NETWORK ANALYSIS ON THE *IN SILICO* ASSIGNED Y CHROMOSOME HAPLOGROUPS IN WESTERN BALKAN POPULATIONS

Emir Šehović[1*], Adna Ašić[1], Mustafa Dogan[1], Ramazan Tunc[1], Damir Marjanovic[1,2], Serkan Dogan[1]

[1]*Department of Genetics and Bioengineering, Faculty of Engineering and natural Sciences, International Burch University (IBU), Sarajevo, Bosnia and Herzegovina*
[2]*Institute for Anthropological Research, Zagreb, Croatia*

***Correspondence**
E-mail:
emir.sehovic@gmail.com

Received
September, 2017
Accepted
November, 2017
Published
December, 2017

Research article

Abstract

The region of Western Balkans has been inhabited since the Paleolithic era and was the route of the spread of farming from the Middle East to Europe during the Neolithic era. In the present study, Y-STR data from European populations have been used to construct median-joining networks. The study was performed using Whit Athey's Haplogroup Predictor, Y Utility and Network 4 software packages to predict Y haplogroups, construct networks, perform clustering of closely related Y chromosomes and calculate time estimates between individual nodes. The results of the study imply that geographically close populations cluster together at both Balkan and European levels. It was observed that an elevated number of study populations and individual haplogroups increases the possibility that individuals of different ethnic background cluster within the same or neighboring clades of network. Subsequent time estimates, performed based on the mutation frequency between the ancestral node and its descendant nodes, revealed that I2a haplogroup within the Western Balkan region has the most compact clustering (age, estimated at 3109 years), followed by Hg E1b1b which has the second most compact clustering (4896 years). The obtained results are nonetheless in accordance with previously published research investigating the frequency of Y haplogroups based on Y-SNP variant frequencies, indicating that Western Balkan countries are mainly represented by I2a subclade (average for six countries 32.3%), followed by E1b1b and R1a (average for six countries of 21.5% and 17%, respectively).

Key words: *Western Balkans, Y haplogroup, Y-STR, Median-joining network analysis, in silico Y haplogroup assignment*

Introduction

The area of the Balkan Peninsula has been occupied since the Paleolithic, and the Neolithic migrations spread across Europe from Anatolia through the Balkans (Bosch et al., 2006). Western Balkan populations contain several major haplogroups that are spread throughout all Balkan countries with some countries having a higher percentage of a certain haplogroup. The four major haplogroups in the Western Balkans are I2a, E1b1b, R1a and R1b, with I2a and E1b1b being the most abundant. The inclusion of a significant percentage of R1a and R1b within the Western Balkan populations confirms their historical intertwining with the other populations of Europe (Rosser et al., 2000; Semino et al., 2000; Barać et al., 2003; Marjanović et al., 2005; Dogan et al., 2016a, 2016b). I2a1 in Southeast

Europe is one of the main male lineages inherited from European Upper Paleolithic Y chromosome that first appeared in the western Mediterranean region. Hg I2 is the most common paternal lineage in the countries that constituted former Yugoslavia, as well as in Romania, Bulgaria Sardinia, and in most other Slavic countries. Lineage E1b1b, representing the migration from Africa into Europe, appeared in Africa and spread to North Africa and the Near East during the late Paleolithic and Mesolithic periods. Hg R1a branched of R1 during or soon after the Last Glacial Maximum (LGM) and contributes to the growing evidence that the area of the Balkans has been inhabited as one of the European refugia during the LGM. It might have originated in Central Asia or Siberia and spread from Eastern Europe to India. Finally, the oldest forms of R1b are found dispersed from Western Europe to India and it is the most common haplogroup in Western Europe. R1b in Balkan probably has a slightly different origin than the one in the rest of the Europe and it is currently considered to be associated with spreading of the West-European or Iberian Y chromosome to the Balkan Peninsula (Semino et al., 2000; Battaglia et al., 2009; Primorac et al., 2011; Varzari et al., 2013; Šarac et al., 2016).

