Forensic features and genetic legacy of the Baloch population of Pakistan and the Hazara population across Durand-line revealed by Y chromosomal STRs

Atif Adnan[1*], Shao-Qing Wen[2], Allah Rakha[3], Rashed Alghafri[4], Shahid Nazir[3], Muhammad Rehman[5], Chuan-Chao Wang[6], and Jie Lu[1*]

1. Department of Human Anatomy, School of Basic Medicine, China Medical University, Shenyang, Liaoning 110122, P.R. China
2. Institute of Archaeological Science, Fudan University, Shanghai 200433, China
3. Department of Forensic Sciences, University of Health Sciences Lahore, 54600 Lahore, Pakistan
4. Dubai Police General Head Quarters, General Department of Forensic Sciences and Criminology, Dubai, United Arab Emirates
5. Forensic Medicine Directorate, Ministry of Public Health, Kabul, Afghanistan
6. Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, PR China

*Corresponding authors:

Jie Lu (lvjie@cmu.edu.cn)
Atif Adnan (mirzaatifadnan@gmail.com)

ABSTRACT:

Hazara population across Durand- line has experienced extensive interaction with Central Asian and East Asian populations. Hazara individuals have typical Mongolian facial appearances and they called themselves descendants of Genghis Khan’s army. The people who speak the Balochi language are called Baloch. Previously, a worldwide analysis of Y-chromosomal haplotype diversity for rapidly mutating (RM) Y-STRs and with PowerPlex Y23 System (Promega Corporation Madison, USA) kit was created with collaborative efforts, but Baloch and Hazara population from Pakistan and Hazara population from Afghanistan were missing. A limited data with limited number of markers and samples is available which poorly define these populations. So, in the current study, Yfiler Plus PCR Amplification Kit loci were examined in 260 unrelated Hazara individuals from Afghanistan, 153 Hazara individuals, and 111 Balochi individuals from Baluchistan Pakistan. For the Hazara population from Afghanistan and Pakistan overall, 380 different haplotypes were observed on these 27 Y-STR loci, gene diversities ranged from 0.51288 (DYS389I) to 0.9257 (DYF387S1) and haplotype diversity was 0.9992 +/- 0.0004. For the Baloch population, every individual was unique at 27 Y-STR loci, gene diversity ranged from 0.5718 (DYS460) to 0.9371(DYF387S1). Twelve haplotypes shared between 178 individuals while only two haplotypes among these twelve were shared between 87 individuals in Hazara populations. Rst and Fst pairwise genetic distance analyses, multidimensional scaling (MDS) plot, Neighbor-joining (NJ) tree, linear discriminatory analysis (LDA), and median-joining network (MJNs) were performed, which shed light on the history of Hazara and Baloch populations. Interestingly null alleles were observed at DYS448 with specific mutation patterns in Hazara populations. The results of our study showed that the Yfiler Plus PCR Amplification Kit marker set provided substantially stronger discriminatory power in the Baloch population of Pakistan and the Hazara population across the Durand-line.

Keywords: Hazara; Pakistan; Afganistan; Baloch; Population history; Forensic Genetics

INTRODUCTION

The variation pattern in Human DNA usually provides a balance between natural selection and neutral processes. Y chromosomal variant analysis for determining the patterns of present and past flow of genes between populations is very helpful [1]. Y-chromosome short tandem repeats (YSTRs) plays an important role in forensic molecular biology [2–5]. Usually, Y-STRs are used for (i) decidedly determine the male component of DNA mixtures under the presence of a high female DNA background as typically confronted with materials from sexual assault cases[6], (ii) to test for paternal relationships between male individuals particularly in deficiency paternity cases with the mother not being available[7], or (iii) for special cases in missing-person or (iv) disaster-victim identification involving males[8], or (v) for evolutionary purposes because male family members share same haplotype distribution which may be different from individual to individual within a population group, or (vi) different geographic regions or in different ethnic groups. Normally, more paternal lineages can be differentiated with an increased number of Y-STRs [9], such as the Powerplex Y Kit (Promega) containing 12 Y-STRs [10], the AmpFlSTR Y-filer PCR Amplification Kit (Life Technologies) (subsequently referred to as Y-filer) containing 17 Y-STRs [11] or Powerplex Y23 Kit (Promega) containing 23 Y-STRs [12], relative to the initially proposed 9-loci haplotype [13]. So, Applied Biosystems have developed Yfiler Plus PCR Amplification Kit [14]. The Yfiler Plus kit provides enhanced discrimination power because it includes the Yfiler loci and 10 additional STRs in which 6 are rapidly mutating (RM) Y STRs. These rapidly mutating Y STRs showed a higher mutation rate of about a few mutations every 100 generations per locus ($\mu > 10^{-2}$) compared with all other commonly used Y-STRs. Molecular biological and cytogenetical studies give us an insight into the presence of many structural variants within the human Y chromosome, which might be deletions [15–17], duplications [18–20], and inversions [19–23]. Null alleles or allele droop-out are well-established factors that can occur with any PCR-based STR typing system. The reason could be the primer binding site problem or deletions within the target region [24,25]. DYS448 lied in the proximal part of the azoospermia factor c (AZFc) region, which is considered important in spermatogenesis and made up of ‘‘ampliconic’’ repeats which act as substrates for nonallelic homologous recombination (NAHR). NAHR could delete larger blocks of the Y chromosome which included DYS448[26]. This null alleles or allelic drop-out phenomenon is more commonly observed in Central Asian and East Asian populations but in the Hazara population of Pakistan, its occurrence was >16% [27].

Durand Line is a boundary established in the Hindu Kush around 1893 running through the tribal lands between Afghanistan and British India (modern-day Pakistan), marking their respective scopes of influence. The recognition of this line, which was named after Sir Mortimer Durand, has settled the Indo-Afghan frontier problem for the rest of the British period. Now, this is an established border between Afghanistan and Pakistan. The origin of the Hazara population is disputed. The Hazara could be of Turko-Mongol ancestry and theorized to be the descendants of an occupying army left in Afghanistan by Genghis Khan in thirteen hundred AD [28]. The Hazara population speaks Persian with some Mongolian words. The total population of Hazaras in the world is 4.5 million. Afghanistan is considered the mainland for the Hazara population (3 million) and they are the third largest ethnic group (9%) after Tajiks (27%) and Pashtuns (42%) [29], while in Pakistan, Hazara is one of the distinct but small groups comprising 0.08% of the total population (http://www.pbscensus.gov.pk). The tribes who speak the Balochi language are called Baloch[30]. Balochi population is 3.6% of total Pakistani population (http://www.pbscensus.gov.pk). They are also found in the neighboring areas of Iran and Afghanistan. Perhaps, the origin of Baloch homeland lay on the Iranian plateau. The Baloch were mentioned in Arabic chronicles of the 10th century. The Seljuq invasion of Kermān in the 11th century started the eastward migration of the Balochi population[30].

In this study, we have investigated the Baloch and Hazara population from Pakistan and the Hazara population from Afghanistan using 27 Y STRs to determine their genetic history and gene diversity. This data has defined the Hazara and Baloch populations better and are supplement to the Y STR haplotype reference database (YHRD).

2. RESULTS AND DISCUSSIONS:

2.1 Allelic frequencies and Forensic parameters

We successfully obtained genotypes of 524 individuals in three ethnic groups (Balochi population, Hazara population from Afghanistan, and Pakistan) (**Supplementary Table 1**). Allelic frequencies of Baloch ethnic group from Baluchistan, Pakistan, and Hazara ethnic groups from Pakistan and Afghanistan along with gene diversity values were shown in **Supplementary Table 2**.

DYF387S1 showed the highest gene diversity/heterozygosity in Baloch and both Hazara populations from Afghanistan and Pakistan with 0.9371, 0.9242, and 0.8792, respectively. Overall DYS570 (0.8624) showed the highest or DYS437 (0.2383) showed the lowest gene diversity/heterozygosity for single Y STR markers. Within three populations, single Y-STR markers DYS570 (0.8624), DYS449 (0.8468), DYS627 (0.7949) showed the highest gene diversity/heterozygosities while DYS460 (0.5718), DYS391 (0.3916), and DYS437 (0.2383) showed the lowest gene diversity/heterozygosities in the Baloch and both the Hazara populations from Afghanistan and Pakistan, respectively. After pooling Hazara populations together DYF387S1, DYS437 showed the highest or lowest gene diversity/heterozygosities with 0.9257 and 0.4053 respectively. The observed numbers of alleles were 222, 240, and 188 for Baloch and both the Hazara populations from Afghanistan and Pakistan, respectively on 27 Y STRs.

Allelic frequencies ranged from 0.0090 to 0.6036 in the Baloch population, 0.0038 to 0.6654 in the Hazara population from Afghanistan, and 0.0065 to 0.8627 in the Pakistani Hazara population.

We evaluated forensic parameters at seven levels (**Table 2**), the minimal 9 Y-STRs loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, and DYS385a/b), the extended 11 Y-STRs loci (MHT+DYS438 and DYS439), PowerPlex Y12 STRs loci (extended 11 Y STRs + DYS437), Y-filer 17 STRs loci (PPY12+DYS448, DYS456, DYS458, DYS635, and Y_GATA_H4), Y21STRs loci(Y-filer + DYS481, DYS533, DYS570, and DYS576), Y27 Yfiler Plus loci (21 STRs + DYF387S1, DYS449, DYS460, DYS518, and DYS627), and 6 rapidly mutating Y STRs loci (DYS570, DYS576, DYF387S1, DYS449, DYS518, and DYS627) which are summarized in Table 2. The discrimination capacity (DC) ranged from 87.38% (the minimal 9 Y-STRs loci) to 100% (Y27 Yfiler Plus loci) with random matching probability from 0.0162 (MHT) to 0.009 (Y27 Yfiler Plus loci) and haplotype diversity (HD) ranged 0.9928 (the minimal 9 Y-STRs loci) to 1.0 (Y27 Yfiler Plus loci) in the Baloch population of Pakistan. The discrimination capacity (DC) ranged from 47.06% (the minimal 9 Y-STRs loci) to 99.35% (Y27 Yfiler Plus loci) with random matching probability from 0.0745 (MHT) to 0.0066 (Y27 Yfiler Plus loci) and haplotype diversity (HD) ranged from 0.9316 (the minimal 9 Y-STRs loci) to 0.9999 (Y27 Yfiler Plus loci) in Pakistani Hazara population while DC ranged 41.15% (the minimal 9 Y-STRs loci) to 88.46% (Y27 Yfiler Plus loci) with random matching probability from 0.0329 (MHT) to 0.0057 (Y27 Yfiler Plus loci) and HD ranged from 0.9708 (the minimal 9 Y-STRs loci) to 0.9937 (Y27 Yfiler Plus loci) for Hazara population from Afghanistan. Pooling both populations together DC ranged 40.19% (the minimal 9 Y-STRs loci) to 92% (Y27 Yfiler Plus loci) with random matching probability from 0.0334 (MHT) to 0.0032 (Y27 Yfiler Plus loci) and HD ranged from 0.9689 (the minimal 9 Y-STRs loci) to 0.9992 (Y27 Yfiler Plus loci). Interestingly six rapidly mutating Y STRs which are included in Yfiler plus kit detects high haplotype diversity (Table 2). We have observed 101 (90.99%) different haplotypes out of 111, among them, 95 (85.58%) were unique in the Baloch population and we have observed 139 (90.84%) different haplotypes out of 153, among them 131 (85.62%) were unique in Pakistani Hazara population while in Afghani Hazara population observed haplotypes were 188 (72.30%) out of 260, among them 152(58.46%) were unique. These six STRs (RM Y STRs) showed the almost same diversity, shown by PPY 23 loci. The above results are showing that Yfiler plus kit loci showed strong discrimination capacity, haplotype diversity, and random mating probabilities which provide utility for forensic identification and paternity testing in three ethnic groups (Baloch and Hazara from Pakistan while Hazara from Afghanistan).

