International Journal of Legal Medicine
https://doi.org/10.1007/s00414-021-02591-2

POPULATION DATA

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] · Allah Rakha[2] · Shahid Nazir[2] · Rashed Alghafri[3] · Qudsia Hassan[4] · Chuan-Chao Wang[5] · Jie Lu[1]

Received: 2 February 2021 / Accepted: 26 March 2021

Abstract

The 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. 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. 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 were 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 plot, neighbor-joining tree, linear discriminatory analysis, and median-joining network were performed, which shed light on the history of Hazara and Baloch 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 · Afghanistan · Baloch · Population history · Forensic genetics

✉ Atif Adnan
mirzaatifadnan@gmail.com

✉ Jie Lu
lvjie@cmu.edu.cn

1 Department of Human Anatomy, School of Basic Medicine, China Medical University, Shenyang, Liaoning 110122, People's Republic of China

2 Department of Forensic Sciences, University of Health Sciences Lahore, Lahore 54600, Pakistan

3 General Department of Forensic Sciences and Criminology, Dubai Police General Head Quarters, Dubai, United Arab Emirates

4 Department of Forensic Medicine & Toxicology, Ziauddin Medical College Clifton, Karachi, Pakistan

5 Department of Anthropology and Ethnology, Institute of Anthropology, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen 361005, People's Republic of China

Y-chromosome short tandem repeats (YSTRs) play an important role in forensic molecular biology [1–3]. Normally, more paternal lineages can be differentiated with an increased number of Y-STRs [4], such as the Powerplex Y Kit (Promega) containing 12 Y-STRs [5], the AmpFlSTR Yfiler PCR Amplification Kit (Life Technologies) (subsequently referred to as Yfiler) containing 17 Y-STRs [6], or Powerplex Y23 Kit (Promega) containing 23 Y-STRs [7], relative to the initially proposed 9-loci haplotype [8]. So, Applied Biosystems has developed the Yfiler Plus PCR Amplification Kit [9]. 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, duplications, and inversions [10]. Null alleles or allele drop-out are well-established factors that can occur with any PCR-based STR typing system. DYS448 lay in the proximal part of the azoospermia factor c (AZFc)

Published online: 05 April 2021

region, which is considered important in spermatogenesis and made up of ''ampliconic'' repeats which act as substrates for nonallelic homologous recombination (NAHR). 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% [11].

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 1300 AD [12]. 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%) [13], 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 [14]. The 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 the Arabic chronicles of the tenth century. The Seljuq invasion of Kermān in the eleventh century started the eastward migration of the Balochi population [14]. 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 supplementary to the Y-STR haplotype reference database (YHRD).

A total of 524 blood samples were collected, in which 111 were Balochi individuals from Baluchistan Pakistan, 153 from Hazara Town Quetta, Baluchistan Pakistan (participants were part of an earlier study 27 and agreed to the secondary use of their DNA samples), and 260 from Bamyan, Afghanistan. The Axygen AxyPrep Blood Genomic DNA Miniprep Kit was used to extract genomic DNA according to the manufacturer's protocol (Axygen Biosciences; CA, USA). 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). 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. The GeneMapper Software version 4.0 (Life Technologies) was used for the genotype assignment. 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 being confirmed with two different kits (Yfiler Plus and GoldenEye 20Y), these samples were amplified and sequenced as described elsewhere [11].

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 (MPs) were calculated as $\Sigma\, Pi^2$, where Pi is the frequency of the ith haplotype. Genetic distances were evaluated using the Rst [15] and Fst [16] statistic, between reference populations and currently studied populations. Reduced dimensionality spatial representation of the populations was performed based on Rst values using multidimensional scaling (MDS) with IBM SPSS Statistics for Windows, Version 23.0 (IBM Corp., Armonk, NY, USA). Heatmaps for Rst and Fst values were also generated using the R program. 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 [17]. 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). The R program V3.4.1 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. 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 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 [11].

We successfully obtained genotypes of 524 individuals in three ethnic groups (Balochi population and Hazara

population from Afghanistan and Pakistan) (Supplementary Table 1). Haplotype data were already made accessible via the Y-chromosome Haplotype Reference Database (YHRD) under accession numbers YA004595 (Balochi), YA004312-2 (Hazara Pakistan), and YA004503 (Hazara Afghanistan). Allelic frequencies of Baloch ethnic group from Baluchistan, Pakistan, and Hazara ethnic groups from Pakistan and Afghanistan along with gene diversity values are 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), and 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 or 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: the minimal 9 Y-STRs loci, the extended 11 Y-STRs loci, PowerPlex Y12 STRs loci, Yfiler 17 STRs loci, Y21STRs loci, Y27 Yfiler Plus loci, and 6 rapidly mutating Y-STRs loci which are summarized in Table 1. 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 from 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 from 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 detect high haplotype diversity (Table 1). We have observed 101 (90.99%) different haplotypes out of 111; among them, 95 (85.58%) were unique in the Baloch population. 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, the 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 the same diversity, shown by PPY 23 loci. The above results are showing that the 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).

