Journal of Genetic Genealogy, 3(1):14-18, 2007

A Major Subclade of Haplogroup G2

T. Whit Athey

Abstract

Haplogroup G2 has two well defined subgroups, G2a and G2b, but both groups are extremely small. Most haplotypes within G2 are classified as G2*. The present study provides further characterization of a cluster, and perhaps new subclade, of G2 mentioned briefly by Goff (2006). This cluster has a characteristic repeat value at DYS388 of 13. The age of this cluster is shown to be slightly less than half of the age of Haplogroup G2.

Introduction

Haplogroup G occurs throughout Europe at a low frequency of about 1-10% (Banks, 2007; Barac, 2003; Capelli, 2003; Capelli, 2006, Karlsson, 2006). It occurs at its highest frequency in the Caucasus region, where frequencies of 30% in Georgia and 70% in North Ossetia have been observed (Nasidze, 2003; Nasidze 2003). The high frequencies in the Caucasus region suggest that the haplogroup had its origin there.

The major subgroup of Haplogroup G in Europe is G2, defined by P15. Within Haplogroup G2, two minor subgroups, G2a and G2b, have been described (Cinnioglu 2004; Hammer 2000). More than 90% of the members of Haplogroup G2 are not further differentiated and are considered to be G2*. Therefore, there is a need for greater resolution in Haplogroup G2, even by limited Y-STR "types," until more binary markers are discovered.

Recently, Goff (2006) showed how Y-STR marker values can assist in predicting membership in Haplogroup G and its subgroup G2. Goff also reported a possible new subclade of Haplogroup G2 that is characterized by a repeat value of 13 at DYS388. The present article provides additional characteristics of this cluster and calculates its approximate age.

Methods

The database of the Sorenson Molecular Genetics Foundation (hereinafter "SMGF") was searched to identify haplotypes within Haplogroup G2. A total of 159 haplotypes were identified and extracted from the SMGF database, 62 of which had DYS388=12 and 97 of which had DYS388=13. These haplotypes were analyzed for evidence that those with DYS388=13 represent a distinct subclade of Haplogroup G2. To accomplish this task, the Y-STR values that are diagnostic for Haplogroup G2 were used (Goff 2006), along with a few additional marker values that are common in G2, but are not unique to G2. Specifically, the following search criteria were used:

DYS426 = 11
DYS391 = 10
DYS392 = 11
DYS454 = 11
DYS455 = 11
DYS459 = 9-9
DYS446 ≥ 14
DYS452 ≤ 27

Following the extraction of the probable G2 haplotypes, the allele frequency distributions for each Y-STR marker were examined for evidence of a difference between the two G2 populations.

The variances of the allele frequency distribution for each marker were used to estimate the relative age of the cluster.

Results

Characteristics of the Cluster

Normally, the Y-STR marker DYS388 does not vary very much within a haplogroup because it has a low mutation rate. For example, in both Haplogroups R1a and R1b, less than 2% of the haplotypes posted on the public database, Y-Search, have a value other than 12. In contrast, in Haplogroup G, there is almost an even split between values of 12 and 13 at DYS388, suggesting some form of founder effect and population dynamics at work. Since the modal value at DYS388 is 12 for Haplogroups G1, G2a, G2b, and G5, it is reasonable to assume that the founder of Haplogroup G2 would likely have had a value of 12 as well.

Address for correspondence: wathey@hprg.com

Received: Dec 14, 2006; accepted: April 25, 2007.

The present fairly even distribution of values of 12 and 13 could not have occurred through a normal random mutational "walk" from the ancestral value. More likely it occurred as a result of a founder with DYS388=13 and his descendants experiencing unusual reproductive success.

In examining the differences in YCAII for the two populations defined by DYS388=12 and DYS388=13, the group with DYS388=12 contained only 9 haplotypes with YCAII=20-20, while 52 had values other than 20-20. In contrast, in the group with DYS388=13, 81 had YCAII=20-20 and only 14 had different values (a few haplotypes did not have values for YCAII listed). Such a striking difference in marker values strongly suggests that the DYS388=13 group represents a distinct subclade.

Table 1 illustrates the difference in values between the two populations at DYS446:

Table 1 DYS446 Values in Two Populations of Haplogroup G2

Repeat Value	Population with DYS388=12	Population with DYS388=13
14	2	
15	7	
16	14	4
17	23	10
18	9	30
19	5	33
20	1	12
21	1	7
22		1
Total	62	97

The haplotypes with DYS388=13 frequently had higher values at DYS446 than those with DYS388=12, although there is sufficient overlap of the two distributions that any given value can not be unequivocally assigned to one clade or the other.

Furthermore, the allele frequency distribution at DYS448 provides additional support for the proposition that the DYS388=13 group represents a subclade of G2.

In **Table 2,** the distribution for the two groups shows a bimodal distribution in the DYS388=12 group, with a weak peak at DYS448=21. In contrast, this bimodal feature is lacking in the DYS388=13 group. Instead, this group has a strong single peak at DYS448=21.

