H., and W. Parson. 2007. Forensic DNA fingerprinting by liquid chromatography-electrospray ionization mass spectrometry. *BioTechniques* 43:Svii-Sxiii.

Oberacher, H., F. Pitterl, G. Huber, H. Niederstätter, M. Steinlechner, and W. Parson. 2008. Increased forensic efficiency of DNA fingerprints through simultaneous resolution of length and nucleotide variability by high-performance mass spectrometry. *Hum. Mutat.* 29:427-432.

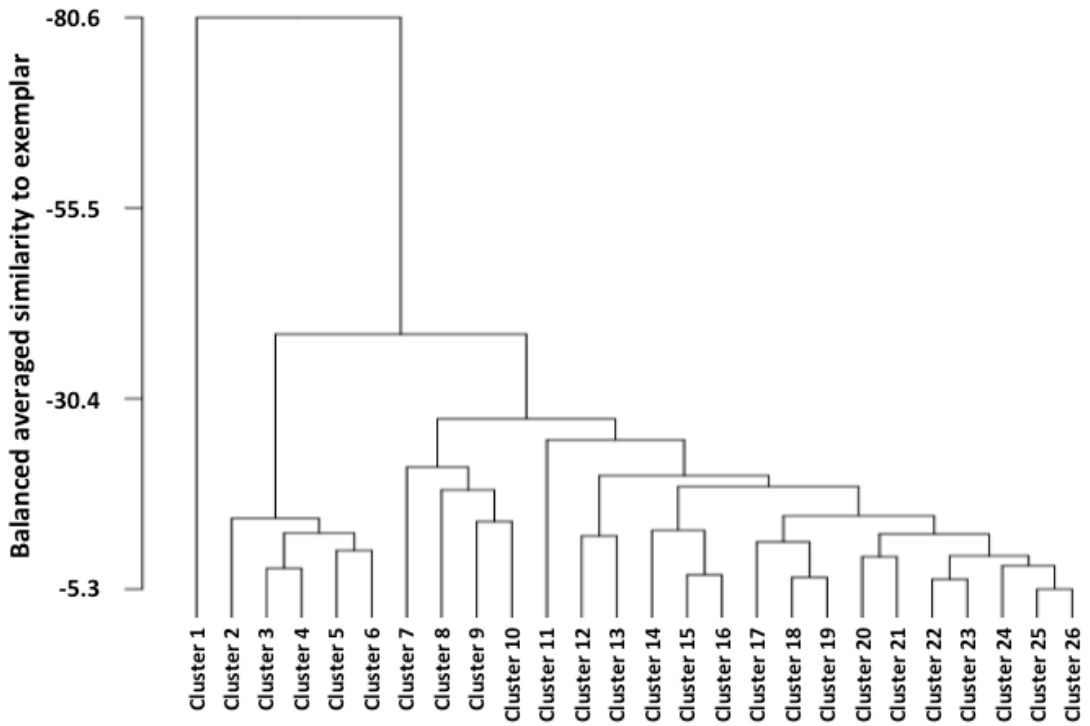
Wenz, H., J.M. Robertson, S. Menchen, F. Oaks, D.M. Demorest, D. Scheibler, B.B. Rosenblum, C. Wike, D.A. Gilbert, and J.W. Efcavitch. 1998. High-precision genotyping by denaturing capillary electrophoresis. *Genome Res.* 8:69-80.

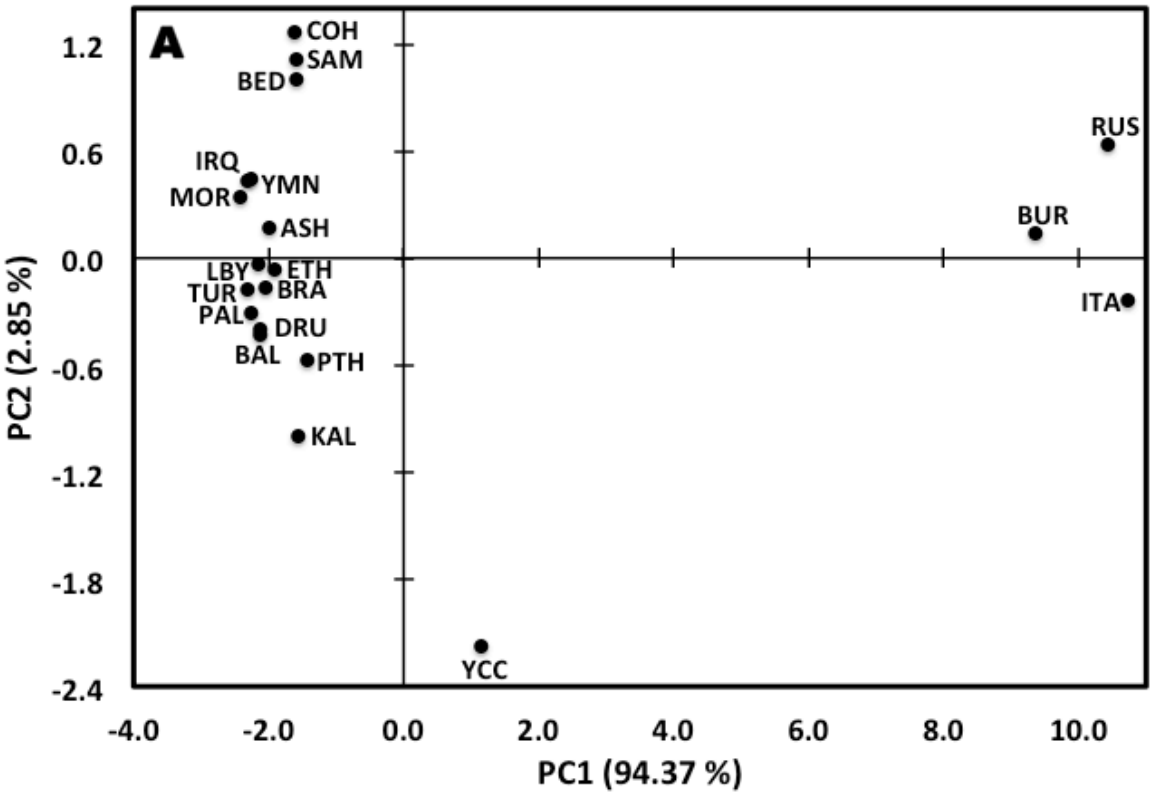
Xiao, W., and P.J. Oefner. 2001. Denaturing high-performance liquid chromatography: A review. *Hum. Mutat.* 17:439-474.