Since the anthropological or ethno-historical relationships between studied populations and reference populations which are included for analysis were poorly 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 by 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 Supplementary Table 3, pairwise genetic distances (Rst and Fst) and their corresponding *p*-values were calculated and were shown in Supplementary Table 4. These Rst and Fst values were visualized using hierarchical clustering heatmap (Supplementary Fig. 1a & b and S Supplementary Text 1A). These results are consistent with our previous study results [18]. 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 (Supplementary Fig. 2) based on pairwise Rst matrix from Supplementary Table 4. In the MDS plot,

Table 1 Forensic parameters on 7 different levels in three ethnic groups

	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
Hazara Pakistan							
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

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 [19], while Pakistani Hazara lined closer to Kazakh and Mongolian population which is similar to our previous study's results [11, 20]. 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 (Supplementary Fig. 3 and S Supplementary Text 1B).

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 (Supplementary Fig. 4) 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 [11, 20, 21]. An explanation about its origin in Mongolia was about ~1000 years ago [21]. The frequency of the 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 [22]. 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 [23]. This is one of the frequent haplogroups in Pakistan and North India. In the Baloch population, the frequency of haplogroup L1 is 22.5% and 1.53% in Afghani Hazara. In sub-continental populations, its frequency is about 7–15% [24, 25]. Genetic studies suggest that this may be one of the original haplogroups of the creators of Indus Valley Civilization [26, 27]. The frequency of L1 is about 28% in Pakistan and Baluchistan, from where the agricultural creators of this civilization emerged [28]. The origins of this haplogroup can be traced to the rugged and mountainous Pamir Knot region in Tajikistan [23]. In an earlier study[21], 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 the 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 was previously observed in Mongols and Kazakhs [29]. Allelic ranges of Kazak [29] population from Kazakhstan Central Asia were similar, while those in Mongol population from Inner Mongolia were almost similar on abovementioned 10 Y-STRs. In our earlier study [18], 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 Hazaras 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 (Supplementary Fig. 5). *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 5). 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. Supplementary Fig. 6 shows all individual samples plotted on the two LDA factors (F1 and F2). The LDA plot showed the association of the Hazara population with East and Central Asian populations.

By using the Yfiler plus kit, we have observed the null allele at DYS448 in 29 individuals in the Hazara population from Afghanistan (Supplementary Fig. 7). Certain factors can cause the phenomena of null alleles, and these are deletions within the target region and primer binding site problems that destabilize hybridization of at least one of the primers flanking the target region [30]. This phenomenon was previously reported, in which other commercial kits were used. The current population study represents the highest frequencies of the null allele at DYS448 when compared with the previously reported population to date (Table 2). The core repeat motif of the DYS448 locus is the hexanucleotide repeat AGAGAT [31]. 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 6) 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

Table 2 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	54	13.08%	Current study
	Korean	708	6	0.85%	Myung Jin Park et al
	Kalmykia	99	7	7.07%	Roewer et al. 2007
	Japan	1079	10	0.92%	Mizuno et al. 2007
	Malaysia	980	3	0.30%	Chang et al., 2007
	Nepal	769	3	0.39%	Parkin et al., 2007
	Tajikistan	124	3	2.41%	Balaresque et al. 2008
	Kyrgyzstan	87	9	10.34	Balaresque et al. 2008
	China	130	3	2.30%	Balaresque et al. 2008
	Asian	330	2	0.61%	AmpFlSTR® Yfiler™ database
Europe	Spain	247	1	0.40%	Sanchez et al., 2007
Africa	Egypt	208	1	0.48%	Omran et al. 2008
Americas	Mexico	326	1	0.30%	Gutierrez-Alarcon et al. 2007
	African American	985	2	0.20%	AmpFlSTR® Yfiler™ database
	Caucasian (USA)	1276	2	0.16%	AmpFlSTR® Yfiler™ database
		7761	108	1.39%	

belongs to haplogroup C2 which is the most frequent haplogroup in Mongol and Kazakh populations. This high frequency of allelic drop-out/mutation is DYS448 in the Hazara population from Pakistan, and Afghanistan strongly supports 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 [32, 33]. The commercial companies should pay special attention while designing DYS448 primers.

