Journal: www.jogg.info
Originally Published: Volume 2, Issue 1 (Spring 2006)
Reference Number: 21.006

DIAGNOSTIC Y-STR MARKERS IN HAPLOGROUP G

Author(s): Phillip G. Goff and T. Whit Athey

DIAGNOSTIC Y-STR MARKERS IN HAPLOGROUP G

Phillip G. Goff and T. Whit Athey

Abstract

Y-Chromosome Haplogroup G reaches its highest frequency in the Caucasus Region (70% in N. Ossetia) and decreases in frequency in Western Europe to about one-to-two percent of the population on the Atlantic coast. Haplogroup G, like its brother haplogroups H, IJ, and K, arose from a mutation from Haplogroup F, M201 in the case of Haplogroup G. Y-STR databases include a limited number of haplotypes for G, H and K*, which have low frequencies in Western Europe, while IJ is well-represented in Western Europe and in Y-STR databases. In this report several Y-STR markers are identified that can distinguish a Haplogroup G haplotype from similar haplotypes in Haplogroups E3b, H, I, and J. The present study identifies four Y-STR markers, DYS425, DYS446, DYS452, and DYF399S1 that are diagnostic for Haplogroup G or one of its subgroups.

Introduction

Interest in Y-chromosome testing for paternal ancestry genealogical research has steadily increased since 2000.[1] As of December 2005, about 40,000 genealogically-relevant haplotypes are available through various online databases.[2] Many who have been tested for their Y-STR haplotype want to know their predicted haplogroup with some level of certainty before taking a SNP-test. The identification of diagnostic Y-STR markers will help to fill this demand.

Members of Y Haplogroup G have repeat values on several Y-STR markers that are distinctively different from those of other haplogroups. These markers include DYS425, DYS452, DYS446, and DYF399S1. In this article we review the available data from public databases on these markers for Haplogroup G.

DYS425 is currently offered by two DNA testing companies: Oxford Ancestors ("OA") as part of its 10-marker Y-STR product and by DNA-Fingerprint (DNAFP), starting in December 2005, as the T-associated allele of the four-copy marker, DYF371. DNA Heritage, another DNA testing company, offered DYS425 from about October 2003 through March 2004.

The study of DYS425 has been limited due to the lack of testing of this marker by more of the DNA testing companies. In addition, a review of comments on the Rootsweb Genealogy-DNA List reveals a widely-held view that DYS425 is of little diagnostic value due to the perception that it always has a repeat value of 12.

Information on the markers DYS452 and DYS446 is available primarily from the Sorenson Molecular Genetics Foundation (SMGF) database and to a lesser extent from Y-Base and Y-Search. DYF399S1 is an unusual three-copy marker that was described by Henson (2005) and recently offered commercially by DNAFP. Information on this marker is available from Y-Match and personal communications from persons who have tested their own samples at DNAFP.

Nomenclature

Some Y-STR markers are reported differently by different companies and by different researchers. In the U. S., the National Institute of Standards and Technology (NIST), and on the international scene, the International Society for Forensic Genetics (ISFG), have published guidelines in an attempt to bring standardized nomenclature conventions to the reporting of Y-STR values. This has only been partially successful, as some companies have been reluctant to change their reporting methods.

Oxford Ancestors (OA) uses a non-ISFG/NIST-standard nomenclature for the marker DYS389i. DYS389I, used in the OA database, is equal to ISFG/NIST-standard DYS389I minus three. DYS389b is reported as ISFG/NIST-standard DYS389II minus ISFG/NIST-standard DYS389I. The non-standard nomenclature must be used when searching the OA database, but the standard nomenclature will be used in discussing DYS389I in this article since it is more familiar.

[1] The Rootsweb Genealogy-DNA List had 202; 326; 681; 1,041; 799 and 2,451 messages in November 2000 (first full month), 2001, 2002, 2003, 2004 and 2005, respectively.

