'SATIABLE CURIOSITY

The Case of the Ubiquitous 16519C

'Satiable Curiosity is a column dedicated to the proposition that genetic genealogists are an untapped resource for resolving questions about DNA behavior -- how DNA changes over the course of a few or many generations and how DNA patterns are distributed around the world. Some questions are so broad that it could take decades to arrive at a conclusion, yet others are narrow enough to answer in a shorter time frame, perhaps even within a semester or two for a student research project. The results may nonetheless be of considerable genealogical utility and scientific interest, worthy of publication in a technical journal.

The first article in the series described the problem of phylogenetically equivalent SNP markers on the Y-chromosome tree, which could be tackled by simultaneously testing each marker on a population selected from genealogical databases.[1] Prescreening samples for diversity of STR results would optimize the chance of finding a distinction. Some individuals have now commissioned tests from commercial companies that offer different markers for the same position on the phylogenetic tree, but there are still many spots awaiting clarification.

We now turn our attention to mitochondrial DNA and the exceedingly common "mutation" reported at 16519C. It is technically more correct to refer to polymorphisms or differences from the Cambridge Reference Sequence (CRS), since the CRS sample itself may have a mutation compared to the ancestral type. However, an individual's test results comparing him to the CRS may refer to mutations, and the word will be used here in that sense.

Mitochondrial DNA (mtDNA) is a circular molecule containing 16569 bases. The numbering system starts at an arbitrary point in the middle of the D-loop (displacement loop, a section where the two DNA strands in the double helix spread apart during replication of the molecule). This region is sometimes called the control region (due to its role in promoting replication), in distinction from the coding region. Because the control region is not responsible for producing proteins, mutations can accumulate without obvious adverse effects. The control region covers base positions 16024 through 16569 and continues around the circle to include bases 1 through 575.

Early research on the control region identified three sections with a particularly high percentage of polymorphic loci: Hypervariable Region 1 (26%, with 88 polymorphic positions in the section covering bases 16024-16365), Hypervariable Region 2 (24%, with 65 polymorphic positions in the section 73-340), and a somewhat less concentrated Hypervariable Region 3 (18%, with 25 polymorphic sites in the section 437-574). Only about 6% of the remaining bases in the control region had been observed to vary (Lutz 1997). By good fortune, the classical definitions for HVR1 and HVR2 covered a range convenient for sequencing studies, which can handle only a few hundred bases at a time. Much of the early research was confined to the most productive portion, HVR1. (Today, HVR1 and HVR2 are often used somewhat more loosely to include mutations in the region 16001-16569 and 1 to 575 respectively.)

As sequencing technology improved and the range expanded to include more bases, one locus in particular, 16519C, made numerous appearances on the GENEALOGY-DNA mailing list, as subscribers reported the results they had obtained from Family Tree DNA or Relative Genetics, commercial companies that included more than the classical HVR1 region. One person would announce that he was in Haplogroup K and had a mutation at 16159C. Another person would chime in, saying that she was Haplogroup T, but she had the same mutation. Reports from people in Haplogroup H were almost comically reminiscent of

[1] http://jogg.info/turner.pdf

fractious children arguing "Do so! Do not! Do so! Do not!" as one person would announce his mutation at 16519C and another would declare that she didn't have it.

This same observation had already confounded researchers who attempted to construct network diagrams or phylogenetic trees, designed to show the relationship between haplogroups as they diverged from mitochondrial Eve. These graphical techniques cluster similar HVR results together, yet 16519C would pop up here and there in branches that were only distantly related. Connecting all the points that matched on 16519C would result in a maze of criss-crossing lines (a reticulation). Many authors (Finnila 2001, Kivisild 2002, Macauley 1999) have explicitly ignored this locus, sometimes sounding mildly vexed when they could construct a network diagram "free of reticulations" only when 16519C was excluded (Maca-Meyer 2001). Later studies confirmed the ubiquitous appearance of 16519C in additional populations, intruding on multiple branches of phylogenetic trees for African and Asian samples (Allard 2004, Allard 2005).