The usage of Y-STR markers to predict Y haplogroups is a relatively novel approach in population genetics (Athey, 2006) that uses one or more haplogroup assignment algorithms to calculate the probability that a certain haplotype corresponds to a previously characterized haplogroup. This approach is faster, less labor-intensive and far less expensive when compared to the standard haplogroup identification procedure utilizing Y-SNP markers (International Society of Genetic Genealogy, 2015).

Therefore, the aim of the present study was to perform *in silico* haplogroup assignment from Y-STR data using Whit Athey's Haplogroup Predictor and to construct median-joining networks for clustering of closely related Y chromosomes. In that way, the present study intends to prove that selected high-quality haplogroup assignment algorithms can reproduce the results generated by Y-SNPs in terms of Y haplogroup identification and that genetic relationships between the populations can be elucidated using a representative set of their Y chromosomes.

Materials and methods

The data used in this study was PowerPlex Y23 allele frequencies for 12 populations including Bosnian-Herzegovinian, Croatian, German (from Bavaria), Irish, Italian, Macedonian, Polish, Slovenian, Spanish, and Swedish (Purps et al., 2014). Furthermore, the haplotypes of Montenegrin and Serbian populations were used over 17 Y-STRs (Mirabal et al., 2010). A total of 1815 samples with 1721 unique haplotypes had their haplogroups predicted from Y-STR data by using Whit Athey's haplogroup predictor (Athey, 2006), which offers results of haplogroup predictions in percentages along with fitness scores and the percentages shown indicate the calculated probability of a haplotype belonging to a certain haplogroup. Out of the total number of samples analyzed, 1127 were from the Western Balkan populations which included 1033 unique haplotypes (Table 1).

The median-joining trees were generated using the Fluxus Network 4.6 program (Bandelt et al., 1999). The post-processing option was utilized as it calculates all the possible variations of the vectors which can be generated in the tree and selects the optimal one, while visualizing all the other possible trees generated by the algorithm (Bandelt et al., 1999). From the mentioned Y-STR datasets, 443 haplotypes were used in the network analysis. The original population datasets contained, on average, 143 unique haplotypes per population, but in order to obtain a clear median-joining tree, 15-20 haplotypes per population which corresponded to a certain haplogroup were selected to be analyzed. Different haplotypes were selected based on the haplogroup being analyzed. The selected individuals had a positive identification of their haplogroup. Also, only major haplogroups that were assigned by the Whit Athey's haplogroup predictor were selected for network analysis while minor ones were excluded. For creating the median-joining tree involving only the Western Balkan populations, ten loci were used, namely DYS393, DYS390, DYS19,

Table 1. Populations used in the present study with their respective sample sizes

Population	Samples	Unique haplotypes
Bosnia and Herzegovina	100	100
Croatia	239	239
Germany (from Bavaria)	195	195
Ireland	31	31
Italy	58	58
Macedonia	101	101
Montenegro	404	318
Poland	102	102
Serbia	179	171
Slovenia	104	104
Spain	251	251
Sweden	51	51
Total	1815	1721

DYS391, DYS439, DYS389I, DYS389II, DYS458, DYS448, and DYS456. The relative time estimates for clusters of interest were also calculated using the Fluxus Network 4.6 program (Bandelt et al., 1999). The relative time estimation was done for the purpose of precise evaluation of network analysis tree compactness which shows the similarity between the haplotypes within that particular cluster when compared to another cluster of interest.

The respective mutation rates for each locus used in network analysis were obtained and their sum was calculated. Finally, the obtained result was multiplied by the number that would generate one mutation per X number of years. The X number of years would then be multiplied by 25 as it represents the generation time. The X years obtained would be inserted in the program which would then calculate the relative time estimates.

Results and Discussion

Haplogroup distribution

According to the Whit Athey's predictor, the countries from the Western Balkans (Bosnia and Herzegovina, Croatia, Macedonia, Montenegro, Serbia and Slovenia) have four major haplogroups, namely I2a, E1b1b, R1a and R1b. The distribution of haplogroups within the Western Balkan countries is relatively uniform with an increased resemblance between Serbia, Montenegro and B&H. The Croatian haplogroup percentages also concur with the Serbian, Montenegrin and B&H haplogroup structure. However, the Slovenian population showed better similarity to the Croatian population than the other populations.