2.2Phylogenetic analyses and Population comparisons

Since the anthropological or ethno-historical relationships between studied populations and reference populations which are included for analysis were already known, so we used two different methods on the basis of their similarity with *a priori* expectations. *Fst* is a standardized variance of haplotype frequency and assumes genetic drift as being the agent that differentiates populations. *Rst* is a standardized variance of haplotype size and takes into account both drift and mutation as causes of population differentiation, assuming a stepwise model in which each mutation creates a new allele either by adding or deleting a single repeat unit. To assess the relationship between these three populations (Baloch, Hazara from Pakistan and Afghani Hazaras), and the other relevant populations which are summarized in **Table 1,** pair-wise genetic distances (Rst and Fst) and their corresponding p-values were calculated and were shown in **Supplementary Table 3**. These Rst and Fst values were visualized using hierarchical clustering heat-map (**Supplementary Figure 1 a & b**). Dendrograms give us a clear picture about the organization of the data which can be compared with NJ trees or MDS plots. The utlization of mean-linkage dendrograms to Y STR data gives us a consistent basis of comparison. Heat-map matrix based on Rst values showed that Hazara from Pakistan were clustered more closely to Central and East Asian (i.e. Kazakh and Mongols) populations while the Baloch population was clustered with other Pakistani (i.e. Pathan and Sindhi) populations and Hazara from Afghanistan were clustered with local Afghan populations. On another hand, the heat-map matrix based on Fst values showed that the Hazara population from Pakistan was tightly clustered with local (i.e. Baloch, Arain, and Pathans) populations while the Hazara population from Afghanistan was clustered with Afghanistan Pathan and Northern Talysh population. The observed pattern of inter-population diversity from Rst was in support of anthropological knowledge, while that based on Fst revealed unexpected and unconvincing population affinities. These results are consistent with our previous study results [31]. The pairwise Rst genetic distances values between Baloch and other relevant populations ranged from -0.0402 to 0.1417. According to Rst values, the Baloch population of Pakistan showed the closest genetic distance to Turks (-0.0402) from Ardabil, Iran while Kazakh (0.1417) from Gansu, China showed the greatest genetic distance. For the Afghan Hazara population, the Afghan population (0.0009) from Afghanistan showed the closest genetic distance and for the Pakistani Hazara group, the Afghan population (0.0381) from Afghanistan showed the closest genetic distance To investigate the paternal relationship among these three and other reference populations, we have generated the MDS plot (figure 1) based on pairwise Rst matrix from supplementary table 3. In the MDS plot, we have seen that the Hazara population from Afghanistan is located closer to the Afghan population from Afghanistan and the Pathan population from northern Afghanistan which is similar to the results of another study [32], while Pakistani Hazara lined closer to Kazakh and Mongolian population which is similar to our previous study's results[27,33].

According to Fst values, the Afghan Hazara population is closest to the Afghan population (0.0053) followed by the Hazara population from Balochistan, Pakistan (0.0057), and Iranian population from Mashhad, Iran (0.0077). Evolutionary relationships between the Baloch and Hazara population of Pakistan, the Hazara population from Afghanistan, and other reference populations were inferred from the Neighbor-joining tree based on Fst values (**Figure 2**). In neighbor-joining trees, an admixed population will always lie on the path between the source populations[34]. In total, we have observed 14 clusters for 62 populations in NJ-tree and the Baloch population placed itself in the second cluster along with West-south Asian populations. Hazara populations from Pakistan and Afghanistan came to the fourth cluster along with the Afghani and Iranian populations. The pattern of inter-population diversity based on *Rst* was consistent with ethnohistorical and anthropological knowledge, while that based on *Fst* shown surprising and unaccepted population affinities.

2.3 Inference of ancestry based on Y STRs

The Y haplogroups were predicted using the online Y-haplogroup predictor software (http://www.nevgen.org/). C2 (previously known as C3-Star cluster) was the most frequent haplogroup in Pakistani and Afghan Hazaras.

The median-joining network of haplotypes (**Figure 3**) showed a bulky central star-like cluster which represents predicated haplogroup M217 and another big cluster representing haplogroup M420 and comprises many of the identical or highly similar haplotypes. These types of features are usually inferred as past male-lineage expansions[35]. Star-like features of haplotypes comprising haplogroup M217 (C2) have been reported previously in Hazara, Mongol, and Kazakh populations[27,33,36]. An explanation about its origin in Mongolia was about ~1,000 years ago [36]. The frequency of R haplogroup in the Baloch population is 36.03%, 22.22% in Pakistani Hazara, and 21.15% in Afghani Hazara. This haplogroup originated in north Asia about 27,000 years ago (http://isogg.org/tree/index.html). R is one of the most frequent haplogroups in Europe, with its branches reaching 80% of the population in some regions. One branch is believed to have originated in the Kurgan culture, known to be the first speakers of the Indo-European languages and responsible for the domestication of the horse[37]. From somewhere in Central Asia, some descendants of the man carrying the M207 mutation on the Y chromosome headed south to arrive in India about 10,000 years ago[38]. This is one of the frequent haplogroups in Pakistan and North India. In the Baloch population frequency of haplogroup L1 is 22.5% and 1.53% in Afghani Hazara. In sub-continental populations its frequency is about 7–15%[39,40]. Genetic studies suggest that this may be one of the original haplogroups of the creators of Indus Valley

Civilization[41,42]. The frequency of L1 is about 28% in Pakistan and Baluchistan, from where the agricultural creators of this civilization emerged[43]. The origins of this haplogroup can be traced to the rugged and mountainous Pamir Knot region in Tajikistan[38].

In an earlier study[36], the star-cluster (C3) profile for DYS389I-DYS389b-DYS390-DYS391-DYS392-DYS393-DYS388-DYS425-DYS426-DYS434 -DYS435-DYS436-DYS437-DYS438-DYS439 was 10-16-25-10-11-13-14-12-11-11-11-12-8-10-10. In present study mostly occurring haplotype for loci DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439 was 15-13-29-24-10-11-13-14-11-12 which repeated itself in 43 individuals while 14-13-29-24-8-11-13-14-11-11 repeated in 9 individuals and 15-13-29-24-11-11-13-14-11-12 repeated in 8 individuals in Pakistani Hazara population while in Afghani Hazara 16-13-29-25-10-11-13-14-10-10, 15-13-29-24-10-11-13-14-11-12, 14-12-28-23-10-11-12-15-9-11, 14-13-29-24-11-13-12-15-12-12 and 15-14-32-25-11-11-13-14-9-10 haplotypes were repeated in 30, 17, 15, 12 and 11 individuals, respectively. The occurrence of these haplotypes were previously observed in Mongols and Kazakhs[35]. Allelic ranges of Kazak[35] population from Kazakhstan Central Asia were similar while Mongol population from Inner Mongolia were almost similar on above mentioned 10 Y STRs. In our earlier study[31], results showed that Hazaras have a close genetic affinity with Turkic-speaking (Kazakh, Kyrgyz and Uyghur) and Mongolian people. Admixture and outgroup findings further clarified that Hazara have 57.8% gene pool from Mongolians.

Here we also speculated a hypothesis that is based on hearsay that Hazaras living in Pakistan are more conserved and they only mate with the Hazaras while across the Durand line the Hazaras mate with other ethnic groups in Afghanistan. Results of gene diversity/heterozygosity and F-statistics tests are also supporting this hypothesis. According to results, all loci showed more diversity in the Hazara population from Afghanistan when compared with the Hazara population from Pakistan (**Figure 4**). F-statistics test within Hazara populations showed variations at four loci only (DYS393- 0.05002, DYS449- 0.01694, DYF387S1- 0.00662 and DYS385a/b- 0.00004) (**Supplementary Table 4**). These variations may be the sampling effect, population diversity, or maybe geographical boundaries. LDA is a transformation technique which is commonly used to understand genome diversity and was performed on the Hazara population, Central Asian, South Asian including the Baloch population, East Asian, and Russian population samples to explore their genetic homology. **Figure 5** shows all individual samples plotted on the two LDA factors (F1 and F2). LDA Plot showed the association of the Hazara population with East and Central Asian populations.

2.4 Physical characterization of DYS448 deletions

By using the Yfiler plus kit, we have observed the null allele at DYS448 in 29 individuals in the Hazara population from Afghanistan (**Figure 6**). Certain factors can cause the phenomena of null alleles and these are deletions within the target region, primer binding sites problem that destabilize hybridization of at least one of the primers flanking the target region [44–47]. This phenomenon was previously reported, in which other commercial kits were used [48–53]. The current population study represents the highest frequencies of the null allele at DYS448 when compared with the previously reported population to date (**Table 3**). The core repeat motif of the DYS448 locus is the hexanucleotide repeat AGAGAT[54]. DYS448 has two polymorphic domains separated by an invariant 42-bp region.

