Haplogroup Prediction from Y-STR Values Using an Allele-Frequency Approach

T. Whit Athey

A new approach to predicting the Y-chromosome haplogroup from a set of Y-STR marker values is presented and compared to other approaches. The method has been implemented in an Excel-based program, where an arbitrary number of STR markers may be input and a "goodness of fit" score for 10 haplogroups (E3a, E3b, G, I1a, I1b, I1c, J2, N3, R1a, and R1b) is returned. This method has been applied to 101 R1b haplotypes and 50 I1a haplotypes (all having 37 STR markers available), and the distribution of results is presented. In the case of I1a, the results are compared with the predictions of another method.

Introduction

Many people have taken advantage of the availability of reasonably priced Y-chromosome testing of short tandem repeats (STRs). The resulting data can be useful in confirming genealogical relationships between two or more males. The set of repeat values that is obtained for a set of Y-chromosome markers is called a *haplotype*.

There is also considerable interest in determining the Y-chromosome *haplogroup*, a group or family of Y-chromosomes related by descent. Y haplogroups are determined by the pattern of *single nucleotide polymorphisms* (SNPs), which can also be tested and determined directly. However, the process of determining the haplogroup by direct testing of SNPs can sometimes be a lengthy process. Therefore, there is considerable interest in predicting the haplogroup from a set of STR markers.

One of the major DNA testing companies, Family Tree DNA (FTDNA), in cooperation with the University of Arizona (UAZ), uses a proprietary algorithm to predict the haplogroup for persons who have their Y-STR values tested by FTDNA. The prediction algorithm has not been published, but it appears to be based upon the genetic distance [1] of the haplotype in question to other haplotypes in the University of Arizona database. In this approach, if a haplotype (whose haplogroup is known) exists in the database that is no more than some genetic distance, reportedly a distance of two on the first 12 markers, then the haplogroup of the reference haplotype is assigned to the test haplotype as a prediction or estimation. If there are no haplogroup-confirmed haplotypes in the database within a distance of two, then no estimate of haplogroup is made. The FTDNA/UAZ approach has been fairly successful and probably 80% of customers get a haplogroup prediction.

The disadvantage of this approach is that if no prediction can be made, then the customer gets no information, even if it is very clear that some haplogroups could be ruled out, or that the haplotype is probably in one of a small number of possible haplogroups. Theoretically, the most likely haplogroups could be provided to the customer using this approach, but this is not currently done.

Another approach is based on the allele frequencies for each haplogroup and how well a given test haplotype fits the pattern of alleles in each haplogroup. This approach is outlined below and it has been implemented on a web site since October, 2004, being used by many people. It allows any number of the FTDNA set of 37 markers to be entered, and the program returns a "goodness of fit" score for 10 haplogroups (E3a, E3b, G, I1a, I1b, I1c, J2, N3, R1a, and R1b). More than 98% of people of West European extraction fall into one of these 10 haplogroups. While the program is known as a "predictor" program, it really just provides information of how well the given haplotype fits the pattern of previously reported STR values for a haplogroup.

Received January 30, 2005.

Address for correspondence: T. Whit Athey, wathey@hprg.com.

[1] The genetic distance is just the sum of the differences of the repeat values on each marker.

Nomenclature

In this paper, the order of presentation of Y-STR values is that traditionally employed by FTDNA. The 37 markers presently tested by FTDNA are the only markers for which sufficient allele frequency data are available to make the haplogroup prediction possible. The 37 markers, ordered as per the FTDNA convention, may be seen at the following web site:

http://www.ftdna.com/9markers.html

Rarely, in some haplotypes, there are extra repeat values for markers such as DYS019 (also called DYS394) and DYS464. These were ignored for purposes of the method described in this paper.

Methods

Let your haplotype be represented by the set of values, $\{w_j\}$. This can represent the haplotype for a set of 12, 25, 37 or any arbitrary number of markers up to 37. For example, we could consider the set of values that represent what FTDNA calls the "Western Atlantic Modal Haplotype" (WAMH):

$$\{w\} = \{13, 24, 14, 11, 11, 14, 12, 12, 12, 13, 13, 16\}$$

In this case the index j runs from 1 to 12.

