 Int J Hum Genet, 15(4): 157-171 (2015)

Male Individualization Based on Y-Chromosomal Short Tandem Repeats: A Comparative Information Theoretical Analysis of 16 Y-STR Loci in Central Anatolia and Iraqi Populations

Aysun Baransel Isir[1], Abdulmuttalip Ozkorkmaz[2], Cesur Baransel[3], Ebru Gokalp Ozkorkmaz[4] and Sacide Pehlivan[5]

[1]Gaziantep University, Faculty of Medicine, Department of Forensic Medicine, Gaziantep, Turkey
[2]Ege University, Faculty of Science, Department of Biology, Izmir, Turkey
[3]Department of Computer Engineering, University of Turkish Aeronautical Association, Ankara, Turkey
[4]Yildirim Beyazit University, Faculty of Health, Ankara, Turkey
[5]Istanbul University, Faculty of Medicine, Department of Medical Biology and Genetic, Istanbul, Turkey

KEYWORDS Central Anatolia and Iraqi Populations. Entropy. Pointwise Mutual Information. Population Genetics. Y-STR Polymorphisms.

ABSTRACT The aim of this study is to investigate the discrimination capacity of 16 Y-Chromosomal Short Tandem Repeat markers (Y-STRs) based on their joint entropy for the purpose of male individualization on samples taken from Central Anatolia and Iraqi Populations. The Y-chromosome polymorphism of sixteen STR loci (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, Y-GATA H4) were studied. Genomic DNA was extracted from buccal swabs using the QIAamp Mini kit and was co-amplified by using Applied Biosystems AmpF*l*STR® Yfiler™ PCR Amplification Kit. The Iraqi data set was readily available in the literature which is based on blood samples randomly collected from 100 healthy unrelated males living in middle or south of Iraq. The researchers observed 106 unique haplotypes in Central Anatolia data set. The genetic diversity values across the 16 Y-STR loci ranged from 0.564 (DYS391) to 0.876 (DYS385a/b). The complete male individualization with only 16 Y-STR markers in a genetically diverse local population is possible. In this study, haplotype diversity was 1.0 and discrimination capacity was 100 percent. The high discrimination capacity of the 16 Y-STR markers makes them valuable for male individualization for forensic purposes in Central Anatolia Region of Turkey. The researchers also show that, the *pointwise mutual information* and the *joint entropy* between allele pairs measure the discrimination power of markers more accurately than individual genetic diversity values and provide a better insight into the interaction between the genetic profile of the population and the given Y-STR marker set.

INTRODUCTION

Human genome is composed of *coding* and *noncoding* parts. The coding regions of the DNA contain information related to protein synthesis and genetic variation in these regions is very limited. On the other hand, the mutations in the noncoding regions are usually kept intact and transmitted to the offspring, giving rise to regions which are very informative about the genetic profile of the individual. About 30 percent of the noncoding DNA consists of repetitive sequences which manifest themselves either in tandem or as interspersed elements. The current forensic typing methods are largely based on genetic loci with tandem repetitive sequences. These sequences are further classified into two groups depending on the length of the core repeat unit, namely STR (Short Tandem Repeat) and VNTR (Variable Number of Tandem Repeats) sequences (Carracedo et al. 2005).

The so-called Y–STRs (Y–chromosomal STRs) exist only in human Y chromosome and, by definition, are male-specific. Barring mutations, they remain stable in a given paternal lineage over many generations, therefore are used as powerful genetic indicators in paternity testing, evolutionary and anthropological studies and molecular diagnostics (Asicioglu et al. 2002; Berger et al. 2003; Brion et al. 2003; Butler 2003;

[*]Address for correspondence:
Aysun Baransel Isir
University of Gaziantep, Faculty of Medicine, Department of Forensic Medicine, 27100 Gaziantep, Turkey
Telephone: +903423606060
Fax: +903423385000
E-mail: aybaransel@yahoo.com

Mizuno et al. 2008; Kareem et al. 2015). However, there is a continuing controversy surrounding their use for forensic purposes. The controversy stems from the fact that, Y-STRs are much less polymorphic compared to autosomal STRs due to lack of recombination, therefore they deemed not sufficiently powerful for the identification of the individual. On the other hand, Y-STRs possess unique properties which make them indispensable tools in some situations, such as rape cases where a mixture of male and female DNA has to be analyzed and the problems of mixed stain interpretation must be avoided (Asicioglu et al. 2003). Consequently, research has been carried out to increase the discrimination power of Y-STRs towards the goal of so-called "male individualization". The early efforts in this regard involved the establishment of population-specific genetic databases for the Y-STR haplotypes. Another line of research has proceeded to discover new Y-STR markers with more discriminative power.

Objective

The aim of this study is to investigate the discrimination capacity of 16-Y-STRs based on their joint entropy for the purpose of male individualization on samples taken from Central Anatolia Population. In this paper, the researchers present a case study to contribute to the published results involving specifically Southern Anatolia population. The study presented in this paper is unique to the best of the researchers' knowledge in the sense that it demonstrates the possibility of complete male individualization with only 16 Y-STR markers in a genetically diverse local population. The researchers evaluate and compare the genetic diversity in their Y-STR data in terms information theoretical concepts of *entropy* and *pointwise mutual information* between the alleles. A general introduction to information theory in molecular biology is provided in (Adami 2004).

Y-STR Markers in Forensics

In forensic cases, we are interested in the probability that a forensic sample will match one sample drawn at random from the population. If we can assume that all loci used in the DNA fingerprinting are independent, we can calculate the match probability on the basis of allele frequency data and Hardy-Weinberg expectation for random union of alleles using the so-called product rule (Waits et al. 2001). The complexity of the Y-STR statistical data analysis is mainly due to the fact that this product rule is simply not valid for Y-STR data, and thus, Y-STR results have to be combined into a haplotype for estimating the rarity of a particular haplotype. For estimating the level of significance of the evidence, it is necessary to understand the frequency of occurrence of observed haplotypes within relevant subpopulations (Walsh 2013). Y-STR haplotypes tend to exhibit similar patterns of population structure which coincide well with groupings according to prior information on geographical and ethnic origin (Purps et al. 2014). Consequently, discrimination capacity and genetic diversity of haplotypes can differ significantly across subpopulations, making a decisive impact on the interpretation of the Y-STR haplotype analysis results. When the evidential Y-STR haplotype cannot be excluded, statistics derived from population data would usually be applied for estimating the likelihood of a random match (Butler 2005). In this regard, population surveys and Y-chromosome haplotype reference databases become very important (Carracedo 2015).