We have observed 29 null alleles among these, long deletions were covering at a minimum the N42 region and the core AGAGAT repeats downstream, and small deletions encompassing upstream repeats as well (all alignments were based on allele 20). Observed null alleles at locus DYS448 in 29 individuals from the Hazara population of Afghanistan, which were later confirmed with the GoldenEye Y20 System kit were successfully amplified using self-designed primers and sequenced (**Supplementary Table 5**) which were submitted to genbank under accession numbers MN623385 to MN623413. Overall we have observed 55 null alleles at DYS448 in the Hazara population from Pakistan and Afghanistan. Interestingly, all individuals (55) who showed deletion at DYS448 belongs to haplogroup C2 which is most frequent haplogroup in Mongol and Kazakh populations. This high frequency of allele drop-out / mutation is DYS448 in Hazara population from Pakistan and Afghanistan strongly support the evidence that they have Kazakh and Mongol origin. Whole genome or Y Chromosomal sequencing is required to get more insight of this polymorphism. The frequency of the null allele at DYS448 is more frequent in Asia more specifically in East and Central Asia when compared to the rest of the world[26,49]. The commercial companies should pay special attention while designing DYS448 primers.

2.5 Concluding Remarks

Finally, our study demonstrates that the Yfiler plus kit detects high haplotype diversity in Baloch population from Pakistan and Hazara populations from across the Durand line **(**Pakistan and Afghanistan) of which two (Baloch and Afghani Hazara) were not previously studied at Yfiler plus STR loci, which in general makes it suitable for forensic casework in these groups. The recent inclusion of these data in the YHRD allows widespread use for forensic and other purposes.

3. MATERIALS AND METHODS

3.1 Samples

A total of 524 blood samples were collected, in which 111 Balochi individuals from Baluchistan Pakistan, 153 from Hazara Town Quetta, Baluchistan Pakistan (Participants were part of an earlier study [27] and were agreed to the secondary use of their DNA samples), and 260 from Bamyan, Afghanistan. All participants who were included in this study were unrelated individuals of at least three generations. All participants gave their informed consent either orally and with thumbprint (in case they could not write) or in writing after the study aims and procedures were carefully explained to them. This collaborative study was approved by the ethical review boards of China Medical University, Shenyang, Liaoning Province, People's Republic of China (2019/067-P), University of Health Sciences Lahore Pakistan (2017-CMU-1/14), and Ministry of Public Health, Forensic Medicine Directorate, Kabul, Afghanistan (FC-2017-02). All the experimental procedures were performed in accordance with the standards of the Declaration of Helsinki.

3.2. DNA extraction

Axygen AxyPrep Blood Genomic DNA Miniprep Kit was used to extract genomic DNA according to the manufacturer's protocol (Axygen Biosciences; CA, USA).

3.3 PCR Amplification

DNA was amplified using Yfiler Plus PCR Amplification Kit (Thermo Fisher Scientific) PCR amplification was carried out using the Applied Biosystems GeneAmp PCR System 9700 thermal cyclers. PCR amplifications were performed as recommended by the manufacturer, although using half of the recommended reaction volume (12.5 μl).

3.4. 27Y-STRs genotyping

After successful PCR amplification, The PCR products were analyzed by using an 8 capillary ABI 3500 DNA Genetic Analyzer with POP-4 polymer (Life Technologies) according to the manufacturer's protocol. GeneMapper Software version 4.0 (Life Technologies) was used for the genotype assignment. DNA typing was performed according to the manufacturer's protocol by using the locus panel and allele bins supplied by the manufacturer and allele designations corresponding with the allelic ladder supplied by the manufacturer. Genotype nomenclature was based on the recommendations of the International Society for Forensic Genetics [55].

3.5. Confirmation of Null DYS 448

For the confirmation of samples that showed no allele call at DYS448, they were re-amplified by using the Goldeneye 20Y amplification kit (Goldeneye Technology Ltd.). After confirmed with two different kits (Yfiler Plus and GoldenEye 20Y), these samples were amplified and sequenced as described elsewhere [27].

3.6. Quality control

Our laboratory has participated and passed the YHRD quality assurance exercise 2015. Haplotype data were already made accessible via the Y-chromosome Haplotype Reference Database (YHRD) under accession number YA004595 (Balochi) in 61st release on dated 2019 June 24, YA004312-2 (Hazara Pakistan) and YA004503 (Hazara Afghanistan) in 59th release on dated 2018 November 01. 29 sequenced samples at null allele call at DYS448 were also submitted to genbank under accession numbers MN623385 to MN623413 on dated 2019 october 28.

3.7. Statistical analysis

Allelic and haplotype frequencies were computed by direct counting method and haplotype diversity (HD) was calculated according to:

$$\mathrm{HD} = \frac{n}{n-1}\left(1 - \sum_i p_i^2\right)$$

where n is the male population size and p_i is the frequency of ith haplotype. Discrimination capacity (DC) was calculated as the ratio of unique haplotypes in the samples. Match probabilities (MP) were calculated as $\Sigma\ \mathrm{P}i^2$, where Pi is the frequency of the i-th haplotype. Genetic distances were evaluated using the Rst[56] and Fst[57–59] statistic, between reference populations and currently studied populations on overlapping STRs (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y_GATA_H4) were calculated by using Arlequin Software v3.5[60]. We calculated both Rst and Fst values because in the generalized stepwise mutation model, Rst offers relatively unbiased evaluations of migration rates and times of population divergence while on other hand Fst tends to show too much population similarity, predominantly when migration rates are low or divergence times are long[56]. Reduced dimensionality spatial representation of the populations was performed based on Rst values using multi-dimensional scaling (MDS) with IBM SPSS Statistics for Windows, Version 23.0 (IBM Corp., Armonk, NY, USA). Heatmaps were generated using Rst and Fst values were generated using R program V3.4.1 platform with the help of a ggplot2 module.

Phylogenetic analysis:

A neighbor-joining phylogenetic tree was constructed for the Hazara and the reference populations based on a distance matrix of Fst using the Mega7 software[61]. We also predicted Y-SNP haplogroups in the samples from Y STR haplotypes (Yfiler STRs) using the Y-DNA

Haplogroup Predictor NEVGEN (http://www.nevgen.org). We have used FTDNA order for 17 Y STRs (Yfiler loci). The microvariant alleles were truncated to the next lowest integer value since values in the database were treated similarly. Any haplotypes which have null alleles or duplication variants in the Baloch or Hazara population from Pakistan or Afghanistan were excluded from the analysis. The results of NEVGEN were cross checked with Athey's Haplogroup Predictor (http://www.hprg.com/hapest5/index.html).

Linear discriminant analysis

R program V3.4.1 platform with the help of a ggplot2 module was used to perform linear discriminant analysis (LDA) for Hazara (Pakistan), Hazara (Afghanistan), Central Asia, East Asia, the Middle East, and Southwest Asian (Baloch) samples [62] on overlapping (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y_GATA_H4) STRs . The multi-copy marker like (DYS385ab) and haplotypes that have null alleles or duplication variants in the Baloch or Hazara population or any of the reference populations were excluded from the analysis. For DYS389I and DYS389II, we have subtracted DYS389I from DYS389II and used DYS389II-I for analysis.

The median-joining network

To define the genetic relationships among Balochi and Hazara individuals for 20 Y STRs (DYS19, DYS389II-I, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS448, DYS456, DYS458, DYS635, Y_GATA_H4, DYS549, DYS460, DYS481, DYS533, DYS570, DYS576, DYS627), we used the stepwise mutation model and Median Joining-Maximum

Parsimony algorithm [63] by using the program Network 5 as described at the Fluxus Engineering website (http://www.fluxus-engineering.com), and the weighting criteria for Y-STRs following [27]. Any haplotypes which have null alleles or duplication variants in the Baloch or Hazara population from Pakistan or Afghanistan were excluded from the analysis.

COMPETING FINANCIAL INTERESTS

None.

AUTHOR CONTRIBUTION

J.L. and A.A. designed this study. A.A., A.R., and S.N. and M.R., collected the samples. A.A. experimented and wrote the manuscript. A.A., J.L., A.R., S.N., R.A., S.W., and C.W., analyzed the results. A.A., and J.L., modified the manuscript. All authors reviewed the manuscript.

COMPLIANCE WITH ETHICAL STANDARDS

The study was approved (2019/067-P) by the ethical review board of China Medical University, Shenyang, Liaoning Province, People's Republic of China, and in accordance with the standards of the Declaration of Helsinki. All participants who were included in this study were unrelated individuals of at least three generations. All participants gave their informed consent either orally and with thumbprint (in case they could not write) or in writing after the study aims and procedures were carefully explained to them.

ACKNOWLEDGMENTS

We thank all volunteers who provided material and data for this project, especially Guanglin He and Abul-Hasan Fawad. This study was financially supported by the China Medical University postdoctoral research grant (100/1210619014).

Legends of figures and tables:

Figure 1: Two-dimensional plot from multi-dimensional scaling analysis of R*st*-values based on Yfiler haplotypes for the Baloch population of Pakistan and Hazara populations across the Durand line with reference populations.

Figure 2: Neighbor-joining tree based on the F*st* values between the Baloch population of Pakistan and Hazara populations across the Durand line with reference populations.

Figure 3: The median-joining network of the Baloch population of Pakistan and Hazara populations across the Durand line based on 20 Y STRs.

Figure 4: Heterozygosity scattered plot for three populations

Figure 5: LDA Analysis between the Baloch population of Pakistan and Hazara populations across the Durand line, Central Asia, South Asia, Russia, and East Asian populations.

Figure 6: Electropherogram of an individual showing null type at DYS448.

Table 1: Reference Populations from Central, Eastern and South Asia populations selected as reference populations used in LDA, NJ tree and multidimensional scaling (MDS) analysis.

Table 2: Forensic parameters on 7 different levels in three ethnic groups

Table 3: Frequencies of the null allele at DYS448 in various ethnic groups across continents

Electronic Supplementary Materials (ESM):

Supplementary Figure 1: Heatmap generated using Rst and Fst values.

Supplementary Table 1: Raw genotypic data of 3 ethnic groups typed with Yfiler plus

Supplementary Table 2: Allele Frequencies and Forensic Parameters 3 ethnic groups

Supplementary Table 3: Pairwise R*st* and F*st* values between 3 ethnic groups and other reference populations

Supplementary Table 4: F-statistics analysis between Hazara population from Pakistan and Afghanistan

Supplementary Table 5: Sequence in the relevant flanking and repeat region of the DYS448 locus for null alleles.