[2] As of 21 December 2005, Sorenson Molecular Genealogy Foundation, online http://smgf.org/ (13,489 records), ysearch, online http://www.ysearch.org/ (20,068 records), ybase, online http://www.ybase.org/ (5,515 records), Relative Genetics, online http://www.relativegenetics.com/relativegenetics/index.jsp (record count unknown) and Oxford Ancestors, online http://www.oxfordancestors.com/index.html (estimated 4,100 records).

DYS425 is part of a larger marker called DYF371. DYF371 has four alleles, three of which have a C base in a particular location adjacent to the repeat structure, and fourth has a T base in that location. DYS425 is defined as the T-associated allele of DYF371. For example, DNAFP might report the results for DYF371 as "10c-12t-13c-13c", from which the DYS425 value is shown to be 12.

The repeat value for DYS452 is reported differently by various companies. The marker consists of one continuous repeating TATAC structure of about 12 repeats, plus 19 additional contiguous units made up of CATAC, TGTAC, or TATAC units. These 19 repeats are normally invariant. Some companies (DNAH and RG) report only the main (variable) repeat value of TATAC (12 in the above example), while others (SMGF, DNAFP) add the other 19 repeats as well for a total of 31 (and this is also the ISFG/NIST-standard nomenclature). We will use the latter notation here.

There are apparently no differences in nomenclature on DYS446 used by any of the labs or databases that include this marker.

DYS399S1 has three similar alleles that are based upon a repeat unit of 'AAAG." Within the sequence containing each allele there are several extra bases that are not a part of an "AAAN" motif (where N represents any base), and the number of these extra bases, usually 10 or 11, is placed after the number of full repeats as a decimal quantity. For example, if there were 24 full repeats plus 11 extra bases, the value on the allele would be reported as 24.11. DNA Fingerprint, the company that developed the test for this marker, has followed the ISFG/NIST guidelines, but has adopted a shorter notation for convenience by subtracting 10 from the number of extra bases. Using this convention, the value of 24.11 would be reported as 24.1 by the company.

Normally, only the overall PCR length is used in routine tests of Y-STR markers, and the known structure allows an unambiguous value to be inferred from that overall length. However, for all members of Haplogroup G so far tested, the overall PCR length for the shortest allele of DYF399S1 has been such that there must be either 8 or 12 extra bases. This causes an ambiguity in interpretation because an allele with a value of 17.12 has exactly the same PCR product length as an allele with a value of 18.8. Only direct sequencing of the PCR product can distinguish these possibilities and this has not yet been done. Tentatively, the convention has been adopted that the extra bases total 12 instead of 8, so that the allele values 17.12/18.8 are reported, for example, as 17.12 (or 17.2 in the short form notation).

Methods

To test the diagnostic value of DYS425, the public repositories were searched for haplotypes with DYS425. This search included ysearch.com, ybase.com, the Rootsweb Genealogy-DNA List and websites of private surname DNA studies. In addition, academic papers were reviewed to find examples of DYS425=14. This initial review revealed multiple examples of DYS425=14 in haplotypes predicted as Haplogroup G.

To determine the degree of correlation between DYS425=14 and Haplogroup G, an effort was made to identify all Haplogroup G haplotypes in the OA database. To ensure completeness, the number of Haplogroup G results in the OA database was estimated. First, the SMGF, Y-Search and Y-Base databases were searched to determine the frequency of 9-marker modal haplotypes for Haplogroups E3b, G, I1a and R1b.[3] Next, the OA database was searched for counts of these same 9-marker modal haplotypes plus DYS425 at each of its possible values (10 through 15, plus M*--designating a missing t-associated allele). The counts in OA were divided by the weighted average frequencies in the other public databases to develop four estimates of the total records in the OA database. The average of these four estimates accurately reflects that the OA database contains about 4,108 records (November, 2005).