The analyzed Croatian, Serbian, Montenegrin and Bosnian-Herzegovinian haplotypes mainly belong to the I2a haplogroup while the majority of Macedonian and Slovenian haplotypes belong to the E1b1b and R1a haplogroups, respectively. The average haplogroup distribution of I2a for the six studied in the Western Balkan is 32.3%, followed by 21.5% and 17% for E1b1b and R1a, respectively (Figure 1).

The results are in concordance with the work published by Battaglia and colleagues (2009) which showed a larger inclusion of Croatian and Slovenian haplotypes in the R1a and R1b haplogroup compared to the B&H and Macedonian haplotypes. According to this paper based on Y-SNP analysis, 38.7% and 27% of Slovenians and Croats belong to the R1a haplogroup, respectively, while 13.8% of B&H haplotypes belong to this same lineage (Table 2; Battaglia et al., 2009). In addition, the percentage of Bosnian-Herzegovinian haplotypes in the E1b1b haplogroup is 13.9%, as compared to 6.7% for

Croatian and 2.7% for Slovenian population (Battaglia et al., 2009), which is concordant with the present results. Battaglia et al. (2009) give information on I2a haplogroup, where 50.3% of B&H Y chromsomes belong to this haplogroup, followed by 32.6% Croats and 20% Slovenians.

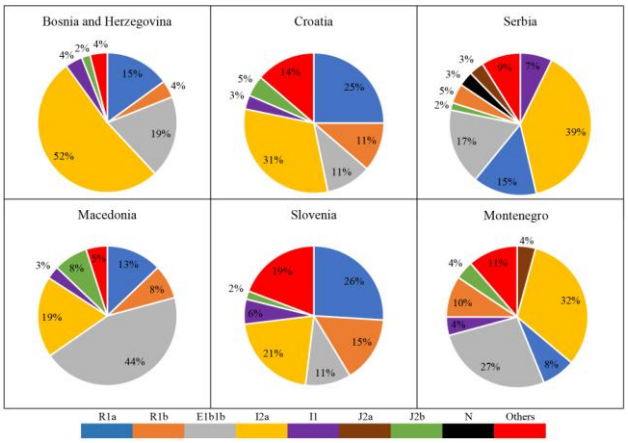

Figure 1. Haplogroup distribution among the Western Balkan populations. The most abundant haplogroup among the Western Balkan populations is the I2a haplogroup followed by the E1b1b and R1a

Lastly, among the four haplogroups analyzed in this study, the least abundant haplogroup R1b, has the following percentage values for the Western Balkan populations based on Y-SNP typing (Battaglia et al., 2009): Bosnia and Herzegovina 4%, Croatia 12.4% and Slovenia 21.3%. According to the present haplogroup assignment study (Figure 1), comparable results were obtained.

Western Balkan network analysis of haplotypes from the four major haplogroups

The clustering in the Western Balkan network analysis tree is gradual, indicating a degree of similarity between the haplotypes despite prediction that they belong to different haplogroups. The clusters of I2a and R1a haplogroups shiow the most proper organization, which can be expected due to them being the most prevalent haplogroups in the Western Balkan region. Haplotypes predicted to be E1b1b haplogroup have also clustered with the haplotypes from the same predicted haplogroup. However, according to Figure 2, there are instances where E1b1b predicted haplotypes do have instances where they cluster with the other major haplogroup clusters, such as I2a cluster. The R1b predicted

Table 2. Haplogroup percentages in three Western Balkan countries based on Y-SNP data (adapted from Battaglia et al., 2009)

	Bosnia and Herzegovina	**Croatia**	**Slovenia**
E1b1b	13.9%	6.7%	2.7%
R1a	13.8%	27%	38%
R1b	4%	12.4%	21.3%
I2a	50.3%	32.6%	20%

haplotypes have the lowest compactness in their clusters which is to be expected as this is the least abundant predicted haplogroup within the Western Balkan populations (Figure 2B). On the other hand, clustering pattern of Y chromosomes on the basis of the population to which they belong, is far less definitive and sharp transitions between the countries, as well as population-specific clusters, cannot be observed (Figure 2A).