References:

1. Oppenheimer, S. Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 770–784 (2012).
2. Adnan, A., Ralf, A., Rakha, A., Kousouri, N. & Kayser, M. Improving empirical evidence on differentiating closely related men with RM Y-STRs: A comprehensive pedigree study from Pakistan. *Forensic Sci Int Genet* **25**, 45–51 (2016).
3. Adnan, A. *et al.* Population data of 17 Y-STRs (Yfiler) from Punjabis and Kashmiris of Pakistan. *International Journal of Legal Medicine* (2017) doi:10.1007/s00414-017-1611-9.
4. Adnan, A., Rakha, A., Lao, O. & Kayser, M. Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan. *Forensic Science International: Genetics* (2018) doi:10.1016/j.fsigen.2018.07.001.
5. Ballantyne, K. N. *et al.* Toward male individualization with rapidly mutating y-chromosomal short tandem repeats. *Hum. Mutat.* **35**, 1021–1032 (2014).
6. Prinz, M., Boll, K., Baum, H. & Shaler, B. Multiplexing of Y chromosome specific STRs and performance for mixed samples. *Forensic Science International* **85**, 209–218 (1997).
7. Ballantyne, K. N. & Kayser, M. Additional Y-STRs in Forensics: Why, Which, and When. *Forensic Sci Rev* **24**, 63–78 (2012).
8. Calacal, G. C. *et al.* Identification of Exhumed Remains of Fire Tragedy Victims Using Conventional Methods and Autosomal/Y-Chromosomal Short Tandem Repeat DNA Profiling: *The American Journal of Forensic Medicine and Pathology* **26**, 285–291 (2005).
9. Vermeulen, M. *et al.* Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. *Forensic Science International: Genetics* **3**, 205–213 (2009).

10. Krenke, B. E. *et al.* "Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex" [Forensic Sci. Int. 148 (1) (2005) 1–14]. *Forensic Science International* **151**, 111–124 (2005).

11. Mulero, J. J. *et al.* Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system. *J. Forensic Sci.* **51**, 64–75 (2006).

12. Thompson, J. M. *et al.* Developmental validation of the PowerPlex® Y23 System: A single multiplex Y-STR analysis system for casework and database samples. *Forensic Science International: Genetics* **7**, 240–250 (2013).

13. Kayser, M. *et al.* Evaluation of Y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine* **110**, 125–133 (1997).

14. Gopinath, S. *et al.* Developmental validation of the Yfiler ® Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications. *Forensic Science International: Genetics* **24**, 164–175 (2016).

15. Jobling, M. A. *et al.* Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. *Hum. Mol. Genet.* **5**, 1767–1775 (1996).

16. Jobling, M. A. *et al.* Structural variation on the short arm of the human Y chromosome: recurrent multigene deletions encompassing Amelogenin Y. *Human Molecular Genetics* **16**, 307–316 (2007).

17. Repping, S. *et al.* Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nature Genetics* **35**, 247–251 (2003).

18. Bosch, E. & Jobling, M. A. Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. *Hum. Mol. Genet.* **12**, 341–347 (2003).

19. Repping, S. *et al.* High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nature Genetics* **38**, 463–467 (2006).

20. Verma, R. S., Rodriguez, J. & Dosik, H. The clinical significance of pericentric inversion of the human Y chromosome: a rare 'third' type of heteromorphism. *J. Hered.* **73**, 236–238 (1982).

21. Affara, N. A. *et al.* Variable transfer of Y-specific sequences in XX males. *Nucleic Acids Res.* **14**, 5375–5387 (1986).

22. Bernstein, R., Wadee, A., Rosendorff, J., Wessels, A. & Jenkins, T. Inverted Y chromosome polymorphism in the Gujerati Muslim Indian population of South Africa. *Hum. Genet.* **74**, 223–229 (1986).

23. Page, D. C. Sex reversal: deletion mapping the male-determining function of the human Y chromosome. *Cold Spring Harb. Symp. Quant. Biol.* **51 Pt 1**, 229–235 (1986).

24. Budowle, B. *et al.* Null allele sequence structure at the DYS448 locus and implications for profile interpretation. *Int. J. Legal Med.* **122**, 421–427 (2008).

25. Westen, A. A. *et al.* Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males. *Forensic Science International: Genetics* **14**, 174–181 (2015).

26. Balaresque, P. *et al.* Dynamic nature of the proximal *AZFc* region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis. *Human Mutation* **29**, 1171–1180 (2008).

27. Adnan, A. *et al.* Genetic characterization of Y-chromosomal STRs in Hazara ethnic group of Pakistan and confirmation of DYS448 null allele. *International Journal of Legal Medicine* (2018) doi:10.1007/s00414-018-1962-x.

28. Siddique, A. *Afghanistan's Ethnic Divides*. www.cidobafpakproject.com (2012).

29. *Afghanistan: a country study*. (Claitor's Pub. Division, c2001).

30. Dashti, N. *The Baloch and Balochistan: a historical account from the beginning to the fall of the Baloch state*. (Trafford, 2012).

31. He, G. *et al.* A comprehensive exploration of the genetic legacy and forensic features of Afghanistan and Pakistan Mongolian-descent Hazara. *Forensic Science International: Genetics* **42**, e1–e12 (2019).

32. Haber, M. *et al.* Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. *PLoS ONE* **7**, e34288 (2012).

33. Adnan, A. *et al.* Phylogenetic relationship and genetic history of Central Asian Kazakhs inferred from Y-chromosome and autosomal variations. *Mol Genet Genomics* (2019) doi:10.1007/s00438-019-01617-0.

34. Kopelman, N. M., Stone, L., Gascuel, O. & Rosenberg, N. A. The behavior of admixed populations in neighbor-joining inference of population trees. *Pac Symp Biocomput* 273–284 (2013).

35. Tarlykov, P. V. *et al.* Mitochondrial and Y-chromosomal profile of the Kazakh population from East Kazakhstan. *Croat. Med. J.* **54**, 17–24 (2013).

36. Zerjal, T. *et al.* The Genetic Legacy of the Mongols. *The American Journal of Human Genetics* **72**, 717–721 (2003).

37. Smolenyak, M. & Turner, A. *Trace your roots with DNA: using genetic tests to explore your family tree*. (Rodale; Distributed to the trade by Holtzbrinck Publishers, 2004).

38. Wells, S. *Deep ancestry: inside the genographic project*. (National Geographic, 2007).

39. Basu, A., Sarkar-Roy, N. & Majumder, P. P. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc Natl Acad Sci USA* **113**, 1594–1599 (2016).

40. Di Cristofaro, J. *et al.* Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. *PLoS ONE* **8**, e76748 (2013).

41. Mcelreavey, K. & Quintana-Murci, L. A population genetics perspective of the Indus Valley through uniparentally-inherited markers. *Annals of Human Biology* **32**, 154–162 (2005).

42. Sengupta, S. *et al.* Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists. *The American Journal of Human Genetics* **78**, 202–221 (2006).

43. Qamar, R. *et al.* Y-Chromosomal DNA Variation in Pakistan. *The American Journal of Human Genetics* **70**, 1107–1124 (2002).

44. Chang, C.-W., Mulero, J. J., Budowle, B., Calandro, L. M. & Hennessy, L. K. Identification of a Novel Polymorphism in the X-Chromosome Region Homologous to the DYS456 Locus. *Journal of Forensic Sciences* **51**, 344–348 (2006).

45. Collins, F. S., Brooks, L. D. & Chakravarti, A. A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.* **8**, 1229–1231 (1998).

46. Fredman, D. HGVbase: a curated resource describing human DNA variation and phenotype relationships. *Nucleic Acids Research* **32**, 516D – 519 (2004).

47. NCBI Resource Coordinators *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **46**, D8–D13 (2018).

48. Chang, Y. M., Perumal, R., Keat, P. Y. & Kuehn, D. L. C. Haplotype diversity of 16 Y-chromosomal STRs in three main ethnic populations (Malays, Chinese and Indians) in Malaysia. *Forensic Sci. Int.* **167**, 70–76 (2007).

49. Park, M. J. *et al.* Characterization of Deletions in the DYS385 Flanking Region and Null Alleles Associated with AZFc Microdeletions in Koreans. *Journal of Forensic Sciences* **53**, 331–334 (2008).

50. Parkin, E. J. *et al.* Diversity of 26-locus Y-STR haplotypes in a Nepalese population sample: Isolation and drift in the Himalayas. *Forensic Science International* **166**, 176–181 (2007).

51. Mizuno, N. *et al.* 16 Y chromosomal STR haplotypes in Japanese. *Forensic Science International* **174**, 71–76 (2008).

52. Roewer, L. *et al.* Y-chromosomal STR haplotypes in Kalmyk population samples. *Forensic Science International* **173**, 204–209 (2007).

53. Sánchez, C. *et al.* Haplotype frequencies of 16 Y-chromosome STR loci in the Barcelona metropolitan area population using Y-Filer^{TM} kit. *Forensic Science International* **172**, 211–217 (2007).

54. Redd, A. J. *et al.* Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci. Int.* **130**, 97–111 (2002).

55. Roewer, L. *et al.* DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR results in forensic analysis. *Forensic Science International: Genetics* **48**, 102308 (2020).

56. Slatkin, M. A measure of population subdivision based on microsatellite allele frequencies. *Genetics* **139**, 457–462 (1995).

57. Weir, B. S. & Cockerham, C. C. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* **38**, 1358 (1984).

58. Michalakis, Y. & Excoffier, L. A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* **142**, 1061–1064 (1996).

59. Reynolds, J., Weir, B. S. & Cockerham, C. C. Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* **105**, 767–779 (1983).

60. Excoffier, L. & Lischer, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources* **10**, 564–567 (2010).

61. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**, 1870–1874 (2016).

62. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2015).

63. Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution* **16**, 37–48 (1999).

64.? Igen?s, C. & Tillmar, A. O. Population genetics of 29 autosomal STRs and 17 Y-chromosomal STRs in a population sample from Afghanistan. *International Journal of Legal Medicine* **128**, 279–280 (2014).

65. Achakzai, N. M. *et al.* Y-chromosomal STR analysis in the Pashtun population of Southern Afghanistan. *Forensic Sci Int Genet* **6**, e103-105 (2012).

66. Lacau, H. *et al.* Y-STR profiling in two Afghanistan populations. *Leg Med (Tokyo)* **13**, 103–108 (2011).

67. Nothnagel, M. *et al.* Revisiting the male genetic landscape of China: a multi-center study of almost 38,000 Y-STR haplotypes. *Human Genetics* **136**, 485–497 (2017).

68. Kwak, K. D. *et al.* Y-chromosomal STR haplotypes and their applications to forensic and population studies in east Asia. *International Journal of Legal Medicine* **119**, 195–201 (2005).

