Int J Legal Med (2011) 125:841–848
DOI 10.1007/s00414-010-0540-7

ORIGINAL ARTICLE

Forensic and genetic characterization of mtDNA from Pathans of Pakistan

Allah Rakha · Kyoung-Jin Shin · Jung Ah Yoon · Na Young Kim · Muhammad Hassan Siddique · In Seok Yang · Woo Ick Yang · Hwan Young Lee

Received: 8 October 2010 / Accepted: 9 December 2010 / Published online: 24 December 2010

Abstract Complete mitochondrial control region data were generated for 230 unrelated Pathans from North West Frontier Province and Federally Administered Tribal Areas of Pakistan. To confirm data quality and to explore the genetic structure of Pathans, mitochondrial DNA haplogroup affiliation was determined by shared haplogroup-specific polymorphisms in the control region and by the analysis of diagnostic coding region single-nucleotide polymorphisms using a multiplex system for the assignment of eight haplogroups: M, N1′5, W, R, R0, T, J, and U. Sequence comparison revealed that 193 haplotypes were defined by 215 variable sites when major insertions were ignored at nucleotide positions 16193, 309, and 573. From a phylogenetic perspective, Pathans have a heterogeneous origin, displaying a high percentage of West Eurasian haplogroups followed by haplogroups native to South Asia and a small fraction from East Asian lineages. In population comparisons, this ethnic group differed significantly from several other ethnic groups from Pakistan and surrounding countries. These results suggest that frequency estimates for mtDNA haplotypes should be determined for endogamous ethnic groups individually instead of pooling data for these subpopulations into a single dataset for the Pakistani population. Data presented here may contribute to the accuracy of forensic mtDNA comparisons in the Pathans of Pakistan.

Keywords Mitochondrial DNA · Control region · Haplogroup · Pakistan · Pathan

Allah Rakha and Kyoung-Jin Shin equally contributed to this work

Electronic supplementary material The online version of this article (doi:10.1007/s00414-010-0540-7) contains supplementary material, which is available to authorized users.

A. Rakha · K.-J. Shin · J. A. Yoon · N. Y. Kim · I. S. Yang · W. I. Yang · H. Y. Lee (✉)
Department of Forensic Medicine,
Yonsei University College of Medicine,
250 Seongsanno, Seodaemun-Gu,
120-752, Seoul, South Korea
e-mail: hylee192@yuhs.ac

K.-J. Shin · H. Y. Lee
Human Identification Research Center, Yonsei University,
250 Seongsanno, Seodaemun-Gu,
120-752, Seoul, South Korea

M. H. Siddique
Department of Zoology, University of the Punjab,
Lahore, Pakistan

Present Address:
A. Rakha
National Forensic Science Agency, National Police Bureau,
Islamabad, Pakistan

Introduction

Pakistan lies on the postulated coastal route that modern humans followed out of Africa and may therefore be one of the first geographical regions that modern humans inhabited [1, 2]. Cultural and linguistic affiliations divide the people of Pakistan into 16 ethnic groups with diverse origins, among whom endogamy is widely practiced [3]. The evolutionary antiquity and endogamy of Pakistani populations generate a high degree of genetic differentiation and structuring [4]. Hence, to obtain the most reliable and conservative frequency estimates for forensic purposes requires that regional or ethnic databases be established.

Major ethnic groups of Pakistan include the Punjabis, Pathans, Sindhis, Seraikis, Muhajirs, Balochis, Hindkowans, and Chitralis. The Pathans represent the tribes who speak Pashto (Eastern Iranian branch of the Indo-Iranian language

family) and inhabit mainly the North West Frontier Province (NWFP), adjoining tribal areas of Pakistan, and southern and eastern parts of Afghanistan. Pathans are the second-largest ethnic group in Pakistan and have reigned as the dominant ethnic group in Afghanistan for over 300 years [5]. From the 1980s, Pathans gained worldwide attention during the Soviet war in Afghanistan and the rise and fall of the Taliban because they were the main ethnic contingent in the movement.

