DIGITAL ANALYSIS OF GENETIC PROXIMITY OF THE BALKAN POPULATION

Georgi Gachev, Jordan Tabov

***Abstract**: According to EUPEDIA the 12 basic haplogroups specific to the European population are:*

I1 I2*/I2a I2b R1a R1b G J2 J*/J1 E1b1b T Q N

Among the Bulgarians their distribution in percentage is:

4 20 2 17 11 5 11 3 23.5 1.5 0.5 0.5

Thus, it may be said that "the average Bulgarian" has "compound haplogroup" which consists of weighted partial inclusion of all basic haplogroups. Most common haplogroups for Bulgarian population are: E1b1b, I2/I2a, R1a, R1b and J2.*

Tabov and Koleva proposed a method for comparing "genetic closeness" between the nations: Each nation is presented as a point in 12-dimensional Euclidean space

Analysis of the "genetic" distances between the regional population groups of Europe makes it possible to associate these groups in "genetically close" clusters. One such cluster is the "Balkan".

From our analysis we conclude that the countries in the Balkan cluster are genetically closer to each other than all European countries among themselves. In fact the average distance in the Balkan cluster is 23 which is twice less than the average e-distance of 46 among European countries; and e.g. the e-genetic distance between North and South Germany is 21. Our results suggest also a comparison of the "genetic proximity" of the Balkan nations among themselves.

***Keywords**: haplogroups, European population, genetic distance, cluster analysis, Balkan population, genetic center*

This paper is partially supported by the Task 1.2.5 of the Bulgarian National Scientific Program "ICT in Science, Education and Security", funded by the Ministry of Education and Science (Contract MES DOI-205/23.11.2018).

Introduction

The DNA analysis is a well-established scientific method that is increasingly used in practice. One of its popular applications is to resolve complicated family cases where the paternity is uncertain. Recently, the attention of both the general public and professionals has been focused on its use to study the genetic relations and origins of people and to classify people according to their genes and origins.

Most often the results of such a comparative analysis are presented in the form of a "tree diagram" similar to the family trees (**Error! Reference source not found.**), or as planar points, grouped "by adjacency" (**Error! Reference source not found.**), or in tabular form with percentage data of the predominant haplogroups (**Error! Reference source not found.**).

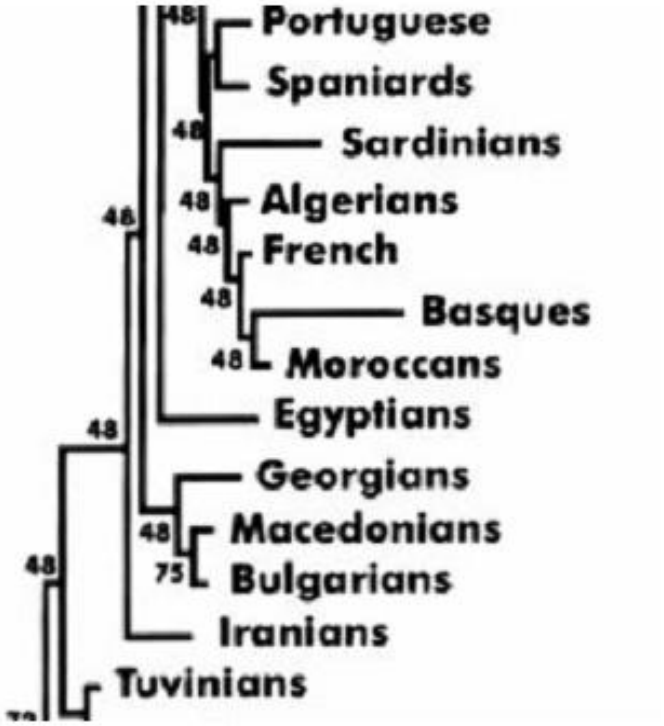

Figure 1. Part of the "Genetic Family Tree" of nations as result of a genetic research [Arnaiz-Villena et al., 2003].

