OVERVIEW OF EUROPEAN POPULATION CLUSTERING BASED ON 23 Y-STR LOCI

Serkan DOGAN1*, Adna AŠIĆ1, Sanida BULJUBAŠIĆ1, Larisa BEŠIĆ1, Monia AVDIĆ1, Elma FERIĆ1, Mirsada HUKIĆ1, Yusuf TURAN1 and Damir MARJANOVIĆ1

1Department of Genetics and Bioengineering, Faculty of Engineering and IT, International Burch University, Sarajevo, Bosnia and Herzegovina


Short tandem repeats (STRs) located on the Y-chromosome are a useful tool for various scientific fields, such as forensic investigation, but also for the investigation of population structure and molecular history. In this study, population data based on 23 Y-STR loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, GATAH4, DYS481, DYS533, DYS549, DYS570, DYS576, and DYS643) from 23 European human populations were compared. All haplotype data for this research were gathered from previously published articles. Arlequin v3.5.1.2, POPTREE2, and MEGA 5.1 software packages were used for the calculation of allelic frequencies and genetic distance, and the construction of the European, as well as worldwide phylogenetic trees. Obtained results indicate a formation of several distinct sub-clusters within European population cluster. Observed sub-clusters were mostly recognized within geographically closer populations, meaning that neighboring populations were a part of the same sub-cluster in most of the cases. Compared with the previously published results obtained using autosomal STR markers, a significant level of concordance was detected. However, it seems that Y-STRs analyzed in this study are more informative since they enabled regional clustering in addition to continental clustering. Also, the use of a larger number of loci yielded clustering that is more specific than what has been calculated to date. Finally, it can be concluded that this study has shown that the application of a larger number of loci enables the more detailed insight into the relationships between European populations, compared to what has been published before.

Key words: Y-chromosome, Y-STRs, PowerPlex Y23, European populations, population clustering, phylogenetic tree, population genetics

Corresponding author: Serkan Dogan, International Burch University, Francuske revolucije bb, Ilidža, 71210 Sarajevo, Bosnia and Herzegovina. Tel: +387 33 944 400 Fax: +387 33 944 500 E-mail: serkan.dogan@ibu.edu.ba
INTRODUCTION

Satellite DNA is a type of a DNA length polymorphism which is characterized by frequently repeating DNA sequences. If these repeat sequences or units are adjacent to one another, they are termed tandem repeats (BUTLER, 2011; HOUCK and SIEGEL, 2010). Furthermore, short repetitive units that are 2-7 bp long are commonly referred to as microsatellites (HOUCK and SIEGEL, 2010), simple sequence repeats (SSRs) (GUNN, 2006), or short tandem repeats (STRs) (HOUCK and SIEGEL, 2010). STRs are found on all chromosomes, including the X and Y chromosomes, and are an exceptionally useful forensic tool since they vary from individual to individual (BUTLER, 2005). Y-chromosome STR molecular markers are repetitive units that are inherited from father to son. Y-chromosome markers are usually chosen because of their high level of diversity that enables simple genotyping (MARJANOVIĆ and PRIMORAC, 2013). The genetic information obtained from lineage markers such as Y-STR is referred to as a haplotype (BUTLER, 2012). Y-STRs have a wide range of applications in forensics. On the other hand, they are widely accepted genetic markers for tracing of human history and migrations as they are passed only from father to son (BUTLER, 2011). The phylogenetic analysis of Y-STR data enables the generation of large phylogenetic trees by sorting various Y-chromosomes by lineage and relatedness (UNDERHILL and KIVISILD, 2007). Those studies are usually conducted in order to identify the possible origin as well as to propose the migration pattern of different human populations, and accordingly to indicate the relatedness among examined groups (SHI et al., 2008; ZHANG et al., 2007).

In this paper, population data on 23 Y-STR loci from 23 European populations were compared in order to gather previously published results in one descriptive framework and to compare them with the results which were previously obtained using autosomal STR markers. Additionally, representative populations from the other parts of the world were analyzed in order to examine their possible connection with the European populations and to place the cluster of European populations within a global picture. Furthermore, the aims of this study include additional clarification of the previously obtained results by increasing the number of analyzed loci, and thereby a demonstration of the significance of Y-STR markers in population studies.