In practice, queries against the YHRD database may return a single match or no match at all. In fact, as more Y-STR loci are used to describe a haplotype, the capability to discriminate more effectively amongst the haplotypes will improve and increasingly higher number of searches will result in no matches. In that case, only some reasonable estimate of the rarity of the given haplotype is possible. The upper bound on the confidence interval of such an estimate is $(1-\alpha^{1/N})$ where the confidence coefficient α is 0.05 for a 95 percent confidence interval and N is the number of individuals in the database. However, this formula does not take the population substructure into account and, thus, its use is problematic. A satisfactory solution to the problem has been elusive over the years (Budowle et al. 2007; Ballantyne et al. 2014). A recent study, based on the markers of the PowerPlex®Y23 system, states that the pattern of interdependence between the allelic states of PPY23 markers is too complex to allow decomposition of the marker set into (quasi) independent subsets and calls for further work on the development of sensible and efficient methods

for match probability calculation (Caliebe et al. 2015). Currently, YHRD offers *Discrete Laplace* and *Kappa* methods for estimations, the former is being only available for adequately sized metapopulations and datasets (Willuweit et al. 2015).

Since the selection of the core Y-STR loci by SWGDAM (Scientific Working Group on DNA Analysis Methods) in 2003, commercial Y-STR kits are continuously augmented by additional loci beyond SWGDAM recommendations. In the same year, the full Y–chromosome sequence became available in which over 400 Y-STR loci identified (Hanson et al. 2006). The widely used sets of Y–STRs have low to midrange mutation rates and they show reduced diversity in certain populations that have experienced population bottlenecks or sex-biased migration, such as Finns, Xhosa, and Polynesians. More recently, a set of 13 Y–STRs characterized by high mutation rates (1×10^{-2} or higher), called rapidly mutating (RM) Y-STRs, have been introduced. The increase in resolution provided by the (RM) Y–STR loci is further discussed in (Ballantyne et al. 2014).

Statistical Methods for Y-STR Data Analysis

A widely used approach for statistical analysis of a sample set of Y-STR haplotypes can be outlined as follows: First, allele frequencies are calculated by direct counting. Then, *single-marker gene diversity* (GD) and *haplotype diversity* (HD) are calculated, using the following formulas (Nei 1987):

$$GD = \frac{n}{n-1}\left[1-\sum_{i=1}^{n} p_i^2\right]$$

$$HD = \frac{n}{n-1}\left[1-\sum_{i=1}^{n} X_i^2\right]$$

where n is the number of samples, p_i is the frequency of the ith allele and X_i is the frequency of the ith haplotype. Match probability (MP) is the sum of squared haplotype frequencies ($\Sigma^n_{i=1}X^2_i$). The discrimination capacity (DC) is defined as the ratio between the number of different haplotypes and the total number of haplotypes (Purps et al. 2014).

Analysis of molecular variance (AMOVA) is a method for analyzing population variation using molecular data, such as Y-STR haplotypes. In order to perform AMOVA analysis, it is necessary to define a genetic distance metric between haplotypes. Then, AMOVA analysis partitions the total variance in allele frequencies across multiple loci within and different strata. The so-called ϕ statistics of AMOVA, which are defined in terms of the additive variances, can be stated as follows;

$$\sigma^2 = \sigma_a^2 + \sigma_b^2 + \sigma_c^2$$

$$\phi_{ST} = \frac{\sigma_a^2}{\sigma^2}, \quad \phi_{IS} = \frac{\sigma_b^2}{\sigma^2}, \quad \phi_{IT} = \frac{\sigma_c^2}{\sigma^2}$$

where σ_a is the variance between subpopulations, σ_b is the variance between individuals within the subpopulations, and σ_c is the variance within an individual in the total population. Consequently, ϕ_{ST} is the correlation between genotypes within a subpopulation relative to the total population, ϕ_{IS} is the correlation between genotypes within subpopulations, and ϕ_{IT} is the correlation between genotypes of individuals relative to the total population. An online tool in the YRHD database is available which performs AMOVA analysis and returns pair-wise F_{ST} or ϕ_{ST} plus p values as a significance test. In addition, an MDS (Multi-Dimensional Scaling) plot is generated to illustrate the genetic distance between the analyzed populations graphically (Roewer et al. 2013).

Note that AMOVA technique requires *a priori* knowledge of the subpopulation structure. Otherwise, a clustering analysis, the results of which are then used as a basis for the AMOVA, should be performed first. Meirmans proposes a way to link the widely used method of *K-means clustering* to the AMOVA framework so that this two-step process, clustering and calculation of ϕ-statistics, could be greatly simplified. He points out that we can also test how well the expected population structure matches the structure observed in the data with the help of the clustering mechanism (Meirmans 2012). At this point, it should be stated that, each contribution to the YHRD is assigned to a sub-population cluster (meta population) according to the clusters defined by the YHRD curators (YHRD 2014).

In this paper, the researchers use some information theoretical concepts in order to discuss the genetic diversity of a given population and the discriminatory power of Y-STR markers. These concepts, namely, *entropy* and *mutual information* are briefly discussed below.

Consider a specific Y-STR marker (for example, DYS 456) which can have N possible allele

values (for example, alleles 13, 14, 15, 16, 17, 18) with probabilities $p_1,\ldots,p_N$. It is known that these probabilities can differ across populations. Assume that for an hypothetical population A and population B, these probabilities are given as {1/6,1/6,1/6,1/6,1/6,1/6} and {1.0,0,0,0,0,0}. Given this information, the researchers have no uncertainty about the value of the marker for a sample from population B; it will be the allele 13 with probability 1.0. For population A, every option is equally likely and the researchers have an uncertainty regarding the value of the marker. *Entropy* is a concept that is used to *quantify the uncertainty* about state of a given system (difficulty involved in predicting the value of the allele of the DYS 456). Higher the entropy, higher the uncertainty, therefore, it is intuitively obvious that genetically more diverse populations have higher entropies. In information theory, so-called *Shannon Entropy* is defined as follows;

$$H(X)= -\sum_{i=1}^{N} p_i \log_2 p_i$$

According to this formula, the researchers have (*H(X) = 2.5850*) for population A, and (*H(X) = 0*) for population B in this example.

Now consider two Y–STR markers (for example, DYS 456 and DYS 392). If we know that allele value for DYS 456, will this knowledge make it less difficult to predict the allele value for DYS 392 marker? The answer depends on the association between two markers. The information that one marker has about the other is given by,

$$I(X:Y) = H(X) + H(Y) - H(X,Y)$$

where *H(X), H(Y)* are the marginal entropies and *H(X,Y)* is the joint entropy. This expression could also be represented using probabilities as,

$$I(X,Y)= \sum_{x\varepsilon X} \sum_{y\varepsilon Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

where *p(x)* and *p(x, y)* are the marginal and joint probabilities respectively. The expression,

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

is called the *pointwise mutual information* between two terms *x* and *y* (Cover 2006). In the discussions section, the researchers will use pointwise mutual information and joint entropy (to be defined later) between allele pairs to evaluate and compare the genetic diversities of two populations.

The Allele Frequency Distribution of 16 Y-chromosomal STR Loci in Central Anatolia Population

In this study, the researchers analyzed the allele frequency distribution of 16 Y chromosome specific STR loci (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, Y-GATA H4) in Central Anatolia population using Applied Biosystems AmpF*l*STR® Yfiler™ PCR Amplification Kit.