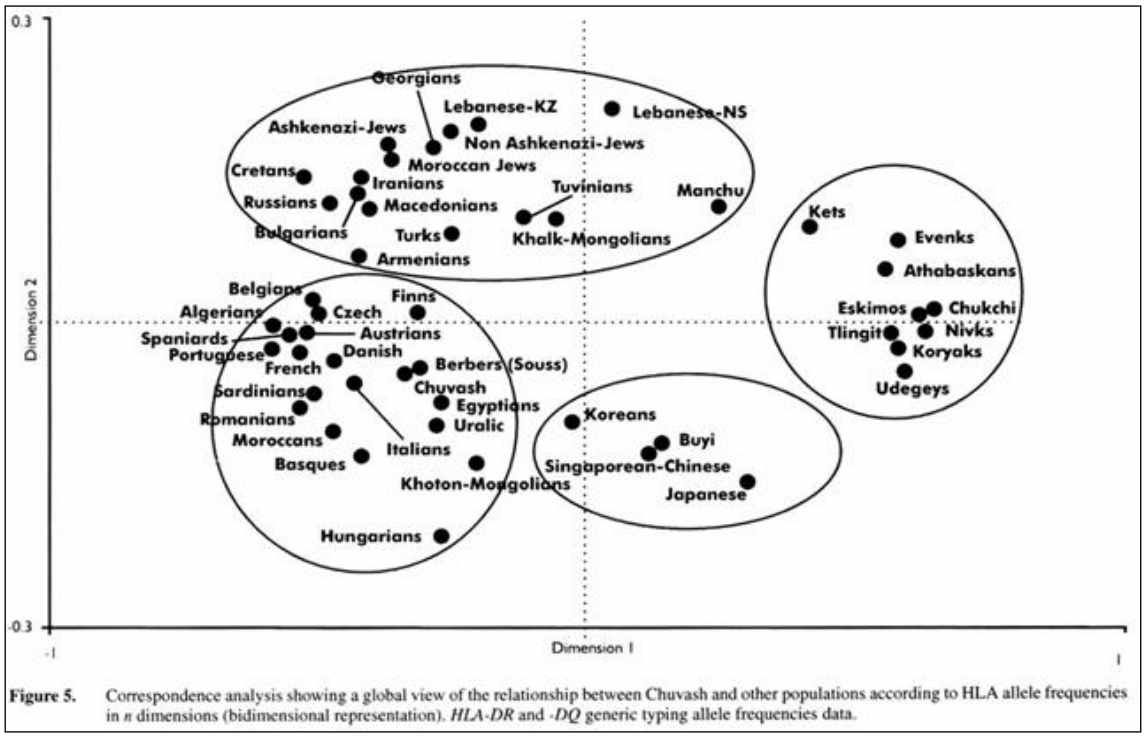

Figure 5. Correspondence analysis showing a global view of the relationship between Chuvash and other populations according to HLA allele frequencies in *n* dimensions (bidimensional representation). *HLA-DR* and *-DQ* generic typing allele frequencies data.

Figure 2. Grouping of people represented as points in the plane, according to their genetic adjacency [Arnaiz-Villena et al., 2003].

From the numerous publications in this field let us mention [Nazarova, 2009] [Balanovsky, 2012], [E., 2018], [Tabov & Sabeva-Koleva, 2018] [Chiaroni, Underhill, & Cavalli-Sforza, 2012] [Delev, 2018], [Eupedia, 2018], [Wikipedia, 2018a], [Chiaroni, Underhill, & Cavalli-Sforza, 2012], [FamilyTreeDNA, 2018], [Karmin, 2015], [Karlsson, Wallerström, Götherström, & Holmlund, 2006], [Wikipedia, 2018b] , [Trombetta B, 2011].