MATERIALS AND METHODS

Study populations

The European populations compared in the present study include the following: Belgian, Bosnian and Herzegovinian, British, Croatian, Czech, Dutch, English (Southern and Central England), Estonian, Finnish, German, Hungarian, Irish, Italian (Ravenna and Brescia), Latvian, Lithuanian, Macedonian, Northern Portuguese, Polish, Slovenian, Spanish, and Welsh. In order to compare European populations with the populations worldwide, the following populations were analyzed additionally: Argentinian, Athapaskan (Central Alaska), Beninese (North Benin), Brazilian, Chinese, Filipinos, Greek, Han (Singapore), Indian (Southern India), Iraqi, Jamaican, Japanese, Kenyan, Lebanese, Mestizo (Bolivia), Mestizo (Costa Rica), Nigerian, Notch, Peruvian, South African, South Korean, Turkish, USA (African American, Asian American, European American, Hispanic American), and Zimbabwean (Figure 1).
Figure 1: World map showing geographical positions of populations analyzed in this study.

The Y-STR data based on 23 loci were obtained from previously published articles, and the number of samples analyzed in each population ranged from 43 to 100. Chromosomes were typed for 23 Y-STR loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, GATAH4, DYS481, DYS533, DYS549, DYS570, DYS576, and DYS643) using the PowerPlex Y23 System (PPY23, Promega Corporation, Madison, WI, USA).

Statistical analyses

The allele frequencies of 21 Y-STR loci were calculated using Arlequin v3.5.1.2, while those of DYS385a/b were estimated using the direct counting method. Phylogenetic comparison of the allelic frequencies over 23 Y-STR loci in the tested populations was performed using the POPTREE2 and MEGA5.21. The genetic distances among populations were calculated using POPTREE2, whereby the lowest distance for each population was determined.

RESULTS

The results of this study considering sub-clustering of European populations based on 23 Y-STR loci are illustrated in the circular phylogenetic tree (Figure 2). The formation of several distinct sub-clusters, all of them belonging to the European continent, can be observed in the tree. This phylogenetic tree highlighted the role of geographical proximity on genetic similarity between populations as it had several sub-clusters, namely the Balkans (full circles), North and Central Europe (hollow squares), Western Europe (full squares), and Southern Europe (hollow triangles). Out of all these sub-clusters, populations from the Balkans seem to be most distant from all other European populations, since they show statistically significant differences on a few loci. At the same time, Balkan populations exhibit almost negligible genetic distances among each other without statistically significant differences on any of the loci tested.
European populations were also compared with the other worldwide populations in order to examine their relations in terms of geographical proximity. The results are presented in the rectangular phylogenetic tree showing 50 global populations (Figure 3). The top of the tree is marked by the cluster of populations from Africa (hollow circles), and, as expected, all, including African Americans living in the USA, were grouped within a single cluster. The Jamaican population was found in the same cluster as well.

On the other hand, Caucasian population from the USA is sharing a high degree of genetic similarity with the European populations, being grouped in the sub-cluster corresponding to the Western Europe. Also, populations from the Latin America (full triangles) clustered closely to the
European populations. The end of the rectangular phylogenetic tree is marked by grouping of Asian populations into Near East (hollow diamonds) and Far East (full diamonds) sub-clusters, the latter of which encompasses Asians living in the USA.

Figure 3: Rectangular phylogenetic tree based on 23 Y-STRs in compared worldwide populations.
DISCUSSION

Current literature lacks continental sub-clustering population studies and phylogenetic tree generation based on Y-STRs, especially when it comes to 23 Y-STR loci. Therefore, current results have been compared to the population studies based on both autosomal and Y-STRs. According to the results of a study conducted among European populations based on Y-STR haplotypes, two sub-clusters, namely Eastern European and Western European, were identified (ROEWER et al., 2005). The authors of this study used seven Y-STR loci and their results imply that the importance of these genetic markers in continental sub-clustering was recognized 10 years ago. However, our results are based on 23 Y-STRs and are, thus, giving more informative results with four different sub-clusters identified in the European continent.

On the global point of view, the results of this research are easily explainable by historical and migration circumstances, as well as by the rule that geographical proximity of populations causes high level of genetic similarity between them. In the first cluster identified in the global phylogenetic tree, strong clustering of populations is related to their common origin, the African continent. Additionally, geographical proximity of all populations, with the exception of Jamaica and USA, is the second factor causing such a strong clustering. It can also be concluded that African populations do not share a high degree of genetic similarity with European populations, since they were grouped as a completely separate cluster.

Caucasian American clustered with the European populations. This observation is expected since the origins of Caucasian population on the American continent are directly related to migrations during the colonization time. Mainly people from the United Kingdom were moving to the American continent, which is also confirmed in our phylogenetic tree. When it comes to the populations from the Latin America, it is obvious that they share a high degree of genetic similarity with the populations from the Southern Europe which is why they are part of the neighboring clusters. This can also be explained by the migration of people from the Southern Europe to the countries of Latin America during the colonization time. The same is valid for the Hispanic American population. Again, sub-clustering of the Asian population cluster reflects geographical position of the countries studied in this research with a high degree of precision and accuracy. Considering their connection with the European populations, it is easily observable that the populations from the Far East do not share genetic similarities with the European populations, while those from the Near East tend to cluster closely to the European populations. This is especially true in the case of Turkey, which is a Eurasian population, positioned at the border of the two continents and has an identical position in the phylogenetic tree presented here.