This erratic behavior has some compensatory advantages in forensic applications, however. Mitochondrial DNA can sometimes be used to identify remains of individuals recovered from old military operations or mass graves. Although many mtDNA haplotypes are quite rare, a few are moderately common. Coble (2004a) investigated the eighteen most common haplotypes in the classical HVR1/HVR2 range, collectively accounting for 21% of a set of European samples. His approach focused on coding region differences to discriminate similar samples, but he noted that both 16519C and 16519T could be found in nine of the eighteen common haplotypes.

Where does that leave the genetic genealogist? We are stranded somewhere between the phylogenetic applications spanning hundreds of generations and forensic applications seeking to distinguish individuals. Should we regard this locus as significant in matching distant cousins, or should we disregard it? A direct study of the mutation rate in close and distant relatives could shed some light on the question, and genetic genealogists are ideally positioned to assemble willing volunteers for testing.

With the ready availability of thousands of samples covering 16519, perhaps it will come back into fashion as an intriguing phylogenetic marker. Before 16519 was abandoned by many authors, Helgason included it in his network diagram of 410 individuals from Iceland (Helgason 2000). Iceland has a limited but diverse subset of European haplotypes. He observed

> *A number of researchers have identified the 16519site as being too hypervariable to include in phylogeneticanalyses... Although our data suggest that site 16519 has undergone multiple mutations, it nonetheless seems to be largely fixed for most haplogroups. Thus, with only a few exceptions, all sequences belonging to haplogroups K, U4, U3, I, X, and T in the Icelanders have 16519C, whereas haplogroups V, J, and U5 have 16519T. The only haplogroup with which site 16519 seems to be at variance is H, where roughly half of the lineages have 16519C.*

In the phylogenetic tree, haplogroups J and T have a common root, as do K and U, and H and V (see Figure 1). If some branches exhibit 16519C but others do not, it is clear that separate mutations have occurred after the sister haplogroups JT, UK, and HV have split off from the main trunk. However, these mutations must have occurred long enough ago to encompass most people inone of the smaller branches. Thus 16519C is not merely a difference from the CRS, which crops up because the CRS has a rare value, but evidence of an actual mutation that has occurred multiple times.

In accordance with Helgason's observations, Bill Hurst, who has collected data on haplogroup K haplotypes, noted that 100% of a well-screened set of 104 samples were 16519C.[2] Other haplogroups can be investigated at Mitosearch, a public access database sponsored by Family Tree DNA.[3] How do Helgason's dichotomies, developed from an isolated population, hold up when the Mitosearch database is consulted? Table 1 shows a sampling of these haplogroups.

Table 1
16519C in Selected Haplogroups

	# 16519C	# samples	% 16519C
T1	78	80	98
U3	12	17	71
U4	79	86	92
I	112	120	93
V	32	122	26
J	13	51	25
U5	4	50	08

[2] http://archiver.rootsweb.com/th/read/GENEALOGY-DNA/2005-09/1126979802
[3] http://www.mitosearch.org, accessed 11/30/05

The distinctions between haplogroups with a high or low percentage of 16519C are still prominent. Studying the exceptions should prove most interesting. Are mutations a one-way street? If so, 16519C samples from haplogroups with a low percentage might be recent mutations, prime hunting ground for placing the mutation within a genealogical time frame. Distant cousins could be recruited to enhance the number of transmission events. Conversely, 16519T samples from haplogroups with a high percentage of 16519C could help date a split within the haplogroup by assessing the amount of HVR variability within each type. Or perhaps the answer is less absolute than a one-way street – more like an expressway with extra lanes going in the direction of rush hour traffic. In fact, Coble did find one example in his study of a back mutation from 16519C to 16519T (Coble 2004b).