Y-Search and Y-Base estimate that 1.6% and 1.0% of their records (in November, 2005), respectively, are in Haplogroup G. If the OA database contains the same proportion of Haplogroup G, results, it was predicted that there would be between 41 and 68 Haplogroup G records in the OA database. The OA database was interrogated with SNP-tested Haplogroup G 9-marker haplotypes (DYS19, DYS388, DYS390, DYS391, DYS392, DYS393, DYS389i, DYS389ii-i and DYS426), from the initial Internet search and academic papers (Butler et al 2002; Behar et al 2004). DYS425 was varied from 10 through 15, plus M*, searching for exact matches. This resulted in 97 estimated Haplogroup G records. The OA database was also interrogated for DYS425 repeats in SNP-tested haplogroups other than G. Approximately 20% of the estimated number of haplotypes in the OA database were captured. For those haplotypes that did not match SNP-tested results, haplogroups were assigned using the Y-Haplogroup Predictor (Athey, 2005; see Electronic Database Information). In cases of multiple SNP-tested

[3] Using the order of DYS19, DYS388, DYS390, DYS391, DYS392, DYS393, DYS389i, DYS389i-ii and DYS426,
E3b=13,12,24,10,11,13,13,17,11;
G=15,12,22,10,11,14,12,17,11;
I1a=14,14,22,10,11,13,12,16,11; and
R1b=14,12,24,11,13,13,13,16,12.

haplogroup designations or ambiguous results from the Y-Haplogroup Predictor, other steps were taken to determine the haplogroup, such as the origin of the family in the OA record.

Until recently, the markers DYS452 and DYS446 were tested only by Sorenson Genetics and its resellers DNAH and RG. Now these markers are also available from DNAFP. Since the SMGF database covers both of these markers, it was used as the primary source of information on these markers.

Candidate Haplogroup G haplotypes were extracted from the SMGF database using somewhat different search criteria[4] from those used for the OA database. Candidate haplotypes were tested using the Haplogroup Predictor Program (Athey 2005) and only those with a score exceeding 50 for Haplogroup G were used. Multiple haplotypes with the same surname listed were deleted, retaining only one haplotype per surname (except where the haplotypes were clearly unrelated).

The marker DYF399S1 is only available from DNAFP, and none of the public databases (except DNAFP's own Y-Match) currently accept data on this marker. Therefore, all of the data for members of Haplogroup G were sent to the authors in private communications (n=5), was commissioned for the present study (n=1), or was found in Y-Match (n=1, but some of the results we received in private communications are now also in Y-Match).

Results

DYS425

DYS425=14 was found to be strongly, but not exclusively, associated with Haplogroup G in the OA database (Table 1). About 88% of the OA Haplogroup G results had 14 repeats at DYS425. Outside of Haplogroup G, 14 repeats at DYS425 was observed in one of 152 estimated Haplogroup I1a records in the OA database and in two of 69 results in Haplogroup Q in an academic study (Seielstad et al 2004). While the present study was focused on Haplogroup G, the results indicate that DYS425 may also have diagnostic value in Haplogroups E3b, H, I1b, and J.

In **Table 1**, the columns labeled G1a and G2 had SNP information that confirmed those designations. The column labeled simply G did not have SNP information but was predicted to be in G using the Haplogroup Predictor program.

Table 1 Allele Frequencies for DYS425 by Haplogroup

Repeats	E3a	E3b	F*	G	G1a	G2	H	I1a	I-P37 (pka I1b)
10									0.500
11									
12	**1.00**	0.028	**1.00**	0.115	**1.00**	0.072		**0.974**	
13						0.139	**1.00**		
14				**0.885**		**0.841**		0.007	
Missing		**0.972**				0.058		0.020	0.500
N	7	36	2	26	1	69	1	151	2

Repeats	I-M223 (pka I1c)	J	J2	K	K2	N	Q	R1a	R1b
10		0.100							
11									0.002
12	**0.875**	**0.900**	**1.00**	**1.00**	**1.00**	**1.00**	**0.812**	**1.00**	**0.981**
13									0.013
14									
Missing	0.125						0.187		0.004
N	8	10	1	1	6	3	16	15	474

[4] The search criteria were DYS388=(12 or 13), DYS391=10, DYS392=11, DYS426=11, DYS454=11, DYS455=11, DYS459=(9-9).