Recent genetic analyses of historical population movements indicate that Pathans are related mainly to the Iranians and to the Hunza Burusho, who speak Burushaski, a language isolate of uncertain origin [4]. The Pathans of Pakistan also show a small Greek influence, which supports a claim that they are descendants of Greeks soldiers who invaded the Indian subcontinent [6]. Because numerous invasions from Central Asia and Afghanistan cross the northern and northwestern regions of Pakistan into India, the Pathans inhabiting these regions display the characters of many races, including Greeks, Sakas, Parthians, Persians, and Turks [5, 7, 8]. Although many have traced the impact of this mixing in the human genetic structure of northern Pakistan, forensic mitochondrial DNA (mtDNA) data for Pathans are limited to the mitochondrial hypervariable region 1 (HV1) and obtained from studies with small numbers of samples [4, 9].

In this study, we obtained data for the entire control region of mtDNA from 230 unrelated Pathans living in the NWFP and Federally Administered Tribal Areas (FATA) of Pakistan. We determined their mtDNA haplogroup affiliations by analyzing diagnostic coding region single-nucleotide polymorphisms (SNPs) using a multiplex single base extension (SBE) system to assign eight haplogroups: M, N1′5, W, R, R0, T, J, and U. Finally, we compared the mtDNA distribution among the various subpopulations, including regional ethnic groups from Pakistan and neighboring countries.

Materials and methods

DNA samples

Blood samples were collected from 230 unrelated male Pathan volunteers in the NWFP and FATA of Pakistan (Fig. 1). All participants gave their informed consent orally or in writing after we explained the aims and procedures of the study. The Institutional Review Board of Severance Hospital, Yonsei University in Seoul, Korea approved this study. DNA was isolated from blood using QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

Fig. 1 Map of Pakistan. *Shaded areas* represent sampling areas (*black* NWFP, *gray* FATA)

Mitochondrial DNA control region sequence analysis

The entire mtDNA control region was amplified in a 25-μL assay containing 1–2 ng of template DNA, 1.5 U AmpliTaq Gold DNA Polymerase (Applied Biosystems, Foster City, CA, USA), 2.5 μL of Gold ST*R 10× buffer (Promega, Madison, WI, USA), and 0.6 μM of F15975 and R635 as primers (Table S1). Thermal cycling was conducted using a PTC-200 DNA engine (MJ Research, Waltham, MA, USA) under the following conditions: 95°C for 11 min; 35 cycles of 95°C for 30 s, 56°C for 30 s, and 72°C for 90 s; and a final extension at 72°C for 7 min. PCR products were purified with ExoSAP-IT® (USB, Cleveland, OH, USA) and were sequenced using a Big Dye Terminator Cycle Sequencing v2.0 Ready Reaction Kit (Applied Biosystems). Sequencing reactions were analyzed using an ABI 3730 DNA Analyzer and/or an ABI 310 Genetic Analyzer (Applied Biosystems). Primers used for sequencing are given in Table S1 (see details at http://forensic.yonsei.ac.kr/protocols). Sequences were analyzed using Sequencing Analysis Software Version 3.4 (Applied Biosystems) and aligned using Sequence Navigator 1.01 (Applied Biosystems) and/or Geneious v4.7 (available from http://www.geneious.com) [10]. The nucleotide positions considered for analysis were 16024–16569 and 1–576, as defined by the revised Cambridge reference sequence (rCRS) [11, 12]. Each sequence is identified as a list of polymorphisms that differ from the rCRS. Observed point heteroplasmies were denoted by the appropriate International Union of Biochemistry code, and observed length heteroplasmies were treated consistently with the predominant molecule observed. To ensure data quality by a redundant approach to data generation and analysis [13], duplicate amplifications were sequenced in both the forward and reverse directions, and the resultant consensus sequences were analyzed using mtDNAmanager (available from http://mtmanager.yonsei.ac.kr/) to estimate the most probable mtDNA haplogroup based on the patterns of shared haplogroup-specific or haplogroup-associated polymorphisms in the control region [14]. Sequences have been submitted and will be searchable via the EMPOP database (http://www.empop.org) under accession numbers EMP00287–EMP00288.