Family Tree DNA offers subclade testing for haplogroup H, with a diagram showing 16519C as a defining event joining several subclades, H1, H2, H3, H7, H9, H10, H12, and H13[4]. The diagram is based on 61 individuals, primarily from Italy and Spain (Achilli 2004). Considering the ubiquity of 16519C and the small sample size, it may be premature to root these subclades together, especially given the fact that the samples closest to CRS (in H2) would be expected to have 16519T. Table 2 shows results that have been uploaded to Mitosearch, with none of the H2 samples showing 16519C but most other samples conforming to expectations. Amusingly, the overall results are still close to Helgason's "roughly half" and perpetuate the "Do so! Do not!" refrain.

The availability of many mtDNA sequences obtained by genealogists with extensive family records should prove valuable in studying some aspects of this curious locus, even if the ultimate question can not be answered: "What is it about 16519C that makes it such a target for mutations?"

Ann Turner
DNACousins@aol.com

[4] http://www.familytreedna.com/hclade.html

Table 2
Haplogroup H Subclades

	# 16519C	# samples	% 16519C
H*	30	43	70
H1	19	21	90
H2	0	0	
H2a	0	0	
H2b	0	6	0
H3	17	17	100
H4	2	5	40
H4a	1	5	20
H5	0	8	0
H5a	1	3	33
H5a1	1	12	8
H6	0	1	0
H6a	0	7	0
H6a1	0	5	0
H6b	0	5	0
H7	15	15	100
H8	0	1	0
H9	0	0	
H10	4	5	80
H11	2	4	50
H11a	0	4	0
H13	1	3	33
Sum	93	170	55

References

Achilli A, Rengo C, Magri C, Battaglia V, Olivieri A, Scozzari R, Cruciani F, Zeviani M, Briem E, Carelli V, Moral P, Dugoujon JM, Roostalu U, Loogvali EL, Kivisild T, Bandelt HJ, Richards M, Villems R, Santachiara-Benerecetti AS, Semino O, Torroni A (2004) The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool. Am J Hum Genet 75:910-918.

Allard MW, Wilson MR, Monson KL, Budowle B (2004) Control region sequences for East Asian individuals in the Scientific Working Group on DNA Analysis Methods forensic mtDNA data set. Leg Med (Tokyo) 6:11-24.

Allard MW, Polanskey D, Miller K, Wilson MR, Monson KL, Budowle B (2005) Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. Forensic Sci Int 148:169-179.

Coble MD, Just RS, O'Callaghan JE, Letmanyi IH, Peterson CT, Irwin JA, Parsons TJ (2004a) Single nucleotide polymorphisms over the entire mtDNA genome that increase the power of forensic testing in Caucasians. Int J Legal Med 118:137-146.

Coble, MDW (2004b) The Identification of Single Nucleotide Polymorphisms in the Entire Mitochondrial Genome to Increase the Forensic Discrimination of Common HV1/HV2 Types in the Caucasian Population. Dissertation submitted to George Washington University, www.cstl.nist.gov/div831/strbase/pub_pres/Coble2004dis.pdf

Finnila S, Lehtonen MS, Majamaa K (2001) Phylogenetic network for European mtDNA. Am J Hum Genet 68:1475-1484.

Helgason A,Sigurðardóttir S,Gulcher JR, Ward R,Stefánsson K (2000) mtDNA and the origin of the Icelanders: deciphering signals of recent population history. Am J Hum Genet 66:999-1016.

Kivisild T, Tolk HV, Parik J, Wang Y, Papiha SS, Bandelt HJ, Villems R (2002) The emerging limbs and twigs of the East Asian mtDNA tree. Mol Biol Evol 19:1737-1751.

Lutz S, Weisser H-J, Heizmann J, Pollak S (1997) A third Hypervariable region in the human mitochondrial D-loop. Hum Genet 101:384.

Maca-Meyer N, Gonzalez A, Larruga JM, Flores C, Cabrera VM (2001) Major genomic mitochondrial lineages delineate early human expansions. BMC Genetics 2:13.

Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, Bonne-Tamir B, Sykes B, Torroni A (1999) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. Am J Hum Genet 64:232-49.

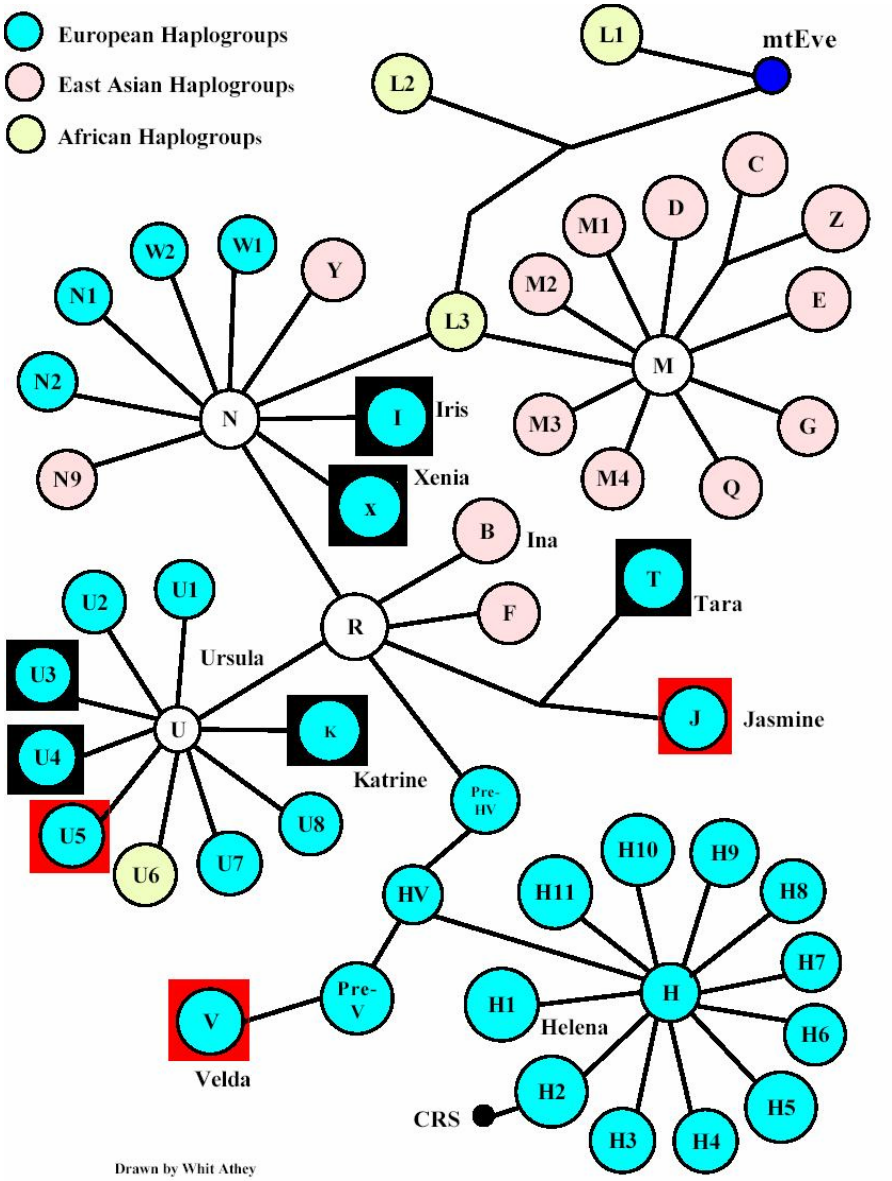

Figure 1: Phylogenetic diagram of mtDNA haplogroups, with observations by Helgason added as colored background squares. Haplogroups with black backgrounds are predominantly 16519C, while haplogroups with red backgrounds are predominantly 16519T. Haplogroup H is evenly split between 16519C and 16519T. Other haplogroups without squares were not observed in Helgason's study.