	У-хромозомни хаплогрупи											
Country/Region	I1	I2*/I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Bulgaria	4	20	2	17	11	5	11	3	23,5	1,5	0,5	0,5
Macedonia	3	23	1,5	13,5	12,5	4	14	2	21,5	1,5	0,5	0,5
Greece	3,5	9,5	1,5	11,5	15,5	6,5	23	3	21	4,5	0	0
Romania	4,5	26	2,5	17,5	12	5	13,5	1,5	15	0,5	0,5	0,5
Spain	1,5	4,5	1	2	69	3	8	1,5	7	2,5	0	0
Portugal	2	1,5	3	1,5	56	6,5	9,5	3	14	2,5	0,5	0
Belgium	12	3	4,5	4	61	4	4	1	5	1	0,5	0
France	8,5	3	3,5	3	58,5	5,5	6	1,5	7,5	1	0,5	0
Italy	4,5	3	2,5	4	39	9	15,5	3	13,5	2,5	0	0
Switzerland	14	1,5	8	3,5	50	7,5	3	0,5	7,5	0,5	1,5	1
South Germany	10,5	4,5	3	9,5	48,5	8	5,5	1	8	1,5	0,5	0,5
Austria	12	7	2,5	19	32	7,5	9	1	8	1	0,5	0,5
Czech Republic	11	9	4	34	22	5	6	0	6	1	1,5	0,5
Montenegro	6	29,5	1,5	7,5	9,5	2,5	9	0,5	27	0	2	1,5
Serbia	8,5	33	0,5	16	8	2	8	0,5	18	1	1,5	2
Hungary	8,5	16	2	29,5	18,5	3,5	6,5	3	8	0	0	0,5

Figure 3. Table presenting the content (in %) of twelve Y-DNA haplogoups typical for the respective European region [Tabov & Sabeva-Koleva, 2018] the data is from EUPEDIA project [Eupedia 2018].

Haplogroups of European nations

According to EUPEDIA the 12 basic haplogroups specific to the European population are:

I1 I2*/I2a I2b R1a R1b G J2 J*/J1 E1b1b T Q N

Among the Bulgarians their distribution in percentage is (Eupedia, 2018):

4 20 2 17 11 5 11 3 23.5 1.5 0.5 0.5

Thus, it may be said that "the average Bulgarian" has "compound haplogroup" which consists of weighted partial inclusion of all basic haplogroups. Most common haplogroups for Bulgarian population are: E1b1b, I2*/I2a, R1a, R1b and J2. For Serbia the corresponding percentages are:

8,5 33 0,5 16 8 2 8 0,5 18 1 1,5 2

These complex data show that the problem of genetic relation among nations is more complicated than the same problem but for individuals.

Genetic Distance

Tabov and Koleva (2018) propose a method for comparing "genetic closeness" between the nations: Each nation is presented as a point in 12-dimensional Euclidean space. Percentage content of each haplogroup specific to the nation is one of the 12 coordinates of that point. The distance between the points is defined as **genetic distance**:

$$Distance(CountryA, CountryB) = \sqrt{\sum_{i=1}^{12} (CountryA_i - CountryB_i)^2} \quad (1)$$

This definition is **European specific**: for population in which the sum of these 12 Y-chromosome haplogroups is less than 95%, we obtain greater deviations compared to the values of most Y- chromosome haplogroups, therefore when we apply this definition for genetic distance we ignore them.

The genetic distances between European nations, calculated by formula (1), further will be called "e-genetic" (i.e. European genetic).

Two-dimensional visualization of genetic distances for four countries is not always appropriate. The reason is that the respective four points may not be in the same plane so the planar projection could be skewed. This may lead to speculations when such results are interpreted and discussed – see **Error! Reference source not found.** and **Error! Reference source not found.**.

Although the points of Bulgaria, Macedonia, Greece and Serbia are almost in the same plane so the planar view is acceptable approximation. The projection is shown in **Figure 4**.