Figure Legends

Fig. 1 Affinity Propagation-based cluster dendrogram of 472 Y-chromosomal short tandem repeat (STR) haplotypes from 20 extant Jewish and non-Jewish populations grouped into 26 clusters.

Fig. 2 Principal component analysis of all pairwise Jewish and non-Jewish population using (A) $(\delta\mu)^2$ and (B) normalized F_{ST} values for 13 Y-chromosome microsatellite loci.


Balanced averaged similarity to exemplar
-80.6
-55.5
-30.4
-5.3
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 14
Cluster 15
Cluster 16
Cluster 17
Cluster 18
Cluster 19
Cluster 20
Cluster 21
Cluster 22
Cluster 23
Cluster 24
Cluster 25
Cluster 26


A
COH
SAM
BED
RUS
IRQ
YMN
MOR
BUR
ASH
LBY
ETH
TUR
BRA
ITA
PAL
DRU
BAL
PTH
KAL
YCC
1.2
0.6
0.0
-0.6
-1.2
-1.8
-2.4
-4.0
-2.0
0.0
2.0
4.0
6.0
8.0
10.0
PC2 (2.85 %)
PC1 (94.37 %)

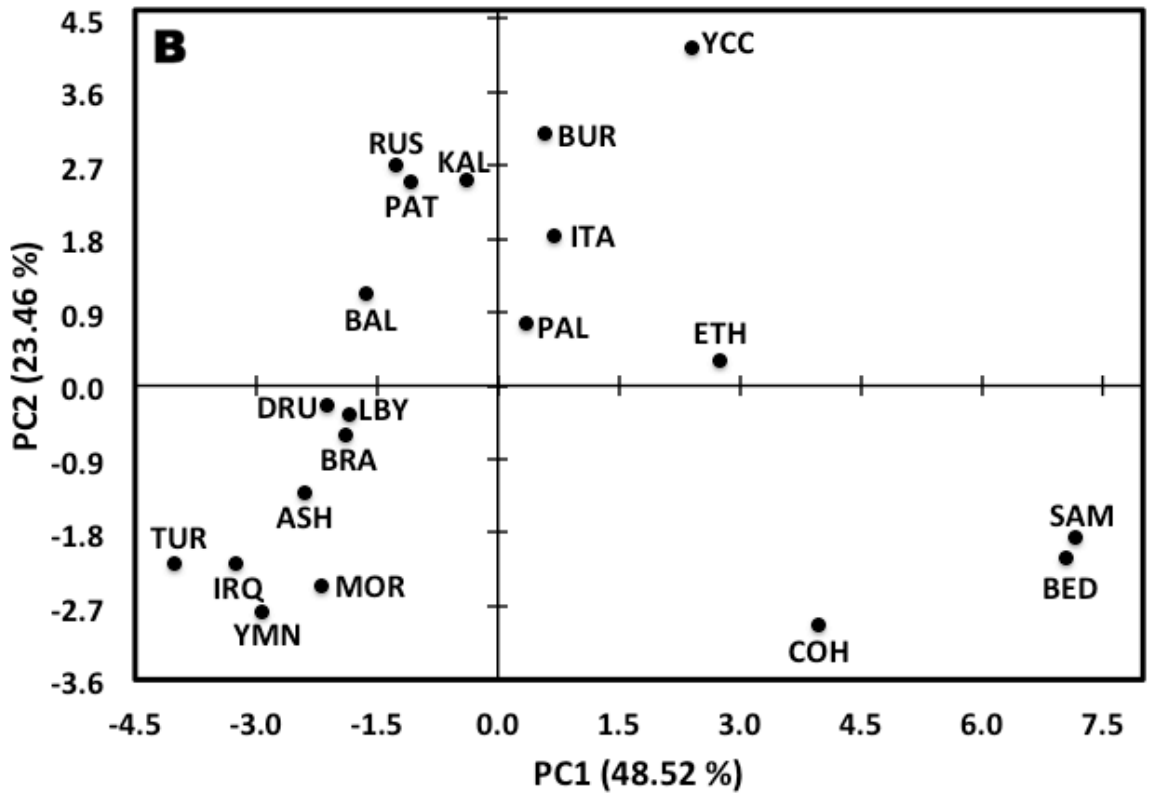

B
YCC
BUR
RUS
KAL
PAT
ITA
BAL
PAL
ETH
DRU
LBY
BRA
ASH
SAM
TUR
IRQ
MOR
BED
YMN
COH
4.5
3.6
2.7
1.8
0.9
0.0
-0.9
-1.8
-2.7
-3.6
-4.5
-3.0
-1.5
0.0
1.5
3.0
4.5
6.0
7.5
PC2 (23.46 %)
PC1 (48.52 %)

Table 1. Samaritan and Cohen Modal Y-chromosome STR haplotypes, using typing nomenclature of Kayser et al. (1997).

	Y chromosome marker (DYS prefix)													
Family	19	388	389-I	389-II	390	391	392	393	426	438	439	YCAIIa	YCAIIb	Haplogroup[3]
Cohen-1	14	12	13	18	24	10	11	13	11	10	19	19	22	E3b (M78)
Cohen-2	14	12	13	18	24	11	11	13	11	10	19	19	22	E3b (M78)
Danfi-1	**14**	15	14	16	24	**10**	**11**	**12**	11	9	18	19	22	J2f (M172, M67)
Danfi-2	**14**	15	14	16	**23**	**10**	**11**	**12**	11	9	18	19	22	J2f (M172, M67)
Joshua-Marhiv-1	**14**	**16**	13	17	**23**	11	**11**	**12**	11	10	17	22	22	J1 (M267)
Joshua-Marhiv-2	**14**	**16**	13	17	**23**	11	**11**	**12**	11	10	17	22	22	J1 (M267)
Joshua-Marhiv-3	**14**	**16**	13	17	**23**	11	**11**	**12**	11	10	17	22	22	J1 (M267)
Joshua-Marhiv-4	**14**	**16**	13	17	**23**	11	**11**	**12**	11	10	17	22	22	J1 (M267)
Tsedaka-1	**14**	15	13	16	**23**	**10**	**11**	**12**	12	8	20	19	22	J2* (M172)
Tsedaka-2	**14**	15	13	16	**23**	**10**	**11**	**12**	12	8	20	19	22	J2* (M172)
Tsedaka-3	**14**	15	13	16	**23**	**10**	**11**	**12**	12	8	20	19	22	J2* (M172)
Tsedaka-4	**14**	15	13	16	**23**	**10**	**11**	**12**	12	8	20	19	22	J2* (M172)
CMH[1]	**14**	**16**	13	16	**23**	**10**	**11**	**12**	11	10	19	22	22	J1 (M267)
CMH[2]	14/15	15/16	14	16	**23**	**10**	**11**	**12**	11	9	19	19	22/23	J2* (M172)

