Available online at www.sciencedirect.com

ScienceDirect

Journal of Genetics and Genomics 42 (2015) 403−407

JGG

LETTER TO THE EDITOR

Convergence of Y Chromosome STR Haplotypes from Different SNP Haplogroups Compromises Accuracy of Haplogroup Prediction

The paternally inherited Y chromosome has been widely used in forensics for personal identification, in anthropology and population genetics to understand origin and migration of human populations, and also in medical and clinical studies (Wang and Li, 2013; Wang et al., 2014). There are two kinds of extremely useful markers in Y chromosome, single nucleotide polymorphism (SNP) and short tandem repeats (STRs). With a very low mutation rate on the order of 3.0×10^{-8} mutations/nucleotide/generation (Xue et al., 2009), SNP markers have been used in constructing a robust phylogeny tree linking all the Y chromosome lineages from world populations (Karafet et al., 2008). Those lineages determined by the pattern of SNPs are called haplogroups. That is to say, we have to genotype an appropriate number of SNPs in order to assign a given Y chromosome to a haplogroup. Compared with SNPs, the mutation rates of STR markers are about four to five orders of magnitude higher (Gusmão et al., 2005; Ballantyne et al., 2010). Typing STR has advantages of saving time and cost compared with typing SNPs in phylogenetic assignment of a Y chromosome (Wang et al., 2010). A set of STR values for an individual is called a haplotype. Because of the disparity in mutation rates between SNP and STR, one SNP haplogroup could actually comprise many STR haplotypes (Wang et al., 2010). It is most interesting that STR variability is clustered more by haplogroups than by populations (Bosch et al., 1999; Behar et al., 2004), which indicates that STR haplotypes could be used to infer the haplogroup information of a given Y chromosome. There has been increasing interest in this cost-effective strategy for predicting the haplogroup from a given STR haplotype when SNP data are unavailable. For instance, Vadim Urasin's YPredictor (http://predictor.ydna.ru/), Whit Atheys' haplogroup predictor (http://www.hprg.com/hapest5/) (Athey, 2005, 2006), and haplogroup classifier of Arizona University (Schlecht et al., 2008) have been widely employed in previous studies for haplogroup prediction (Larmuseau et al., 2010; Bembea et al., 2011; Larmuseau et al., 2012; Tarlykov et al., 2013).

YPredictor is based on the phylogenetic trees of each haplogroup and uses the difference in marker values, marker mutation rates and age of parent node to calculate prediction probability. Whit Atheys' haplogroup predictor is based on genetic-distance approach (Athey, 2005) and Bayesian allele-frequency approach (Athey, 2006). The haplogroup classifier of Arizona University is based on machine-learning approaches, which implements a collection of algorithms including naive Bayes, support vector machines, and decision tree classifiers (Schlecht et al., 2008). Although these approaches for haplogroup prediction are widely used by genealogists, there are still many ongoing debates about the accuracy of using STR haplotypes in haplogroup assigning (Athey, 2011; Muzzio et al., 2011). In particular, it has been reported that the number of STRs used in prediction and the available STR-SNP associated reference data have a significant impact on the accuracy of haplogroup prediction (Schlecht et al., 2008). Furthermore, taking the mutation rates of the STRs and the time depth of the haplogroup ramifications into consideration, it is possible to find the same or similar haplotypes from different haplogroups (Muzzio et al., 2011). However, the possible bias caused by the convergence of STR haplotypes in haplogroup prediction has not been discussed before. In addition, the prediction programs are based on SNP/STR datasets; however, most datasets are inaccessible, intransparent and biased towards certain Y chromosomes due to insufficient sampling. Here, we created a database with a large amount of worldwide Y chromosome SNP and STR data and used this database to address the question about prediction accuracy.