Finally, our study demonstrates that the Yfiler Plus Kit detects high haplotype diversity in the 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 the 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.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00414-021-02591-2.

Acknowledgements We thank all volunteers who provided material and data for this project, especially Muhammad Rehman and Abulhasan Fawad.

Funding This study was financially supported by the China Medical University postdoctoral research grant (100/1210619014).

Declarations

All participants who were included in this study were unrelated individuals of at least three generations.

Informed consent 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.

Ethical approval 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.

References

1. Adnan A, Ralf A, Rakha A, Kousouri N, Kayser M (2016) 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. https://doi.org/10.1016/j.fsigen.2016.07.005
2. Adnan A, Rakha A, Noor A, van Oven M, Ralf A, Kayser M (2017) Population data of 17 Y-STRs (Yfiler) from Punjabis and Kashmiris of Pakistan. Int J Legal Med. https://doi.org/10.1007/s00414-017-1611-9
3. Adnan A, Rakha A, Lao O, Kayser M (2018) Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan. Forensic Sci Int Genet. https://doi.org/10.1016/j.fsigen.2018.07.001
4. Vermeulen M, Wollstein A, van der Gaag K, Lao O, Xue Y, Wang Q, Roewer L, Knoblauch H, Tyler-Smith C, de Knijff P, Kayser M (2009) Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. Forensic Sci Int Genet 3:205–213. https://doi.org/10.1016/j.fsigen.2009.01.009