[1]Original Cohen Modal Haplotype (CMH) allelotypes printed in bold (Thomas et al. 1998). The allelotypes of DYS389I&II, DYS426, DYS438, DYS439, and YCAIIa&b are the consensus observed in five Samaritan and tweleve Cohen haplogroup J1 sequences.

[2]Consensus CMH STR haplotypes associated with haplogroup J2 sequences of six Samaritans and nine Cohanim.

[3]Haplogroup assignment based on single-nucleotide polymorphisms given in parentheses (Shen et al. 2004)

Table 2. Y chromosome haplotype distances among Samaritan families.

Tribe													
		Levi		**Ephraim**						**Manasseh**			
Lineage	**Family**	C1	C2	JM	JM	JM	JM	D1	D2	TS1	TS1	TS1	TS2
Cohen	C1		1	13	13	13	13	9	10	11	11	11	11
	C2			12	12	12	12	10	11	12	12	12	12
Joshua-Marhiv	JM				0	0	0	10	9	12	12	12	12
	JM					0	0	10	9	12	12	12	12
	JM						0	10	9	12	12	12	12
	JM							10	9	12	12	12	12
Danfi	D1								1	6	6	6	6
	D2									5	5	5	5
Tsedaka	TS1										0	0	0
	TS1											0	0
	TS1												0
	TS2												

Entries in the table are the total number of single-step repeat mutations between two corresponding chromosomes. Tribes may include more than one lineage as defined by family name. Family names are annotated as in Table 1.

Table 3. Within-population variation for 13 Y-chromosome microsatellites

	Expected Heterozygosity*	Gene Diversity**	Number of Alleles
Samaritans	0.801± 0.106 (0.616 ± 0.273)***	0.818 ± 0.084	2.5 ± 0.707 (2.15 ± 0.899)***
Libyan Jews	0.796 ± 0.176	0.974 ± 0.025	3.62 ± 0.870
Moroccan Jews	0.822 ± 0.139	0.984 ± 0.024	3.77 ± 1.013
Cohanim	0.747± 0.169	0.993 ± 0.014	3.54 ± 0.660
Druze	0.834 ± 0.096	0.941 ± 0.042	3.69 ± 0.947
Bedouins	0.671 ± 0.257	0.931 ± 0.030	3.83 ± 1.387
Iraqi Jews	0.860 ± 0.057	1.000 ± 0.016	4.00 ± 1.000
Ethiopian Jews	0.818 ± 0.109	0.978 ± 0.027	3.46 ± 1.127
Ashkenazi Jews	0.801 ± 0.173	0.979 ± 0.021	3.54 ± 0.877
Palestinians	0.783 ± 0.155	0.935 ± 0.033	4.39 ± 1.044
Yemeni Jews	0.849 ± 0.055	0.995 ± 0.018	3.54 ± 0.877

* Heterozygosity is corrected to be comparable to autosomal values using the formula $H_{corr} = 4H_{uncorr}/(3H_{uncorr} +1)$ for each locus; means and standard deviations are taken across corrected locus values.

** Sample-size corrected value ± standard deviation

*** Average over 13 markers including three monomorphic markers

Table 4. Within-population variation for 15 autosomal microsatellites*

	Expected Heterozygosity	Number of Alleles
Samaritans	0.616 ± 0.174	4.067 ± 1.552
Libyan Jews	0.702 ± 0.075	5.267 ± 2.738
Moroccan Jews	0.738 ± 0.063	5.667 ± 1.320
Cohanim	0.714 ± 0.056	5.333 ± 1.792
Druze	0.714 ± 0.079	5.333 ± 2.469
Bedouins	0.726 ± 0.059	6.200 ± 2.631
Iraqi Jews	0.719 ± 0.065	4.933 ± 2.336
Ethiopian Jews	0.763 ± 0.063	5.867 ± 2.446
Ashkenazi Jews	0.724 ± 0.088	5.467 ± 1.821
Palestinians	0.730 ± 0.054	6.333 ± 2.789
Yemeni Jews	0.697 ± 0.092	5.333 ± 2.658

* Allowable level of missing data was set to 0.09 to allow 15 rather than 13 loci to be included for calculations; estimates ± standard deviation; sample-size corrected values ± standard deviation

Table 5. Genetic distances of Samaritans from other populations

	F_{ST}				Nei's D (D corrected for sample size)			$(\delta\mu)^2$	
	Y*	Y Haplotypes*	Autosomes	Autosomal SNPSTRs	Y	Autosomes	Autosomes SNPSTRs	Y	Autosomes**
Libyan Jews	0.050	0.027	0.047	0.047	0.227 (0.160)	0.065 (0.01)	0.072 (0.011)	0.292	0.510
Moroccan Jews	0.038	0.025	0.045	0.039	0.172 (0.102)	0.056 (-0.003)	0.049 (-0.016)	0.422	0.493
Cohanim	0.021	0.024	0.054	0.057	0.072 (0.021)	0.078 (0.029)	0.096 (0.041)	0.076	0.378
Druze	0.055	0.032	0.056	0.050	0.260 (0.185)	0.081 (0.022)	0.078 (0.012)	0.651	0.441
Bedouin	0.041	0.033	0.036	0.045	0.128 (0.083)	0.049 (0.006)	0.072 (0.025)	0.208	0.366
Iraqi Jews	0.031	0.023	0.054	0.050	0.136 (0.059)	0.079 (0.025)	0.079 (0.020)	0.397	0.540
Ethiopian Jews	0.072	0.027	0.061	0.067	0.349 (0.275)	0.076 (0.0)	0.103 (0.018)	0.957	0.906
Askenazi Jews	0.034	0.026	0.058	0.052	0.143 (0.076)	0.086 (0.037)	0.084 (0.031)	0.425	0.507
Palestinian	0.074	0.032	0.043	0.041	0.355 (0.308)	0.057 (0.014)	0.059 (0.011)	0.599	0.414
Yemeni Jews	0.025	0.024	0.072	0.069	0.106 (0.033)	0.114 (0.063)	0.120 (0.064)	0.315	0.517

* Y F_{ST} values are corrected to be comparable to autosomal values using the formula $F_{STcorr} = F_{STuncorr}/(4-3F_{STuncorr})$.

