Int J Legal Med (2011) 125:143–147
DOI 10.1007/s00414-009-0404-1

SHORT COMMUNICATION

Software for Y-haplogroup predictions: a word of caution

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

Received: 13 May 2009 / Accepted: 10 December 2009 / Published online: 15 January 2010

Abstract The development of online software designed for genetic studies has been exponentially growing, providing numerous benefits to the scientific community. However, they should be used with care, since some require adjustments. The efficiency of two programs for haplogroup prediction was tested with 119 samples of known haplotypes and haplogroups from Argentine populations. Quantitative estimates of the predictive quality of both software systems were computed with the uncertainty coefficient; and sensitivity, specificity, positive, and negative likelihood ratios were also calculated to assert the reliability of both programs, showing high probabilities of assigning an incorrect haplogroup.

Keywords Molecular anthropology · Population genetics · Human Y-chromosome · SNP/STR · Bioinformatic approaches

Introduction

Y-chromosomal lineages are established by single nucleotide polymorphism (SNP) and short tandem repeats (STRs), which provide the corresponding haplogroup and haplotype, respectively. In the study of human population genetics, haplogroup determination is of great interest, as it reveals the phylogenetic relationships by descent.

Considering the findings of Bosch et al. [1] and Behar et al. [2], where the STR variability is partitioned by haplogroups to a greater extent than by populations, there has been increasing interest in unifying these sources and finding further ways of predicting the haplogroup of a given haplotype when SNP data are unavailable. One of them is Whit Atheys' haplogroup predictor [3, 4] (https://home.comcast.net/~hapest5/index.html) which has been employed in previous studies [5–7] to estimate the ethnic composition of different populations and diagnostic STR values of a given haplogroup. The other is the haplogroup classifier [8] (http://bcf.arl.arizona.edu/haplo), consisting of machine-learning algorithms that require previous models and haplotypes with a known haplogroup for training the software.

The purpose of this study was to establish the accuracy of both software systems: the haplogroup predictor and the haplogroup classifier.

Materials and methods

We analyzed a sample of 119 males from four provinces of the Northwest of Argentina (Jujuy, Salta, Catamarca, and Tucumán), all of them with the informed consent of donors.

Haplogroups were determined in a previous report [9]; the nomenclature used followed YCC [10] recommendations, and haplotypes were defined by the amplification of DYS19, DYS389 I and II, DYS390, DYS391, DYS392, and DYS393, according to methods previously published [11–15].

Electronic supplementary material The online version of this article (doi:10.1007/s00414-009-0404-1) contains supplementary material, which is available to authorized users.

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, 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

These haplotypes were submitted to the haplogroup predictor, with equal priors, obtaining probabilities for inferred haplogroups. In the case of the haplogroup classifier, we followed the models, tree files, and public data provided by the authors in the downloadable version of the software; this data set consisted of 1,527 Y-chromosome profiles with haplogroup and haplotype gathered from published data [16, 17]. SNP-determined haplogroups were compared with those provided by the software. It was not possible to evaluate the sample analyzed by Schlecht et al. [8] given the availability of their data.

Sensitivity (s), specificity (e), positive likelihood ratio (LR+=$s/(1-e)$, and negative likelihood ratio (LR−=$(1-s)/e$) of the haplogroup predictor, and the haplogroup classifier were calculated per haplogroup and total [18]: "s" represents the probability of corroboration for a predicted haplogroup when the haplotype was determined of that haplogroup by SNP analysis, while "e" stands for the probability of confirmation of another haplogroup when the haplotype corresponded to a different haplogroup by our typing. It has been stated that a test is adequate if it has both high sensitivity and specificity [19] and a LR+ value of at least 10; this is the reason why we followed these criteria.

In the case of the predictor, we considered different precision-of-assignment categories: by being the first in the haplogroup ranking and 50–95% cut-off points by intervals of five, whereas for the classifier, we only counted those cases showing an agreed haplogroup, thus reducing the number of cases considered in the calculations (N=62; 52.1% of the original sample).