Let $f_{ij}(x)$ represent the allele frequency at the jth marker for the ith haplogroup, where x represents the value (repeat count) of the allele. These allele frequencies are simply determined empirically from public databases and published haplotypes.[2] $f_{ij}(x)$ will form a table of values where the rows are labeled with the repeat values and the columns are labeled with the DYS marker names. Table 1 represents an example for the R1b haplogroup, using only the first 12 markers for simplicity. In practice, there will be many more columns of markers, 37 in the present implementation, and there will also be more rows required for many of the other markers. There will be a table like this for each haplogroup in the prediction program, the haplogroups being labeled with the index i, and the markers being labeled with the index j.

In Table 1, the values in the column labeled with DYS426, for example, show the frequency of occurrence of the repeat values 10, 11, 12, 13, and 14, where we see that almost all (98%) R1b haplotypes have a repeat value of 12, with small percentages for the other four closest values. Note also that the great majority of the table is "empty," or that most cells contain a frequency of zero (showing that no haplotype has been found yet with those repeat values on those markers).

Next we compute for the test haplotype, the "goodness of fit" parameter for the ith haplogroup. This calculation is straightforward, but complicated. The approach first calculates, for a given test haplotype, the following ratio for each marker:

$$f_{ij}(w_j)/ f_{ij}(w_{i,max})$$

where the f represents the table of allele frequencies. That is, for the jth marker and the ith haplogroup, we calculate the frequency from the table for the test haplotype's repeat value for that marker, and divide by the frequency of the modal value for that marker (in that particular haplogroup). As an example, let's calculate this ratio for the fourth marker (DYS391) in the haplogroup R1b for the following test haplotype, which has a value for DYS391 of 10:

$$\{w\} = \{13, 24, 14, 10, 11, 14, 12, 12, 13, 13, 13, 16\}$$

We look at the column in Table 1 labeled with DYS391 and go down the column to the row corresponding to repeat value of 10, and here we find the frequency of .318. We see that this is not the modal value for this haplogroup—11 is the modal value. For the denominator of the ratio we are calculating, we take the frequency of the modal value—the frequency for a value of 11, which we see is .628. Then our ratio becomes:

$$f_{ij}(w_j)/ f_{ij}(w_{i,max}) = 0.318/0.628 = 0.506$$

The overall "goodness of fit" parameter for that haplogroup, is simply the geometric mean[3] of all of these ratios (one for each marker). The calculation of the "goodness of fit" parameter is illustrated in detail in Table 2 for the test haplotype above and haplogroup R1b.

[2] Note that a substantial portion of the haplogroup identifications that are reported by the contributor to Y-Search and Y-Base probably came originally from the haplogroup prediction algorithm by Family Tree DNA and the University of Arizona.

[3] The geometric mean of a set of N numbers is the Nth root of the product of the N numbers.

Table 1
Allele Frequencies for Haplogroup R1b

REPEAT	DYS Marker Number											
	393	390	019	391	385a	385b	426	388	439	389a	392	389b
7	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
8	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
9	0.0%	0.0%	0.0%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
10	0.0%	0.0%	0.0%	31.8%	2.8%	0.0%	0.1%	0.0%	0.3%	0.1%	0.0%	0.0%
11	0.0%	0.0%	0.0%	62.8%	89.7%	1.6%	0.5%	0.3%	14.6%	0.4%	0.0%	0.0%
12	2.0%	0.0%	0.0%	4.9%	5.9%	1.7%	98.0%	98.4%	74.1%	3.6%	0.6%	0.0%
13	95.4%	0.0%	0.5%	0.1%	0.5%	8.7%	1.0%	1.1%	9.5%	85.8%	90.2%	0.1%
14	2.5%	0.0%	93.2%	0.0%	0.6%	69.2%	0.4%	0.2%	1.3%	9.8%	8.8%	0.0%
15	0.0%	0.0%	5.7%	0.0%	0.3%	16.5%	0.0%	0.0%	0.1%	0.3%	0.5%	5.0%
16	0.0%	0.0%	0.4%	0.0%	0.0%	2.2%	0.0%	0.0%	0.1%	0.0%	0.0%	79.3%
17	0.0%	0.0%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	13.8%
18	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.6%
19	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.2%
20	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
21	0.0%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
22	0.0%	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
23	0.0%	28.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
24	0.0%	55.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
25	0.0%	14.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
26	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 2
Example Calculation