Wilcoxon signed-rank test comparing F_{ST} values across populations (excluding Ethiopians) for autosomes vs. Y (based on separate microsatellite loci): p-value = 0.25.

Table 6. F_{ST} genetic distances per locus for the indicated population comparisons[a].

	Samaritans vs. Jews*		Samaritans vs. Non-Jews	
Y chromosome marker	**F_{ST} uncorrected**	**F_{ST} corrected****	**F_{ST} uncorrected**	**F_{ST} corrected****
DYS19	0.234	0.071	0.281	0.089
DYS388	0.174	0.050	0.327	0.108
DYS389I	0.036	0.009	0.143	0.040
DYS389II	0.002	0.001	0.072	0.019
DYS390	0.080	0.021	0.154	0.044
DYS391	0.086	0.023	-0.021	-0.005
DYS392	0.091	0.024	0.131	0.0363
DYS393	0.100	0.027	0.300	0.097
DYS426	0.043	0.011	0.189	0.055
DYS438	0.081	0.022	0.212	0.063
DYS439	0.117	0.032	0.210	0.0621
YCAII/1	0.051	0.013	0.071	0.019
YCAII/2	0.206	0.061	0.302	0.098
Mean		*0.028*		*0.056*
Autosomal Marker				
F13B		0.183		0.146
TPOX		0.148		0.149
D2S1400		0.079		0.077
D3S1358		0.031		0.031
D4S2361		0.030		0.025
D5S1456		0.031		0.024
5SR1		0.051		0.029
D7S2846		0.014		-0.0005
D8S1179		0.025		0.009
D10S1426		0.016		0.020
GATA48		0.141		0.150
D13S317		0.113		0.052
FES		0.025		0.010
D16S539		0.004		0.050
D17S1298		-0.010		-0.017
Mean		*0.059*		*0.050*

*Ethiopian Jews were excluded for this analysis.

**F_{ST} values for Y chromosomes corrected for comparison to autosomes as in Table 5. Two-sided Wilcoxon signed-rank tests (setting negative values to 0): Autosomes p = 0.103, Y chromosome p = 0.003.

a. All F_{ST} distances from Arlequin (3.5).

Table 7. Genetic distances of Samaritans from other populations, grouped into Jewish[a] and Non-Jewish subsets.

	F_{ST}				Nei's *D* (D corrected for sample		
	Y*	Y haplotypes*	Autosomes**	Autosomes SNPSTRs	Y	Autosomes	A S
Samaritans vs. Jewish	0.023	0.021	0.044	0.040	0.112 (0.073)	0.069 (0.04)	
Samaritans vs. Non-Jewish	0.041	0.025	0.039	0.037	0.198 (0.158)	0.056 (0.025)	
Jewish vs. Non-Jewish	0.009	0.003	0.003	0.003	0.048 (0.035)	0.006 (-0.005)	

a. Ethiopian Jews are excluded from this analysis.

* Y F_{ST} values are corrected to be comparable to autosomal values as in Table 5.

Table 8. Assignment of Affinity Propagation based clusters derived from 13 Y-chromosomal short tandem repeat loci and depicted in Figure 1 to their respective Jewish and non-Jewish populations Y-SNP based haplogroups.

Population	C-1	C-2	C-3	C-4	C-5	C-6	C-7	C-8	C-9	C-10	C-11	C-12	C-13	C-14	C-15	C-16	C-17	C-18	C-19	C-20	C-21	C-22	C-23	C-24	C-25	C-26
Samaritans (12)																		4	6							2
Cohanim (25)																	1	11	9	1	1	1				1
Bedouins (28)																		17	1	2		2	1	2	3	
Ashkenazi Jews (20)														2		1	1	2	4	4			1	1	4	
Iraqi Jews (20)													4	2	1			3	3	1	1	1	1		3	
Moroccan Jews (20)													2	2				4	2			2	4		3	1
Libyan Jews (20)														1		1	6		3	1			2		4	2
Yemeni Jews (20)												1		5			3	5				1	1		3	1
Ethiopian Jews (17)												7	1				1								3	5
Palestinians (39)													3	2				6		2		15	2		2	7
Druze (18)													1	1	7			2	2	1	1					3
Turks (50)												1	2	4		6	2	3	11	3	2	5	3	1	2	5
Baluch (23)														2	5	1		1	3	6	3			1	1	
Pathan (20)														2	1	3				9		1	1	1	2	
Kalash (20)									1							5	2			4		4		4		
Brahui (24)									1						1	1		3	4	8	3	1	1	1		
Burusho (20)		2	2		11	5																				
Italians (29)	10	3		10	1	5																				
Russians (23)		4	8	2	3	6																				
Yoruba (24)							4	2	5	8	3											1				1
Haplogroup (n)	I(10)	J(6)	R(10)	R(11)	R(6)	G(8)	A(4)	A(2)	E(3)	E(7)	B(3)	A(8)	T(13)	R(22)	L(14)	L(10)	J(16)	J(62)	J(46)	R(38)	R(10)	G(31)	G(14)	H(7)	E(28)	E(20)
		I(2)		Q(1)	N(3)	H(2)			B(2)	B(1)		R(1)		Q(1)	Q(1)	Q(4)			I(2)	C(2)	C(1)	E(2)	I(2)	G(1)	R(1)	DE(4)
		R(1)			L(3)	I(3)			G(1)							N(3)				B(1)		DE(1)	J(1)	DE(1)	H(1)	G(2)
					H(1)	C(1)			O(1)							J(1)				O(1)				I(1)		B(1)
					I(1)	E(1)																		J(1)		C(1)
					K(1)	O(1)																				

Supplemental Table 1. GenBank accession numbers, nucleotide positions of the 5' ends of the forward primers in GenBank accessions, ranges of observed allele sizes and corresponding numbers of repeats, and sequences of the forward and reverse primers employed for the amplification of the 13 Y-chromosome and 14 autosomal STR loci studied.