In order to get quantitative estimates of the software predictive quality, we computed the uncertainty coefficient of y, $U(y|x)$, with the subroutine cntab2 of Press et al. [20] who also provide a detailed explanation of the meaning and way of computing that coefficient. Let us see what is the meaning of $U(y|x)$ and how it can aid us to estimate the software predictive quality. Suppose that we have a certain sample and that we want to know the result of performing the SNP typing on it, i.e., that result is all the information we want. Let us further suppose that before performing the SNP typing, we get, for that same sample, the prediction of the software. Then, knowing beforehand the results given by the software will produce a loss of the information that we could later obtain from the SNP typing and the better the software, the larger that loss will be. If the software was perfect and could accurately predict the result that we would later obtain from the SNP typing, then knowing the result of the former would make us lose all the information that the result of the latter would provide, and of course, it would be unnecessary to do the SNP typing. On the other hand, if the software were useless and had no predictive value, to know its result beforehand would make us lose no information at all and accordingly, the whole information will have to come from the SNP typing. The uncertainty coefficient of y, $U(y|x)$, quantifies what we have just explained qualitatively. If we take the results from the software systems as the x variable and those from SNP_typing as the y variable, then $U(y|x)$ gives the fraction of the SNP typing information that is lost if the software result is already known. As the two extreme cases described before we would have: (1) $U(y|x)$=1.00 (or 100%), that would imply that the software gives perfect answers (i.e., all the information that would be provided by subsequent SNP typing has already been provided by the software) and (2) $U(y|x)$=0.00 (or 0%), that would imply that the software provides no information (i.e., all the information should be obtained by subsequent SNP typing).

Results

The s, e, LR+, and LR− values per cut-off point for the haplogroup predictor are summarized in Table 1. The classifier software showed the following values: s=0.45, e=0.92, LR+=5.99, and LR−=0.59. It is important to highlight that about half of the haplotypes (57 out of a total of 119) could not be assigned by the classifier to an agreed haplogroup ("Unclassified" category in Table 2). The profiles of the Argentinean population sample are presented as electronic supplementary material in Table 3 to allow the validation of these results.

When we consider LR+ per haplogroup, Q1a3a in both classifier and predictor and DE* in the classifier, values are higher than 10.

Table 1 Total s, e, LR+, and LR− values at each cut-off point for the haplogroup predictor

Cut-off point	s[a]	e[b]	LR+[c]	LR−[d]
RK	0.5	0.88	4.21	0.6
50	0.5	0.89	4.52	0.5
55	0.5	0.89	4.54	0.6
60	0.5	0.89	4.46	0.6
65	0.5	0.9	4.65	0.6
70	0.5	0.91	5.12	0.6
75	0.4	0.91	4.59	0.7
80	0.4	0.92	5.22	0.7
85	0.4	0.93	5.81	0.7
90	0.3	0.94	5.93	0.7
95	0.3	0.96	8.00	0.7

[a] Sensitivity

[b] Specificity

[c] Positive likelihood ratio

[d] Negative likelihood ratio

Table 2 Haplogroup frequencies by SNP typing, haplogroup predictor and haplogroup classifier, and *s*, *e*, LR+, and LR− values

			SNP typing							
			Haplogroup							
			F*	K*	Q1a3a	R*	DE*	Total software	False positives	False positive (%)
Predictor	Haplogroup	F*	9	0	1	4	5	19	10	52.6
		K*	0	0	2	0	0	2	2	100.0
		Q1a3a	0	3	17	0	0	20	3	15.0
		R*	6	12	16	30	1	65	35	53.8
		DE*	2	1	2	3	5	13	8	61.5
	Total by SNP		17	16	38	37	11	119	58	48.7
	Sensitivity		52.9	0.0	44.7	81.1	45.5	51.3		
	Specificity		90.2	98.1	96.3	57.3	92.6			
	LR+[a]		5.4	0.0	12.1	1.9	6.1			
	LR−[b]		0.5	1.0	0.6	0.3	0.6			
Classifier	Haplogroup	F*	4	0	1	4	3	12	8	66.7
		K*	0	1	3	1	0	5	4	80.0
		Q1a3a	0	1	2	0	0	3	1	33.3
		R*	2	8	9	19	0	38	19	50.0
		DE*	0	0	1	1	2	4	2	50.0
	Unclassified		11	6	22	12	6	57		
	Total by SNP		17	16	38	37	11	119	34	54.8
	Sensitivity		66.7	10.0	12.5	76.0	40.0	23.5		
	Specificity		92.9	96.3	99.0	79.8	98.2			
	LR+[a]		9.4	2.7	12.9	3.8	22.8			
	LR−[b]		0.4	0.9	0.9	0.3	0.6			

[a] Positive likelihood ratio

[b] Negative likelihood ratio

R* haplogroup showed the highest false positive proportion, significantly higher than false positive proportions of the remaining haplogroups.