69. Zhang, D. *et al.* RETRACTED ARTICLE: Y Chromosomal STR haplotypes in Chinese Uyghur, Kazakh and Hui ethnic groups and genetic features of DYS448 null allele and DYS19 duplicated allele. *Int J Legal Med* (2019) doi:10.1007/s00414-019-02049-6.

70. Shan, W. *et al.* Genetic polymorphism of 17 Y chromosomal STRs in Kazakh and Uighur populations from Xinjiang, China. *Int J Legal Med* **128**, 743–744 (2014).

71. Wang, C. *et al.* Genetic polymorphisms of 27 Yfiler® Plus loci in the Daur and Mongolian ethnic minorities from Hulunbuir of Inner Mongolia Autonomous Region, China. *Forensic Science International: Genetics* **40**, e252–e255 (2019).

72. Fu, X. *et al.* Genetic polymorphisms of 26 Y-STR loci in the Mongolian minority from Horqin district, China. *International Journal of Legal Medicine* **130**, 941–946 (2016).

73. Ou, X. *et al.* Haplotype analysis of the polymorphic 40 Y-STR markers in Chinese populations. *Forensic Science International: Genetics* **19**, 255–262 (2015).

74. Zhu, B. *et al.* Y-STRs haplotypes of Chinese Mongol ethnic group using Y-PLEX™ 12. *Forensic Science International* **153**, 260–263 (2005).

75. Bian, Y. *et al.* Analysis of genetic admixture in Uyghur using the 26 Y-STR loci system. *Scientific Reports* **6**, (2016).

76. Roewer, L., Willuweit, S., Stoneking, M. & Nasidze, I. A Y-STR database of Iranian and Azerbaijanian minority populations. *Forensic Sci Int Genet* **4**, e53-55 (2009).

77. Sayyari, M., Salehzadeh, A., Tabatabaiefar, M. A. & Abbasi, A. Profiling of 17 Y-STR loci in Mazandaran and Gilan provinces of Iran. *Turk J Med Sci* **49**, 1277–1286 (2019).

78. Sayyari, M., Salehzadeh, A., Tabatabaiefar, M. A. & Abbasi, A. Genetic polymorphisms of Y-chromosome short tandem repeats (Y-STRs) in a male population from Golestan province, Iran. *Mol Biol Res Commun* (2020) doi:10.22099/mbrc.2020.35547.1462.

79. Nasidze, I., Schädlich, H. & Stoneking, M. Haplotypes from the Caucasus, Turkey and Iran for nine Y-STR loci. *Forensic Science International* **137**, 85–93 (2003).

80. Javed, F. *et al.* Male individualization using 12 rapidly mutating Y-STRs in Araein ethnic group and shared paternal lineage of Pakistani population. *Int. J. Legal Med.* **132**, 1621–1624 (2018).

81. Ullah, I. *et al.* High Y-chromosomal Differentiation Among Ethnic Groups of Dir and Swat Districts, Pakistan. *Ann. Hum. Genet.* **81**, 234–248 (2017).

82. Adnan, A. *et al.* Population data of 17 Y-STRs (Yfiler) from Punjabis and Kashmiris of Pakistan. *International Journal of Legal Medicine* (2017) doi:10.1007/s00414-017-1611-9.

83. Lee, E. Y. *et al.* Analysis of 22 Y chromosomal STR haplotypes and Y haplogroup distribution in Pathans of Pakistan. *Forensic Science International: Genetics* **11**, 111–116 (2014).

84. Adnan, A. *et al.* Genetic structure and forensic characteristics of Saraiki population from Southern Punjab, Pakistan, revealed by 20 Y-chromosomal STRs. *Int J Legal Med* **134**, 977–979 (2020).

85. Perveen, R., Shahid, A. A., Shafique, M., Shahzad, M. & Husnain, T. Genetic variations of 15 autosomal and 17 Y-STR markers in Sindhi population of Pakistan. *International Journal of Legal Medicine* **131**, 1239–1240 (2017).

86. Zhabagin, M. *et al.* Development of the Kazakhstan Y-chromosome haplotype reference database: analysis of 27 Y-STR in Kazakh population. *Int J Legal Med* **133**, 1029–1032 (2019).

87. Aliferi, A. *et al.* UK and Irish Y-STR population data—A catalogue of variant alleles. *Forensic Science International: Genetics* **34**, e1–e6 (2018).

88. Ali, N., Coulson-Thomas, Y. M., Norton, A. L., Dixon, R. A. & Williams, D. R. Announcement of population data: genetic data for 17 Y-STR AmpFℓSTR® Yfiler™ markers from an immigrant Pakistani population in the UK (British Pakistanis). *Forensic Sci Int Genet* **7**, e40-42 (2013).

89. Mizuno, N. *et al.* 16 Y chromosomal STR haplotypes in Japanese. *Forensic Science International* **174**, 71–76 (2008).

90. Gutiérrez-Alarcón, A. B., Moguel-Torres, M., León-Jiménez, A. K., Cuéllar-Nevárez, G. E. & Rangel-Villalobos, H. Allele and haplotype distribution for 16 Y-STRs (AmpFlSTR® Y-filer™ kit) in the state of Chihuahua at North Center of Mexico. *Legal Medicine* **9**, 154–157 (2007).

Table 1: Reference Populations from Central, Eastern and South Asia populations selected as reference populations used in LDA, NJ tree and multidimensional scaling (MDS) analysis.

No	Population	Haplotypes	Country
1	Afghan	152	Afghanistan[64]
2	Pathan	125	Afghanistan[65]
3	Pathan	44	North Afghanistan[66]
4	Pathan	142	South Afghanistan[66]
5	Farsi, Arab	35	Ahvaz, Iran (accession no YA004581)
6	Farsi, Azerbaijani	34	Hamedan, Iran (accession no YA004582)
7	Han	934	Beijing, China[67,68]
8	Kazakh	231	Xinjiang, China[69,70]
9	Mongol	272	Hulun Buir, China[71]
10	Mongol	454	Inner Mongolia, China[72–74]
11	Uighur	732	Xinjiang, China[75]
12	Arab	33	Ahvaz, Iran[76]
13	Azerbaijani	39	Tabriz, Iran (accession no YA004586)
14	Azerbaijani	50	Urmia, Iran (accession no YA004587)
15	Bakthiari	45	Izeh, Iran[76]
16	Baloch	19	Balochistan, Iran (accession no YA003794)
17	Baloch	59	Zahedan, Iran (accession no YA004238)
18	Farsi	286	Tehran, Iran (accession no YA004580)
19	Gilak	98	Gilan, Iran[77]
20	Gilaki	42	Rasht, Iran[76]
21	Iranian	27	Birjand, Iran (accession no YA003902)
22	Iranian	152	Central Iran, Iran (accession no YA003782)
23	Iranian	106	Fars, Iran (accession no YA004229)
24	Iranian	106	Golestan, Iran[78]
25	Iranian	94	Iran (accession no YA004237)
26	Iranian	161	Isfahan, Iran[79]
27	Iranian	127	Mashhad, Iran (accession no YA003903)
28	Kurd	51	Iran (accession no YA004244)
29	Kurdish	77	Kermanshah, Iran (accession no YA004584)
30	Kurdish	73	Kurdistan, Iran (accession nos YA003795 and YA004585)
31	Lor	37	Lorestan, Iran (accession nos YA003796 and YA004243)
32	Lurs	9	Kohgiluyeh-Buyer Ahmad, Iran (accession no YA003797)
33	Mazandarani	44	Sari, Iran[76]

34	Mazani	126	Mazan daran, Iran[77]
35	Parsee	17	Fars, Iran (accession no YA003798)
36	Qashqaee	15	Fars, Iran (accession no YA003799)
37	Sistani	64	Zabol, Iran (accession no YA004241)
38	Talysh	15	Masal, IranSouth[76]
39	Turk	11	Ardabil, Iran (accession no YA004240)
40	Zoroastrian	6	Yazd, Iran (accession no YA003800)
41	Luri	60	Ilam, Iran Kurdish (accession no YA004583)
42	Arain	85	Punjab, Pakistan[80]
43	Baloch	98	Balochistan, Pakistan (accession no YA004595)
44	Gujjar	20	Swat and Dir District, Pakistan[81]
45	Hazara	160	Balochistan, Pakistan [27]
46	Kashmiri	175	Azad Kashmir, Pakistan [82]
47	Kohistani	20	Swat and Dir District, Pakistan [81]
48	Pathan	269	Pakistan[83]
49	Punjabi	383	Punjab, Pakistan[3]
50	Roma	278	Punjab, Pakistan (accession no YA004554)
51	Saraiki	51	Southern Punjab, Pakistan (accession no YA004225)
52	Saraki	148	Punjab, Pakistan [84]
53	Sindhi	98	Sindh, Pakistan[85]
54	Yousafzai Pathan	71	KhyberPakhtunkhwa, Pakistan (accession no YA003748)
55	Tharklani, Pashtun	20	Swat and Dir District, Pakistan[81]
56	Uthmankheil, Pashtun	20	Swat and Dir District, Pakistan[81]
57	Yousafzai, Pashtun	20	Swat and Dir District, Pakistan[81]
58	Urdu, Punjabi	241	Punjab, Pakistan (accession no YA004381)
59	Kazakh	305	Kazakhstan [86]
60	Kazakh	67	East Kazakhstan, Kazakhstan (accession no YA003700)
61	Kazakh	99	South Kazakhstan, Kazakhstan (accession no YA003729)
62	Southwest Asian	493	United Kingdom[87]
63	British Pakistani	132	United Kingdom[88]
Total haplotypes		**8457**	