STR	GenBank Accession No.	Position	Allele sizes observed (bp)	No. of repeats	Forward primer (5'-3')	Reverse primer (5'-3')
DYS19	AF 140632	1	151-175	11-17	CTACTGAGTTCTGTTATAGTGTTTTT	ATCTGGGTTAAGGAGAGTGTCAC
DYS388	AC 004810	62380	150-174	10-18	GAATTCATGTGAGTTAGCCGTTTAGC[1]	GAGGCGGAGCTTTTAGTGAG[1]
DYS389-I	AF 140635	1	146-166	10-15	CCAACTCTCATCTGTATTATCTATG[1]	GTAAGAAGACGATGAGTCCCTATTG[1]
DYS389-II			270-290	15-21		
DYS390	AF 140636	19	132-168	17-26	GCCCTGCATTTTGGTAC	CAGAAACAAGGAAAGATAGATAGATG
DYS391	NG 002806.1	24917	136-152	8-12	CTATCATCCATCCTTATCTCTTGT	ATTGCCATAGAGGGATAGGTAGG
DYS392	AF 140638.1	23	133-151	9-16	CAACTAATTTGATTTCAAGTGTTTG	ACCTACCAATCCCATTCCTTAG
DYS393	AF140639	1	115-131	11-15	GTGGTCTTCTACTTGTGTCAATAC[1]	AACTCAAGTCCAAAAAATGAGG[1]
DYS426	AC 007034	133574	88-97	10-13	CTCAAAGTATGAAAGCATGACCA[1]	GTGTTTCAGAGCAGAACAGTGG[1]
DYS438	AC 002531	129799	211-236	8-13	TGGGGAATAGTTGAACGGTAA	GTGGCAGACGCCTATAATCC
DYS439	AC 002992	91172	205-225	16-21	TCGAGTTGTTATGGTTTTAGGTCT[1]	CCCATTTTCTTAAGGTTCGGTC[1]
YCAII a+b	AC015978	79865	144-158	16-23	TGTCAAAATTTAACCCACAATCA[1]	CGATTGGAATACCACTTTCTGACG[1]
F13B	AADC01009526.1	36818	169-185	6-10	TGAGGTGGTGTACTACCATA[2]	GATCATGCCATTGCACTCTAG[2]
TPOX	M68651	1817	114-130	8-12	CACTAGCACCCAGAACCGTCG[2]	GCTGCCAAGACCCACGATCAC[2]
D2S1400	AY083997	358	111-139	7-14	TGGAATCGTTTTACCTCTGCCTGC[3]	GATAGGTCAACGATAACTCATTCG[3]
D3S1358	AC099539.2	77721	119-143	13-19	ACTGCAGTCCAATCTGGGT[2]	ATGAAATCAACAGAGGCTTG[2]
D4S2361	AC079160.5	58789	136-162	7-16	CCACGTGACTTTCATTAGGG[3]	ACACCATCATGGCGCATG[3]
5SR1	AC026743.4	147644	156-174	13-22	CTTAAATAGACTGTGCTACTTTG[3]	ATGCTATGATTAGTAGCTAACTAGG[3]
D5S1456	AC008680.5	172273	182-218	6-15	TATCGAATTGTAACCCCGTT[3]	GCTGGAAAACCCTAATTCTCC[3]
D7S2846	AC073068	93318	170-190	10-15	TCTAAACTCCTTTGCACAGTC[3]	ACATGTGTCCATCAAATGATG[3]
D8S1179	AC100858.3	140061	161-201	8-18	TTTTTGTATTTCATGTGTACATTCG[2]	CGTAGCTATAATTAGTTCATTTTCA[2]
D10S1426	AL360172	131699	146-174	8-15	TTGGTGGTGTCATCCTCTTT[3]	CTCTTAACTGATTTGGCCGA[3]
GATA48E08	AC087783	93228	115-143	7-14	CATCCATCTCATCCCATCATT[4]	TTCACCCTACTGCCAACTTC[4]
D13S317	AL391354.12	16762	173-197	9-15	ACAGAAGTCTGGGATGTGGA[2]	GCCCAAAAAGACAGACAGAA[2]
FES/FPS	AC124248	131152	142-166	8-14	GGAAGATGGAGTGGCTGTTA[2]	CTCCAGCCTGGCGAAAGAAT[2]
D16S539	G07295	224	141-169	7-14	GATCCCAAGCTCTTCCTCTT[2]	ACGTTTGTGTGTGCATCTGT[2]
D17S1298	AADC01128115	48874	128-144	7-11	CCACCCTAGTAACTAGCATGG	GTTTGACTGGGTAGGATGG

Primer sequences were obtained from [1]Butler et al. (2002), [2]http://www.cstl.nist.gov/biotech/strbase/seq–info.htm, [3]http://www.ncbi.nlm.nih.gov, and [4]http://www.genome.ucsc.edu.

Supplemental Table 2. Expected molecular masses of repeat motifs and of the shortest alleles observed for forward and reverse strands, respectively.