	DYS Marker Number												Geom Mean
	393	390	019	391	385 a	385 b	426	388	439	389 a	392	389 b	
Test Haplo-type	13	24	14	10	11	14	12	12	13	13	13	16	
$f_{ij}(w_j)$	.954	.552	.932	.318	.897	.692	.980	.984	.095	.858	.902	.793	
$f_{ij}(w_{i,max})$	.954	.552	.932	.628	.897	.692	.980	.984	.741	.858	.902	.793	
Ratio	1.0	1.0	1.0	.506	1.0	1.0	1.0	1.0	.128	1.0	1.0	1.0	.796

Once the "goodness of fit" parameter is obtained, we simply multiply by 100 to get the final "score" for the haplogroup—in this case a score of about 80. We would expect to get such a high score for a haplotype that was only different from the modal haplotype at two markers. If you have the modal value for a marker in a particular haplogroup, the fraction $[f_{ij}(w_j)/ f_{ij}(w_{i\ max})]$ will equal 1.0. So, if you have all of your marker values on the modal values for a haplogroup, the geometric mean of all those ones will just be one (which when multiplied by 100 will yield a fitness score of 100).

As a practical computational matter, it is more convenient to perform the calculation of the geometric mean by instead taking the arithmetic mean of the logarithms of the ratios, and then raising 10 to the resulting power. Using this approach, the score for the ith haplogroup for a given haplotype can be written as:

$$F_i = 100 \times 10^{\{(1/N) \times \sum_j^N \log [f_{ij}(w_j)/ f_{ij}(w_{i,max})]\}},$$

where N is the number of markers considered.

A large fraction of allele frequencies for any given haplogroup (those several repeat units off of the modal value) will be zero, as we saw in Table 1 above. If any of these zero values were actually used in the calculation, finding a zero on any marker would result in the overall score for that haplogroup being zero, regardless of how well all the other markers might fit. Therefore, these zero allele frequencies were set arbitrarily to 0.000001 (one chance in a million) so that a very rare value would not totally dominate the final score. On the particular marker DYS-455, the "zero" values were set to 10^{-8} rather than 10^{-6}, because that marker is fairly convincingly diagnostic for one haplogroup, I1a, and the value for a "zero ratio" that is assigned in this case effectively weights DYS-455 more highly than the others, at least for discordant values.

It is also possible to weight some markers more heavily than others by simply counting them twice or more in the calculation, but this has not been done in the present implementation. Presumably, it would be the slower mutating markers that would be weighted more heavily, and these would tend to have a sharp and tight distribution about the modal value, resulting in ratios for non-modal values that were very low anyway, effectively weighting more heavily any slow-mutating marker. The optimum way to weight markers remains an open question.

Allele Frequencies

The approach to prediction of haplogroups outlined above requires knowledge (or at least a good estimate) of the allele frequencies for each haplogroup, which constitutes a major obstacle to successful implementation. It was only through the establishment of public Y-STR databases, such as Y-Search and Y-Base, that calculation of the allele frequencies for several haplogroups became possible. These public databases usually have included a field for the haplogroup, which were obtained primarily from the FTDNA/UAZ prediction algorithm. Therefore, a major part of the implementation of the allele frequency approach for haplogroup prediction, must include the development of a database of haplotypes from members of single haplogroups.