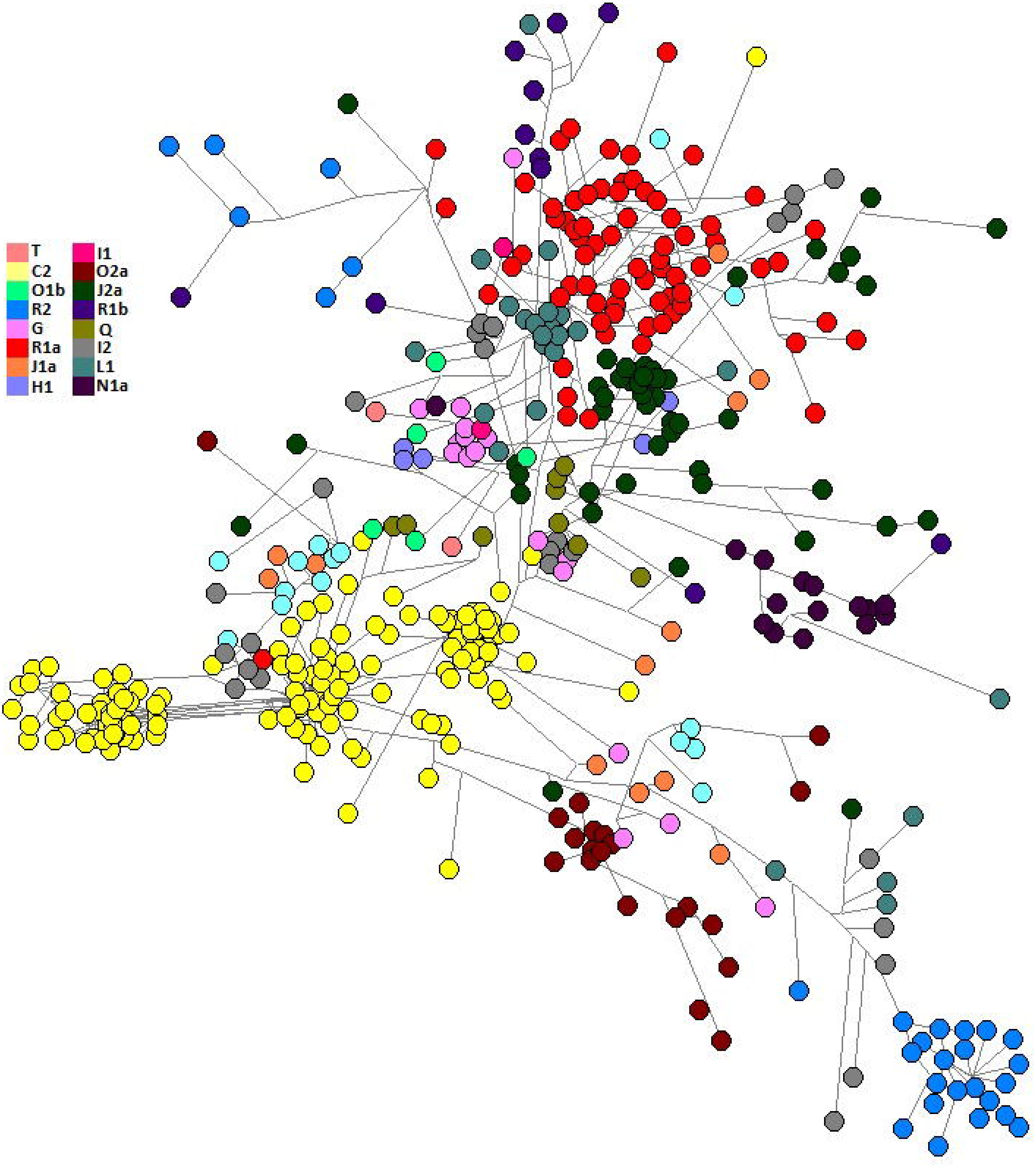
Table 2: Forensic parameters on 7 different levels in three ethnic groups

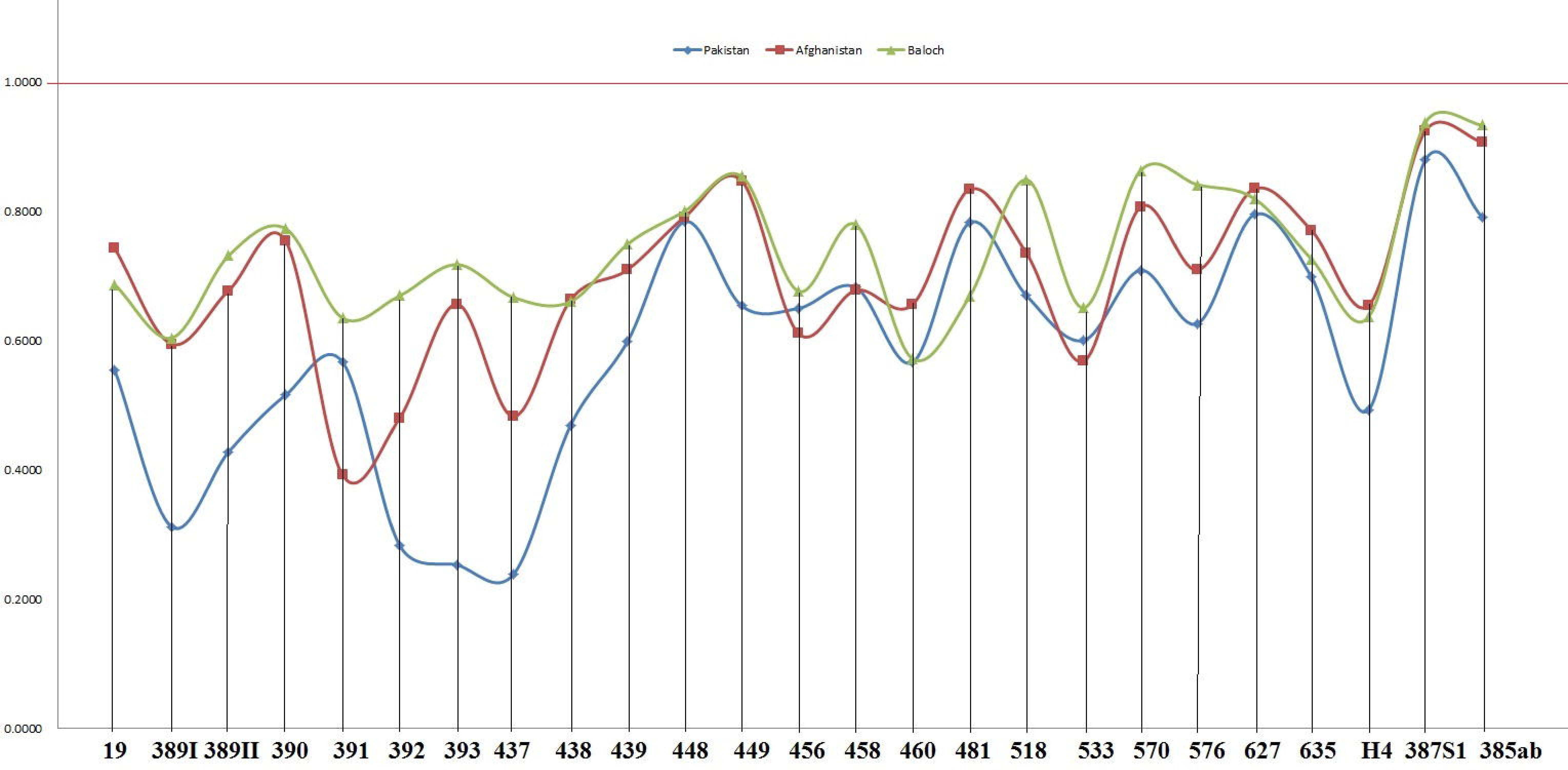
Hazara Pakistan	**MHT 9 Y-STRs**	**EHT 11 Y-STRs**	**PPY-12 Y-STRs**	**Yfiler 17 Y-STRs**	**PPY23 21 Y-STRs**	**Yfiler plus 27 Y-STRs**	**6 RM Y STRs**
No of Samples	**153**	**153**	**153**	**153**	**153**	**153**	**153**
RMP	0.0745	0.0577	0.0577	0.0123	0.0084	0.0066	0.0091
HD	0.9316	0.9485	0.9485	0.9942	0.9981	0.9999	0.9974
No of haplotypes	72	81	81	117	140	152	139
NUH	54	63	63	97	132	151	131
DC	0.4705	0.5294	0.5294	0.7647	0.9150	0.9934	0.9084
% of Unique Haplotypes	0.3529	0.4176	0.4176	0.6339	0.8627	0.9869	0.8562
Hazara Afghanistan							
No of Samples	**260**	**260**	**260**	**260**	**260**	**260**	**260**
RMP	0.0329	0.0285	0.0272	0.0184	0.0129	0.0057	0.0101
HD	0.9708	0.9753	0.9765	0.9854	0.9909	0.9982	0.9937
No of haplotypes	107	122	124	166	190	230	188
NUH	64	81	83	132	157	207	152
DC	0.4115	0.4692	0.4769	0.6384	0.7307	0.8846	0.723
% of Unique haplotypes	0.2461	0.3115	0.3192	0.5076	0.6038	0.7961	0.5846
Pak-Afg Hazara							
No of Samples	**413**	**413**	**413**	**413**	**413**	**413**	**413**
RMP	0.0334	0.0268	0.0262	0.0113	0.007	0.0032	0.0058
HD	0.9689	0.9756	0.9761	0.9911	0.9954	0.9992	0.9966
No. of Haplotypes	166	191	193	273	317	380	320
NUH	109	137	139	223	274	357	274
DC	0.4019	0.4624	0.4673	0.661	0.7675	0.92	0.7748
% of Unique Haplotypes	0.2639	0.3317	0.3365	0.5399	0.6634	0.8644	0.6634
Baloch Pakistan							
No of Samples	**111**	**111**	**111**	**111**	**111**	**111**	**111**
RMP	0.0162	0.0136	0.0136	0.0095	0.0092	0.009	0.0114
HD	0.9928	0.9954	0.9954	0.9995	0.9998	1	0.9975
No of haplotypes	97	100	100	108	110	111	101
NUH	93	96	96	105	109	111	95
DC	0.8738	0.9009	0.9009	0.9729	0.9909	1	0.9099
% of Unique haplotypes	0.83783	0.8648	0.8648	0.9459	0.9819	1	0.8558