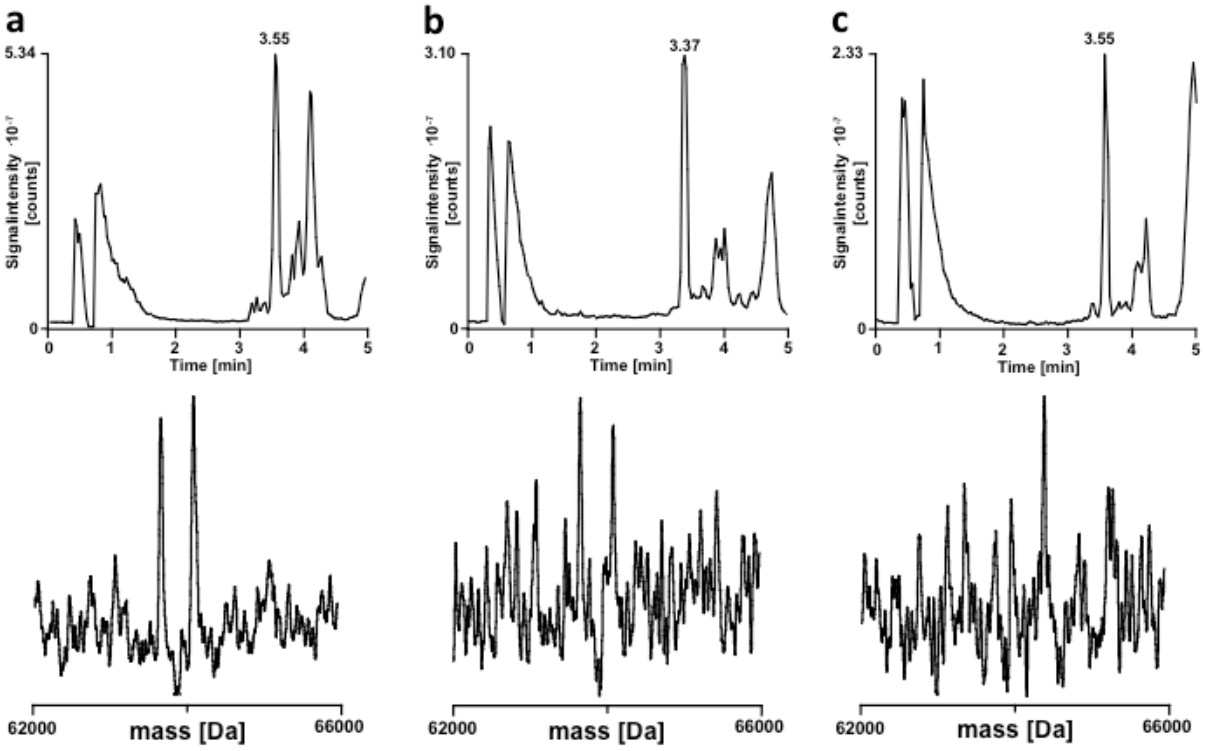
STR	$motif_{for}$	increment [DA]	Mw_{for} [Da][1]	$motif_{rev}$	increment [Da]	Mw_{rev} [Da] [1]
DYS19	TAGA	1259.82	46843.50	ATCT	1210.79	46296.22
DYS388	ATT	921.60	46234.95	TAA	930.62	46319.08
DYS389-I	[TCTG] [TCTA]	[1226.78][1210.78]	44316.34	[AGAC] [AGAT]	[1244.78][1259.82]	45751.80
DYS389-II	[TCTG] [TCTA]	[1226.78][1210.78]	80875.35	[AGAC] [AGAT]	[1244.78][1259.82]	83344.31
DYS390	[TCTG] [TCTA]	[1226.78][1210.78]	41347.72	[AGAC] [AGAT]	[1244.78][1259.82]	41284.97
DYS391	TCTA	1210.78	41180.67	AGAT	1259.82	42713.82
DYS392	TAT	921.60	40156.16	ATA	930.62	40011.19
DYS393	AGAT	1259.82	35665.19	TCTA	1210.78	35247.99
DYS426	GTT	937.60	28057.17	CAA	915.60	28037.31
DYS438	TTTTC	1505.97	64736.81	AAAAG	1582.04	65501.56
DYS439	GATA	1259.84	63989.71	CTAT	1210.79	62514.60
YCA II	CA	602.40	44213.95	GT	633.40	44609.90
F13B	AAAT	1243.83	52230.92	ATTT	1225.80	52040.89
TPOX	AATG	1259.82	35334.99	CATT	1210.78	34976.60
D2S1400	[CCTT][CCTG]	[1186.76][1211.77]	33612.65	[GGAA][GGAC]	[1284.83][1260.81]	34852.65
D3S1358	AGAT	1259.82	36881.99	ATCT	1210.78	36511.68
D4S2361	TAT	921.60	42570.62	ATA	930.62	42554.73
5SR1	CA	602.40	47839.13	GT	633.40	48407.41
D5S1456	GATA	1259.82	61291.91	TATC	1210.78	60886.62
D7S2846	CTAT	1210.78	51863.74	ATAG	1259.82	53017.58
D8S1179	[TCTA][TCTG]	[1210.78][1226.78]	49595.24	[TAGA][CAGA]	[1259.82][1244.78]	49733.37
D10S1426	GATA	1259.82	45270.53	TATC	1210.78	44796.03
GATA48E08	GATA	1259.82	35062.75	TATC	1210.78	35856.36
D13S317	GATA	1259.82	53722.96	TATC	1210.78	53030.32
FES/FPS	ATTT	1225.80	44281.70	AAAT	1243.83	43318.21
D16S539	GATA	1259.82	43629.42	TATC	1210.78	43355.05
D17S1298	[AATG][AACC]	[1259.82][1204.79]	39305.58	[TTCA][GGTT]	[1210.78][1266.80]	39646.70

[1]Theoretical molecular mass of smallest fragment observed with the least number of repeats based on reference sequence found in GenBank.

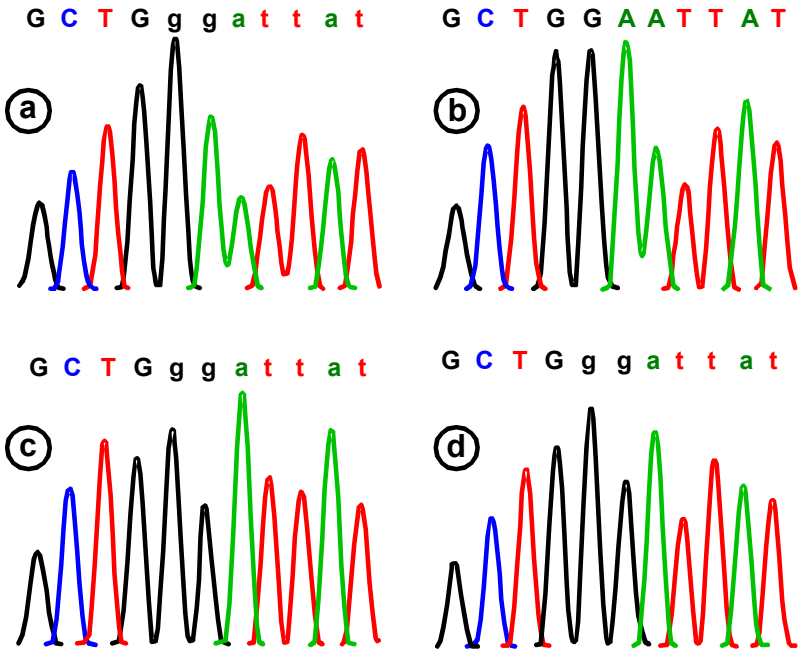
Supplemental Table 3. Nature and genomic location of single-base substitutions observed within or adjacent to short tandem repeats DYS438 and DYS393.