Altogether, 20,403 pieces of Y chromosome data with informative SNP and STR markers have been included in this study (The database is provided in Data S1), including unpublished data of 231 East Asian samples from our lab, unpublished data of 101 samples from Genographic Consortium (haplogroup B), and other 20,071 pieces of data retrieved from the literature. We renewed the haplogroup names according to the nomenclature of Y Chromosome Consortium and the ISOGG Y-DNA Haplogroup Tree 2013 (Karafet et al., 2008; Yan et al., 2011; International Society of Genetic Genealogy, 2013). Different authors typed different STR markers, which

http://dx.doi.org/10.1016/j.jgg.2015.03.008

has reduced the feasibility for STR haplotype references. Here, we used the AmpFlSTR® Yfiler™ seventeen Y chromosomal STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a, DYS385b, DYS438, DYS439, DYS437, DYS448, DYS456, DYS458, DYS635, and YGATAH4) as standard for creating the SNP/STR database (Data S1) and haplogroup prediction test. In order to include more individuals in comparison, we only used 10 commonly used STRs for R_{ST} and structure analysis.

Slatkin's R_{ST}, a linearized F_{ST} suited for the stepwise mutation model that we think, applies to Y-STR data (Slatkin, 1995). R_{ST} matrices for 10 commonly used STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439) among different haplogroups were performed using the Arlequin ver 3.5 software package (arlecore3513_64bit in Linux) (Excoffier and Lischer, 2010). The neighbor-joining tree was constructed in MEGA 5.10 (Tamura et al., 2011) using the *Rst* matrices. A Markov Chain Monte Carlo analysis of haplogroup structure was carried out using the above 10 STRs in program Structure 2.3.4 to give detailed comparisons of each haplotype among haplogroups (Pritchard et al., 2000). To use the Structure program, we first assumed that Y-STR loci are independent within haplogroups. Each haplotype was represented as a single locus with 10 alleles. Although this assumption simplified the mutational mechanism of Y-STR, it doesn't seem to affect haplogroup prediction and haplotype comparison (Athey, 2006). YPredictor by Vadim Urasin v1.5.0 (http://predictor.ydna.ru/) was used for haplogroup prediction. The reason for choosing YPredictor to estimate error rates of prediction is that YPredictor contains almost all the known worldwide haplogroups.

Our dataset has covered all the main haplogroups and almost all their sublineages in the Y chromosome phylogeny tree (Fig. 1A, the dataset is provided in Data S1). This informative database will be very useful in subsequent Y chromosome studies. Although the high mutation rates of STR markers make it difficult to construct phylogeny trees, the neighbor-joining tree of STR data (Fig. 1B, the R_{ST} matrices and their P values are given in Data S2) shows a similar pattern as the trunk of Y chromosome haplogroup tree (Fig. 1A). There are four main branches in the neighbor-joining tree, namely I, II, III, and IV. The three ancient haplogroups C-M130, D-M174, and E-M96 representing the out of Africa migration were clustered in branch I. Middle East and Europe specific haplogroups F-M89, G-M201, H-M69, I-M170, and J-M304 were mainly clustered in branch II. Haplogroups L-M11, T-M272, K-M9, N-M231, and O-M175 representing the peopling of the Far East were clustered in branch III. The youngest haplogroups P-M45, Q-M242, and R-M207 were mainly clustered in branch IV.

Obvious haplogroup divisions were also observed in the neighbor-joining tree (Fig. 1B), in which haplogroups R1b1+s (R1b1-P25) (+s means certain haplogroup and its sublineages), haplogroups Q1+s (Q1-P36.2), haplogroups G2a1+s (G2a1-L293), haplogroups J2+s (J2-M172), haplogroups E1b1b1+s (E1b1b1-M35), haplogroups D+s (D-M174), and haplogroups E1b1a1+s (E1b1a1-M2) were clustered tightly together, demonstrating the specific or even exclusive STR haplotypes of those haplogroups. However, haplogroups A+s (A00-AF4) and B+s (B-M60) were scattered in the tree, probably due to the high diversification among different sublineages of the two oldest haplogroups. Haplogroups C+s (C-M130) were mainly clustered with haplogroups D-M174 and E-M96. Haplogroups C1-M131 and C3-M217 showed strong affinity with haplogroup E1b1b1-M35; however, C2a*-M208 and C2a1-P33 tended to cluster with E1b1a1-M2. The three main subclades of haplogroups O-M175, O1-MSY2.2, O2-M268 and O3-M122, tended to be clustered together. Haplogroups O1a+s (O1a-M119) showed very strong affinity with haplogroups N*-M231 and N1c-Tat. Haplogroups O2a1-M95 and O2a1a-M88 fell out the scale of STR patterns of haplogroup O-M175. Haplogroups O2b*-M176 and O2b1a-47z were clustered with T*-M272 and T1a*-M70, and also showed affinity with O3a*-M324 and O3a1c*-002611. Haplogroups O3a2+s (O3a2-P201) formed a tight cluster, indicating high similarities between those lineages, although haplogroups L+s (L-M11) have also been placed in the O3a2-P201 cluster. Haplogroup P-M45 was clustered with haplogroups Q+s (Q-M242) in a separated small branch. Haplogroup R1a1-SRY10831.2 was clustered with E1b1a1-M2 and C2a-M208 in branch I, while its sister clades R1b1+s (R1b1-P25) were grouped with R2-M479 and M1b-P87 in branch IV.