This clustering, as mentioned above, was done using ten Y-STR loci. The reason for this is that some loci are highly uninformative with low levels of polymorphism in the study populations. Furthermore, using too many loci causes excessive tree branching and the similarities within the haplotypes difficult to visualize.

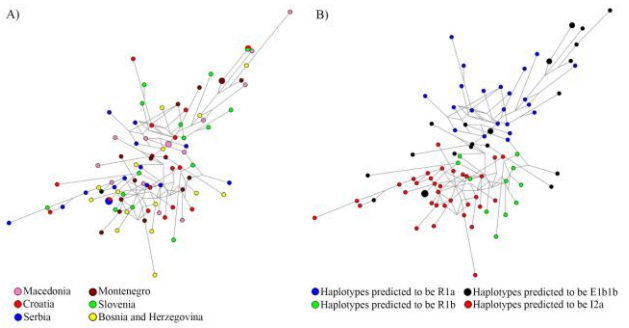

Figure 2. Western Balkan analysis including haplotypes from all four major haplogroups (I2a, E1b1b, R1a and R1b). (A) The population origin of the analyzed haplotypes. (B) The haplogroup prediction and clustering of the same haplotypes

According to the work of Zhivotovsky (2001), the summed mutation rate for the used ten loci for the

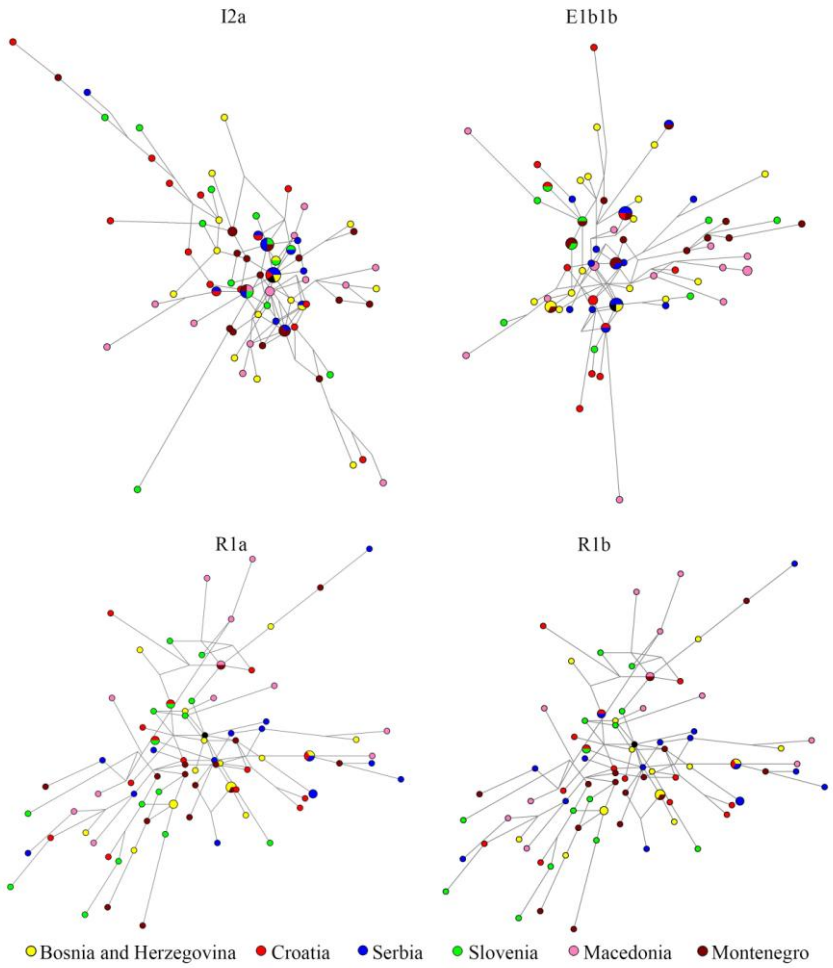

Figure 3. Four median-joining trees generated with haplotypes from the four most abundant haplogroups in the Western Balkan: I2a, E1b1b, R1a and R1b. The I2a and E1b1b haplogroup median-joining trees show the highest degree of compactness