MATERIAL AND METHODS

Population

Buccal swabs were collected from 106 apparently healthy and unrelated males from Central Anatolia Region of Turkey whose ancestors had lived in this region for at least three generations and of the same Turkish ethnic origin.

DNA Extraction

Genomic DNA was extracted from buccal swab samples using the QIAamp Mini kit (Qiagen, Hilden, Germany) and dissolved in TE buffer, pH 8.0.

PCR

The loci (DYS19, DYS385a/b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, Y-GATA H4) were co-amplified using the Applied Biosystems AmpF*l*STR® Yfiler™ PCR Amplification Kit in a final volume of 12.5ìl according to the manufacturer's instructions. PCR was carried out in a thermocycler GeneAmp1 PCR System 9700 (Applied Biosystems, Foster City, CA). The samples were extracted with QIAamp Mini kit (Qiagen, Hilden, Germany) and 5ìl extracted DNA in 7.5ìl PCR amplification reaction were used.

Electrophoresis and Typing

Samples were run in an ABI PRISM1 3100 Genetic Analyzer (Applied Biosystems, USA) with a 36 cm array and POP-4 polymer according to the manufacture's recommended protocols. The sample run data were analyzed together with

an allelic ladder and positive and negative controls using GeneMapper1 ID Software Version 3.2. (Applied Biosystems). The researchers have also included some control samples. Note that allele nomenclature of GATA H4 (GATA H4.1) must be converted by adding 10 repeats, according to ISFG recommendations (Gusmao et al. 2006).

Analysis of Data

Haplotype and allele frequencies were estimated by direct counting method. Haplotype and gene diversities were estimated according to (Nei 1987). Arlequin 2.0 software is used for the evaluation of the Hardy-Weinberg equilibrium expectations and calculation of relevant statistical parameters (Schneider et al. 2000). The joint entropy of allele pairs are calculated by MatLab® scripts.

RESULTS

A total of 106 unique haplotypes were identified in 106 samples, thus, the observed haplotype diversity was 1.00 for Central Anatolia population sample (Table 1). Allelic frequencies of the16 Y-STRs are summarized in Table 2. The allelic number varied from 5 (for DYS391, DYS437, and DYS438) to 39 (for DYS385a/b). The locus DYS385 exhibits the highest gene diversity value (0.876) and locus DYS391, the lowest (0.564). The most frequent allele was *allele-10* in DYS391 with a frequency of 0.594, followed by *allele-14* in DYS437 with a frequency of 0.566 and *allele-11* in DYS392 with a frequency of 0.557. The frequencies ranged from 0.009 to 0.594 (Table 2). In other studies concerning the Y-STR marker distribution in the Turkish population, the reported data also show a high degree of haplotype diversity which makes Y–STRs a very useful tool in forensic cases (Cakir et al. 2004; Ozbas-Gerceker et al. 2013; Rustamov et al. 2004; Rustamov 2006; Serin et al. 2011; Yukseloglu 2003). In the next section, the researchers will discuss how this diversity manifests itself in the Y-STR data in terms of pointwise mutual information and joint entropy between the allele pairs.

DISCUSSION

In this section, the researchers will refer to two data sets. The first one is the Y-STR data presented in this paper. The other set appears in (Kareem et al. 2015), which is chosen because it is a recent study presented in sufficient haplotype-wise detail on a comparable marker set. Blood samples in their study were randomly collected from 100 healthy unrelated males living in middle or south of Iraq. In the first part of the discussion, the researchers use only two markers, namely DYS456 and DYS392, to illustrate how the pointwise mutual information for both sets are calculated and how the results are compared and visualized using heat maps. Note that marker DYS385 is left out of following discussions because Iraqi data were provided separately for a/b, while our data combined them into a single peak.

For the Turkish population, the co-occurrence matrix of the two alleles is provided in Table 3. The entries of the central matrix (the bold rectangle) show how many times two alleles from either marker appeared together in the haplotype data (for example, allele 15 of DYS456 has been observed 10 times in the same haplotype with the allele 11 of DYS392, out of the total 106 unique haplotypes). The column *f(456)* indicates how many times each allele of the DYS456 marker appeared in total (for example, allele 15 of DYS456 has been observed 20 times altogether in the given dataset). The row *f(392)* has the same interpretation for the alleles of marker DYS392. The observation count vectors for all alleles of marker DYS456 are (27, 8, 20, 19, 10, 22) and (1, 7, 61, 23, 4, 0) for Turkish and Iraqi populations, respectively. The observation count vectors for all alleles of marker DYS392 are (5, 59, 23, 15, 2, 2) and (0, 86, 1, 6, 3, 0) for Turkish and Iraqi populations, respectively. Note that, Iraqi population is grouped around relatively small number of alleles, where some alleles are never observed.

There were 106 and 96 unique haplotypes in Turkish and Iraqi populations, respectively. The researchers use these numbers to normalize the co-occurrence matrices of both populations (Tables 4 and 5).

Using Tables 4 and 5, the researchers calculate the pointwise mutual information between the alleles of two markers for both populations according to the previously given formula;

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

The results are provided in Tables 6 and 7, for Turkish and Iraqi populations, respectively.

Table 1: Frequencies of the haplotypes obtained by studying 16 Y-STRs in Turkey population sample (106 individuals)