Mitochondrial DNA coding region SNP analysis

To distinguish the mtDNA haplogroups M, N1′5, W, R, R0, T, J, and U, we analyzed eight coding region SNPs using a multiplex SBE reaction.

Six primer pairs were designed to amplify the eight SNPs using web-based primer designing software Primer3 (available from http://frodo.wi.mit.edu/primer3/) [15] (Table S2). Closely located SNPs at positions 10398 and 10400 and positions 15884 and 15928 were amplified to yield single PCR products. The size of each amplicon was kept under 150 bp to increase success when typing samples that are degraded. Multiplex PCR was performed in 25-μL reactions containing 1 ng of template DNA, 2.0 U AmpliTaq Gold DNA Polymerase, and 2.5 μL of Gold ST*R 10× buffer. Primer sequences and individual final concentrations are given in Table S2. Thermal cycling was conducted on the ABI 2720 Thermal Cycler (Applied Biosystems) using the following conditions: 95°C for 11 min; 33 cycles of 95°C for 20 s, 59°C for 1 min, and 72°C for 30 s; and a final extension at 72°C for 7 min. Amplified PCR products were purified with ExoSAP-IT®.

SBE primers were designed using a web-based primer designing program, Batchprimer3 (available from http://wheat.pw.usda.gov/demos/BatchPrimer3/) [16] (Table S3). A degenerate primer was used to score the T haplogroup-specific SNP at position 15928 due to the presence of the 15924 mutation for the N1e′I haplogroup in the primer sequence. The SBE reaction was carried out using the SNaPshot™ Kit (Applied Biosystems) according to the manufacturer's instructions. Extension products were analyzed by capillary electrophoresis using an ABI PRISM 310 Genetic Analyzer and GeneScan software 3.1 (Applied Biosystems).

Additional coding region SNPs were also scored by direct sequencing or using a SBE reaction to determine exact haplogroups for mtDNA haplotypes with unresolved or ambiguous haplogroup designation. SNPs at positions 14766 and 2706 for haplogroups HV and H, respectively, were typed in a duplex SBE reaction (Tables S4 and S5). Direct sequencing of amplified fragments was performed for the coding region SNPs at positions 14783 (M), 4916 (M4b1), 1888 (M5), 9824 (M7), 7196 (M8), 12973 (M17), 15431 (M30), 15530 (M31), 10556 (M37), 15458 (M71), 4883 (D), 3010 (D4 and H1), 14668 (D4), 4833 (G), 12285 (R6), 8584 (R30), 8281–8289d (B4′5), 4820 (B4b), 14766 (HV), 2706 (H), 8598 (H2b), and 1811 (U2′3′4′7′8′9).

Analysis of population data and inter-population comparisons

Haplotype diversity indices and random match probability were calculated according to Nei [17] and Tajima [18]. Arlequin v. 3.5.1.2 was used to generate Φ_{st} values and pairwise differences between and within populations [19]. An analysis of molecular variance (AMOVA) was conducted using 1,000 permutation replicates and the Kimura two-parameter method for calculating distance using the Arlequin software. Cytosine insertions at positions 16193, 309, and 573 were excluded from statistical analyses and all comparisons.

The Pathan samples from NWFP and FATA were compared with samples from other Pakistani ethnic

groups Baluch, Brahui, Hazara, Hunza Burusho, Kalash, Makrani, Parsi, and Sindhi, and with heterogeneous samples from Karachi in southern Pakistan (Pakistani-Karachi) [4]. The Pathan group was also compared with population groups in nearby countries, including India [20] and Uzbekistan [21], and with several Uzbek subpopulation groups with direct ancestry from Afghanistan, Kazakhstan, Kyrgyzstan, Russia, Tajikistan, and Turkmenistan [21]. All sequences were aligned and trimmed to a greatest common range of 16024–16383 or 16024–16450 (C insertions around 16193 were ignored).

