Int J Legal Med (2011) 125:901–903
DOI 10.1007/s00414-010-0459-z

LETTER TO THE EDITOR

Comments on the article, "Software for Y Haplogroup Predictions, a Word of Caution"

Whit Athey

Received: 6 February 2010 / Accepted: 13 April 2010 / Published online: 27 April 2010

Dear Sirs:
The article by M. Muzzio et al., entitled "Software for Y Haplogroup Predictions, a Word of Caution," purports to evaluate two software programs for predicting Y haplogroups from a set of Y-STR marker values, and the authors conclude that "haplogroup prediction software available does not show adequate accuracy." The authors' title urges "caution" in using these programs. In the interests of disclosure, I am the author of one of those programs, the Haplogroup Predictor.

In fact, the study that was carried out is really more of an evaluation of the minimally informative dataset that was used. As the authors themselves point out, at just seven STR markers, the possibility exists that the same haplotype could occur in more than one haplogroup. In this case no algorithm can provide the answer as to which haplogroup is the "real" one. One would have to readily agree that if using only seven markers, "caution" should indeed be exercised, or better yet, the whole effort should probably be avoided.

If the authors had employed a dataset with only five markers instead of seven, they would have found that the programs performed even less well. If they had used a dataset with 17, 19, or 25 markers, they would have found that the programs perform quite well. With the addition of a sufficient number of markers, the prediction probability for the correct haplogroup can be "driven" past 99% in nearly all cases, and this almost always occurs by the point where 20 markers have been used. The evaluation carried out by the authors does not really address the quality of the software at all—it only addresses the inadequacy of their own dataset.

Very few studies today measure so few Y-STR markers as the one that produced the dataset used in the authors' evaluation; indeed, the seven-marker dataset apparently resulted from a study carried out over 5 years ago and published in 2005. Studies with so few markers were common 10 years ago, but not anymore. Today even forensic and population studies typically employ 16 [1], 17 [2], 19 [3], or more [4] markers. In the genetic genealogy field, where at least one of the evaluated programs has its main application, 37-, 43-, and 67-marker haplotypes are common. My own program can accept up to 76 markers (or in most cases, up to 86).

The authors include the following very curious and incorrect statement in their Discussion section: "An increase in the number of STRs employed to predict the haplogroup would not enhance accuracy, considering the few reference samples available with the standard seven STRs...." Leaving aside the statement that seven STRs are "standard," it is obvious that the authors have not tried very hard to discover the needed reference samples. I have found sufficient data to include 76 markers in 23 haplogroups in my program, though it is true that one or two of these haplogroups barely had sufficient samples for inclusion, but most of the 23 had plenty of reference samples, even out to 76 markers. Compiling the necessary data to support one of these programs is, indeed, the major challenge of implementing it, but it can and has been done. The quoted statement makes one wonder if the authors actually looked at the Haplogroup Predictor program web site, with its data entry section covering 86 markers. How did they think so many markers could be offered for use if no data were available?

The authors' dataset also had very minimal SNP information, only enough to define four broad groupings

W. Athey (✉)
Brookeville, MD, USA
e-mail: wathey@hprg.com

of haplogroups, plus one specific haplogroup, which they name as follows:

> F*—presumably including G, H, I, and J (H would be rare in the population of Argentina), or more accurately, F (xK)
>
> K*—presumably meaning L, M, N, O, S, and T (M and O would be unlikely in this population), or more precisely, K (xQ1a1a, R)
>
> Q1a1a—this is the only specifically defined haplogroup and the name appears to reflect the use of very recent haplogroup nomenclature (probably referring to Q-M3), rather than the YCC-2002 nomenclature as stated, and if Q-M3 was intended, this haplogroup probably occurs only in the Native American population in Argentina.
>
> R*—presumably meaning R1a, R1b, R1 (xR1a, R1b), R2, and R (xR1, R2) (real R*, as opposed to the authors' construct, would be rare in this population)
>
> DE*—presumably meaning DE and including all of D and E (D is probably rare in this population)

There are other more mundane problems with the authors' dataset. Several haplotypes are almost certainly not in the "haplogroup" that is indicated, which can be seen even with only seven markers in a few cases. SNP lab errors occur at a frequency on the order of 1%, and "clerical errors" are even more common. Four problematic haplotypes are listed below as they are shown in the supplementary data file by Muzzio et al.; specifically, they are listed in the marker order DYS019, 389i, 389ii, 390, 391, 392, and 393. In the following, "Ysearch" refers to the public Y-STR database, currently containing about 81,000 Y-STR haplotypes, located at http://www.ysearch.org. To perform a search in Ysearch, it is necessary to use at least eight markers, so a value of DYS426 of 11 or 12 was added to each seven-marker haplotype, since greater than 99% of all samples would have one of those two values (about 99% of samples in F*, K*, and DE would have DYS426=11, while about 99% of R* and Q would have DYS426=12). Here are some examples of problematic haplotype assignments to haplogroups:

> Haplotype 25 (labeled by the authors as "K*") 14-13-31-24-11-13-13

Results from Ysearch using haplotype 25 plus DYS426=11: no matches. Haplogroups L, N, O, and T, presumably the constituents of the authors' "K*", all have 99% of DYS426 values equal to 11.