Western Balkan network analysis trees is 0.031456, which was further used to calculate the time estimate values of haplogroup generation in years. Time estimate calculation has shown that the I2a haplogroup within the Western Balkan region has the most compact clustering which is shown by relative time estimate value represented in age of 3109 years, followed by haplogroup E1b1b which has the second most compact clustering out of the four major haplogroup network analysis trees with the time estimate value of 4896 years.

The study of individual haplogroups in the Western Balkan region

In order to better assess the haplogroup distribution in the Western Balkan region, individual network analysis trees have been generated for every major haplogroup. In Hg I2a network, there are two minor branches which indicate a certain degree of difference between these haplotypes and the ones in the main cluster, despite this haplogroup having a highly compact network and relatively similar haplotypes predicted to belong to this lineage (Figure 3).

The E1b1b network analysis tree does not have as compact cluster as observed in the case of Hg I2a. It has 2 subclusters which contain mainly Montenegrin and Macedonian haplotypes in separate subclusters. The Bosnian-Herzegovinian haplotypes are the most scattered within the E1b1b haplogroup network analysis tree, while the rest of the network can be considered relatively compact (Figure 3). This observation might indicate that the B&H haplotypes predicted to be E1b1b have notable differences

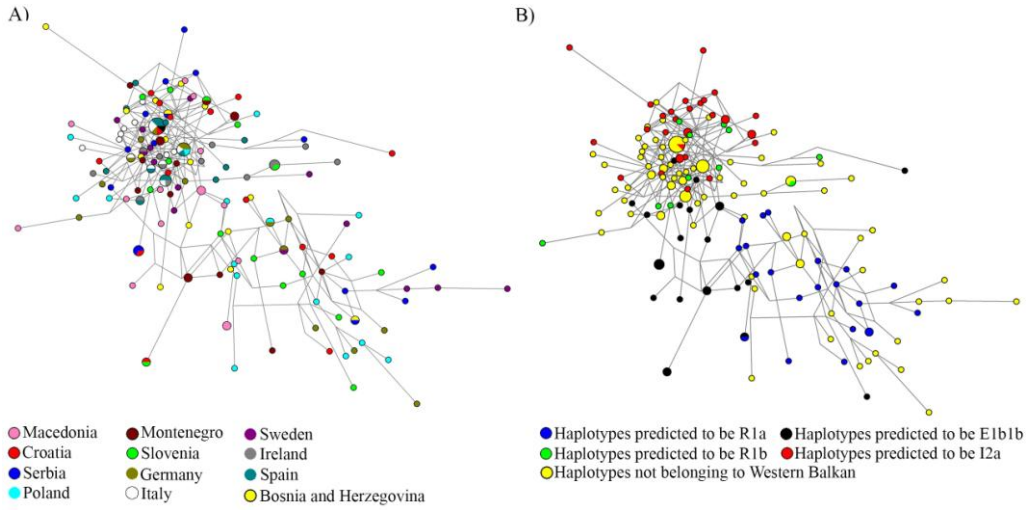

Figure 4. Median-joining tree of the Western Balkan populations analyzed together with a set of six selected European populations. (A) The population origin of the analyzed haplotypes. (B) The haplogroup prediction and clustering of the same haplotypes

between each other when compared to haplotypes from other populations.

The R1a haplogroup cluster has a very high variance rate and does not have a compacted clustering pattern when compared to the previously discussed I2a and E1b1b haplogroup network analysis trees (Figure 3), which was found to have further implications when this haplogroup is analyzed in the broader context. Within the R1a network analysis, the most scattered haplotypes are the ones from Macedonia and the same conclusion can be drawn as with the B&H haplotypes within the E1b1b haplogroup network analysis tree.