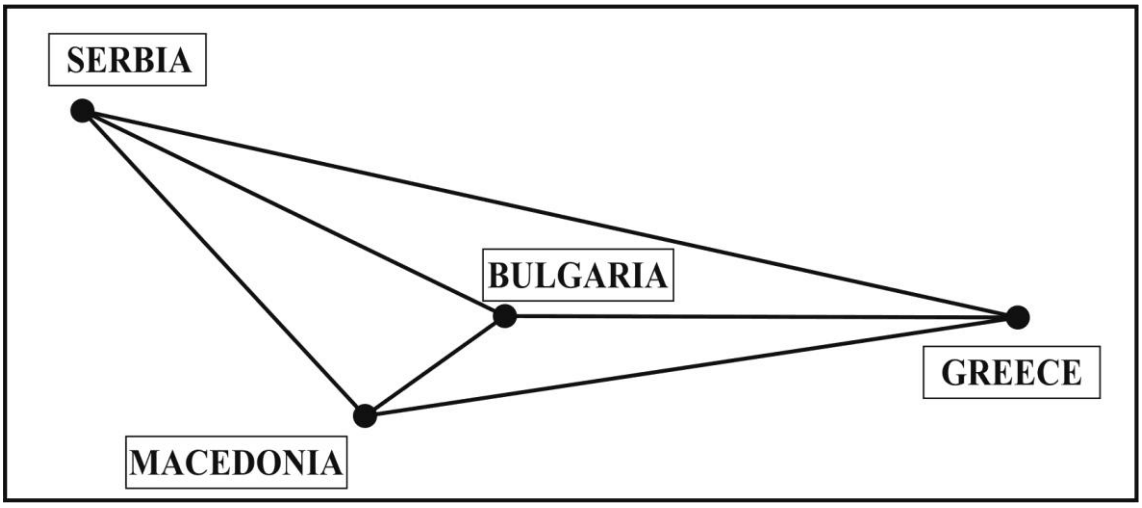

Figure 4. Visualization of genetic similarity of (Bulgaria, Macedonia, Serbia and Greece).

Diffuse distribution of the "European genes"

In the paper [Tabov & Sabeva-Koleva, 2018], the following hypothesis was offered:

Hypothesis: The change of ratio of haplogroups in Europe is a diffusion type process: it may be conjectured that the main part of this process has been played by "exchange of haplogroups" between neighboring nations. Most probably large scale migration has played less influence in this process. The results we are going to present below provide further arguments and clarification of this hypothesis.

In support of this hypothesis the following arguments are used in [Tabov et al., 2019]:

1- Genetic-geographical correlation: The first argument is connected to the correlation when "genetic distance" is compared to "geographical distance". Since such a correlation exists this leads to the conclusion for significant diffusion type exchange of genes. Indeed, for all 38 European nations (we exclude small ones by territory and population) the correlation has coefficient of 0.61.

2- The second argument supporting the diffusion type distribution of male genes in Europe is the qualitative distribution (in percentages) of each of the 12 European haplogroups.

E.g. the percentages of haplogroup I2*/I2a are higher in Bosnia and Herzegovina; they are significant in neighboring countries – Serbia, Croatia, Montenegro, Rumania, Slovenia, Bulgaria and Moldova; they are to be found in most countries, neighbors of the above – Hungary, Slovakia, Ukraine, Belarus, Greece [Wikipedia 2018a].

Cluster analysis of e-genetic distances between the European people

Analysis of the "e-genetic" distances between the regional population groups of Europe makes it possible to associate these groups in "genetically close" clusters. One such cluster is the "Balkan", formed by the peoples of the following 8 Balkan countries: Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia; the table of the e-genetic distances between them is presented in **Figure 5**.