DYS438 (Genbank Accession No. AC002531)		Position	Marker ID*
Ref.	-cr-TTTTCTTTT C $[TTTTC]_6$ -cr- A -cr- G -cr-		
1	-cr-TTTTCTTTT **A** $[TTTTC]_6$ -cr- A -cr- G -cr-	g.129837	
2	-cr-TTTTCTTTT C $[TTTTC]_6$ -cr- **C** -cr- G -cr-	g.129884	M393
3	-cr-TTTTCTTTT C $[TTTTC]_6$ -cr- A -cr- **A** -cr-	g.129884	M391
DYS393 (Genbank Accession No. AF140639)		Position	Marker ID*
Ref.	gtggtcttctacttgtgtcaatac A GAT $(AGAT)_{14}$ -cr-		
1	gtggtcttctacttgtgtcaatac **C** GAT $(AGAT)_{11}$ -cr-	g.26	M380

*Stanford numbering system

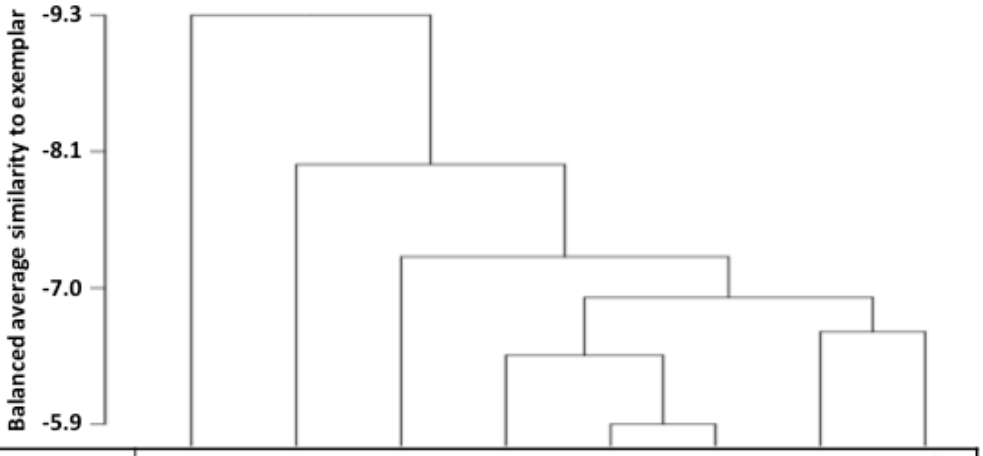


Supplemental Figure 1. Impact of different *Taq* polymerases on the spectral quality of a 184-bp amplicon: (a) Optimase™, Transgenomic, Omaha, NE; (b) Discoverase™ dHPLC DNA Polymerase, Life Technologies, Invitrogen, Carlsbad, CA; (c) AmpliTaq® Gold, Life Technologies. The upper row shows the reconstructed ion chromatograms: within the first 1.5 minutes unincorporated deoxynucleotides and primers elute from the column, followed by a peak at about 3.5 minutes, which contains the two chromatographically not resolved single-stranded components of the PCR amplicon of interest. The lower row shows the deconvoluted mass spectra of the amplicon. The two major signals in the deconvoluted mass spectra represent the mass spectrometrically resolved forward and reverse strands of the amplified DNA and their respective molecular masses in Dalton. The differences in signal-to-noise ratio reflect differences in proofreading capability, absence of 3'-adenylation activity, and polymerase storage buffer composition.



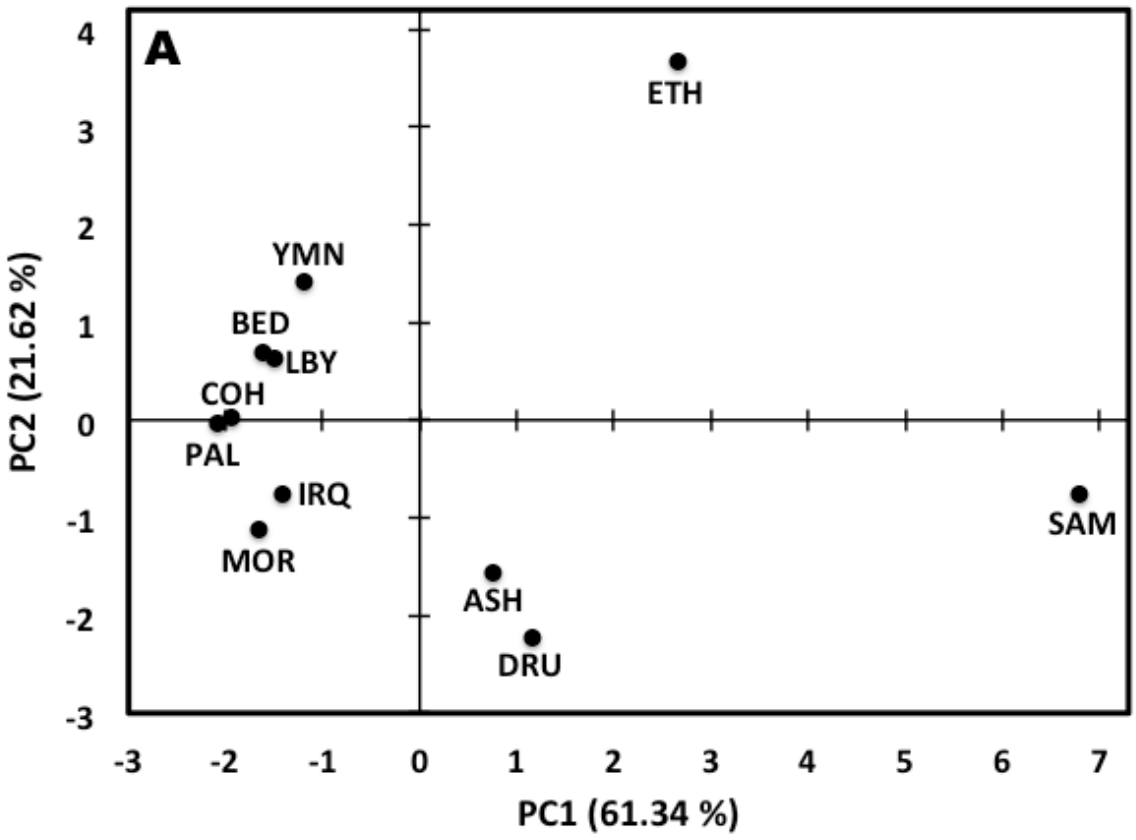
Enzyme		Mw_{for} [Da] (theor.)	Mw_{rev} [Da] (theor.)	Mw_{for} [Da] (meas.)	Mw_{rev} [Da] (meas.)	Dev_{for} [Da]	Dev_{rev} [Da]
Optimase	mutant	69254.71	70247.70	69236	70265	-18.71 (-270 ppm)	17.30 (246 ppm)
	wild type			69249	70248	-5.71 (-82 ppm)	0.30 (4 ppm)
AmpliTaq Gold	mutant	69567.92	70560.91	69565	70557	-2.92 (-4 ppm)	-3.91 (-5 ppm)
	wild type			69573	70565	5.08 (74.09 ppm)	(6 ppm)