Being the haplogroup with the least haplotypes from the Western Balkan populations, the R1b haplogroup network analysis tree has a relatively high variance degree and is not compacted when compared to the I2a and E1b1b trees (Figure 3). One notable characteristic of this median-joining tree is the higher degree of clustering between the B&H, Montenegrin and Serbian haplotypes.

European network analysis trees

When six new European countries (Germany, Ireland, Italy, Poland, Spain and Sweden) are added to the network analysis trees of the Western Balkan populations, a new insight into the study results is obtained (Figure 4). As already seen in the Western Balkan analysis tree, sharp transitions between the haplotypes from individual European populations is not easy to observe (Figure 4A). It is also observable that, out of four major haplogroups in the Western Balkans, I2a and R1b have the most compact clustering with other European populations (Figure 4B). Such clustering pattern of Hg R1b is not surprising, considering that R1b is the most or the second most abundant lineage in all six European populations studied in this analysis tree (Table 3). Furthermore, an R1a cluster with Western Balkan, Polish and German populations was formed (per-country data not shown). This can indicate that there is a significant degree of similarity between the predicted R1a haplotypes from the European and Western Balkan populations. On the other hand, the E1b1b predicted haplotypes were the least likely to cluster with the other European populations (Figure 4B). This analysis was performed using eight Y-STR loci as it is easier for the program to visualize the similarities and differences in the generated tree if a higher number of populations is analyzed.

Semino et al. (2004) performed a similar network analysis using five Y-STR loci and analyzing subclades of haplogroups E and J based on 36 binary markers (Y-SNPs). Their median joining trees showed very compact clustering among most of the subclades, further showing the usefullness of median-joining network analysis for the purpose of haplogroup distribution analysis. Two other studies (Zerjal et al., 2003; Rootsi et al., 2004) also performed network analysis on haplogroups of

interest but with different approaches. Rootsi et al. (2004) focuse on European Hg I and displayed the Balkan clustering among the European populations. The study used six Y-STR loci for generating the network tree with haplotyes with a frequency >1 in that set of populations, which yielded clear trees and an informative network about the relations of different haplogroups but not relations within the haplogroup clusters themselves (Rootsi et al., 2004).

Table 3. List of European populations analyzed in the current study and their major haplogroups as assigned by Whit Athey's Haplogorup Predictor

Population	**Major haplogroups**	
Germany	R1a	R1b
Ireland	R1b	
Italy	R1b	
Poland	R1b	R1a
Spain	R1b	
Sweden	R1b	R1a

The other study (Zerjal et al., 2003) used 15 Y-STR loci for generating a very detailed median-joining tree of the haplogroup C distribution among the Mongolian population. Using the same set of samples, 16 Y-SNP markers were also analyzed to perform initial haplogroup assignment and assess the accuracy of the haplogroup prediction from STR markers. In contrast, the current study used ten loci for the network analysis among the Western Balkan populations and eight loci for Europe-level population clustering and obtained very informative trees. This reduction of the number of loci enables the creation of clearer clustering of major haplogroups at the expense of the intermediate haplotypes clustered between the major clusters being invisible.

Conclusions

The Western Balkan area is found on the crossroad between East and West. Hence, many attributes and aspects of many cultures are blended in this area. As this and many other studies have shown, the Balkan area is genetically very diverse. According to the present results, there are four major haplogroups in the Western Balkan populations, namely I2a, E1b1b, R1a and R1b. Depending on the haplogroup in focus, the level of similarity between the Western Balkan populations is different, but in general, the Western Balkan populations can be considered relatively similar to one another.

The haplogroup assignment algorithm used in the present study has proven its reliability by being in agreement with previously published data on the haplogroup distribution in the Western Balkan countries based on Y-SNPs. However, it should be noted that Y haplogroups are identified and defined according to specific mutations and that Y-SNP analysis is the ultimate and final way to classify a certain Y chromosome to a specific haplogroup. In order to optimize the usage of a particular Y-STR dataset for network analysis, it would be ideal that multiple haplogroup predictors are used and only haplotypes with full concordance between the predictors are used for further studies.