Results and discussion

Control region sequence analysis

Complete mtDNA control region sequences were determined for 230 unrelated Pathan males from NWFP and FATA in Pakistan (Table S6, ESM 2). Based on sequence comparisons, 193 haplotypes, defined by 215 variable sites, were found (disregarding C insertions around positions 16193, 309, and 573). Of the variable sequences, 171 (74.3%) were observed once; 12 (10.4%), twice; six (7.8%), three times; three (5.2%), four times; and one (2.2%), five times. The most frequent haplotype (16092C-16278T-263G-309.1C-315.1C) occurred in 2.2% of the samples belonging to haplogroup HV, but this haplotype showed no random match among 431 Pakistani mtDNAs [4], 472 Indian mtDNAs [20], and 9,294 mtDNAs from around the world [14].

Point heteroplasmies were identified in 14 samples (6.1%) at 15 different positions (Table S6, ESM 2). Except for one sample (PK-229) that had two heteroplasmic sites (16524R and 266Y), all of the 13 heteroplasmic samples in this study possessed molecules differing only at a single site. Length heteroplasmies were observed in 15 (6.5%), 115 (50.0%), and six (2.6%) samples at the homopolymeric C stretches of HV1, HV2, and HV3, respectively. Heteroplasmic mtDNA with length variation at position 71 was also observed in a sample (PK-185) due to the presence of a 71d mutation in almost half of the molecules in the sample (Fig. S1).

Ignoring length variations around positions 16093, 309, and 573, random match probability between two unrelated individuals from the Pathan ethnic group dataset was 1:154 (0.65%), and the corresponding power of discrimination was 0.9978 for the complete control region. The mean number of pairwise differences between individuals was 11.59±5.27 nucleotides.

The relatively low value for random match probability and high value for the average number of nucleotide differences observed in the Pathan population implies that the mtDNA data will be well suited for application to forensic casework. However, the frequent occurrence of point heteroplasmy in this population group must be considered in the forensic interpretation of the mtDNA data.

Coding region SNP analysis and haplogroup designation

A multiplex assay was developed to score eight SNPs for the detection of prevalent haplogroups of South and Central Asia [22–27]. Figure 2 shows the electropherograms used to score the eight coding region SNPs in eight different samples (for more details, see Fig. S2). In almost all samples, the haplogroup information from the multiplex assay confirmed the haplogroup status of control region sequencing, but additional coding region SNPs were scored for samples with unresolved or ambiguous haplogroup designations (Table S6, ESM 2).

In the duplex SBE reaction, we used coding region SNPs at positions 14766 and 2706 to assign haplogroups HV and H, respectively, in the 54 samples that were assigned to haplogroup R0 in the multiplex reaction (Table S6, Fig. S3). During the analysis, seven HV samples with mutations at positions 16319, 143, 152, and 263 were revealed to have a novel coding region mutation at position 14764, and all five samples with mutations at positions 16092 and 16278 were identified as HV haplotypes. Therefore, the control region mutations of 16319A-143A-152C and 16092C-16278T were considered to represent additional mutation motifs for haplogroup HV.

Control region information also proved useful to subdivide 71 samples, assigned to the M haplogroup in the multiplex assay, into various subhaplogroups. However, additional coding region SNP information was required to precisely assign subhaplogroups in a few samples (e.g., those with M, M5, M37, and D4 haplotypes). One sample (PK-222) with the mutation motif 16223T-16362C-489C, initially assigned to the D4 haplogroup, was revealed to belong to the M haplogroup by scoring for SNPs at positions 3010, 4883, 14668, and 14783. Another sample (PK-070), one of four belonging to the M haplogroup, was found to have the 3010 mutation. On the other hand, two samples with H haplotypes (PK-063 and PK-192) displayed the 3010 mutation, and this defined the H1 haplogroup.