Our results for $U(y|x)$ are 0.244 (or 24.4%) for the haplogroup predictor and 0.207 (or 20.7%) for the haplogroup classifier. In other words, only about 20% and 25% of the information on the SNP typing is lost if software results are already known.

Discussion

These results represent a high probability of error and a bias towards the R* haplogroup, so it is most likely that results based on the haplogroup predictions of these software systems are weakened. For cases in which sex bias in multiethnic populations is estimated by this method, an overestimation of the European component is expected. Haplogroup determination by SNP analysis remains the best approach, considering the low reliability of prediction of software available.

The adequate LR+ for the Q1a3a and DE haplogroups could be explained by a lower diversity within each group. Especially in the Q1a3a case, which is a relatively recent haplogroup, the homogeneity is the result of its young evolutionary age, given that the time lapse in which the haplotypes spread away from the haplogroup founder is rather short [14].

Considering that the samples from which the calibration frequencies are estimated belong to European (or of European descent) populations, the high false positive proportion in the R* haplogroup could be a reflection of this sampling error, as this haplogroup is the most common among those populations.

In order to get a better idea of what the $U(y|x)$ values mean, let us assume that we are interested in knowing the result of throwing a die (i.e., the information we want to know is whether we will get 1, 2, …, or 6) and take it as the variable y. Let us further assume that we had some way to know beforehand whether the result would be an odd (1, 3, 5) or an even (2, 4, 6) number and take it as the variable x. It turns out that in this case, we get $U(y|x)=0.387$ (38.7%), that

is, knowing beforehand whether the throw of the die will result in an odd or even number, produces a loss of 38.7% of the information that the actual throw will yield. If we compare this result with the values obtained above (24.4% and 20.7%), we see that, for the knowledge of the result of the SNP typing, knowing in advance the result of any of the software systems provides less information than, for the throwing of the die, knowing beforehand whether the result will be odd or even would give.

For the case of the classifier, in which the user chooses the data to train the software, greater Y-STR profiles with associated haplogroups might improve its accuracy. Even so, this software shows haplotypes without an agreed haplogroup, given their recurrence across haplogroups, which is a better approach than suggesting a haplogroup when there is evidence that the haplotype could belong to different haplogroups. However, this software shows higher accuracy than the predictor, given LR+ values.

Why do these software systems show such low accuracy levels? We propose two explanations: (1) there are not enough Y-STR profiles with associated haplogroups to calibrate the software properly and (2) given the mutation rates of the STRs (available at www.yhrd.org) and the time depth of the haplogroup ramifications [21], it is possible to find the same haplotype in samples from different haplogroups (cases of convergence). For instance, the most recent haplogroup, R1*, has a time to the most recent common ancestor (TMRCA) of 18,500 (12,500–25,700) years, and the most ancient clade is estimated at 70,000 years [20], whereas the Y-chromosome STR mutation rates vary from 6.35 (4.19–9.22 95%CI)$\times 10^{-3}$ for DYS 439 to 0.45 (0.12–1.17 95%CI)$\times 10^{-3}$ for the case of DYS 392, so that the repetition of allele combinations among evolutionary divergent clades are not uncommon, given the meiosis events accumulated along such a vast time depth.

A simple way to confirm this is to check the allele frequencies for each STR locus among the metapopulations of the YHRD database, which show that the same alleles are present. Even though their frequencies can differ, (and taking into account that this database does not provide haplogroup information) given the geographic association of the haplogroups evidenced by other authors [22], if the association between STR alleles and haplogroups was strong enough to allow the prediction of the last based on the haplotypes, these should show similar geographic distribution to the one found on the haplogroups.

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 and associated haplogroup, while these reference samples decrease even more as the amount of STRs demanded increases. Also, there is no homogeneity in the use of more STRs, different authors chose different markers, reducing the haplotype references drastically. For example, Zalloua et al. [23] analyzed the seven standard STRs plus DYS 388, DYS 437, DYS 438, and DYS 439, while Di Gaetano et al. [24] studied the seven standard STRs plus DYS 385 A/B in all samples (in only a fraction of their sample other STRs were also included); these data sets are only comparable with each other on the standard STRs, not on the rest.

At present, haplogroup prediction software available does not show adequate accuracy. Thus, typing a set of SNPs which precisely define a phylogenetic branch, is the only reliable method to establish to which haplogroup a given sample belongs.