Background of the results obtained in this study is also in accordance with the results obtained within several studies conducted on autosomal STRs. A study completed on Asian populations based on 19 autosomal STR loci showed clustering patterns within Asian populations similar to those present in this study (LIU et al., 2013). Also, LIU et al. (2013) generated two clusters by comparing worldwide populations, namely one consisting of populations from Europe, Africa, America, Oceania, and Latin America, and another containing populations living in Asia (Chinese, Japanese, Korean, and Indian). These results are in accordance with the present study in a way that Asian populations are organized as a separate cluster. On the other hand, the present study showed the possibility to separate African and European (together with Latin American) populations into two separate clusters, thus showing once again improved quality of population studies if Y-STRs are used.
As a result of autosomal STR loci analyses in two different studies (DOGAN et al., 2014a; HONG et al., 2013), three overall clusters were formed as a result of autosomal STR loci analyses, namely the European, Asian and African, which is in accordance with the present study. Discrepancies between autosomal STR and Y-STR analyses are highlighted via formations of sub-clusters in the Y-STR analysis. This is confirmed by the results of the present study, since the European cluster is easily divided into four distinct groups represented as four branches of the phylogenetic tree, while the Asian population cluster consists of two sub-clusters. The results of two aforementioned population studies were not clearly detected by autosomal STR analysis and the formation of sub-clusters within the three main continental clusters was not observed. Given the aforementioned clustering trends evident in both phylogenetic trees, it can be concluded that clusters were indeed formed as a consequence of geographical proximity that triggered a mixing of gene pools, which in turn resulted in the formation of neighboring populations that exhibit strong genetic similarities.

Overall, the current study effectively highlights that Y-STRs could be more informative tool in structural population studies as they are more informative than autosomal STRs because they not only enable continental clustering, but are also a great tool for additional regional studies as well. Formation of four sub-clusters of European populations is once again proving the great potential of Y-chromosomal markers in the wide spectrum of genetic analyses.

REFERENCES


GRUPISANJE EVROPSKIH POPULACIJA BAZIRANO NA 23 Y-STR LOKUSU

Serkan DOGAN1*, Adna AŠić1, Sanida BULJUBAŠIĆ1, Larisa BEŠić1, Monia AVDIĆ1, Elma FERIĆ1, Mirsada HUKIĆ1, Yusuf TURAN1 i Damir MARJANOVIĆ1

1Odsjek za genetiku i bioinženjering, Fakultet za inženjering i informacijske tehnologije, Internacionalni Burch Univerzitet, Sarajevo, Bosna i Hercegovina

Izvod

Short tandem repeat (STR) lokusi na Y hromosomu su veoma korisni za izučavanje populacija i istorijska istraživanja. U ovom radu, 23 Y-STR lokusa (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS385a/b, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, GATA-H4, DYS481, DYS533, DYS549, DYS570, DYS576 i DYS643) su upoređena na uzorku od 23 evropske populacije. Alelske frekvencije su izračunate koristeći Arlequin v3.5.1.2, dok su informacije o haplotipovima za ove proračune skupljene iz prethodno objavljenih radova. POPTREE2 i MEGA 5.1 softveri su korišteni za generisanje filogenetskog stabla i računanje genetičkih distanci. Rezultati su prikazani formiranje četiri odvojene podgrupe populacija unutar evropskog kontinenta, a to grupisanje je uglavnom uslovljeno geografskom blizinom, što je rezultiralo time da se susedne populacije nađu u istoj grupi. Dodatno poredjenje sa svjetskim populacijama je moglo da se odrede sličnosti i razlike između evropskih i populacija sa drugih kontinenata, ali i da se Evropa kao kontinent smeti u globalnu sliku nastalu upotrebom populacione genetike. Iako su rezultati ovog istraživanja u saglasnosti sa rezultatima prethodno objavljenih istraživanja autozomalnih STR lokusa, Y-STR lokusi iz ovog istraživanja su, uz kontinentalno, ponudili regionalno grupisanje populacija. Konačno, upotreba velikog broja lokusa u ovom istraživanju je dala grupisanje koje je mnogo specifičnije od onog ponuđenog u dosadašnjim istraživanjima.

Odobreno 05. IX. 2015.