	Bulgaria	Croatia	Greece	Macedonia	Montenegro	Romania	Serbia
Bosnia-Herzegovina	39	22	53	37	32	33	25
Bulgaria		24	18	6	15	11	16
Croatia			38	23	25	16	13
Greece				17	27	22	31
Macedonia					13	9	15
Montenegro						17	13
Romania							12

Figure 5. Table of e-genetic distances between 8 Balkan countries: Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia.

Specific aspect of Slovenia and Albania

Analysis of the data for another two Balkan countries, Slovenia and Albania, shows that:

1 Slovenia is closer to another cluster of countries to which belong Slovakia and Ukraine. The e-genetic distances from Slovenia to Slovakia and Ukraine are respectively 8 and 14, which for example is less than the distances to Bulgaria – 31 and Greece - 40.

2 The result for Albania calculated by formula (1) is not reliable as the sum of percentages of all 12 European haplogroups is about 92%; therefore Albania is not included in the table in **Figure 4**. To have better approximation and improved view of Albania the "13th coordinate" is introduced. These are missed percentage values to sum of 100%. Thus for each country is added 13th "haplogroup" of unknown type just to fill the gap. Also this extends the formula (1) with another, 13^{th} coordinate; the distance modified this way will be called b-genetic (i.e. Balkan genetic) distance.

Thus, the countries in the Balkan cluster become 9: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia; the table of the b-genetic distances between countries in the cluster is presented in **Figure 6**.

"Genetic Center" of the Balkan population

Virtual "genetic center" could be defined for the 9 countries from the "Balkan genetic cluster". This "genetic center" is geometric median of the cluster points, i.e the point which minimizes the sum of distances to all other 9 points representing corresponding countries. The B-genetic distance of a country to the "genetic center" characterizes its stronger or weaker belonging to the cluster.

The table with these b-genetic distances to the Balkan Genetic Center and b-genetic distances between the countries is presented in **Figure 7**.

	Bosnia-Herzegovina	Bulgaria	Croatia	Greece	Macedonia	Montenegro	Romania	Serbia
Albania	51	18	38	13	16	23	23	30
Bosnia-Herzegovina		39	22	53	37	32	33	25
Bulgaria			24	18	6	15	11	16
Croatia				38	23	25	16	13
Greece					17	27	22	31
Macedonia						13	9	15
Montenegro							17	14
Romania								12

Figure 6. Table of the b-genetic distances between countries in the Balkan cluster: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia

	Albania	Bosnia-Herzegovina	Bulgaria	Croatia	Greece	Macedonia	Montenegro	Romania	Serbia
Balkans genetic Center	21	31	9	18	22	6	11	6	9
Albania		51	18	38	13	16	23	23	30
Bosnia-Herzegovina			39	22	53	37	32	33	25
Bulgaria				24	18	6	15	11	16
Croatia					38	23	25	16	13
Greece						17	27	22	31
Macedonia							13	9	15
Montenegro								17	14
Romania									12

Figure 7. Table of the b-genetic distances between 9 Balkan countries: Albania, Bosnia-Herzegovina, Bulgaria, Croatia, Greece, Macedonia, Montenegro, Romania and Serbia and their b-genetic distances to the "Balkan genetic center".

The table in **Figure 7** shows that closest countries to the Balkan Genetic Center are: Macedonia, Romania, Bulgaria and Serbia (with b-genetic distances to it, respectively 6, 6, 9, 9). Of these four, Bulgaria and Serbia are adjacent to each of the other two - Macedonia and Romania. It may be said that the population of these four countries together form the "closest b-genetic neighborhood" of the Balkan Genetic Center.

Conclusion

From the above analysis we may conclude that the countries in the Balkan cluster are genetically closer to each other than all European countries among themselves. In fact the average distance in the Balkan cluster is 23 which is twice less than the average e-distance of 46 among European countries; and e.g. the e-genetic distance between North and South Germany is 21.