Haplotype	*N*	*DYS 456*	*DYS 389I*	*DYS 390*	*DYS 389II*	*DYS 458*	*DYS 19*	*DYS 385*	*DYS 393*	*DYS 391*	*DYS 439*	*DYS 635*	*DYS 392*	*DYS GATA H4*	*DYS 437*	*DYS 438*	*DYS 448*
1	1	16	14	24	30	14	16	15-19	13	10	13	24	12	12	15	10	21
2	1	13	13	21	30	15	16	18-20	12	11	12	21	11	12	14	11	21
3	1	18	14	24	31	16	15	13-19	9	11	11	20	11	12	14	10	20
4	1	18	12	23	29	20	13	13-15	9	9	11	23	11	8	16	9	21
5	1	16	15	24	32	18	15	14-19	13	11	12	21	12	10	14	10	23
6	1	15	14	25	31	17	18	12-12	13	9	11	23	11	13	14	10	20
7	1	15	14	22	31	18	15	13-14	13	10	8	21	11	8	16	10	21
8	1	13	12	21	31	15	16	19-20	14	10	12	21	11	10	14	11	19
9	1	16	12	21	28	17	16	14-15	15	11	12	21	12	12	15	10	23
10	1	15	14	22	30	16	13	13-16	13	10	11	22	13	12	16	9	19
11	1	15	13	24	30	16	15	13-17	9	9	8	20	11	10	15	9	20
12	1	13	13	24	29	18	14	14-19	8	10	11	23	11	10	14	11	21
13	1	13	14	25	31	14	15	11-14	10	10	11	23	11	13	14	11	20
14	1	13	13	22	30	17	16	13-15	10	9	12	21	13	10	14	9	19
15	1	16	13	25	30	15	15	12-14	9	11	10	24	11	11	14	12	20
16	1	17	15	23	31	16	17	11-14	8	10	12	23	11	13	14	11	20
17	1	15	13	24	29	17	14	11-13	8	10	13	23	12	12	14	12	20
18	1	15	13	24	30	16	14	14-16	13	10	12	21	11	12	15	9	19
19	1	14	14	25	32	20	13	19-20	10	10	12	22	11	12	14	10	20
20	1	13	13	24	30	14	13	16-17	8	10	11	21	10	10	14	11	19
21	1	18	14	22	30	16	13	14-16	12	10	8	22	15	9	13	11	19
22	1	13	13	23	30	18	14	14-15	8	10	12	20	11	12	14	11	20
23	1	13	13	24	31	19	15	14-18	13	11	11	23	12	10	15	10	22
24	1	15	12	23	30	19	13	13-17	12	10	12	22	11	10	14	11	20
25	1	15	13	23	29	14	15	13-16	12	9	13	22	11	12	14	9	21
26	1	18	12	23	30	17	14	13-19	14	10	11	21	12	9	14	10	20
27	1	18	12	24	29	19	15	14-17	16	7	13	23	11	8	14	9	19
28	1	16	14	26	30	17	17	13-16	14	12	11	24	12	12	14	12	19
29	1	16	14	24	31	19	15	15-20	13	12	12	23	12	12	15	11	20
30	1	13	10	25	29	19	16	12-16	8	10	11	22	11	10	14	10	21
31	1	13	13	24	31	20	14	15-17	14	10	12	24	13	9	16	10	19
32	1	15	11	26	30	19	15	13-20	13	11	11	22	12	10	14	11	20
33	1	13	12	23	28	16	15	13-14	9	10	11	23	11	12	14	10	20
34	1	15	12	23	28	14	14	13-18	12	10	13	23	11	11	14	10	21
35	1	17	13	24	29	15	15	12-12	13	10	12	21	11	11	14	11	21
36	1	18	12	24	28	19	15	12-13	9	10	12	26	12	12	14	11	17
37	1	18	15	24	30	18	14	13-16	8	10	10	20	11	10	15	13	21
38	1	13	12	22	30	19	16	14-14	11	10	8	22	10	12	16	10	19

Table 1: Contd...

Haplotype	*N*	*DYS 456*	*DYS 389I*	*DYS 390*	*DYS 389II*	*DYS 458*	*DYS 19*	*DYS 385*	*DYS 393*	*DYS 391*	*DYS 439*	*DYS 635*	*DYS 392*	*DYS GATA H4*	*DYS 437*	*DYS 438*	*DYS 448*
39	1	17	13	24	29	16	14	12-15	9	11	12	23	13	13	15	12	20
40	1	13	12	21	29	19	14	13-16	10	10	11	21	11	9	16	10	23
41	1	13	12	24	31	19	16	11-14	9	10	10	23	11	10	14	11	20
42	1	13	13	24	31	16	13	11-13	9	11	10	23	11	12	14	11	20
43	1	15	12	23	28	18	14	13-17	12	10	12	21	11	11	14	9	20
44	1	13	12	23	28	16	13	13-15	9	10	11	20	13	11	16	10	19
45	1	18	14	26	30	16	17	11-13	9	9	10	23	10	12	14	11	20
46	1	17	15	24	29	14	15	11-14	9	10	12	23	11	12	14	11	20
47	1	13	13	23	29	16	15	13-17	9	9	11	23	11	11	14	9	21
48	1	18	13	23	30	15	14	12-20	8	10	10	22	11	12	15	9	19
49	1	13	12	23	28	19	14	13-18	10	10	12	22	11	9	15	9	21
50	1	18	10	23	30	17	16	15-16	9	10	13	21	12	13	15	10	19
51	1	16	12	23	29	19	15	13-16	15	10	11	21	11	9	15	9	22
52	1	15	11	24	31	18	15	13-21	12	9	11	21	11	8	14	10	19
53	1	15	11	26	32	18	17	10-13	12	10	11	24	12	8	14	12	20
54	1	17	10	19	33	19	11	15-16	13	10	12	20	12	8	17	13	20
55	1	16	14	24	31	16	14	11-14	9	11	12	23	13	13	14	12	19
56	1	18	13	24	29	16	14	12-16	8	11	12	23	13	12	15	12	19
57	1	16	14	24	31	15	15	15-19	13	11	11	22	12	12	15	10	20
58	1	16	14	24	30	14	15	13-14	13	10	11	24	11	11	14	10	21
59	1	15	13	23	30	19	15	13-16	15	10	11	24	10	9	16	11	19
60	1	13	13	25	31	20	15	15-19	14	11	11	22	12	12	14	11	20
61	1	13	13	24	29	17	13	11-13	10	11	12	23	11	12	14	11	20
62	1	17	13	24	31	17	14	14-16	9	11	13	23	11	12	15	10	20
63	1	14	13	24	31	19	15	14-18	13	10	11	23	12	10	15	10	22
64	1	18	10	22	31	17	16	12-13	14	9	12	20	11	10	16	10	21
65	1	18	13	25	29	17	16	11-14	13	10	12	23	11	12	14	12	20
66	1	17	13	23	30	20	14	13-18	9	11	12	20	11	12	14	10	20
67	1	13	14	23	29	16	13	13-16	9	10	11	21	13	12	15	9	19
68	1	13	13	24	31	14	15	11-14	9	11	12	23	11	12	14	11	20
69	1	13	13	23	30	17	15	13-17	9	10	13	22	13	13	16	10	19
70	1	13	13	23	30	15	15	12-20	9	10	10	21	11	8	16	9	19
71	1	15	11	26	31	18	16	12-16	12	9	12	23	11	8	14	12	20
72	1	14	13	23	30	16	15	15-16	9	10	10	21	12	10	14	10	19
73	1	16	15	24	32	14	13	16-19	13	10	12	22	11	12	14	10	20
74	1	16	13	23	29	15	15	13-17	12	9	11	22	11	11	14	9	21
75	1	18	13	24	29	16	14	13-16	8	11	12	23	13	12	15	12	19
76	1	13	13	21	32	17	15	16-18	10	10	11	22	11	11	16	10	20
77	1	16	13	23	29	17	14	11-15	9	10	12	23	13	12	15	12	19

Table 1: Contd...