In identifying and collecting haplotypes for a single haplogroup, it has sometimes been possible to collect haplotypes by searching the public databases using a minimal modal haplotype for a haplogroup (obtained from published studies). This approach can identify haplotypes that match the minimal search criteria, but which also contain 25 or 37 markers. The Y-STR (minimal) haplotypes that have been published (Behar et al. 2000; Behar et al. 2004; Bosch et al. 2001; Butler et al. 2002; Capelli et al. 2003; Cinnioglu et al. 2004; DiGiacomo et al. 2004; Kivisild et al. 2003; Rootsi et al. 2004) as belonging to a particular haplogroup, confirmed by SNP testing, were also added to the database and these contributed to the allele frequencies for those few markers.

For the special case of haplogroup I1a, every haplotype so far identified as I1a, has had a DYS455 value of 8 (or rarely, 7 or 9, but never 10 or higher). This allowed a convenient method for identifying I1a haplotypes in the databases.

When compiling a set of haplotypes from one haplogroup for use in determining the allele frequencies, one is likely to find that there are several haplotypes that have the same surname. This is because a large fraction of people who are tested for Y-STR values are tested through a surname project, and individuals tested may share a common ancestor within the last few hundred years. Such haplotypes from the same surname will likely be much more similar (or may even be identical)

than two haplotypes of different surnames. Therefore, in compiling the database for determining the allele frequencies, an effort was made to avoid including haplotypes from the same surname, except when there were several differences. Within a surname project, the matching haplotypes were compared and a single representative one was selected for the database. Variant values for the cluster of haplotypes were included as a single partial "haplotype" that contained only the variant values and not the matching values. In this way, the full extent of variation could be included without overweighting the haplotypes in the cluster.

For some haplogroups such as N3, there have been very few 37-marker haplotypes reported. Therefore, the allele frequency distributions for some of the markers are very rough. If a new N3 haplotype is tested, it may have a value that has not been previously reported, causing abnormally low scores to be reported for the N3 haplogroup. Some manual "smoothing" of the allele frequency distributions at the edges was applied to help avoid this problem. One method for such "edge smoothing" uses a Gaussian curve fit to the existing data. However, many of the allele frequency distributions are obviously not Gaussian. If the approach is applied to each "tail" of the distribution independently, however, the approximate frequencies at the extremes can be estimated satisfactorily.

The haplogroup predictor algorithm has been implemented in a web-based Excel program at the following web site:

http://www.hprg.com/hapest5/

The initial version of the program is limited to the 10 most common haplogroups in Europe and to the 37 markers that most often appear in Y-Search. The lack of Y-STR data for haplogroups other than these 10 and on markers other than these 37, prevents the addition of those rarer haplogroups and other markers to the program at present. However, there is no reason why more haplogroups or markers could not be incorporated into the program as more data becomes available.

Results

Several tests were carried out using the haplogroup predictor on sets of haplotypes with known or predicted haplogroups. It is important to test the program with haplotypes that were not used in determining the allele frequencies. This constrains the extent of such validation testing because nearly all of the available haplotypes were used to determine the allele frequencies. Recently, however, additional haplotypes have been added to Y-Search that were not available at the time of compiling of the allele frequencies. There were sufficient numbers of haplotypes for testing purposes only for the most common haplogroups such as R1b and I1a.

Results From Testing R1b Haplotypes

The prediction algorithm was applied to 101 haplotypes from Y-Search where the haplogroup had been indicated as R1b. Probably, nearly all of these predicted haplogroup assignments had been provided by FTDNA. All had been tested to 37 STR markers by FTDNA, and none of the surnames associated with these haplotypes were among those used to calculate the allele frequencies used in the predictor program. Nearly all were recent additions to Y-Search where the submission occurred after the original collection of haplotypes for the calculation of allele frequencies.

The scores from the haplogroup prediction algorithm for the R1b set of haplotypes ranged from 40 to 85 with one exception. This one exception resulted in a score of 4 for R1b and raises questions about whether or not the haplotype is really in R1b. The mean of the scores is about 65. Figure 1 shows a frequency histogram for the 101 scores.