The tight genetic relations are possible motivation for further improvement of the collaboration on the Balkans. The increasing of the number and the development of new genetic research methods will allow for more accurate comparisons and conclusions.

Appendix

Percentage of the quantity of haplogroups measured up to 20.11.2018 in population of 39 European nations/countries [Eupedia 2018]:

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Albania	2	12	1.5	9	16	1.5	19.5	2	27.5	1	0	0
Austria	12	7	2.5	19	32	7.5	9	1	8	1	0.5	0.5
Belarus	5.5	17.5	1	51	5.5	1.5	2.5	1	4	0	0	10
Belgium	12	3	4.5	4	61	4	4	1	5	1	0.5	0
Bosnia-Herzegovina	3	55.5	0	15	3.5	1.5	4	0.5	12	1	2	2
Bulgaria	4	20	2	17	11	5	11	3	23.5	1.5	0.5	0.5
Croatia	5.5	37	1	24	8.5	2.5	6	1	10	0.5	1	0.5
Czech Republic	11	9	4	34	22	5	6	0	6	1	1.5	0.5

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Denmark	34	2	5.5	15	33	2.5	3	0	2.5	0	1	1
England	14	2.5	4.5	4.5	67	1.5	3.5	0	2	0.5	0.5	0
Estonia	15	3	0.5	32	8	0	1	0	2.5	3.5	0.5	34
Finland	28	0	0.5	5	3.5	0	0	0	0.5	0	0	61.5
France	8.5	3	3.5	3	58.5	5.5	6	1.5	7.5	1	0.5	0
Germany	16	1.5	4.5	16	44.5	5	4.5	0	5.5	1	0.5	1
Greece	3.5	9.5	1.5	11.5	15.5	6.5	23	3	21	4.5	0	0
Hungary	8.5	16	2	29.5	18.5	3.5	6.5	3	8	0	0	0.5
Iceland	29	0	4	23	42	0	0	0	0	0	1	1
Ireland	6	1	5	2.5	81	1	1	0	2	0	0	0
Italy	4.5	3	2.5	4	39	9	15.5	3	13.5	2.5	0	0
Latvia	6	1	1	40	12	0	0.5	0	0.5	0.5	0.5	38

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Lithuania	6	6	1	38	5	0	0	0	1	0.5	0.5	42
Macedonia	3	23	1.5	13.5	12.5	4	14	2	21.5	1.5	0.5	0.5
Malta	1	10	1	3.5	32.5	6.5	21	8	9	4.5	1	0
Moldova	5	21	3	30.5	16	1	4	4	13	1	0	1.5
Montenegro	6	29.5	1.5	7.5	9.5	2.5	9	0.5	27	0	2	1.5
Netherlands	16.5	1	6.5	4	49	4.5	3.5	0.5	3.5	1	0	0
Norway	31.5	0	4.5	25.5	32	1	0.5	0	1	0	1	2.5
Poland	8.5	5.5	2	57.5	12.5	1.5	2.5	0	3.5	0.5	0.5	4
Portugal	2	1.5	3	1.5	56	6.5	9.5	3	14	2.5	0.5	0
Romania	4.5	26	2.5	17.5	12	5	13.5	1.5	15	0.5	0.5	0.5
Russia	5	10.5	0	46	6	1	3	0	2.5	1.5	1.5	23
Serbia	8.5	33	0.5	16	8	2	8	0.5	18	1	1.5	2

	I1	I2*/ I2a	I2b	R1a	R1b	G	J2	J*/J1	E1b1b	T	Q	N
Slovakia	6.5	16	1.5	41.5	14.5	4	2	1	6.5	0.5	0.5	3
Slovenia	9	20.5	1.5	38	18	1.5	2.5	0	5	1	0	0
Spain	1.5	4.5	1	2	69	3	8	1.5	7	2.5	0	0
Sweden	37	1.5	3.5	16	21.5	1	2.5	0	3	0	2.5	7
Switzerland	14	1.5	8	3.5	50	7.5	3	0.5	7.5	0.5	1.5	1

Bibliography

[Arnaiz-Villena et al., 2003] A. Arnaiz-Villena, J. Martinez-Laso, J. Moscoso, G. Livshits, J. Zamora, E. Gomez- Casado. HLA Genes in the Chuvashian Population from European Russia: Admixture of Central. European and Mediterranean Populations. *Human Biology, 75*(3), 375-392.