Haplotype	*N*	*DYS 456*	*DYS 389I*	*DYS 390*	*DYS 389II*	*DYS 458*	*DYS 19*	*DYS 385*	*DYS 393*	*DYS 391*	*DYS 439*	*DYS 635*	*DYS 392*	*DYS GATA H4*	*DYS 437*	*DYS 438*	*DYS 448*
78	1	17	13	23	30	16	14	11-13	12	10	12	23	13	12	14	12	19
79	1	17	11	26	27	17	16	12-17	14	11	12	24	12	12	16	12	20
80	1	18	14	25	31	15	14	18-20	14	11	12	22	12	12	15	11	20
81	1	13	13	23	30	17	13	13-18	9	9	10	21	11	12	14	9	21
82	1	18	13	24	30	18	13	13-15	8	10	13	24	11	12	15	9	20
83	1	15	13	19	29	17	14	13-13	13	11	12	24	13	9	15	10	19
84	1	18	13	23	30	18	13	13-16	8	10	10	21	11	12	14	13	20
85	1	18	14	22	30	17	13	13-14	9	10	11	22	13	13	15	9	18
86	1	15	12	24	31	19	14	19-21	13	10	11	22	12	10	14	10	19
87	1	18	12	23	28	20	14	13-18	9	10	10	21	11	12	15	9	21
88	1	18	13	23	30	19	14	12-15	8	10	12	22	11	12	14	10	19
89	1	14	13	25	29	16	13	11-15	10	10	12	23	11	12	14	11	20
90	1	14	12	24	28	15	10	16-16	12	10	10	21	11	11	14	9	19
91	1	18	13	23	30	17	14	13-16	8	10	11	22	11	12	15	9	21
92	1	16	14	23	30	14	15	13-16	9	10	12	23	12	13	14	10	20
93	1	15	15	22	26	17	13	14-16	13	10	11	22	15	12	14	11	20
94	1	16	13	22	30	15	16	12-16	16	11	11	22	12	11	17	11	22
95	1	14	14	23	30	18	14	13-18	8	11	9	20	11	10	14	10	20
96	1	15	13	23	29	17	14	12-14	13	10	10	25	14	12	14	10	18
97	1	14	14	23	30	18	13	15-20	12	10	11	23	11	13	14	9	20
98	1	15	15	24	28	16	13	14-17	13	10	8	21	14	13	13	10	19
99	1	17	15	24	30	16	13	18-19	13	10	11	22	11	13	14	10	21
100	1	18	14	22	31	15	15	13-19	9	10	11	22	11	12	14	11	20
101	1	14	14	23	31	17	15	14-16	10	11	11	22	13	12	14	9	19
102	1	13	11	24	29	20	14	13-18	8	11	10	21	11	8	15	9	19
103	1	16	11	22	32	19	15	15-18	13	11	13	23	12	12	16	12	20
104	1	16	13	25	31	17	17	19-21	14	11	11	23	11	10	15	10	21
105	1	16	12	24	29	15	15	14-17	12	11	11	21	10	13	16	9	19
106	1	16	15	23	29	18	10	14-19	8	10	8	20	11	13	14	10	19

Table 2: Allele frequencies and gene diversity values of 16 Y-STR loci in Turkey.

Haplotype	*N*	*DYS 456*	*DYS 389I*	*DYS 390*	*DYS 389II*	*DYS 458*	*DYS 19*	*DYS 385*	*DYS 393*	*DYS 391*	*DYS 439*	*DYS 635*	*DYS 392*	*DYS GATA H4*	*DYS 437*	*DYS 438*	*DYS 448*
7								0.009								10-13	0.0094
8							0.161		0.057			0.085				11-13	0.0471
9							0.264	0.114	0.009			0.075		0.236		11-14	0.0660
10		0.038				0.019	0.085	0.594	0.313		0.047	0.170		0.358		11-15	0.0188
11		0.066				0.009	0.009	0.264	0.358		0.557	0.104		0.245		12-12	0.0188
12		0.189				-	0.132	0.019	0.349		0.217	0.434		0.133		12-13	0.0188
13	0.254	0.415				0.189	0.217		0.094		0.141	0.132	0.019	0.028		12-14	0.0188
14	0.075	0.207			0.094	0.274	0.085				0.019		0.566			12-15	0.0188
15	0.189	0.085			0.123	0.321	0.028				0.019		0.255			12-16	0.0377
16	0.179				0.199	0.132							0.141			12-17	0.0094
17	0.055				0.217	0.047							0.019		0.009	12-20	0.0188
18	0.248				0.132	0.009									0.019	13-13	0.0094
19			0.019		0.170										0.311	13-14	0.0377
20			-		0.067					0.094					0.406	13-15	0.0377
21			0.047							0.237					0.189	13-16	0..1132
22			0.104							0.246					0.038	13-17	0.0576
23			0.330							0.311					0.028	13-18	0.0660
24			0.350							0.094						13-19	0.0283
25			0.094							0.009						13-20	0.0094
26			0.056	0.009						0.009						13-21	0.0094
27				0.009												14-14	0.0094
28				0.094												14-15	0.0188
29				0.236												14-16	0.0471
30				0.350												14-17	0.0283
31				0.236												14-18	0.0188
32				0.057												14-19	0.0283
33				0.009												15-16	0.0283
34																15-17	0.0094
																15-18	0.0094
																15-19	0.0283
																15-20	0.0188
																16-16	0.0094
																16-17	0.0094
																16-18	0.0094
																16-19	0.0094
																18-19	0.0094
																18-20	0.0188
																19-20	0.0188
																19-21	0.0188
GD	0.810	0.736	0.744	0.755	0.839	0.766	0.824	0.564	0.720	0.769	0.620	0.742	0.594	0.737	0.700		0.876

Table 3: Co-occurrence counts of alleles for DYS456 and DYS392 markers in Turkish Central Anatolia Population

Turkish	*DYS 392*							
	Allele	*10*	*11*	*12*	*13*	*14*	*15*	*f(456)*
DYS456	13	2	18	2	5	0	0	27
	14	0	5	2	1	0	0	8
	15	1	10	4	2	2	1	20
	16	1	7	9	2	0	0	19
	17	0	6	2	2	0	0	10
	18	1	13	4	3	0	1	22
	f(392)	5	59	23	15	2	2	106

Table 4: Normalized co-occurrence counts of alleles for DYS456 and DYS392 markers in Turkish Population

	DYS 392							
Turkish	*Allele*	*10*	*11*	*12*	*13*	*14*	*15*	*f(456)*
DYS456	13	0.018868	0.16981	0.018868	0.04717	0	0	0.25472
	14	0	0.04717	0.018868	0.009434	0	0	0.075472
	15	0.009434	0.09434	0.037736	0.018868	0.018868	0.009434	0.18868
	16	0.009434	0.066038	0.084906	0.018868	0	0	0.17925
	17	0	0.056604	0.018868	0.018868	0	0	0.09434
	18	0.009434	0.12264	0.037736	0.028302	0	0.009434	0.20755
	p(392)	0.04717	0.5566	0.21698	0.14151	0.018868	0.018868	1.00

Table 5: Normalized co-occurrence counts of alleles for DYS456 and DYS392 markers in Iraqi population