In all but three of the 101 cases, the second-highest score was for Haplogroup J2, with the three other cases having as second-highest scores, scores for R1a. In a few cases, the score for Haplogroup N3 came close to being the second-highest score. In no case did the second-highest score exceed 26, so there was no overlap in the distribution of scores, except for the one score of 4 on the one R1b haplotype.

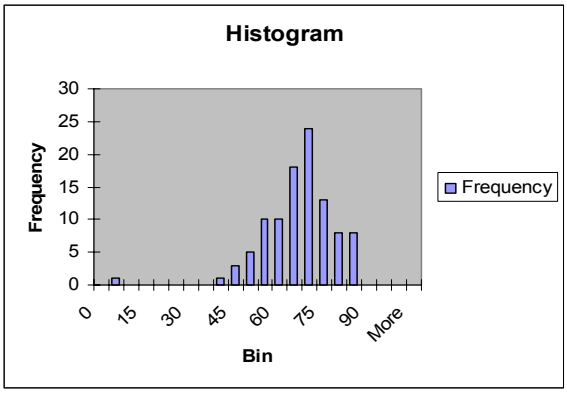

Figure 1 Distribution of R1b Scores for 101 R1b Haplotypes

Table 3
Unusual Values for an R1b Haplotype

	393	019	385a	426	388	392	389b	459a	459b	448	464				YCA2b	570	438
Unusual Haplotype	15	15	15	11	13	12	18	8	8	21	14	14	14	14	21	20	10
Frequency In R1b (%)	<1	6	<1	<1	1	1	2	2	<1	<1	13	1	<1	<1	<1	<1	<1
Modal R1b	13	14	11	12	12	13	16	9	10	19	15	15	17	17	23		11

It is instructive to examine the R1b haplotype that gave the low score of 4, in order to understand which marker values are contributing the most to the low score. Table 3 shows that this haplotype has several unusual values for the R1b haplogroup.

If there were only a few values that were off the modal values for R1b, one could still allow the possibility that the haplotype is R1b. The pattern of multiple discordant values, compared to the typical values of R1b, suggests that this haplotype might have been mislabeled.

Results from Testing Fifty I1a Haplotypes

50 haplotypes were identified on Y-Search that had a DYS455 repeat value of 7, 8, or 9 (generally considered to indicate membership in Haplogroup I1a), all with different surnames, and all with surnames different from the set of haplotypes used to compute the original allele frequencies used in the predictor program. One haplotype had a value of 9 for DYS455 and the other 49 had the value of 8. These were mostly recent additions to Y-Search. The haplogroup prediction program was applied to each of these 50 haplotypes and the resulting scores for ten haplogroups were compiled.

The I1a scores for the 50 haplotypes ranged from 31 to 89, with an average score of about 65. Only four of the haplotypes had scores less than 50. Figure 2 shows the distribution of scores.

In all but two cases, the score for I1a was at least twice that for any other haplogroup. The haplogroup with the second highest score was most often J2, with I1b, I1c, and G close behind. The scores for E3a, E3b, N3, R1a and R1b were much smaller and did not exceed 7. The single haplotype with the value of 9 on DYS455 got a rather low score of 34, accurately reporting that this haplotype has unusual values compared to typical I1a haplotype.

There is considerable overlap of the allele frequency distribution on many of the DYS markers between I1a and J2. However, the highest score observed for Haplogroup J2 on any of these 50 haplotypes was 34, and the mean value was about 23.