References

Athey TW (2006) Haplogroup prediction from Y-STR values using an allele-frequency approach. J Genet Geneal, 1:1-7.

Bandelt HJ, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evolution, 16(1):37-48.

Barać L, Peričić M, Klarić IM, Rootsi S, Janićijević B, Kivisild T, Parik J, Rudan I, Villems R, Rudan P (2003) Y chromosomal heritage of Croatian population and its island isolates. Eur J Hum Genet, 11(7):535-542.

Battaglia V, Fornarino S, Al-Zahery N, Olivieri A, Pala M, Myres NM, ..., Semino O (2009) Y-chromosomal evidence of the cultural diffusion of agriculture in Southeast Europe. Eur J Hum Genet, 17(6):820-830.

Bosch E, Calafell F, González-Neira A, Flaiz C, Mateu E, Scheil HG, ..., Comas D (2006) Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. Ann Hum Genet, 70(4):459-487.

Dogan S, Ašić A, Doğan G, Bešić L, Marjanović D (2016a) Y-Chromosome Haplogroups in the Bosnian-Herzegovinian Population Based on 23 Y-STR Loci. Hum Biol, 88(3):201-209.

Dogan S, Babic N, Gurkan C, Goksu A, Marjanovic D, Hadziavdic V (2016b) Y-chromosomal haplogroup distribution in the Tuzla Canton of Bosnia and Herzegovina: A concordance study using four different in silico assignment algorithms based on Y-STR data. HOMO, 67(6):471-483.

International Society of Genetic Genealogy (2015) Y-DNA Haplogroup Tree 2015, Version: 10.120, Date: 27 December 2015. Retrieved 19 November 2017, from http://www.isogg.org/tree/.

Marjanovic D, Fornarino S, Montagna S, Primorac D, Hadziselimovic R, Vidovic S, ..., Semino O (2005) The peopling of modern Bosnia-Herzegovina: Y-chromosome haplogroups in the three main ethnic groups. Ann Hum Genet, 69(6):757-763.

Mirabal S, Varljen T, Gayden T, Regueiro M, Vujovic S, Popovic D, Djuric M, Stojkovic O, Herrera RJ (2010) Human Y-chromosome short tandem repeats: A tale of acculturation and migrations as mechanisms for the diffusion of agriculture in the Balkan Peninsula. Am J Phys Anthropol, 142(3):380-390.

Primorac D, Marjanović D, Rudan P, Villems R, Underhill PA (2011) Croatian genetic heritage: Y-chromosome story. Croat Med J, 52(3):225-234.

Purps J, Siegert S, Willuweit S, Nagy M, Alves C, Salazar R, ..., Roewer L (2014) A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. Forensic Sci Int Genet, 12:12-23.

Rootsi S, Kivisild T, Benuzzi G, Help H, Bermisheva M, Kutuev I, ..., Semino O (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. Am J Hum Genet, 75(1):128-137.

Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, ..., Jobling MA (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. Am J Hum Genet, 67(6):1526-1543.

Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, …, Santachiara-Benerecetti AS (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. Am J Hum Genet, 74(5):1023-1034.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, ..., Underhill PA (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: A Y chromosome perspective. Science, 290(5494):1155-1159.

Šarac J, Šarić T, Havaš Auguštin D, Novokmet N, Vekarić N, Mustać M, ..., Rudan P (2016) Genetic heritage of Croatians in the Southeastern European gene pool - Y chromosome analysis of the Croatian continental and Island population. Am J Hum Biol, 28(6):837-845.

Varzari A, Kharkov V, Nikitin AG, Raicu F, Simonova K, Stephan W, Weiss EH, Stepanov V (2013) Paleo-Balkan and Slavic contributions to the genetic pool of Moldavians: insights from the Y chromosome. PLOS ONE, 8(1):e53731.

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, ..., Tyler-Smith, C (2003) The genetic legacy of the Mongols. Am J Hum Genet, 72(3):717-721.

Zhivotovsky LA (2001) Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow. Mol Biol Evol, 18(5):700-709.