The neighbor-joining tree based on pairwise comparisons gives an overall clustering pattern of the worldwide haplogroups. However, the results at haplogroup level could be misleading because of the highly diversified STR haplotypes within haplogroups, especially in basal lineages. Here, we used Structure software to show the STR haplotype patterns among haplogroups at individual level (Fig. S1). We also used YPredictor to infer haplogroup for each haplotype and then compared the inferred haplogroups with the genotyped haplogroups to estimate the error rates (Data S3). The basal branch A00-AF4 also has the exclusive STR haplotypes. The haplotypes of haplogroups A1a-M31 and A1b1b2b-M13 show similarities with haplogroups DE-YAP and E1b1a1-M2, and thus about 30% of A1a-M31 and A1b1b2b-M13 samples were mistaken as DE-YAP in YPredictor. Haplogroups B+s (B-M60) are probably the most diverse clades, sharing similar haplotypes with various haplogroups, such as haplogroups I2a1-P37.2, R1a1-SRY10831.2, D2a-M55, E1b1b1-M35, and L-M11. Actually, only 18% of haplogroup B-M60 samples could be successfully inferred, 26% were mistaken as I2-M438 or IJ-M429, and 21% were assigned as haplogroup R-M207 in YPredictor. Similar to haplogroup B-M60, the haplotypes of paragroups F*-M89, H*-M69, and K*-M9 are also too diverse to be used in haplogroup prediction. Most haplotypes of haplogroup C1-M131 are similar to those of E1b1b1-M35 and 22% of C1-M131 samples were mistaken as E1b1b1-M35 and its subhaplogroups in prediction. Similarly, haplogroup C2-M38 and its sublineages C2a-M208 and C2a1-P33 shared most haplotypes with E1b1a1-M2 and therefore 37% of those C2-M38 samples were mistaken as E1b1a1-M2 and E1b1a1g-U175. The haplotype pattern of haplogroups H1-