5. Krenke BE, Viculis L, Richard ML, Prinz M, Milne SC, Ladd C, Gross AM, Gornall T, Frappier JRH, Eisenberg AJ, Barna C, Aranda XG, Adamowicz MS, Budowle B (2005) "Validation of a male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex". Forensic Sci Int 148(1): 1–14. Forensic Sci. Int. 151 (2005) 111–124. https://doi.org/10.1016/j.forsciint.2005.02.008
6. Mulero JJ, Chang CW, Calandro LM, Green RL, Li Y, Johnson CL, Hennessy LK (2006) 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. https://doi.org/10.1111/j.1556-4029.2005.00016.x
7. Thompson JM, Ewing MM, Frank WE, Pogemiller JJ, Nolde CA, Koehler DJ, Shaffer AM, Rabbach DR, Fulmer PM, Sprecher CJ, Storts DR (2013) Developmental validation of the PowerPlex® Y23 System: a single multiplex Y-STR analysis system for casework and database samples. Forensic Sci Int Genet 7:240–250. https://doi.org/10.1016/j.fsigen.2012.10.013
8. Kayser M, Caglià A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Perez-Lezaun A, Piccinini A, Prinz M, Schmitt C, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, de Knijff P, Roewer L (1997) Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med 110:125–133. https://doi.org/10.1007/s004140050051
9. Gopinath S, Zhong C, Nguyen V, Ge J, Lagacé RE, Short ML, Mulero JJ (2016) Developmental validation of the Yfiler ® Plus PCR Amplification Kit: an enhanced Y-STR multiplex for casework and database applications. Forensic Sci Int Genet 24:164–175. https://doi.org/10.1016/j.fsigen.2016.07.006
10. Jobling MA, Samara V, Pandya A, Fretwell N, Bernasconi B, Mitchell RJ, Gerelsaikhan T, Dashnyam B, Sajantila A, Salo PJ, Nakahori Y, Disteche CM, Thangaraj K, Singh L, Crawford MH, Tyler-Smith C (1996) Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. Hum Mol Genet 5:1767–1775
11. Adnan A, Rakha A, Kasim K, Noor A, Nazir S, Hadi S, Pang H (2018) Genetic characterization of Y-chromosomal STRs in Hazara ethnic group of Pakistan and confirmation of DYS448 null allele. Int J Legal Med. https://doi.org/10.1007/s00414-018-1962-x
12. Siddique A (2012) Afghanistan's Ethnic Divides. www.cidobafpakproject.com
13. Library of Congress ed. (n.d.) Afghanistan: a country study, Claitor's Pub. Division, c2001, Baton Rouge
14. Dashti N (2012) The Baloch and Balochistan: a historical account from the beginning to the fall of the Baloch State, Trafford Sl
15. Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. Genetics 139:457–462
16. Michalakis Y, Excoffier L (1996) A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. Genetics 142:1061–1064
17. Kumar S, Stecher G, Tamura K (2016) MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. MolBiolEvol 33:1870–1874. https://doi.org/10.1093/molbev/msw054
18. He G, Adnan A, Rakha A, Yeh H-Y, Wang M, Zou X, Guo J, Rehman M, Fawad A, Chen P, Wang C-C (2019) A comprehensive exploration of the genetic legacy and forensic features of Afghanistan and Pakistan Mongolian-descent Hazara. Forensic Sci Int Genet 42:e1–e12. https://doi.org/10.1016/j.fsigen.2019.06.018
19. Haber M, Platt DE, AshrafianBonab M, Youhanna SC, Soria-Hernanz DF, Martínez-Cruz B, Douaihy B, Ghassibe-Sabbagh M, Rafatpanah H, Ghanbari M, Whale J, Balanovsky O, Wells RS, Comas D, Tyler-Smith C, Zalloua PA (2012) The genographic consortium, Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. PLoS ONE 7:e34288. https://doi.org/10.1371/journal.pone.0034288
20. Adnan A, He G, Rakha A, Kasimu K, Guo J, Hassan SE, Hadi S, Wang C-C, Xuan J (2019) Phylogenetic relationship and genetic history of Central Asian Kazakhs inferred from Y-chromosome and autosomal variations. Mol Genet Genomics. https://doi.org/10.1007/s00438-019-01617-0
21. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, Qamar R, Ayub Q, Mohyuddin A, Fu S, Li P, Yuldasheva N, Ruzibakiev R, Xu J, Shu Q, Du R, Yang H, Hurles ME, Robinson E, Gerelsaikhan T, Dashnyam B, Mehdi SQ, Tyler-Smith C (2003) The genetic legacy of the Mongols. Am J Hum Genet 72:717–721. https://doi.org/10.1086/367774
22. Smolenyak MA (2004) Turner, Trace your roots with DNA: using genetic tests to explore your family tree, Rodale ; Distributed to the trade by Holtzbrinck Publishers, Emmaus, Pa.: New York
23. Wells S (2007) Deep ancestry: inside the genographic project. Washington, D.C.: National Geographic
24. Basu A, Sarkar-Roy N, Majumder PP (2016) Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci 113:1594–1599. https://doi.org/10.1073/pnas.1513197113
25. Di Cristofaro J, Pennarun E, Mazières S, Myres NM, Lin AA, Temori SA, Metspalu M, Metspalu E, Witzel M, King RJ, Underhill PA, Villems R, Chiaroni J (2013) Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. PLoS ONE 8:e76748. https://doi.org/10.1371/journal.pone.0076748
26. Mcelreavey K, Quintana-Murci L (2005) A population genetics perspective of the Indus Valley through uniparentally-inherited markers. Ann Hum Biol 32:154–162. https://doi.org/10.1080/03014460500076223
27. Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow C-ET, Lin AA, Mitra M, Sil SK, Ramesh A, Usha Rani MV, Thakur CM, Cavalli-Sforza LL, Majumder PP, Underhill PA (2006) 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. Am J Hum Genet 78:202–221. https://doi.org/10.1086/499411
28. Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, Tyler-Smith C, Mehdi SQ (2002) Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 70:1107–1124. https://doi.org/10.1086/339929
29. Tarlykov PV, Zholdybayeva EV, Akilzhanova AR, Nurkina ZM, Sabitov ZM, Rakhypbekov TK, Ramanculov EM (2013) Mitochondrial and Y-chromosomal profile of the Kazakh population from East Kazakhstan. Croat Med J 54:17–24
30. Chang C-W, Mulero JJ, Budowle B, Calandro LM, Hennessy LK (2006) Identification of a novel polymorphism in the X-chromosome region homologous to the DYS456 locus. J Forensic Sci 51:344–348. https://doi.org/10.1111/j.1556-4029.2006.00052.x
31. Redd AJ, Agellon AB, Kearney VA, Contreras VA, Karafet T, Park H, de Knijff P, Butler JM, Hammer MF (2002) Forensic value of 14 novel STRs on the human Y chromosome. Forensic Sci Int 130:97–111
32. Balaresque P, Bowden GR, Parkin EJ, Omran GA, Heyer E, Quintana-Murci L, Roewer L, Stoneking M, Nasidze I, Carvalho-Silva DR, Tyler-Smith C, de Knijff P, Jobling MA (2008) Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication

events revealed by microsatellite analysis. Hum Mutat 29:1171–1180. https://doi.org/10.1002/humu.20757

33. Park MJ, Shin K-J, Kim NY, Yang WI, Cho S-H, Lee HY (2008) Characterization of deletions in the DYS385 flanking region and null alleles associated with AZFc microdeletions in Koreans. J Forensic Sci 53:331–334. https://doi.org/10.1111/j.1556-4029.2008.00660.x