FTDNA does not estimate the membership in the subgroups of Haplogroup I, but only predicts overall Haplogroup I. In Y-Search, the 50 haplotypes had been labeled (by the submitter) with a haplogroup in about half of the cases. Four had been labeled with "I1a," implying that the submitter had information beyond what he may have obtained from FTDNA (or perhaps from SNP testing), 25 had been labeled with "I" (implying that those predictions came from FTDNA), and 21 had been

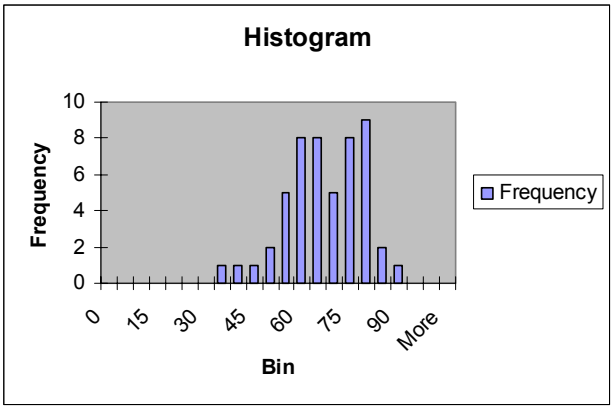

Figure 2 Distribution of I1a Scores for 50 I1a Haplotypes

labeled as "Unknown." In five of the "Unknown" cases, the submitter confirmed that FTDNA had not predicted a haplogroup for his haplotype, and in seven cases, the submitter replied that FTDNA had predicted I, but for various reasons, he had not added that information to Y-Search. In the remaining nine cases, the submitter did not respond to an inquiry.

For all but five of the 50 haplotypes, the haplogroup predictor program reported a score above 50 (one of the five just missed with a score of 49) and would have been predicted I1a (not just I) by the haplogroup predictor program.

Conclusion

The allele-frequency approach to haplogroup prediction appears to provide a powerful and robust alternative to genetic-distance approaches.

Electronic-Database Information

www.ysearch.org	genetic genealogy database
www.ybase.org	genetic genealogy database
http://www.hprg.com/hapest5/	haplogroup predictor

References

Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D, Bradman N, Weale ME (2003) Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. Am J Hum Genet 73:768–779

Behar DM, Garrigan D, Kaplan ME, Mobasher Z, Rosengarten D, Karafet TM, Quintana-Murci, Ostrer H, Skorecki K, Hammer MF (2004) Contrasting patterns of Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum Genet 114:354-365

Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. Am J Hum Genet 68:1019–1029

Butler JM, Schoske R, Vallone PM, Kline MC, Redd, AJ, Hammer MF (2002). A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. Foren Sci Int 129:10-24

Capelli C, Redhead N, Abernethy JK, Gratrix F, Wilson JF, Moen T, Hervig T, Richards M, Stumpf MP, Underhill PA, Bradshaw P, Shaha A, Thomas MG, Bradman N, Goldstein DB (2003) A Y chromosome census of the British Isles. Curr Biol 13:979–984

Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, Lillie AS, Roseman CC, Lin AA, Prince K, Oefner PJ, Shen P, Semino O, Cavalli-Sforza LL, Underhill PA (2004) Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet 114:127–148

Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, Brdicka R, Barbujani G, Papola F, Ciavarella G, Cucci F, Di Stasi L, Gavrila L, Kerimova MG, Kovatchev D, Kozlov AI, Loutradis A, Mandarino V, C. Mammi C, Michalodimitrakis EN, Paoli G, Pappa KI, Pedicini G, Terrenato I, Tofanelli S, Malaspina P, Novelletto A (2004). Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet 115:357-371

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Golge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72: 313-332

Rootsi S, Magri C, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, Barac L, Pericic M, Balanovsky O, Pshenichnov A, Dion D, Grobei M, Zhivotovsky LA, Battaglia V, Achilli A, Al-Zahery N, Parik J, King R, Cinnioglu C, Khusnutdinova E, Rudan P, Balanovska E, Scheffrahn W, Simonescu M, Brehm A, Goncalves R, Rosa A, Moisan JP, Chaventre A, Ferak V, Furedi S, Oefner PJ, Shen P, Beckman L, Mikerezi I, Terzic R, Primorac D, Cambon-Thomsen A, Krumina A, Torroni A, Underhill PA, Santachiara-Benerecetti AS, Villems R, Semino O.(2004). Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet 75:128-137