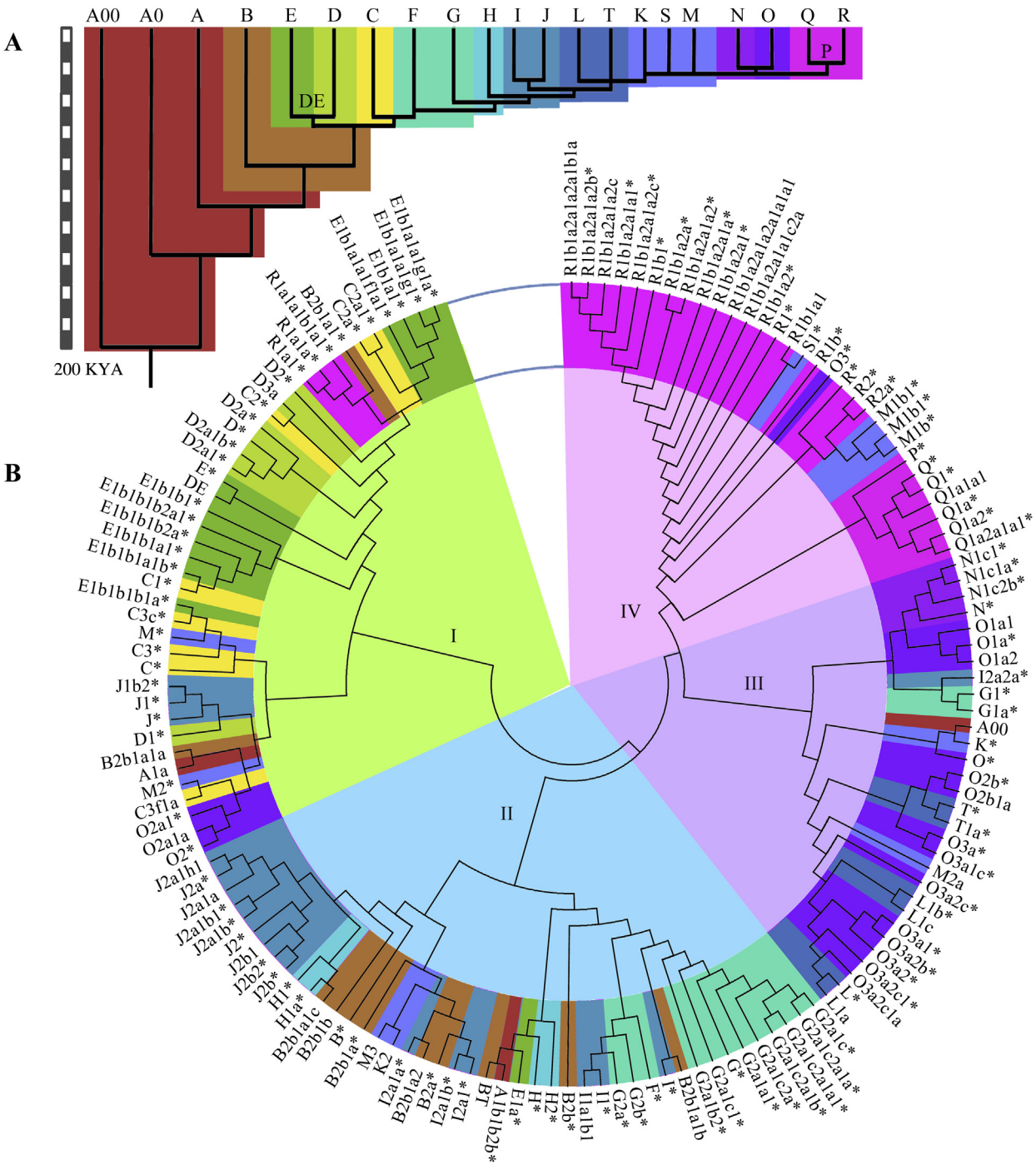

Fig. 1. Neighbor-joining tree and STR structure of 146 Y chromosome haplogroups.

A: The trunk of the Y chromosome haplogroup tree. The X-axis is the Y chromosomal haplogroup names and the Y-axis is the TMRCA (time to the most recent common ancestor) for the haplogroups. KYA, kilo years ago. **B:** Neighbor-joining tree of 146 Y chromosome haplogroups based on *Rst* distance of 10 commonly used STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439).

M52 and H1a-M82 is very similar to that of haplogroup J-M304, resulting in erred assigning of about 20% of H1-M52 and H1a-M82 samples as J-M304 in prediction. Haplogroups I*-M170, I1*-M253, and I1a1b1-P109 share some haplotypes with haplogroups G+s (G-M201). The haplotypes of haplogroups L+s (L-M11) are similar to those of haplogroup O3a2c1*-M134 and O3a2c1a-M117. The haplotypes of O1a+s (O1a-M119) bear some similarity to those of haplogroup N-M231. Similarly, haplotypes of haplogroup O3a1c-002611 show some similarity to O3a2b-M7, and M1a-P34 and M1b-P87 are similar to O3a2+s (O3a2-P201). Those haplotype sharing similarities among different haplogroups often mislead us in haplogroup prediction. On the contrary, haplogroups D2a-M55, D3a-P47, O2a1-M95, O2b-M176, R-

M207, and Q-M242 have haplogroup specific haplotypes and could be predicted by STR with high accuracy. It is still worthy to note that the affinity between haplogroups D-M174 and E-M96 might confuse the two in some cases.