One sample (PK-115) with an apparent partial haplogroup B4-specific control region mutation motif (16183C-16189C-16217C) was assigned to the N haplogroup based on additional coding region information. It is interesting that this haplotype showed very high sequence similarity to a sample from Macedonia (MC2F10—16086C, 16172C, 16187T, 16189C, 16217C, 16223T, 73G, 146C, 210G, 263G, 315.1C, 315.2C, 534T, 571T) [14, 28]. Based only on control region sequence information in the previous

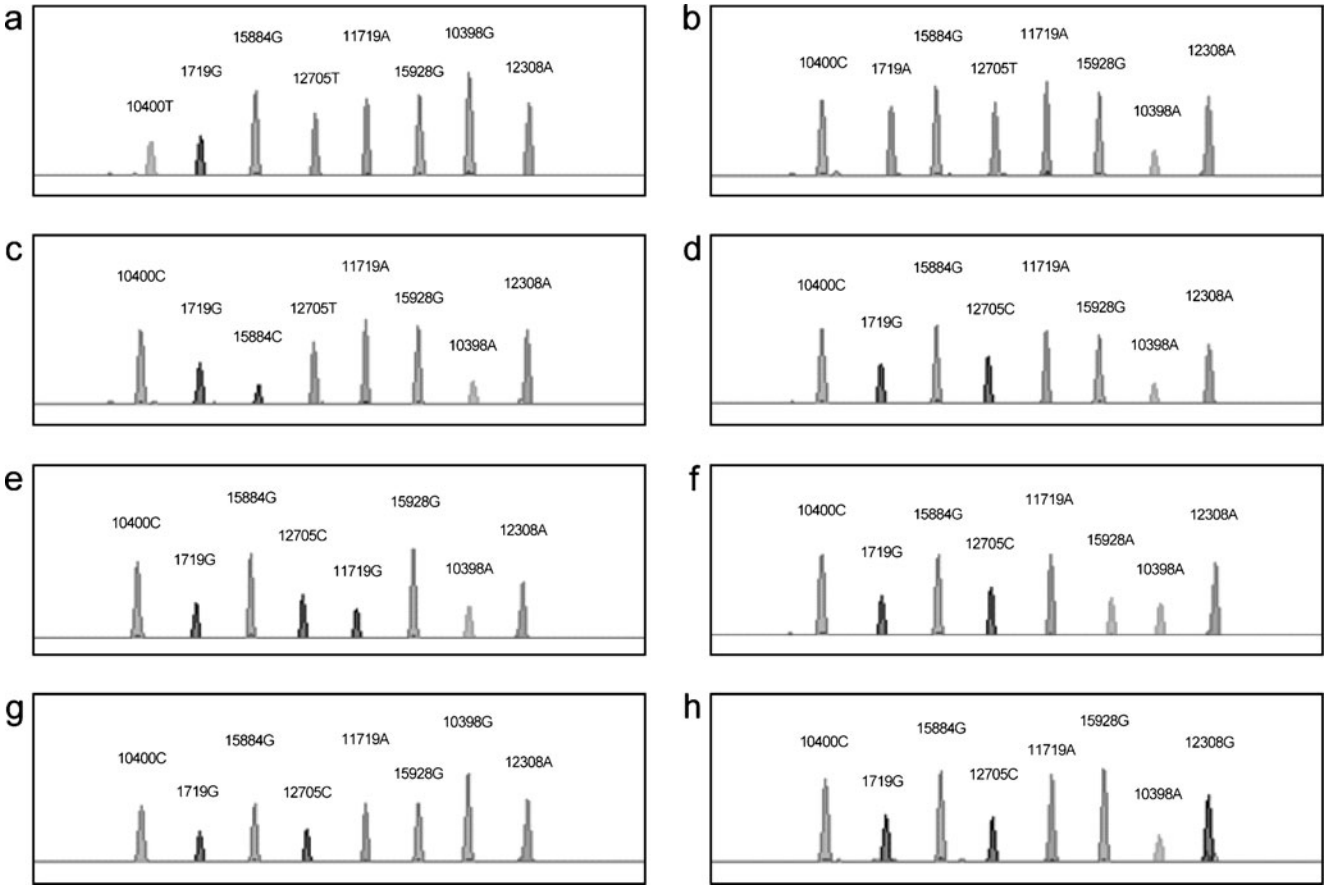

Fig. 2 Electropherograms of the multiplex SBE obtained from samples belonging to the haplogroups M (**a**), N1′5 (**b**), W (**c**), R (**d**), R0 (**e**), T (**f**), J (**g**), and U (**h**). Labeled positions are variations with reference to rCRS. SBE primers used for scoring SNPs at positions 10400, 1719, 11719, and 12308 were in reverse orientation to the L strand

report [28], the sample (MC2F10) was presumed to be of the B4 haplotype, which suggested that it may belong to the N haplogroup.

