

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 ana-

lyzed 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