The purpose of our study is not to address the quality of the haplogroup prediction software, as no algorithms could be powerful enough to distinguish the same or very similar haplotypes and assign them into different haplogroups. The convergence of Y chromosome STR haplotypes among different haplogroups has compromised the accuracy of haplogroup prediction. For samples with ambiguous STR haplotypes, typing SNPs is the only reliable method to determine the haplogroups. Furthermore, we propose two possible explanations for the observed convergence. For basal lineages, the time is long enough for their STR haplotypes to mutate to high resemblance since they branched out. For Neolithic expanded lineages, such as R1b1-P25 and E1b1b1-M35, the high resemblance of haplotypes belonging to several subhaplogroups is probably the results of recent radiations as discussed by Larmuseau et al (2014).

ACKNOWLEDGMENTS

This work was supported by the National Excellent Youth Science Foundation of China (No. 31222030), the National Natural Science Foundation of China (No. 91131002), the Shanghai Rising-Star Program (No. 12QA1400300), the China Ministry of Education Scientific Research Major Project (Nos. 311016 and 113022A), the MOE University Doctoral Research Supervisor's Funds (No. 20120071110021), and the Shanghai Professional Development Funding (No. 2010001).

SUPPLEMENTARY DATA

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jgg.2015.03.008.

Chuan-Chao Wang[a], *Ling-Xiang Wang*[a], *Rukesh Shrestha*[a], *Shaoqing Wen*[a], *Manfei Zhang*[a], *Xinzhu Tong*[a], *Li Jin*[a,b], *Hui Li*[a,*]

[a]*State Key Laboratory of Genetic Engineering and MOE Key Laboratory of Contemporary Anthropology, School of Life Sciences and Institutes of Biomedical Sciences, Fudan University, Shanghai 200433, China*

[b]*CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences of Chinese Academy of Sciences, Shanghai 200031, China*

*Corresponding author. Tel: +86 21 5163 0427, fax: +86 21 5163 0607.
E-mail address: lihui.fudan@gmail.com (H. Li)

Received 30 November 2014
Revised 21 March 2015
Accepted 24 March 2015
Available online 1 April 2015

REFERENCES

Athey, W.T., 2005. Haplogroup prediction from Y-STR values using an allele-frequency approach. J. Genet. Geneal. 1, 1−7.

Athey, W.T., 2006. Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach. J. Genet. Geneal. 2, 34−39.

Athey, W., 2011. Comments on the article, "Software for Y haplogroup predictions, a word of caution". Int. J. Legal. Med. 125, 901−903.

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., Brauer, S., Decorte, R., Poetsch, M., von Wurmb-Schwark, N., de Knijff, P., Labuda, D., Vézina, H., Knoblauch, H., Lessig, R., Roewer, L., Ploski, R., Dobosz, T., Henke, L., Henke, J., Furtado, M.R., Kayser, M., 2010. Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. Am. J. Hum. Genet. 87, 341−353.

Behar, D.M., Garrigan, D., Kaplan, M., Mobasher, Z., Rosengarten, D., Karafet, T.M., Quintana-Murci, L., Ostrer, H., Skorecki, K., Hammer, M.F., 2004. Contrasting patterns of the Y chromosome variation in Ashkenazi Jewish and host non-Jewish European populations. Hum. Genet. 114, 354−365.

Bembea, M., Patocs, A., Kozma, K., Jurca, C., Skrypnyk, C., 2011. Y-chromosome STR haplotype diversity in three ethnically isolated population from North-Western Romania. Forensic. Sci. Int. Genet. 5, e99−e100.

Bosch, E., Calafell, F., Santos, F., Pérez-Lezaun, A., Comas, D., Benchemsi, N., Tyler-Smith, C., Bertranpetit, J., 1999. Variation in short tandem repeats is deeply structured by genetic background on the human Y chromosome. Am. J. Hum. Genet. 65, 1623−1638.

Excoffier, L., Lischer, H.E., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol. Ecol. Resour. 10, 564−567.