Using the multiplex assay and additional control region and coding region information, all of the 230 Pathan subjects were successfully assigned to relevant haplogroups and subhaplogroups. These methods may increase the accuracy in mtDNA haplogroup determination and help to ensure data quality and the absence of errors in the 230 Pathan mtDNA control region sequences produced by redundancy analysis.

Haplogroup distribution

Within this study group of 230 Pathans, the proportions of macrohaplogroups M, N, and R were 30.9%, 7.8%, and 61.3%, respectively (Fig. 3). Haplogroup U7 was observed in the highest percentage (11.3%) of the total dataset followed by HV (10.4%) and M3 (8.7%). Neither African lineages nor their internal derivatives were detected among these Pathans. A very limited contribution from East Asian haplogroups (5.2%) was observed, which was lower than those from other northern Pakistani ethnic groups, i.e., Hazara (34.7%) and Hunza Burusho (6.9%) [4]. Using the clustering method proposed by Quintana-Murci et al. [4], about one half of the haplogroup constituents derive from West Eurasia (55.6%) and one half from South Asia (39.1%) and East Asia (5.2%) (Fig. S4). This haplogroup distribution reveals the genetic mosaic of Pathan ancestry, with components from West Eurasia, South Asia, and East Asia. These Pathans included several deeply rooted local lineages, i.e., U7, HV2, R2, R5, and U2 (28,000–52,000 YBP), which suggests that they descended from the first inhabitants of the south-western Asian region through the Paleolithic population expansion [29]. The very small haplogroup contribution from East Eurasians is surprising in view of the east-to-west movements of Mongol, Altaic-speaking populations which are known to have occurred. The high frequency (55.6%) of West Eurasian lineages may correspond to major historical movements from Central Asia and Europe (e.g., invasion by the armies of Alexander the Great, the Arab and Muslim conquests, and the rise of the British Indian Empire) [29].