,

The testing of one G2-P15 subject for DYF371 was carried out to estimate whether or not the value of 14 on DYS425 was present from the beginning of Haplogroup G2. The subject was from a tribal area of India and his G2 lineage has likely been separated from the lineage that led to most European G2's from the earliest history of G2. This subject was found to have repeat values on DYF371 of 10c-12t-13c-13c, so the DYS425 value (associated with the "T" allele) was 12.

Therefore, it appears that the two repeats were added in a G2 individual at some early time after the founding of G2. Therefore, we would not normally expect to find DYS425=14 in a member of G1 or G*, and this conclusion is supported by the single example in Table 1 of a Haplogroup G1a individual, plus the single example of a GxG2 individual reported to us in a private communication.

DYS452

DYS452 is a complex marker with several sets of repeats on the main pentabase motif, TATAC, the longest of which contains about 11-14 repeats. Here is an example of the sequence for one of the YCC samples, YCC33, which is a member of Haplogroup E3a:

. . . . .
GGTGTTCTGATGAGGATAATT/TATAC/TATAC/
TGTAC/TGTAC/TATAC/TATAC/TATAC/TATAC/
TATAC/TATAC/TATAC/TATAC/TATAC/TATAC/
TATAC/TATAC/TATAC/TATAC/CATAC/TATAC/
CATAC/TATAC/TATAC/TATAC/CATAC/CATAC/
TATAC/TATAC/TATAC/CATAC/TATAC/TATAC/
TATAC/AACCAATTAATTAGCTGAGTATAATAA
. . . . .

From the sequence, we see that this example has the following repeat structure (Redd 2002):

$(TATAC)_2(TGTAC)_2(TATAC)_{14}(CATAC)_1(TATAC)_1$
$(CATAC)_1(TATAC)_3(CATAC)_2$
$(TATAC)_3(CATAC)_1(TATAC)_3$

Some commercial labs (e.g., DNA Heritage, Relative Genetics) report just the main repeat section, which would give a value of 14 in the above example. Normally, it is only this part of the marker that is variable. However, the guidelines of the International Society for Forensic Genetics (ISFG) and also the guidelines of the U. S. National Institute of Standards and Technology (NIST) suggest that all of the similar penta-base repeats in this marker should be counted, resulting in a value of 33 for YCC33, and this is how it is reported by DNA Fingerprint and SMGF (their reported values for DYS452 are 19 repeats greater than those reported by DNAH and RG).

By fortunate coincidence, one of the sequences for DYS452 that was reported by Redd (2002) is for YCC24, a member of Haplogroup G2a1-P18. The published PCR sequence actually shows the deletion. Here is the repeat structure shown by Redd for YCC24:

$(TATAC)_2(TGTAC)_2(TATAC)_{14}(CATAC)_1(TATAC)_1$
$(CATAC)_1(TATAC)_3$.....$(TATAC)_1$
$(CATAC)_1(TATAC)_3$

Here we see that 20 bases of the form

$(CATAC)_2(TATAC)_2$

have been deleted. Since the deletion occurred in a normally invariant part of the marker, it should be considered as a Unique Event Polymorphism (UEP). Interestingly, the companies reporting only what they believe to be the main repeat section on this marker, would report a value of 10 for YCC24, whereas this is not the actual number of repeats (14) of that structure.

