**Phylotree / Haplogrep Data Validation**

Phylotree version 15 is a phylogenetic tree of mitochondrial DNA and the standard for calling haplotypes. The XML file available for this tree gives both the mutations present for each haplotype and for many, an example genbank accession file with a complete mtDNA sequence.

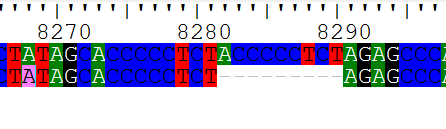
I am currently calling haplotypes based on whole genome and exome sequencing data, and wanted to be sure that the phylotree data files were being interpreted correctly and that mtDNA haplotype calls made with “new” data would match how they should be called in a “phylotree” style. To do this, I downloaded all the accession numbers available from phylotree, realigned them to the rCRS sequence, and called variants in the phylotree notation (e.g. 151C). I then checked how often all of the calls made by my method matched the mutations present in the phylotree XML file for a given haplotype.

The results showed that phylotree data was often unreliable, and it was very easy to get discordance.

A total of 2,736 haplotype/genbank file combinations were tested for concordance (after ignoring the 533 genbank sequences with ambiguous bases). Across all of these haplotype/genbank file combinations, 72,015 mutations were found in both the genbank file and the respective haplotype, but 3,197 mutations were called either in just the genbank file or were expected from the haplotype file but did not exist in the genbank file (this is after masking certain locations known to be problematic).

There were three main reasons for the discord variant calls. First it appears that the genbank files often had many more mutations than the haplotype they were describing. 1,926 file/haplotype combinations had mutations present in the file but not in the haplotype, while only 545 had mutations suggested by the haplotype that were not found in the genbank file.

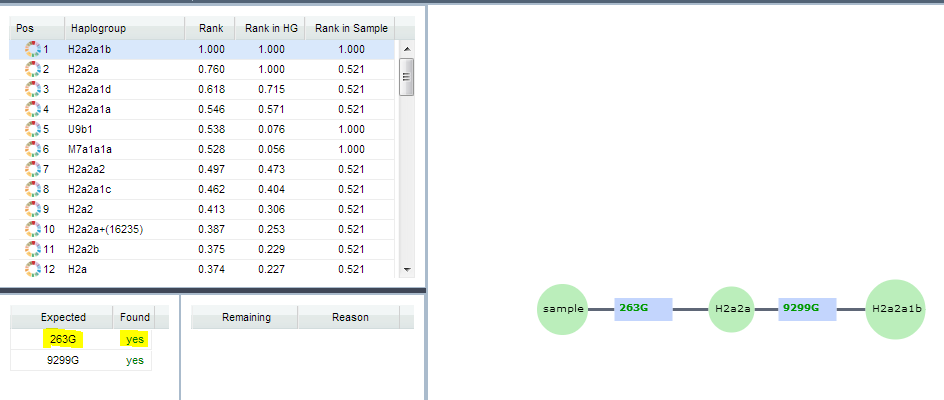
Many of the discordant variant calls were due to alignment differences. While most variant calling software packages typically “left-align” the variants, it appears HaploGrep/Phylotree typically “right-aligns”. For example, the alignment below shows an mtDNA molecule aligned relative to the rCRS, where a tandem repeat of the “ACCCCCTCT” sequence has been deleted. Phylotree calls this as a deletion of the first repeat, while most software would call it a deletion of the second.



A simple way to avoid this is to never treat indels as informative for assigning haplotypes.

Perhaps most concerning, there were often times when Haplogrep expected a polymorphism in a haplotype, but one was not found in the file. For example, Haplogrep thinks that the haplotypes H2a2a1[a, b and d] should all have a 263G mutation. However, examining the example file for haplotype H2a2a1a ([EU795361](http://www.ncbi.nlm.nih.gov/nuccore/EU795361)), it is clear that there is no mutation relative to the reference at this location, and that the base is an “A” (as in CRS) rather than a “G”. Why this discrepancy? Looking at the original Phylotree website, it appears the A mutation is expected at this position for all of these haplotypes:



However, Haplogrep thinks that the H2a2a1b genotype is expected to have this mutation (screen shot from web tool below), which is incorrect.  


This is likely due to an error made when transposing phylotree to be relative to the rCRS sequence. But these mismatches will require further investigation.