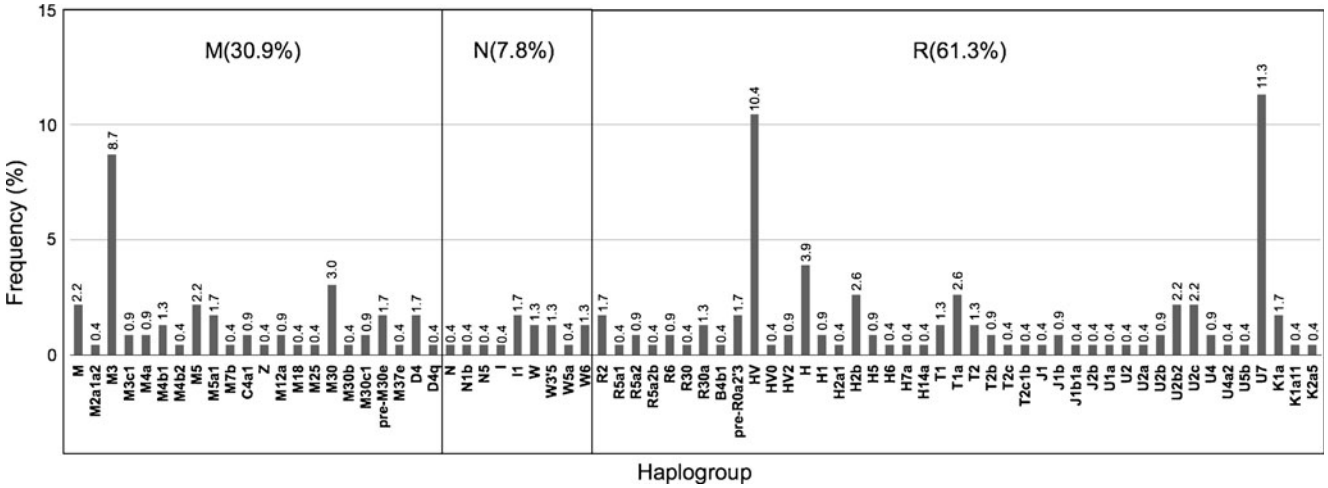

Fig. 3 Haplogroup frequencies among 230 Pathans from the NWFP and FATA of Pakistan

Very rare haplogroups were also found in the present study of Pathans. PK-216 belongs to the new D4 subbranch of D4q [30], and PK-115 belongs to the N haplogroup. Considering that a small proportion of the Pathan Y chromosomes showed evidence of Greek heritage [6], we could not ignore the very high sequence similarity between the rare N haplotype (PK-115) in the present study and a sample from Macedonia [28]. However, since present-day Macedonians share a mixed ancestry, Greek contribution to the Pathan gene pool awaits further investigation using more samples from both Pathans and Greeks.

Inter-population comparisons

To explore the population genetic structure in Pakistan, the present dataset for Pathans from the NWFP and FATA of Pakistan was compared with data from a previous study which included Pathans and distinct ethnic groups of Pakistan [4] (Table 1). The haplotype diversities of Pathans from present and previous studies (0.993) were the same and were highest among the 10 Pakistani groups. The Kalash ethnic group showed the lowest haplotype diversity with an average of 3.92 pairwise differences.

AMOVA analysis was used to test for significance variation among the various ethnic and regional groups in Pakistan (Table S7). More than 2.95% of the total genetic variation could be attributed to inter-population differences, with the remaining variation attributable to differences within populations. Pairwise Φ_{st} values for Pathans and each of the other groups ranged from −0.001 (Sindhi and Pakistani-Karachi) to 0.149 (Kalash and Sindhi) (Table S7b, Fig. S5). Pathans in the present and previous studies [4] and some of the population pairs displayed genetic homogeneity ($p>0.05$). Most of the pairwise Φ_{st} values were significant, however, and the Kalash showed the largest genetic distances with respect to the other ethnic groups. Previous reports have also noted the outlying genetic position of the Kalash [31, 32], which may be explained by the strong effects of genetic drift in this group; the western Eurasian presence in the Kalash population approaches 100%, with no detectable contribution from East or South Asian lineages [4, 29]. On the

Table 1 Diversity measures for various ethnic groups in Pakistan

Parameters	Pathan	Baluch [4]	Brahui [4]	Hazara [4]	Hunza Burusho [4]	Kalash [4]	Makrani [4]	Parsi [4]	Pathan-P [4]	Sindhi [4]	Pakistani-Karachi [4]
No. of samples	230	39	38	23	44	44	33	44	44	23	100
No. of haplotypes	157	26	22	21	32	12	24	22	39	21	77
No. of unique haplotypes	128	18	15	19	25	5	18	12	35	19	63
Genetic diversity	0.993	0.974	0.952	0.992	0.980	0.851	0.975	0.950	0.993	0.992	0.992
Mean number of pairwise differences	5.60	4.27	4.91	6.32	6.59	3.92	6.77	4.74	5.67	5.66	5.65

Investigated region, 16024–16383; Pathans from a previous study [4] are represented by Pathan-P; Pakistani-Karachi represents heterogeneous samples collected in Karachi, mainly comprised of Sindhis

other hand, the closest genetic distance between Sindhi and Pakistani-Karachi may be explained by the fact that the heterogeneous samples collected in Karachi were comprised mainly Sindhis [4].