[Balanovsky, 2012] O. Balanovsky.Variability of the gene pool in space and time: the synthesis of data on the mitochondrial DNA genogeography and Y-chromosome. In B. O., *Abstract of dissertation for the degree of Doctor of Biological Sciences.* Moscow.

[Chiaroni et al., 2012] J. Chiaroni, P. Underhill, & L.Cavalli-Sforza. L Y chromosome diversity, human expansion, drift, and cultural evolution. *Proceedings of the National Academy of Sciences of the United States of America.*

[Delev, 2018] E. Delev. Population genetics of the Bulgarians, founders of the European civilization. *BOMI*.

[Eupedia, 2018] *Distribution of European Y-chromosome DNA (Y-DNA) haplogroups by country in percent.* Retrieved 11 20, 2018, from EUPEDIA: https://www.eupedia.com/europe/european_y-dna_haplogroups.shtml

[FamilyTreeDNA, 2012] *Understanding Haplogroups (2012). Understanding Haplogroups: How are the haplogroups named? Family Tree DNA. Archived from the original on June 21, 2012*. Retrieved 11 11, 2018, from https://web.archive.org/web/20120621052330/ http:/www.familytreedna.com/understanding-haplogroups.aspx

[Karlsson et al., 2006] A. Karlsson, T. Wallerström, A. Götherström, & G. Holmlund. Y-chromosome diversity in Sweden - A long-time perspective. *European Journal of Human Genetics,* 963-970.

[Karmin et al. 2015] M. Karmin et al. A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research*, 459-466.

[Nazarova, 2009] F. Nazarova. The closeness of the Finnish Ugric, Slav and Germanic populations according to anthropological and genetic data. *Journal of Evolutionary Biology Research 2009, 1*(1), 18-26.

[Tabov & Sabeva-Koleva, 2018] J. Tabov & N. Sabeva-Koleva. A method for assessing the genetic proximity of neighboring peoples. *Bulgarian Historical Readings 2018*.

[Tabov et al., 2019]. J. Tabov, G. Gachev, N. Sabeva-Koleva. On the genetic ties between European nations. *Khimiya, 28*, 615-627.

[Trombetta, 2011]. B. Trombetta. New Topology of the Human Y Chromosome Haplogroup E1b1 (E-P2) Revealed through the Use of Newly Characterized Binary Polymorphisms. *PLoS ONE*.

[Wikipedia, 2018a] *Haplogroup I2*/I2a (Y-DNA). From Wikipedia.* Retrieved 11 11, 2018, from https://ru.wikipedia.org/w/index.php?title=Гаплогруппа_E1b1b_(Y-ДНК)&oldid=97696115

[Wikipedia, 2018b]. *List of Y-chromosome in the population of the world.* Wikipedia. Retrieved 11 11, 2018, from https://en.wikipedia.org/wiki/List_of_Y-chromosome_haplogroups_in_populations_of_the_world

Authors' Information

Georgi Gachev *- Assistant researcher; Institute of Mathematics and Informatics, BAS, Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria;*

Major Fields of Scientific Research: Mathematics Education, Modeling, Software dvelopment

Jordan Tabov *– Professor (retired); Institute of Mathematics and Informatics, BAS, Acad. G.Bonthev St., bl.8, Sofia-1113, Bulgaria; tabov@math.bas.bg*

Major Fields of Scientific Research: Applications of mathrmatics and informatics in the Humanities, Didactics of mathematics and informatics