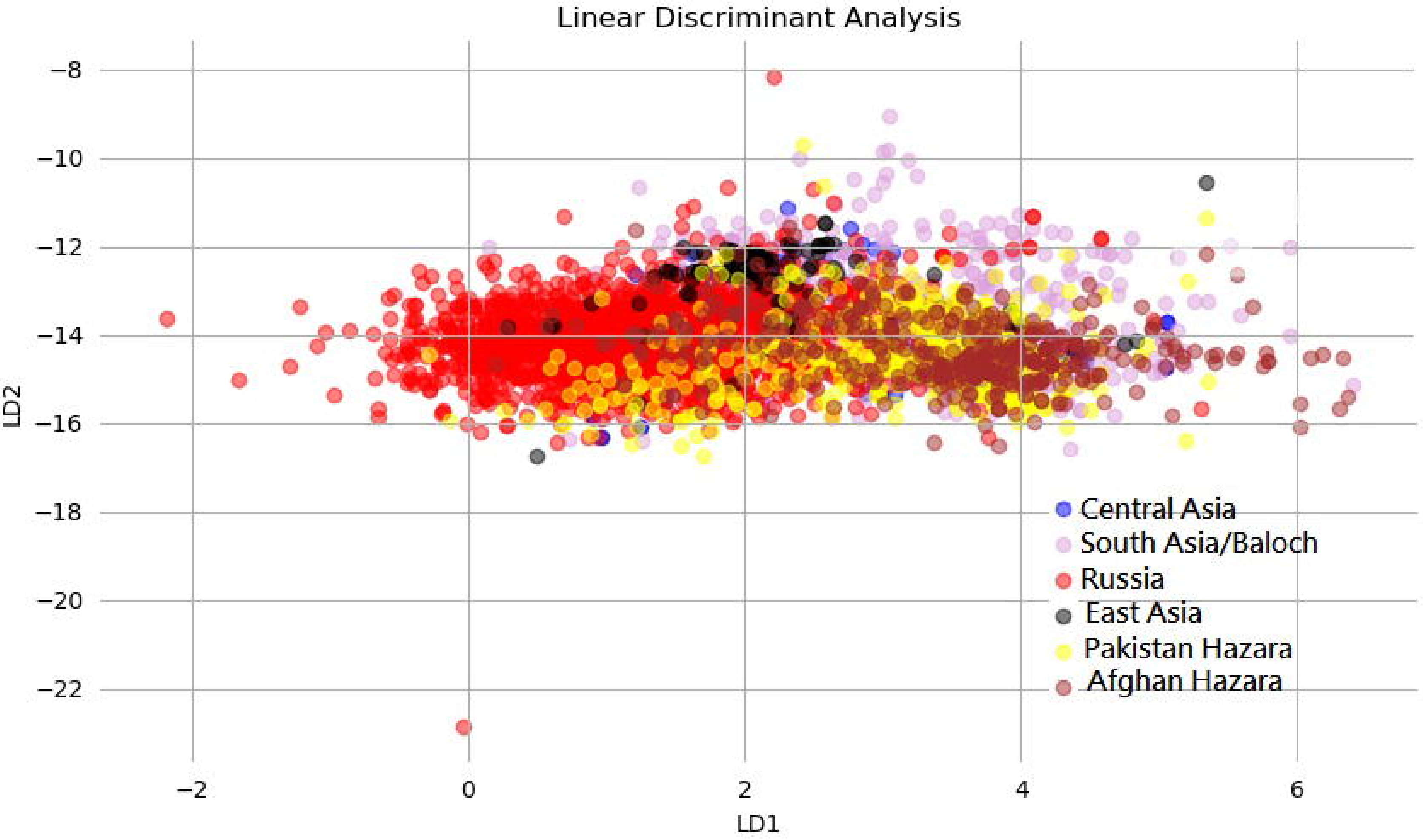
RMP= Random Matching Probability ; HD= Haplotype Diversity ; NUH= No. of unique haplotypes; DC = Discrimination Capacity

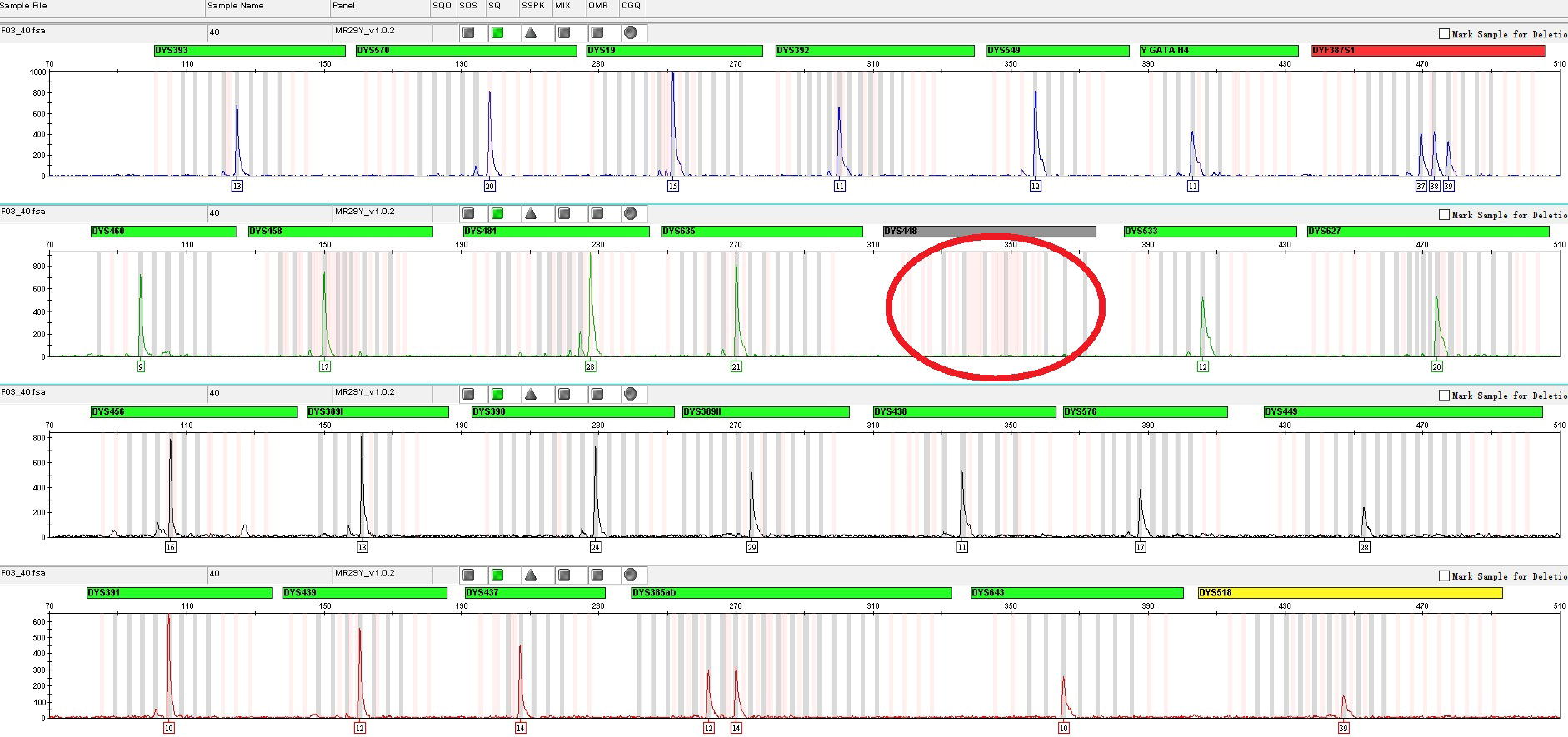
Table 3: Frequencies of the null allele at DYS448 in various ethnic groups across continents

Continent	**Population**	**Number of samples**	**No of del**	**%**	**Reference**
Asia	Hazara (Pak & Afg)	413	55	13.31%	Current Study
	Korean	708	6	0.85%	Park et al[49]
	Kalmykia	99	7	7.07%	Roewer et al [52]
	Japan	1079	10	0.92%	Mizuno et al[89]
	Malaysia	980	3	0.30%	Chang et al[48]
	Nepal	769	3	0.39%	Parkin et al[50]
	Tajikistan	124	3	2.41%	Balaresque et al[26]
	Kyrgyzstan	87	9	10.34	Balaresque et al[26]
	China	130	3	2.30%	Balaresque et al[26]
	Asian	330	2	0.61%	AmpFlSTR Yfiler™ database
Europe	Spain	247	1	0.40%	Sanchez et al[53]
Africa	Egypt	208	1	0.48%	Balaresque et al [26]
Americas	Mexico	326	1	0.30%	Gutierrez-Alarcon et al [90]
	African American	985	2	0.20%	AmpFlSTR Yfiler™ database
	Caucasian (USA)	1276	2	0.16%	AmpFlSTR Yfiler™ database
		7761	**109**	**1.40%**	