Results from Ysearch using haplotype 25 plus DYS426=12: there are 127 exact matches in Ysearch, almost all of which could be identified as R1b, either as a result of SNP tests, or from the extended Y-STR haplotype characteristics. The value of DYS426=12 alone would distinguish "K*" from R* and Q in 99% of cases. Since haplotype 25 plus DYS426=11 produces no matches, whereas haplotype 25 plus DYS426=12 produces numerous matches in Haplogroup R1b, the assignment of the haplotype by the authors to "K*" appears to be incorrect.

> Haplotype 31 (labeled by the authors as "K*") 14-13-30-24-11-13-12

Results from Ysearch using haplotype 31 plus DYS426=11: one match to a 37-marker haplotype, which, other than the unusual DYS426=11 value, has values quite typical of R1b.

Results from Ysearch using haplotype 31 plus DYS426=12: 75 exact matches in Ysearch, all of which could be readily identified as R1b. Therefore, the assignment by the authors to "K*" appears to be incorrect.

> Haplotype 74 (labeled by the authors as "R*") 13-13-30-24-10-11-13

This haplotype has the exact modal values for Haplogroup E1b1b, but the modal R1a seven-marker haplotype, 15-13-30-25-10-11-13, is different on just two markers. The allele frequency in R1a for DYS390=24 is 18%, so that is possible, but the frequency in R1a for DYS019=13 is only 0.3%, which is rather unlikely. The probability that this haplotype could be in R1a is very low, though it is not impossible, while it would fit perfectly into E1b1b. It would fit even less well in R1b and R2 than R1a.

> Haplotype 77 (labeled by the authors as "R*") 14-13-28-24-11-11-16

The value DYS393=16 does not occur in my collection of 2,000 R1a haplotypes or 1,600 R1b haplotypes (a value of 15 occurs only three times in each in these R1a and R1b datasets), or in 83 R2 haplotypes. The value is unusual in other haplogroups too, but does occur at just under 1% in Haplogroups C3, E1b1a, G2a, I2b1, and N. It is much more likely to be in one of these five haplogroups than R1a, R1b, or R2.

There are other haplotypes that appear to have improbable assignments, but with only seven markers, only the four above are sufficiently improbable to be singled out here. It is likely that several of the discrepant results found in the study by Muzzio et al. result from incorrect haplogroup assignments by the authors, rather than problems in the haplogroup programs. Before one embarks on a validation study, one must be sure of the validity of the data that will be used.

In summary, the study by Muzzio et al. has only demonstrated that their Y-STR database is unsuited for the validation of any haplogroup prediction program. If one uses a set of seven-marker haplotypes with a haplogroup prediction program, then the authors' warning that one should "use caution" in accepting the results, is entirely appropriate. Their assessment does not, however, address the capabilities of the two programs when they are used with an adequate dataset.

Sincerely yours,
Whit Athey

Conflicts of interest The author of this Letter to the Editor is also the author of the Haplogroup Predictor software. There are no financial conflicts of interest since the program is freely available.

References

1. Nonaka I, Minaguchi K, Takezaki N (2007) Y-chromosomal binary haplogroups in the Japanese population and their relationship to 16 Y-STR polymorphisms. Ann Hum Genet 71:480–495. doi:10.1111/j.1469-1809.2006.00343.x
2. McEvoy B, Simms K, Bradley DG (2008) Genetic investigation of the patrilineal kinship structure of early medieval Ireland. Am J Phys Anthropol 136:415–422. doi:10.1002/ajpa.20823
3. Adams SM, Bosch P, Balaresque PL, Ballereau SJ, Lee AC, Arroyo E, López-Parra AM, Aler M, Gisbert-Grifo MS, Brion M, Carracedo A, Lavinha J, Martínez-Jarreta B, Quintana-Murci L, Picornell A, Ramon M, Skorecki K, Behar DM, Calafell F, Jobling MA (2008) The genetic legacy of religious diversity and intolerance: paternal lineages of Christians, Jews, and Muslims in the Iberian Peninsula. Am J Hum Genet 83:725–736. doi:10.1016/j.ajhg.2008.11.007
4. Balaresque P, Bowden GR, Parkin EJ, Omran GA, Heyer E, Quintana-Murci L, Roewer L, Stoneking M, Nasidze I, Carvalho-Silva DR, Tyler-Smith C, de Knijff P, Jobling MA (2008) Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis. Hum Mut 29:1171–1180. doi:10.1002/humu.20757