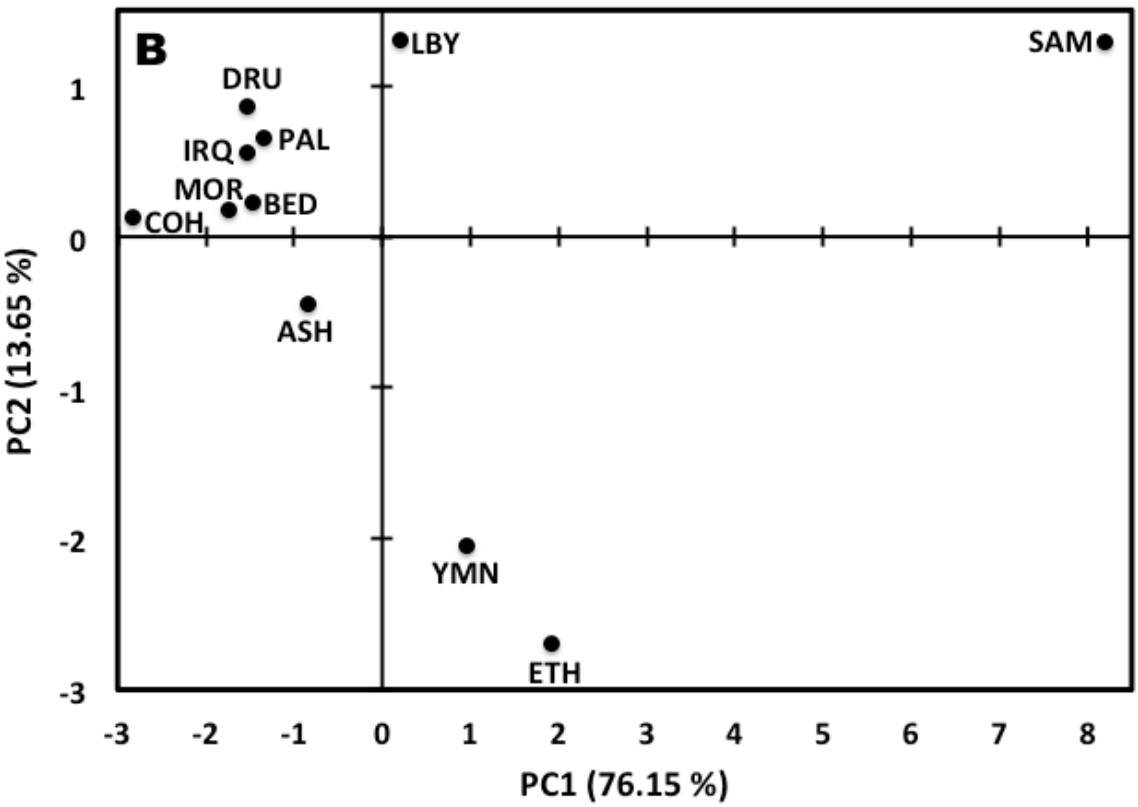
Supplemental Figure 2. Sequence traces confirm the presence of an A>G transversion next to the 3′ terminus of the primer detected after amplification with Optimase™, a proofreading enzyme with 3′-5′ exonuclease activity (a, b, d), while the single nucleotide polymorphism went undetected after amplification with AmpliTaq Gold, that lacks 3′-5′ exonuclease activity (c). (a) Optimase™, mutant, short product, (b) Optimase™, mutant, long product, (c) AmpliTaq® Gold, mutant, short product, and (d) Optimase™, wildtype, short product.



Population	Cl-1	Cl-2	Cl-3	Cl-4	Cl-5	Cl-6	Cl-7	Cl-8
Samaritans (12)				2	3	5	2	
Cohanim (24)		1	2	5	3	5	3	5
Ashkenazi Jews (20)		2	2	4	1	7	3	1
Libyan Jews (20)	2	1	3	3	4	6	1	
Yemeni Jews (20)	2		3	5	2	1	5	2
Moroccan Jews (20)	1	1	2	3	6	1	3	3
Iraqi Jews (20)		1	2	1	4	4	5	3
Palestinians (39)		2	2	6	10	3	11	5
Druze (18)	1		2	4	3	2	4	2
Bedouins (28)	1	5	4	6	5	1	3	3
Ethiopian Jews (17)		5	2	5	1	1	1	2
TOTAL (238)	7	18	24	44	42	36	41	26

Supplemental Figure 3. Affinity Propagation based clustering of 238 Jewish and non-Jewish individuals based on 15 autosomal STR loci, and respective assignment of the 8 clusters generated to the 11 Israeli populations studied.





Supplemental Figure 4. Principal Component Analysis of all pairwise Jewish and non-Jewish Israeli populations based on 15 autosomal STRs: (A) $(\delta\mu)^2$ and (B) F_{ST} values.