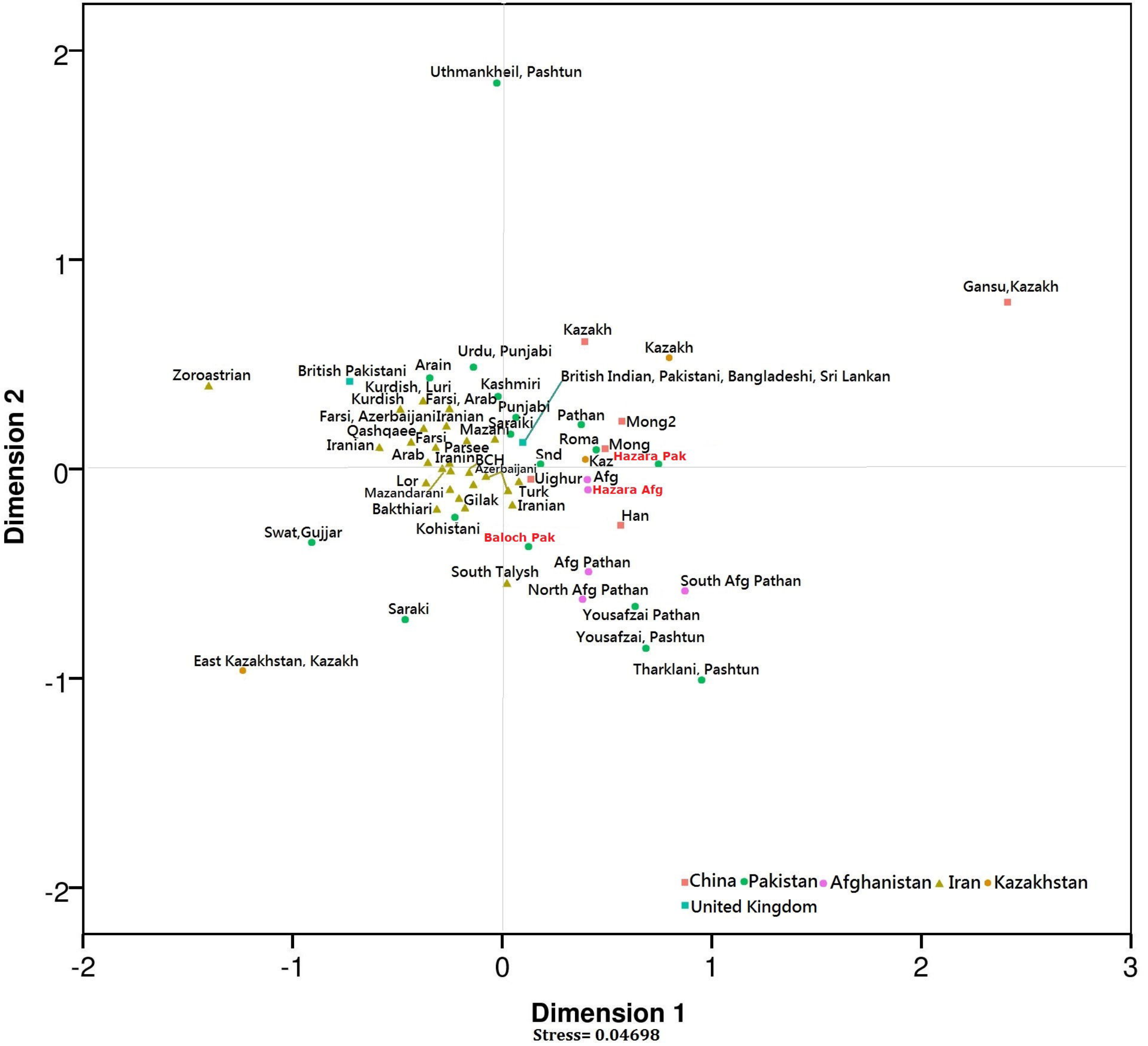

T
C2
O1b
R2
G
R1a
J1a
H1
I1
O2a
J2a
R1b
Q
I2
L1
N1a

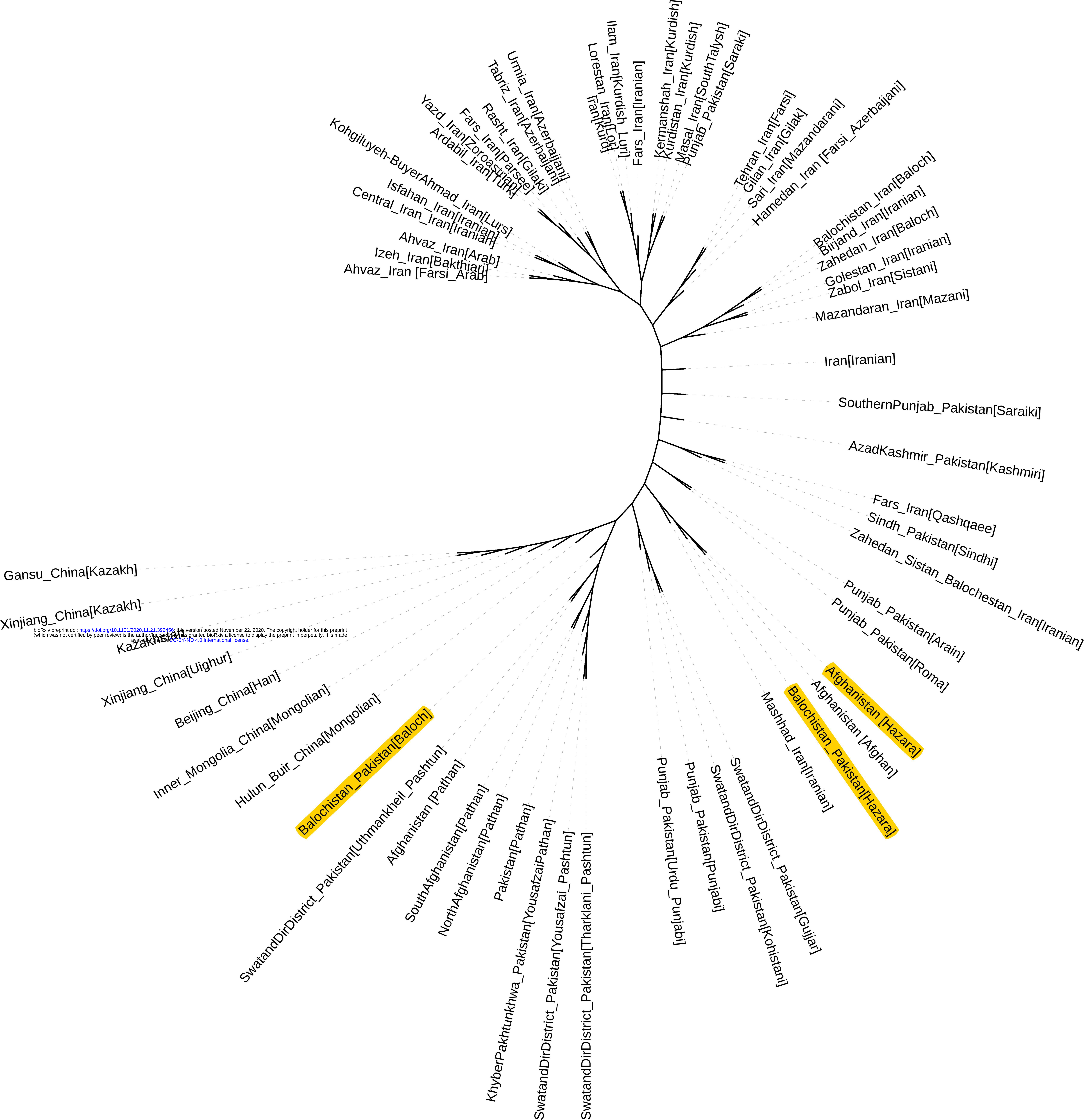

Pakistan
Afghanistan
Baloch
1.0000
0.8000
0.6000
0.4000
0.2000
0.0000
19
389I
389II
390
391
392
393
437
438
439
448
449
456
458
460
481
518
533
570
576
627
635
H4
387S1
385ab


Linear Discriminant Analysis
−8
−10
−12
−14
−16
−18
−20
−22
LD2
−2
0
2
4
6
LD1
Central Asia
South Asia/Baloch
Russia
East Asia
Pakistan Hazara
Afghan Hazara



Sample File	Sample Name	Panel	SQO	SOS	SQ	SSPK	MIX	OMR	CGQ	
F03_40.fsa	40	MR29Y_v1.0.2	■	🟩	▲	■	■	●		□ Mark Sample for Deletio
F03_40.fsa	40	MR29Y_v1.0.2	■	🟩	▲	■	■	●		□ Mark Sample for Deletio
F03_40.fsa	40	MR29Y_v1.0.2	■	🟩	▲	■	■	●		□ Mark Sample for Deletio
F03_40.fsa	40	MR29Y_v1.0.2	■	🟩	▲	■	■	●		□ Mark Sample for Deletio


Uthmankheil, Pashtun
Gansu,Kazakh
Kazakh
Kazakh
Urdu, Punjabi
Zoroastrian
British Pakistani
Arain
British Indian, Pakistani, Bangladeshi, Sri Lankan
Kurdish, Luri
Kashmiri
Kurdish
Farsi, Arab
Punjabi
Farsi, Azerbaijani
Iranian
Pathan
Mong2
Saraiki
Qashqaee
Mazani
Farsi
Roma
Iranian
Parsee
Mong
Arab
Snd
Kaz
Hazara Pak
IraninBCH
Azerbaijani
Lor
Uighur
Afg
Mazandarani
Turk
Hazara Afg
Gilak
Bakthiari
Iranian
Han
Kohistani
Swat,Gujjar
Baloch Pak
Afg Pathan
South Talysh
South Afg Pathan
North Afg Pathan
Saraki
Yousafzai Pathan
Yousafzai, Pashtun
East Kazakhstan, Kazakh
Tharklani, Pashtun
■ China ● Pakistan ● Afghanistan ▲ Iran ● Kazakhstan
■ United Kingdom
2
1
0
-1
-2
Dimension 2
-2
-1
0
1
2
3
Dimension 1
Stress= 0.04698


Ilam_Iran[Kurdish_Luri]
Lorestan_Iran[Lor]
Iran[Kurd]
Fars_Iran[Iranian]
Kermanshah_Iran[Kurdish]
Kurdistan_Iran[Kurdish]
Masal_Iran[SouthTalysh]
Punjab_Pakistan[Saraki]
Urmia_Iran[Azerbaijani]
Tabriz_Iran[Azerbaijani]
Rasht_Iran[Gilaki]
Fars_Iran[Parsee]
Yazd_Iran[Zoroastrian]
Ardabil_Iran[Turk]
Kohgiluyeh-BuyerAhmad_Iran[Lurs]
Isfahan_Iran[Iranian]
Central_Iran_Iran[Iranian]
Ahvaz_Iran[Arab]
Izeh_Iran[Bakthiari]
Ahvaz_Iran [Farsi_Arab]
Tehran_Iran[Farsi]
Gilan_Iran[Gilak]
Sari_Iran[Mazandarani]
Hamedan_Iran [Farsi_Azerbaijani]
Balochistan_Iran[Baloch]
Birjand_Iran[Iranian]
Zahedan_Iran[Baloch]
Golestan_Iran[Iranian]
Zabol_Iran[Sistani]
Mazandaran_Iran[Mazani]
Iran[Iranian]
SouthernPunjab_Pakistan[Saraiki]
AzadKashmir_Pakistan[Kashmiri]
Fars_Iran[Qashqaee]
Sindh_Pakistan[Sindhi]
Zahedan_Sistan_Balochestan_Iran[Iranian]
Punjab_Pakistan[Arain]
Punjab_Pakistan[Roma]
Afghanistan [Hazara]
Afghanistan [Afghan]
Balochistan_Pakistan[Hazara]
Mashhad_Iran[Iranian]
SwatandDirDistrict_Pakistan[Gujjar]
SwatandDirDistrict_Pakistan[Kohistani]
Punjab_Pakistan[Punjabi]
Punjab_Pakistan[Urdu_Punjabi]
Gansu_China[Kazakh]
Xinjiang_China[Kazakh]
Kazakhstan
Xinjiang_China[Uighur]
Beijing_China[Han]
Inner_Mongolia_China[Mongolian]
Hulun_Buir_China[Mongolian]
Balochistan_Pakistan[Baloch]
SwatandDirDistrict_Pakistan[Uthmankheil_Pashtun]
Afghanistan [Pathan]
SouthAfghanistan[Pathan]
NorthAfghanistan[Pathan]
Pakistan[Pathan]
KhyberPakhtunkhwa_Pakistan[YousafzaiPathan]
SwatandDirDistrict_Pakistan[Yousafzai_Pashtun]
SwatandDirDistrict_Pakistan[Tharklani_Pashtun]