Table 2 DYS448 Values in Two Populations of Haplogroup G2

Repeat Value	Population with DYS388=12	Population with DYS388=13
19	1	0
20	16	12
21	29	80
22	5	5
23	11	0
Total	62	97

Age of the Cluster

The average squared difference (ASD) or variance of the allele values for each marker is proportional to the time since the founder lived (Jobling, 2004). It appears that the cluster defined by DYS388=13 is younger than the DYS388=12 group because the ASD of the allele values for each marker is generally less in the population with DYS388=13.

Table 3 illustrates the ratios of the variance in the two populations on each of the 29 DYS markers. Because of the random nature of mutations, the following ratios show considerable variation, but the average of the ratio over all the markers helps determine the relative ages of the two groups.

In averaging the ASD ratios, the highest and lowest values were discarded as outliers. The average over all remaining markers (2.27) for the ASD ratio implies that the DYS388=13 clade is only about 1/2.27 or 44% of the age of G2 as a whole. Even if the outliers are included in the calculation, the average is only slightly higher—2.74 (resulting in a slightly younger age).

This approach for calculating the relative age does not require the assumption of any mutation rates, in contrast to the calculation of absolute ages. In one study, the absolute age of Haplogroup G2 was calculated as 12,500 years (Cinnioglu, 2004), which if correct, would make the age of the new clade about 5500 years old. However, the calculation of absolute ages based on STR variance for time scales larger than a few thousand years remains controversial because of the unknown effect of population dynamics.

Geographic Distribution of the Cluster

Many of the haplotypes in the SMGF database include information on the country of origin, but all such indications are self-identified by the participant. **Table 4** shows the geographic distribution of those haplotypes

that reported a country of origin. **Table 5** shows similar data from a few published studies.

Table 3 Variance in Repeat Values for 29 Y-STR Markers

Marker	Variance for Population with DYS388=12	Variance for Population with DYS388=13	Ratio of Variances
DYS019	.2097	.1667	1.258
DYS385a	.8167	.4639	1.760
DYS385b	.9167	.4330	2.117
DYS389i	.2903	.1250	2.322
DYS389ii	.5161	.3077	1.677
DYS390	.4355	.1146	3.800
DYS393	.2951	.2708	1.090
DYS437	.0806	.1134	0.711
DYS438	.0484	.0526	0.919
DYS439	1.0161	.2632	3.861
DYS441	.1475	.1340	1.101
DYS442	.5000	.1538	3.250
DYS444	.7258	.4433	1.637
DYS445	.3871	.0206	18.774*
DYS446	1.5806	1.5979	0.989
DYS447	.5323	.2917	1.825
DYS448	1.1129	.1753	6.349
DYS449	2.5674	1.7113	1.500
DYS452	.5082	.2954	1.720
DYS456	.7097	.2990	2.374
DYS458	2.065	.5625	3.671
DYS460	.5484	.3299	1.662
DYS461	.3387	.1443	2.347
DYS462	.0164	.0515	0.318*
DYS463	.1667	.1333	1.250
1B07	.1452	.1031	1.408
GATA A-10	.5645	.1735	3.254
GATA C-4 (DYS635)	.6290	.4948	1.271
GATA H-4	.5968	.2680	2.227
YCAIIa	.6613	.1546	4.276
YCAIIb	.3871	.0928	4.172
Average			**2.27**

*Extreme values of the ratio, omitted in the main calculation of age.

The Haplogroup G Project is a voluntary association of people who have been tested or predicted to be in Haplogroup G. Participants self-identify their country or region of origin. For those tested and confirmed to be in G2 (P15+), or for those predicted by Family Tree DNA to be in Haplogroup G2, **Table 6** summarizes information from the Haplogroup G Project similar to that in **Tables 4 and 5** (Christy, 2007). All of those haplotypes not explicitly identified or predicted as G1, G2a, G2b, or G5[1] were included; however, the resulting set of haplotypes probably includes a few that are not G2*. The effect of this would be to slightly increase the numbers in the DYS388=12 column.

It should be emphasized that no firm conclusions can be drawn based upon a comparison of numbers in **Table 4** or **Table 6** between different regions for the two clades. There is too much uncertainty about the degree of bias toward northwest Europe among the people tested by SMGF and self-selected for the Haplogroup G Project. However, within a country or region the *relative* frequency in the two clades can provide insight regarding their distribution in various regions of the world. While the small samples limit the conclusions that can be drawn, it appears that the DYS388=13 cluster occurs less frequently in the east (Russia/Ukraine) and south (Mediterranean Region) of Europe, and more frequently in Britain, and in central Europe from Denmark south to Switzerland. Also, as illustrated in **Tables 5 and 6**, it occurs less frequently among populations in Turkey, and India.

Table 4 Country of Origin for Two Populations of Haplogroup G2 in SMGF

Country or Region of Origin	Population with DYS388=12	Population with DYS388=13
Britain/Ireland	6	17
Spain/Portugal/ Latin America	11	4
Germany/Austria/ Hungary/Switzerland/ Slovakia/Slovania	13	22
Denmark/Norway*	3	6
Italy	5	5
Balkans	1	0
Russia/Ukraine	5	1
U. S. State (probably mostly Britain/Ireland)	14	37
No Information or Other	4	5
Total SMGF	62	97

* All but one of the "Denmark/Norway" group were from Denmark

[1] Family Tree DNA only recently began testing for G5, so all of the haplotypes were checked with the Haplogroup Predictor (Athey, 2005, 2006). Those that were predicted as G5 were excluded from the analysis.