Table 2 Allele Frequency Distribution for DYS452 by Haplogroup

Repeats	G2	Gx G2	E3a	E3b	I1a	I-P37	I-M223	J1	J2	N	Q	R1a	R1b
25	0.115												
26	**0.541**												
27	0.331												
28	0.010					0.001			0.052			0.042	
29			0.088			0.034		**0.780**	0.139		0.029	0.011	0.068
30			**0.647**	0.15	0.003	0.138	0.081	0.195	**0.671**	0.069	**0.286**	**0.800**	**0.808**
31		**1.0**	0.176	**0.81**	**0.952**	**0.724**	**0.459**	0.024	0.134	**0.897**	**0.314**	0.147	0.110
32			0.088	0.04	0.043	0.103	0.378		0.004	0.034	**0.286**		
33					0.002	0.005	0.081				0.086		
34													
n	148	4	34	27	588	29	37	41	231	29	35	95	73

For members of Haplogroup G2, they are reporting a value that is four repeats less than what is actually present. This is a good reason for using the ISFG/NIST standard nomenclature.

Allele frequencies on DYS452 for the most common European haplogroups are shown in **Table 2**. The values for Haplogroup G2 are smaller than for most haplogroups.

The limited data for GxG2 suggests that the deletion event in DYS452 occurred in a Haplogroup G2 individual or else was present in the founder of G2. The value of 27 on this marker for the G2 individual from a tribal area of India supports the idea that the deletion occurred in a person who was already G2, or that it occurred very early in the history of G2. Probably, the deletion was present in the founder of G2, but has become extinct outside G2, but this observation is based on limited data.

DYS446

The structure of DYS446 has also been reported by Redd (2002). Here is the sequence for the PCR product reported by Redd for YCC33:

$$24bp\ (TCTCT)_{13}\ 214bp$$

The structure for YCC24 (Haplogroup G2a1) is:

$$24bp\ (TCTCT)_{16}\ 214bp$$

This shows that the relatively high number of repeats (16 in this case, though 16 is actually low for G2) is simply a result of extra repeats of the usual type.

The allele frequency distribution for DYS446 in Haplogroup G2 is quite different depending on whether the value on DYS388 is 12 or 13. Possibly, there are two previously undescribed subclades of G2, one with a modal value of 12 and the other with 13 on DYS388. However, with only one repeat difference on DYS388, there is likely some overlap of values (due to normal mutations) on DYS388 from the two subclades.

Allele frequencies on DYS446 for the most common European haplogroups are shown in **Table 3**. The values for Haplogroup G are larger than for most European haplogroups, though there is a small overlap in a few cases. Only a small amount of data is so far available on Haplogroup L, but this haplogroup may have values almost as large as those of Haplogroup G.

DYF399S1

One of the alleles of DYF399S1 (probably the shortest allele in most people) has the structure:

. . .
ggttttcaccagtttgcataggtagagggaggccaaaagcccaacagg

AAA/aaat/A/aaag/aaag/aaag/AA/aaag/A/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/aaag/AAC

ttttacccttttgacagcatatgagactt . . . .

The main part of this allele (the central section above) can be written more compactly as:

$$AAA(aaat)A(aaag)_3AA(aaag)A(aaag)_{18}AAC$$

Table 3 Allele Frequency Distributions for DYS446 by Haplogroup

Repeats	G2 DYS-388=12	G2 DYS-388=13	Gx G2	E3a	E3b	I1a	I-P37	I-M223	J1	J2	N	Q	R1a	R1b
8														
9						0				0.007				
10						0.001		0.135		0.049				
11				0.025	0.02	0.023	0.158	**0.712**		0.285			0.035	0.009
12				0.146	**0.58**	0.089	0.079	0.115	0.024	0.251		0.051	**0.617**	0.070
13				**0.462**	0.22	0.642	**0.658**	0.019	0.238	0.135		**0.538**	0.348	**0.683**
14	0.032			0.196	0.12	0.205	0.105	0.019	**0.476**	0.225	0.188	0.359		0.189
15	0.113			0.089	0.03	0.036			0.214	0.049	**0.406**	0.026		0.048
16	0.226	0.046		0.044		0.004			0.048		0.344	0.026		
17	**0.371**	0.103	.2	0.025							0.063			
18	0.145	**0.322**	**.8**	0.013										
19	0.081	**0.310**												
20	0.016	0.126												
21	0.016	0.080												
22		0.011												
n	62	87	5	157	60	687	38	52	42	267	32	39	115	227