	DYS 392							
Iraqi	*Allele*	*10*	*11*	*12*	*13*	*14*	*15*	*f(456)*
DYS456	13	0	0.010417	0	0	0	0	0.010417
	14	0	0.052083	0.010417	0.0104171	0	0	0.072917
	15	0	0.57292	0	0.03125	0.03125	0	0.63542
	16	0	0.23958	0	0	0	0	0.23958
	17	0	0.020833	0	0.020833	0	0	0.041667
	18	0	0	0	0	0	0	0
	p(392)	0	0.89583	0.010417	0.0625	0.03125	0	1.00

Table 6: Pointwise mutual information between the alleles of DYS456 and DYS392 markers in Turkish population

	DYS 392						
Turkish	*Allele*	*10*	*11*	*12*	*13*	*14*	*15*
DYS456	13	0.6511	0.26031	-1.5505	0.38807	0	0
	14	0	0.16721	0.20436	-0.17897	0	0
	15	0.084064	-0.15472	-0.11757	-0.5009	2.406	1.406
	16	0.15806	-0.5953	1.1264	-0.4269	0	0
	17	0	0.10831	-0.11757	0.4991	0	0
	18	-0.053439	0.086286	-0.25507	-0.053439	0	1.2685

A closer inspection of the PMI formula might be helpful for interpreting the results provided in these tables. If there is a strong association between *x* and *y* (that is, the alleles x and y appear often together in the same haplotype), then $p(x,y) >> p(x)p(y)$. If the association between *x* and *y* is weak, then $p(x,y) \approx p(x)p(y)$ (Chen 2011). In that case, the researchers will be taking the

Table 7: Point wise mutual information between the alleles of DYS456 and DYS392 markers in Iraqi population

		DYS 392					
Iraqi	*Allele*	*10*	*11*	*12*	*13*	*14*	*15*
DYS456	13	0	0.1587	0	0	0	0
	14	0	-0.32673	3.7776	1.1926	0	0
	15	0	0.0093201	0	-0.34577	0.65423	0
	16	0	0.1587	0	0	0	0
	17	0	-0.8413	0	3	0	0
	18	0	0	0	0	0	0

logarithm of a number which is very close to 1 *($log_2(1)=0$)*and the researchers can say that these alleles behave independently each other. In other words, neither allele has any knowledge about the other, thus, mutual information between them is zero. It is also possible that the alleles x and y appear never together, then *p (x, y)=0*, and the pointwise mutual information will be minus infinity *($log_2(0)=-\infty$)*. In this research, the researchers interpret these two cases as equivalent, in the sense that there is no useful information available, and set the corresponding entries in tables to zero.

An interesting case occurs when *($0<p(x,y)/(p(x)\,p(y))<1$)*, and *PMI* assumes a negative value. In other words, the researchers have *negative* mutual information (also called *misinformation*) between two alleles. Negative values of mutual information rise when the given two alleles occur very rarely together although each one is observed very frequently in combination with other alleles. Therefore, they can potentially serve as a differentiating factor between the genetic profiles of two populations, and the negative values are preserved in the tables.

The co-occurrence and *PMI* tables reveal that fewer different allele combinations occur in the Iraqi population. As can be seen in the co-occurrence tables, 25 different allele pairs are observed in the Turkish population, compared to 10, out of theoretically possible 36. Furthermore, a single allele pair (DYS456-Allele15, DYS392-Allele11) occurs 55 times in 96 haplotypes (57.29%), compared to the most frequent allele pair in Turkish population (DYS456-Allele13, DYS392-Allele11) which occurs 18 times in 106 haplotypes (16.98%). It is intuitively obvious that Turkish population is genetically more diverse of the two, as far as these two markers are concerned. *PMI* in Turkish and Iraqi Populations for two markers are presented in Figure 1, using 3-dimensional bar graphs and heat maps. The 3-dimensional bar graphs get difficult to interpret as the number of alleles increases, so the researchers prefer to use heat maps which convey the same information in a visually more appealing manner.

At this point, consider what can be inferred from GD (genetic diversity) values of markers. For Iraqi population, GD(*DYS392*)=0.185, GD(*DYS456*)=0.752, and for Turkish population GD(*DYS392*)=0.620, GD(*DYS456*)=0.810. Comparing these values, it is easy to conclude that, especially the marker DYS392 should have much higher discrimination capacity in Iraqi population. Yet, the overall results belie this conclusion. The reason for this misevaluation is provided by the *PMI* and the joint entropy between allele pairs, as further discussed in the following.

Now, the researchers will consider how much information these two markers have about each other. Note that information here is a symmetrical concept. In other words, each marker has the same amount of information about the other. If one marker knows everything about another marker, there is no additional information that the measurement of the other can provide and one of the markers becomes redundant. Since uncertainty increases with decreasing information, the researchers expect the joint entropy of two markers become larger as one marker knows less about the other. Therefore, the researchers interpret a marker pair with higher *joint entropy* as having a higher discrimination capacity. The joint entropy of two markers (*X,Y*) is given by:

$$H(X, Y) = -\sum_{x\epsilon X} \sum_{y\epsilon Y} (p(x,y) log_2 p(x,y)$$

where the sum is taken over all alleles x of marker X and all alleles y of marker Y. As expected, the joint entropy *H*(*DYS456,DYS292*) is calculated as *1.9273* for Iraqi population and *4.086* for Turkish Central Anatolia population.

The heat maps of joint entropies for all marker pairs in both populations are provided in Figures 2 and 3. These figures show that a larger number of marker pairs have higher joint entro-