,

Table 5 Country of Origin for Two Populations of Haplogroup G2 in Published Studies

Country/Region of Origin or Ethnic Group	Population with DYS388=12	Population with DYS388=13
Anatolia (Cinnioglu, 2004)	32	4
India (Sengupta, 2006)	17	0
Ashkenazi Jews (Behar 2004)	9	0
Mediterranean Basin[2] F (xI,J,K) (Capelli, 2005) – May not all be Haplogroup G2	35	4

Table 6 Country of Origin for Two Populations of Haplogroup G2 from the Haplogroup G Project

Country or Region of Origin	Population with DYS388=12	Population with DYS388=13
Britain—All	47	75
England	38	57
Scotland	6	9
Wales	3	9
Ireland	3	5
Belgium/ Netherlands/ Denmark	9	9
Germany	41	41
Austria/Hungary Czechoslovakia	14	8
Switzerland	5	16
Poland/Prussia/ Belarus/Moldovia Lithouania/ Russia/Ukraine	35	3
Spain/Galicia/ Portugal/Azores/ Latin America	20	13
France	11	5
Italy/Sicily	29	8
Greece/Romania	7	0
India	4	0
Turkey/Lebanon	5	1
Morocco	2	0

[2] Populations studied by Capelli included new samples from Sicily, Italy, Sardinia, Malta, and Cyprus. Data from previous studies were included in the analysis from Turkey, Iberia, Oman, Syria, Yemen, UAE, Iraq, and Palestine, plus samples from Kurdish, Jewish, and Beduin ethnic populations.

Thus, the DYS388=13 clade may have arisen somewhere in Central Europe and spread from there. The clade's high frequency in Britain and Ireland as compared to the whole of Europe suggests a possible founder effect.

Additionally, while the mutation rate for DYS388 is low, a small percentage of members of Haplogroup G2 may have DYS388=13 by independent mutation, but do not belong to the clade described above. Similarly, there may be a small percentage of members of the DYS388=13 clade whose repeat value on DYS388 has reverted back to 12. Because of this probable small overlap of DYS388 values, the DYS388=13 cluster should not be used to construct a formal phylogeny. However, it can be useful for informal classification, as well as suggesting the existence of further SNPs that, if discovered, would define the new G2 clade more definitively.

Acknowledgement

I wish to thank Ken Nordtvedt for his insights and helpful suggestions. However, the responsibility for the final content and conclusions contained in this article is mine alone.

Electronic-Database Information

https://home.comcast.net/~hapest5/index.html
Haplogroup Predictor Program

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

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

References

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

Athey TW (2006) Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. J Genetic Genealogy, 2:34-39.

Banks R (2007) Haplogroup G, country by country and region by region [web site]
http://www.members.cox.net/morebanks/MoreG2.html

Barac L, Pericic M, Klaric IM, Rootsi S, Janicijevic B, Kivisild T, Parik J, Rudan L, Vellems R, Rudan P (2003) Y chromosomal heritage of Croatian population and its island isolates. Eur J Hum Genet, 11:535-542.

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.

Capelli C, Redhead N, Romano, Cal`ı VF, Lefranc G, Delague V, Megarbane A, Felice AE, Pascali VL, Neophytou PI, Poulli Z, Novelletto A, Malaspina P, Terrenato L, Berebbi A, Fellous M, Thomas MG, Goldstein DB (2006) Population Structure in the Mediterranean Basin: A Y Chromosome Perspective. Ann Hum Genet, 70:207–225.

Christy P (2007) Family Tree DNA Haplogroup G project [web site]: http://www.familytreedna.com/public/G%2DYDNA/

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

Goff PG, Athey TW (2006) Diagnostic Y-STR markers in Haplogroup G. Journal of Genetic Genealogy, 2:12-17.

Jobling MA, Hurles ME, Tyler-Smith C. Human Evolutionary Biology—Origins, Peoples, and Disease (2004). Garland Science, New York, p. 180.

Karlsson AO, Wallerström T, Götherström A, Holmlund G (2006) Y-chromosome diversity in Sweden – A long-time perspective. Eur J Hum Genet, 14:963-970.

Nasidze I, Quinque D, Dupanloup I, Rychkov S, Naumova O, Zhukova O, Stoneking M (2004) Genetic Evidence Concerning the Origins of South and North Ossetians. Ann Hum Genet, 68:588–599.

Nasidze I, Ling EYS, Quinque D, Dupanloup I, Cordaux R, Rychkov S, Naumova O, Zhukova O, Sarraf-Zadegan N, Naderi GA, Asgary S, Sardas S, Farhud DD, Sarkisian T, Asadov C, Kerimov A, Stoneking M (2004) Mitochondrial DNA and Y-Chromosome Variation in the Caucasus. Ann Hum Genet, 68:205-221.

,