Acknowledgments The authors thank Dr. J.E. Dipierri and E.L. Alfaro for their valuable contribution, Dr. J.C. Muzzio for his help with the uncertainty coefficient calculations, and Dr. A.N. Califano and C.M. Bravi for their suggestions.

Grant sponsorship: CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), CICPBA (Comisión de Investigaciones Científicas de la Provincia de Buenos Aires), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica), and Antorchas Foundation of Argentina.

G. Bailliet and J.S. López Camelo are members of the CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Argentina.

References

1. Bosch E, Calafell F, Santos F et al (1999) Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am J Hum Genet 65:1623–1638
2. Behar DM, Garrigan D, Kaplan M et al (2004) Contrasting patterns of the Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum Genet 114:354–365
3. Athey WT (2005) Haplogroup prediction from Y-STR values using an allele-frequency approach. J Genet Geneal 1:1–7
4. Athey WT (2006) Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. J Genet Geneal 2:34–39
5. 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
6. Mertens G (2007) Y-Haplogroup frequencies in the Flemish population. J Genet Geneal 3(2):19–25
7. Goff PG, Athey TW (2006) Diagnostic STR values for haplogroup G. J Genet Geneal 2(1):12–17
8. Schlecht J, Kaplan ME, Barnard K, Karafet T, Hammer MF, Merchant NC (2008) Machine-Learning approaches for classifying Haplogroup from Y Chromosome STR data. PloS Comput Biol 4(6):e1000093. doi:10.1371/journal.pcbi1000093
9. Ramallo V, Alfaro EL, Dipierri JE, Bianchi NO, Bailliet G (2005) Caracterización de linajes paternos en muestras provenientes de tres provincias del NOA. Revista de la Sociedad Argentina de Genética. Journal or Basic & Applied Genetics Actas XXXIV Congreso Argentino de Genética XVII(Suplement):181, Septiembre
10. YCC (The Y Chromosome Consortium) (2002) A nomenclature system for the tree of Y chromosomal binary haplogroups. Genome Res 12:339–348

11. de Knijff P, Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G et al (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med 110:134–140
12. Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133
13. Bravi CM, Sans M, Bailliet G, Martinez-Marignac VL, Bianchi NO (1997) Characterization of mitochondrial and Y-chromosome haplotypes in an Uruguayan population of African ancestry. Hum Biol 69(5):641–652
14. Bianchi NO, Catanesi CI, Bailliet G, Martinez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera RJ, Lopez Camelo JS (1998) Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet 63:1862–1871
15. Bailliet G, Castilla EE, Adams JP, Orioli IM, Martínez-Marignac VL, Richard SM, Bianchi NO (2001) Correlation between molecular and conventional genealogies in Aicuña: a rural population from Northwestern Argentina. Hum Hered 51:150–159
16. Sengupta S, Zhivotovsky LA, King R, Mehdi SQ et al (2006) Polarity and temporality of high-resolution ychromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. Am J Hum Genet 78:202–221
17. Cinnioglu C, King R, Kivisild T, Kalfoglu E et al (2004) Excavating y-chromosome haplotype strata in Anatolia. Hum Genet 114:127–148
18. McAlister FA, Straus SE, Sackett DL (1999) Why we need large, simple studies of the clinical examination: the problem and a proposed solution. Lancet 354:1721–24
19. Greenhalgh T (1997) How to read a paper: Papers that report diagnostic or screening tests. BMJ 315:540–543
20. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1994) Numerical recipes in FORTRAN. The art of scientific computing, 2nd edn. Cambridge University Press
21. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF (2008) New binary polimorphisms reshape and increase the resolution of the Human Y chromosomal haplogroup tree. Genome Res 18:830–838
22. Underhill P, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza L (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Anal Hum Genet 65:43–62
23. Zalloua PA, Xue Y, Khalife J, Makhoul N, Debiane L, Platt DE, Royyuru AK, Herrera RJ, Soria Hernanz DF, Blue-Smith J, Spencer Wells R, Comas D, Bertranpetit J, Tyler-Smith C, The Genographic Consortium. 2008. Am J Hum Genet. April 2008. Supplementary Data
24. Di Gaetano C, Cerutti N, Crobu F, Robino C, Inturri S, Gino S, Guarrera S, Underhill PA, King RJ, Romano V, Cali F, Gasparini N, Matullo G, Salerno A, Torre C, Piazza A (2009) Differential Greek and northen African migrations to Sicily are supported by genetic evidence from the Y chromosome. Eur J Hum Gen 17(1):91–99