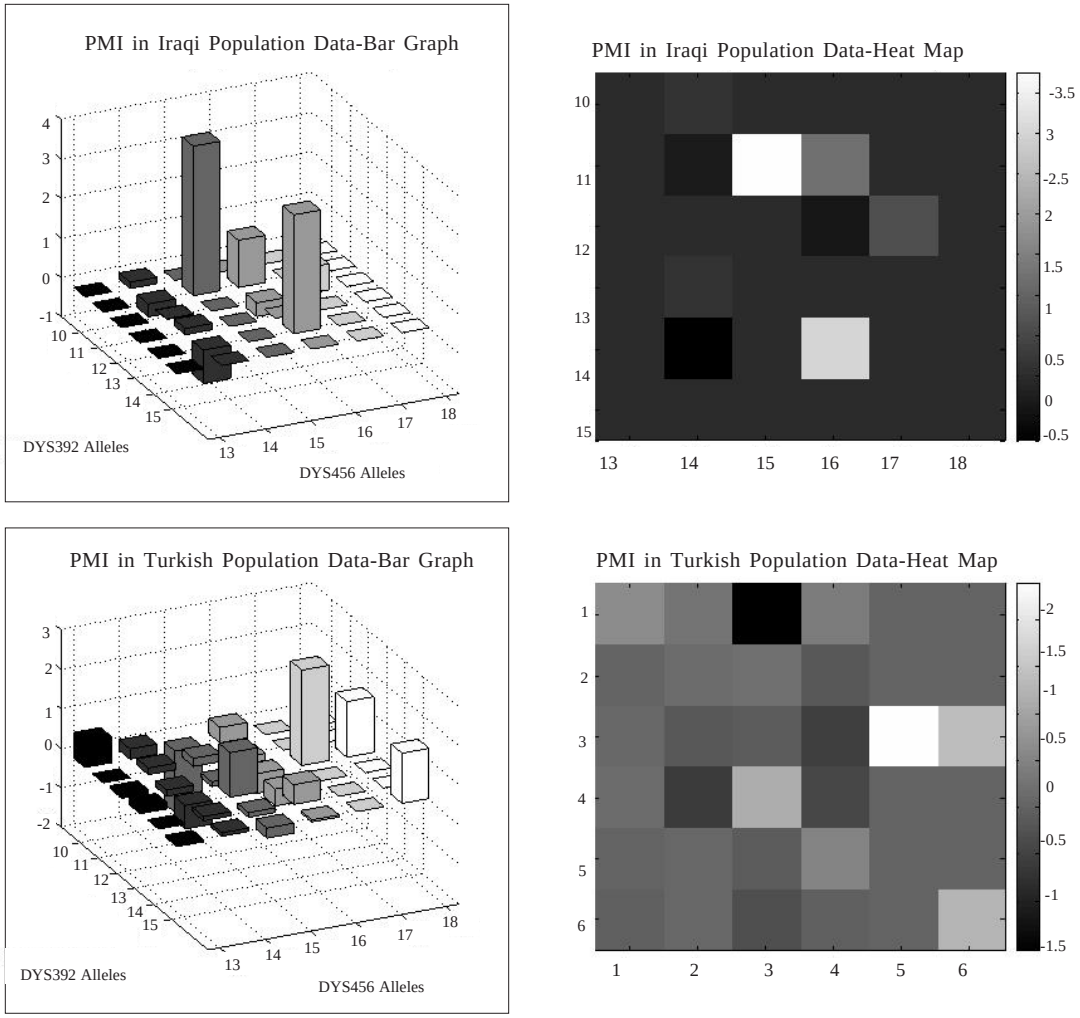

Fig. 1. Pointwise Mutual Information in Turkish and Iraqi Populations – 2 Markers

pies in Turkish population compared to Iraqi population, which, in turn, explains why all haplotypes set were unique and why the same set of markers have a higher male individualization capacity in Turkish data set.

CONCLUSION

The 16 Y–STR loci markers set does not have the same discrimination power for the purpose of male individualization across different populations. While the traditional genetic diversity values (for example, *single-marker gene diversity, haplotype diversity*) fail to properly assess the suitability of a marker set for a given population data, information theory offers more elucidating measures to this end and provides helpful arguments to explain the higher likelihood of complete male individualization with these 16 Y–STR markers in Turkish population. In this paper, the researchers show that the *pointwise mutual information* and the *joint entropy* between marker/allele pairs measure the discrimination power of markers more accurately than individual genetic diversity values and provide a better insight into the interaction between the

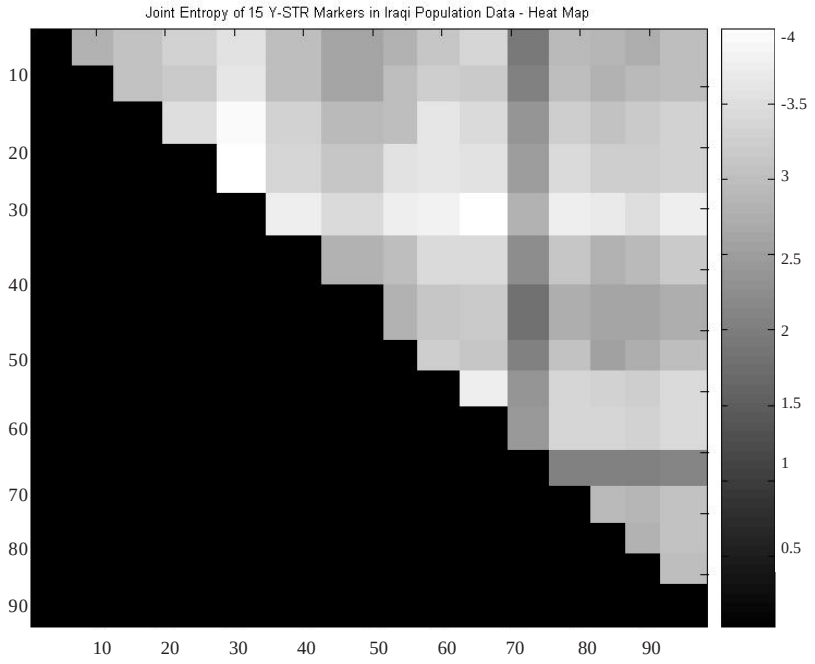

Fig. 2. Joint entropy of Y-STR markers in Iraqi population – 15 Markers

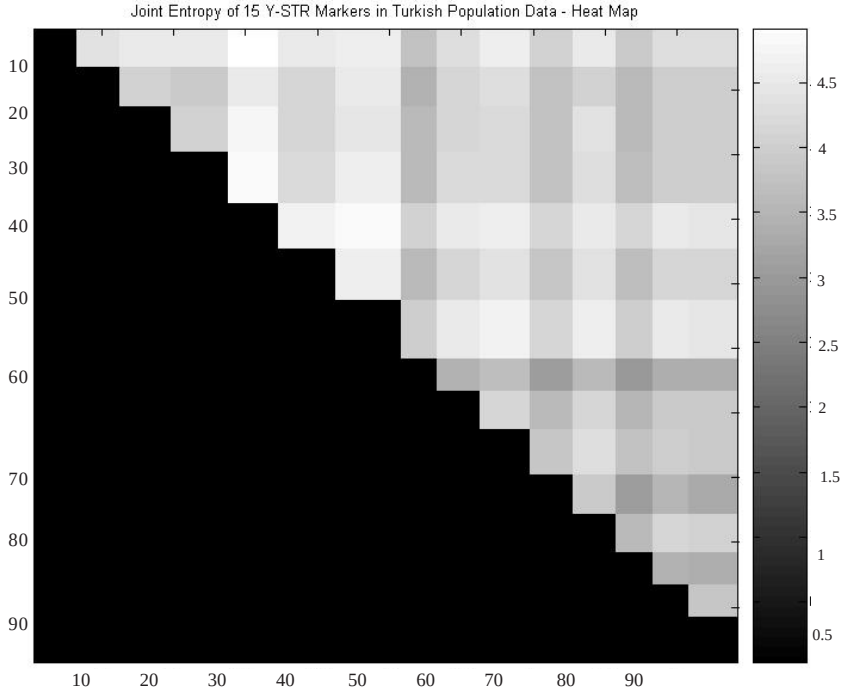

Fig. 3. Joint entropy of Y-STR markers in Turkish population – 15 Markers

genetic profile of the population and the given Y–STR marker set. It is obvious that, the presented heat maps provide a better insight into the interaction between the genetic profile of the population and the given Y-STR marker set which goes beyond what possibly can be conveyed through marker-wise genetic diversity values.