Gusmão, L., Sanchez-Diz, P., Calafell, F., Martin, P., Alonso, C.A., Alvarez-Fernandez, F., Alves, C., Borjas-Fajardo, L., Bozzo, W.R., Bravo, M.L., Builes, J.J., Capilla, J., Carvalho, M., Castillo, C., Catanesi, C.I., Corach, D., Di Lonardo, A.M., Espinheira, R., Fagundes de Carvalho, E., Farfán, M.J., Figueiredo, H.P., Gomes, I., Lojo, M.M., Marino, M., Pinheiro, M.F., Pontes, M.L., Prieto, V., Ramos-Luis, E., Riancho, J.A., Souza Góes, A.C., Santapa, O.A., Sumita, D.R., Vallejo, G., Vidal Rioja, L., Vide, M.C., Vieira da Silva, C.I., Whittle, M.R., Zabala, W., Zarrabeitia, M.T., Alonso, A., Carracedo, A., Amorim, A., 2005. Mutation rates at Y chromosome specific microsatellites. Hum. Mutat. 26, 520−528.

International Society of Genetic Genealogy, 2013. Y-DNA Haplogroup Tree 2013, Version: 8.76. Date: 3 October 2013. http://www.isogg.org/tree/ [Date of access: 15 October 2013].

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L., Hammer, M.F., 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 18, 830−838.

Larmuseau, M.H., Ottoni, C., Raeymaekers, J.A., Vanderheyden, N., Larmuseau, H.F., Decorte, R., 2012. Temporal differentiation across a West-European Y-chromosomal cline: genealogy as a tool in human population genetics. Eur. J. Hum. Genet. 20, 434−440.

Larmuseau, M.H., Vanderheyden, N., Jacobs, M., Coomans, M., Larno, L., Decorte, R., 2010. Micro-geographic distribution of Y-chromosomal variation in the central-western European region Brabant. Forensic. Sci. Int. Genet. 5, 95−99.

Larmuseau, M.H., Vanderheyden, N., Van Geystelen, A., van Oven, M., de Knijff, P., Decorte, R., 2014. Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. Ann. Hum. Genet. 78, 92−103.

Muzzio, M., Ramallo, V., Motti, J.M., Santos, M.R., López, Camelo, J.S., Bailliet, G., 2011. Software for Y-haplogroup predictions: a word of caution. Int. J. Legal. Med. 125, 143−147.

Pritchard, J.K., Stephens, M., Donnelly, P., 2000. Inference of population structure using multilocus genotype data. Genetics 155, 945−959.

Schlecht, J., Kaplan, M.E., Barnard, K., Karafet, T., Hammer, M.F., Merchant, N.C., 2008. Machine-Learning approaches for classifying Haplogroup from Y Chromosome STR data. PLoS Comput. Biol. 4, e1000093.

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics 139, 457−462.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731−2739.

Tarlykov, P.V., Zholdybayeva, E.V., Akilzhanova, A.R., Nurkina, Z.M., Sabitov, Z.M., Rakhypbekov, T.K., Ramanculov, E.M., 2013. Mitochondrial and Y-chromosomal profile of the Kazakh population from East Kazakhstan. Croat. Med. J. 54, 17−24.

Wang, C.C., Li, H., 2013. Inferring human history in East Asia from Y chromosomes. Investig. Genet. 4, 1−10.

Wang, C.C., Jin, L., Li, H., 2014. Natural selection on human Y chromosomes. J. Genet. Genomics 41, 47−52.

Wang, C.C., Yan, S., Li, H., 2010. Surnames and the Y chromosomes. Commun. Contemp. Anthropol. 4, 26−33.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., Asan, MacArthur, D.G., Quail, M.A., Carter, N.P., Yang, H., Tyler-Smith, C., 2009. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr. Biol. 19, 1453−1457.

Yan, S., Wang, C.C., Li, H., Li, S.L., Jin, L., Genographic Consortium, 2011. An updated tree of Y-chromosome haplogroup O and revised phylogenetic positions of mutations P164 and PK4. Eur. J. Hum. Genet. 19, 1013−1015.