LETTER TO THE EDITOR

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

Marina Muzzio · Virginia Ramallo ·
Josefina M. B. Motti · María R. Santos ·
Jorge S. López Camelo · Graciela Bailliet

Dear Sirs:

The letter by Dr. Athey entitled "Comments on the article, "Software for Y-Haplogroup Predictions, a Word of Caution" discusses a paper of our authoring in the present Journal [1]. Though at first glance his opinion may seem reasonable, a careful analysis reveals a series of mistakes committed by Dr. Athey.

In the mentioned letter, one of the main concerns of Dr. Athey is the amount of short tandem repeats (STRs) that were analyzed. He comments that seven markers should be avoided, explaining that an increase of that amount augments the probability of assignment, stating that "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." Nonetheless, he does not make reference to any sort of validation study, so said 99% must only refer to the probability of assignment his own software provides. What is more, Dr. Athey seems oblivious that our paper examined this software's

M. Muzzio · V. Ramallo · J. M. B. Motti · M. R. Santos ·
J. S. López Camelo · G. Bailliet
Laboratorio de Genética Molecular Poblacional, Instituto
Multidisciplinario de Biología Celular (IMBICE), CICPBA, CCT,
La Plata–CONICET,
La Plata, Argentina

M. Muzzio (✉)
Instituto Multidisciplinario de Biología Celular (IMBICE),
526 e/10 y 11, P.O. Box C.C. 403, 1900 La Plata, Argentina
e-mail: marinamuzzio@yahoo.com.ar

predictive values in different cutoff points, up to 95% of probability of assignment, still obtaining inadequate predictive values on said cutoff point. Twelve of our haplotypes proposed from 100% to 99.6% probability of assignment to the R1b haplogroup in the *Haplogroup Predictor*, while none of those samples belong to the said haplogroup. In the R haplogroup, one haplotype gives 99.8% probability to an erred haplogroup (E1b1b), while 14 give from 100% to 99.4%, assigning these samples to the R1b haplogroup, and were considered as correct predictions. Note that in these cases, the predicted haplogroup follows the nomenclature of the software, and the probability was not pooled by major branches. If the seven markers of the minimal haplotype were not a proper set, Dr. Athey's software should not provide such high probabilities of assignment in these cases, since it misleads the user. Unfortunately, Dr. Athey's mention that seven Y-STRs are too few occurs only in his letter; neither of the two papers of his authoring [2, 3] available at the *Haplogroup Predictor's* website (http://www.hprg.com/hapest5/) nor the software instructions (http://www.hprg.com/hapest5/page4.html) explain that the user should employ more than seven markers. Otherwise, we would not have attempted to use it. Moreover, a simple glance at recent literature where the *Haplogroup Predictor* was employed clearly shows that researchers do not use such high amounts of Y-STRs when they rely on the *Haplogroup Predictor*; for instance, Salas et al. [4] used the seven Y-STRs we did, plus DYS 385 and Petrejcikova et al. [5] only 12 Y-STRs.

His statement "…indeed, the seven-marker dataset apparently resulted from a study carried out over five

years ago and published in 2005" shows a misunder-standing of our paper, since in it we clearly expressed that "Haplogroups were determined in a previous report" with the corresponding citation of a paper that includes only haplogroup information, without Y-STR data.

Advancing to his enumeration of papers where high amounts of Y-STRs were analyzed, Dr. Athey does not realize that, for a validation study, we needed to employ an independent dataset, one that was not used by him to calibrate his software. This is quite difficult when, in the *Haplogroup Predictor*'s webpage, the only information about the samples used for calibration remains the 2005 paper; although it mentions several updates, it is not explicit which haplotypes were included, from what other sources besides *Y search*.

When Dr. Athey comments that the Q1a3a, nomenclature from Karafet et al. [6] (or QM3 in the nomenclature anterior to the cited paper), "probably occurs only in the Native American population in Argentina," he contradicts the literature regarding the said haplogroup, which is the most frequent in all Native Americans and also present in admixed populations (a few classic papers are [7–9]).

Concerning his discussion about four haplotypes that he does not find trustworthy, four wrongly assigned haplotypes do not change significantly our likelihood ratio results nor the uncertainty coefficient calculations, given that the whole sample was 119 haplotypes. Secondly, his search on *Y search* (http://www.ysearch.org.) is not enough to invalidate our single nucleotide polymorphism (SNP) results, especially when one considers that this database was employed by Dr. Athey to calibrate his software [2], so any attempt to employ it to question our work is circular reasoning. Furthermore, one of the haplotypes (H77) that he questions was assigned by his software to the same haplogroup we defined by SNP.

It is a standard procedure in statistics to introduce new samples to verify a given set of classifying tests. Further, validation analyses, employing samples and populations that were not employed to calibrate the *Haplogroup*

Predictor, will allow even more precise estimates of its predictive value.

Yours sincerely,

Marina Muzzio
Virginia Ramallo
Josefina M.B. Motti
M. Rita Santos
Jorge S. López Camelo
Graciela Bailliet

Conflicts of interest   The authors of this letter are also the authors of "Software for Y-Haplogroup Predictions: a Word of Caution," published on this journal.

References

1. Muzzio M, Ramallo V, Motti JMB, Santos MR, López-Camelo JS, Bailliet G (2010) Software for Y-haplogroup predictions: a word of caution. Int J Leg Med. doi:10.1007/s00414-009-0404-1
2. Athey WT (2005) Haplogroup prediction from Y-STR values using an allele- frequency approach. J Genet Geneal 1:1–7
3. Athey WT (2006) Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. J Genet Geneal 2:34–39
4. Salas A, Jaime JC, Álvarez-Iglesias V, Carracedo Á (2008) Gender bias in the multiethnic genetic composition of Central Argentina. J Hum Genet 53:662–674
5. Petrejcikova E, Sivakova D, Sotak M et al (2009) Y-STR haplotypes and predicted haplogroups in the Slovak Haban population. J Genet Geneal 5(2):69–74
6. Karafet TM, Mendez FL, Meillerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polymorphisms reshape and increase the resolution of the human Y chromosomal haplogroup tree. Genom Res 18:830–838
7. Underhill P, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996) A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. Proc Natl Acad Sci USA 93:196–200
8. Seielstad M, Yuldasheva N, Singh N, Underhill P, Oefner P, Shen P, Spencer Wells R (2003) A novel Y-chromosome variant puts an upper limit on the timing of first entry into the Americas. Am J Hum Genet 73:700–705
9. Zegura SL, Karafet TM, Zhivotovsky LA, Hammer MF (2004) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y-chromosomes into the Americas. Mol Biol Evol 21:164–175