ACKNOWLEDGMENTS

This study is financially supported by University of Gaziantep Research Fund (TF.08.19). The Y-STR data that appears in this manuscript were presented as a poster at 4th Mediterranean Academy of Forensic Sciences Meeting held in Antalya, Turkey from 14 to 18 October 2009.

REFERENCES

Adami C 2004. Information theory in molecular biology. *Physics of Life Reviews*, 1: 3-22.

Asicioglu F, Akyuz F, Cetinkaya U, Ozbek U 2002. Allele distribution data of nine short tandem repeat loci for Turkish population: D3S1358, vWa, FGA, D8S1179, D21S11, D18S51, D13S317, D7S820. *Forensic Sci Int*, 129: 75-77.

Asicioglu F, Akyuz F, Cetinkaya U, Canli MA 2003. Allele and haplotype frequencies of Y-Short tandem repeat loci in Turkey. *Croat Med J*, 44: 310-314.

Brion M, Quintans B, Gonzalez-Neira A, Zarrabeitia M, Salas A et al. 2014. Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats. *Human Mutation*, 35: 1021-1032.

Berger B, Niederstatter H, Kochl S, Steinlechner M, Parson W 2003. Male/female DNA mixtures: A challenge for Y-STR analysis. *Int Cong Series*, 1239: 295-299.

Brion M, Quintans B, Gonzalez-Neira A, Zarrabeitia M, Salas A et al. 2003. Microgeographic patterns of highly informative Y-chromosome haplotypes (using biallelic markers and STRs) in Galicia (NW Spain): Forensic and anthropological implications. *Int Cong Series*, 1239: 61-66.

Budowle B, Ge J, Chakraborty R 2007. Basic Principles for Estimating the Rarity of Y-STR Haplotypes Derived from Forensic Evidence. *Paper presented in the 18th International Symposium on Human Identification* in Hollywood, CA, 1-4, October 2007.

Butler JM 2003. Recent developments in Y-Single tandem repeat and Y-Single nucleotide polymorphism analysis. *Forensic Sci Rev,* 15: 91-111.

Butler JM 2005. *Forensic DNA Typing - Biology Technology and Genetics of STR Markers*. 2nd Edition. Burlington, MA, USA: Elsevier Academic Press.

Caliebe A, Jochens A, Willuweit S, Roewer L, Krawczak M 2015. No shortcut solution to the problem of Y-STR match probability calculation. *Forensic Sci Int Genet*, 15: 69-75.

Carracedo A, Sánchez-Diz P 2005. Forensic DNA-typing technologies-A review. In: A Carracedo (Ed.): *Forensic DNA Typing Protocols*. Totowa, New Jersey, USA: Humana Press, pp. 1-11.

Carracedo A 2015. Forensic genetics: History. In: Max H Houck (Ed.): *Forensic Biology.* San Diego (CA): Academic Press, pp. 19-22.

Cakir AH, Celebioglu A, Yardimci E 2004. Y-STR haplotypes in Central Anatolia region of Turkey. *Forensic Sci Int*, 144: 59-64.

Cover TM, Thomas JA 2006. *Elements of Information Theory*. 2nd Edition. Hoboken, New Jersey, USA: John Wiley & Sons.

Chen Z, Lu Y 2011. A word co-occurrence matrix based method for relevance feedback. *J Comput Inf Sys*, 7: 17-24.

Gusmao L, Butler JM, Carracedo A, Gill P, Kayser M et al. 2006. DNA commission of the international society for forensic genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci Int*, 157: 187-197.

Hanson EK, Ballantyne JM 2006. Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Legal Med*, 8: 110-120.

Kareem MA, Jebor MA, Hameed IH 2015. Allele frequency present within the DYS635, DYS437, DYS448, DYS456, DYS458, YGATA H4, DYS389I, DYS389II, DYS19, DYS391, DYS438, DYS390, DYS439, DYS392, DYS393, DYS385a and DYS385b of unrelated individuals in Iraq. *African J Biotec*, 14: 851-858.

Kayser M, Sajantila A 2001. Mutations at Y-STR loci: Implications for paternity testing and forensic analysis. *Forensic Sci Int*, 118: 116-121.

Meirmans PG 2012. AMOVA-based clustering of population genetic data. *J of Heredity*, 103: 744-750.

Mizuno N, Nakahara H, Sekiguchi K, Yoshida K, Nakano M, Kasai K 2008. 16 Y chromosomal STR haplotypes in Japanese. *Forensic Sci Int*, 174: 70-75.

Nei M 1987. *Molecular Evolutionary Genetics*. New York: Columbia University Press.

Ozbas-Gerceker F, Bozman N, Arslan A, Serin A 2013. Population data for 17 Y-STRs in samples from southeastern Anatolia region of Turkey. *Int J Hum Genet*, 13: 105-111.

Purps J, Siegert S, Willuweit S, Nagy M, Alves C et al. 2014. A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Sci Int Genet*, 12: 12-23.

Roewer L, Parson W 2013. Internet accessible population databases: YHRD and EMPOP. In: JA Siegel, PJ Saukko (Eds.): *Encyclopedia of Forensic Sciences*. 2nd Edition. Waltham, MA, USA: Elsevier Academic Press, pp. 357-364.

Rustamov A, Gumus G, Karabulut HG, Elhan AH, Kadikiran A et al. 2004. Y-STR polymorphism in Central Anatolian region of Turkey. *Forensic Sci Int*, 139: 227-230.

Rustamov A 2006. *12 Y-STR Polymorphism and Haplotypes Frequencies Investigation in Turkey*. PhD Thesis, Unpublished. Ankara: University of Ankara.

Schneider S, Roessli SD, Exoffier L 2000. *A Software for Population Genetics Data Analysis, Arlequin Version 2.0*. Genetics and Biometry Laboratory. Geneva: University of Geneva.

Serin A, Canan H, Alper B, Sertdemir Y 2011. Haplotype frequencies of 17 Y-chromosomal STR loci from the Cukurova region of Turkey. *Croat Med J*, 52: 703-708.

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

Waits LP, Luikart G, Taberlet P 2001. Estimating the probability of identity among genotypes in natural populations: Cautions and guidelines. *Molecular Ecology*, 10: 249-256.

Walsh SJ 2013. Significance. In: JA Siegel, PJ Saukko (Eds.): *Encyclopedia of Forensic Sciences*. 2nd Edition. Waltham, MA, USA: Elsevier Academic Press, pp. 295-299.

Willuweit S, Roewer L 2015. The new Y chromosome haplotype reference database. *Forensic Sci Int Genet*, 15: 43-48.

Y-Chromosome Haplotype Reference Database (YHRD) 2014. From <www.yhrd.org> (Retrieved on 9 September 2014).

Yukseloglu HE 2003. *Y Kromozomu STR Polimorfizminin Babalik Tayini Ve Adli Identifikasyonda Kullanimi.* PhD Thesis, Unpublished. Istanbul: University of Istanbul.