,

Table 4 Allele Frequencies for DYF399S1 in Haplogroup G

DYS399S1a			DYS399S1b			DYS399S1c			DYS399S1d		
Repeats	Count	Freq.	Repeats	Count	Freq.	Repeats	Count	Freq.	Repeats	Count	Freq.
17.2	4	.571	(missing)	1	.143	21	2	.286	(missing)	5	.714
18.2	2	.286	20.1	5	.714	22	3	.429	24	1	.143
19.2	1	.143	21.1	0		23	0		25	1	.143
			22.1	1	.143	24	2	.286			

where the lower case letters are part of a countable repeat motif and the upper case letters are "extra" bases (10 of them in the above example). This example would be scored as 18.10 (or 18.0 in the short notation).

There are as yet only a few results available for DYF399S1, but those available for Haplogroup G (all but one are G2 and the remaining one is GxG2) are shown in **Table 4**. Beside the odd structure for the shortest allele in Haplogroup G2, the whole number of repeats in the shortest allele is lower than in Haplogroups R1b, I, and J, although one example in Haplogroup I is the lowest so far found (16). Each G person represented in Table 4 has one allele with the "half" repeat (a .2 following the main number), one allele with a .1 following the main number, and one allele with a whole number. Even though the three alleles are reported in numerical order, the alleles can be distinguished for members of Haplogroup G. One person in the table had four allele values (a member of G2), apparently representing a doubling of the allele with the whole number of repeats.

Discussion

The modal value for DYS452 for Haplogroup G2 was found to be 26, lower than for any other haplogroup. The modal value for DYS446 in Haplogroup G2, in contrast, is the highest for any haplogroup. Therefore, the difference in values on these two markers would be particularly diagnostic of Haplogroup G2. The difference (DYS452 – DYS446) will typically have a value of 11 or less in Haplogroup G2, but a value of 17 or higher in other haplogroups. Interestingly, members of Haplogroup GxG2 appear to have similarly large values on DYS446, but do not have low values on DYS452. Therefore, DYS452 can serve to distinguish G2 from other parts of G.

The shortest allele of the marker, DYG399S1 has a small number of whole repeats in Haplogroup G, and also has a fractional repeat value that has only been found so far in Haplogroup G.

Members of Haplogroup G are fortunate to have several Y-STR markers that are either diagnostic or strongly suggestive of membership in G. However, these diagnostic markers are not tested as often as other STRs. With these markers now generally available, they will be of value in predicting Haplogroup G.

Electronic-Database Information

http://www.hprg.com/hapest5/
Haplogroup Predictor Program

http://www.oxfordancestors.com/index.html
Oxford Ancestors Database

http://www.ysearch.org
Y-Search Y-STR Public Database

http://www.ybase.org
Y-Base Y-STR Public Database

http://www.smgf.org
Database of Sorenson Molecular Genetics Foundation

http://www.dna-fingerprint.com/modules.php?op=modload&name=ymatch
Y-Match Y-STR Public Database

References

Athey TW (2005) Haplogroup prediction using an allele-frequency approach. J Genetic Genealogy, 1:1-7.

Behar DM, Garrigan D, Kaplan ME, Mobasher Z, Rosengarten D, Karafet TM, Quintana-Murci L, 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.

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.

Henson G (2005) DYF399S1: A unique three-copy short tandem repeat on the human Y chromosome. J Genetic Genealogy, 1:8-11.

Seielstad M, Yuldasheva N, Singh N, Underhill P, Oefner P, Shen P, Wells RS (2003) A novel Y-chromosome variant puts an upper limit on the timing of first entry into the Americas. Am J Hum Genet 73:700-705.