Population comparisons were also performed for Pathans, Indians [20], Uzbeks [21], and several Uzbek subpopulation groups with direct ancestry from Afghanistan, Kazakhstan, Kyrgyzstan, Russia, Tajikistan, and Turkmenistan [21]. AMOVA results attributed most of the observed variance (97.69%) to differences within populations and 2.31%, to differences between populations (Table S8a). All of the populations compared in the present study differed significantly ($p<0.001$ in almost all cases) (Table S8b, Fig. S6), but the genetic distance revealed between the Pathans and the Uzbeks, a group with direct Afghan ancestry, was not expected. However, the Uzbek samples were obtained from regions adjacent to northern Afghanistan, while Pathans, the largest ethnic group in Afghanistan, inhabit mainly the southern and eastern parts of the country. Hence, the main genetic constituents of the Uzbeks with direct Afghan ancestry may not be Pathan, which could explain the genetic distance between the two groups. This further demonstrates the need to establish ethnic databases especially for this region to ensure reliable frequency estimates for forensic analysis.

Conclusions

The unique cultural and genetic heritage of the Pathan population provided strong incentive to generate an independent mtDNA reference database. The Pakistani Pathan group displays a heterogeneous origin, with a high percentage of West Eurasian haplogroups followed by haplogroups native to South Asia and a small percentage from East Asian lineages. Mitochondrial DNA variations and admixture levels varied between ethnic groups from the North to the South of Pakistan. Datasets for the Pakistani Pathan group differed significantly from datasets available for other Pakistani groups (except for the Hunza Burusho and Sindhi) and groups from surrounding countries. These results suggest that frequency estimates for mtDNA haplotypes in these endogamous ethnic groups should be determined independently instead of pooling them into a single dataset for the Pakistani population. Further sampling with larger sample numbers from other Pakistani ethnic groups will provide more detailed information on the extent of mtDNA variation and haplotype representation.

This paper follows the recommendations of the ISFG on the use of mtDNA in forensic analysis and the guidelines for publication of population data requested by the journal [33, 34].

Acknowledgments This work was supported by a faculty research grant of Yonsei University College of Medicine for 2009 and the National Research Foundation (NRF) of Korea grant funded by the Korea government (MEST) (No. 2009-0084144).

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Wolpert S (2000) A new history of India. Oxford University Press, New York
2. Qamar R, Ayub Q, Khaliq S, Mansoor A, Karafet T, Mehdi SQ, Hammer MF (1999) African and Levantine origins of Pakistani YAP+Y chromosomes. Hum Biol 71:745–755
3. Ayub Q, Tyler-Smith C (2009) Genetic variation in South Asia: assessing the influences of geography, language and ethnicity for understanding history and disease risk. Brief Funct Genom Proteom 8:395–404
4. Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, Rengo C, Al-Zahery N, Semino O, Santachiara-Benerecetti AS, Coppa A, Ayub Q, Mohyuddin A, Tyler-Smith C, Qasim Mehdi S, Torroni A, McElreavey K (2004) Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet 74:827–845
5. Caroe O (1992) The Pathans. Oxford University Press, Karachi
6. Firasat S, Khaliq S, Mohyuddin A, Papaioannou M, Tyler-Smith C, Underhill PA, Ayub Q (2007) Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan. Eur J Hum Genet 15:121–126
7. Bellew HW (1979) The races of Afghanistan. Sang-e-Meel, Lahore
8. Grimes BF (1992) Ethnologue: languages of the world, 12th edn. Summer Institute of Linguistics, Dallas
9. Rose HA (1991) Imperial gazetteer of India. Provincial Series, North-West Frontier Province. Sang-e-Meel, Lahore
10. Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2009) Geneious v4.7. Available at http://www.geneious.com
11. Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465
12. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23:147
13. Parson W, Bandelt HJ (2007) Extended guidelines for mtDNA typing of population data in forensic science. Forensic Sci Int Genet 1:13–19
14. Lee HY, Song I, Ha E, Cho SB, Yang WI, Shin KJ (2008) mtDNAmanager: a web-based tool for the management and quality analysis of mitochondrial DNA control-region sequences. BMC Bioinform 9:483
15. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Meth Mol Biol 132:365–386
16. You FM, Huo N, Gu YQ, Luo MC, Ma Y, Hane D, Lazo GR, Dvorak J, Anderson OD (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinform 9:253
17. Nie M (1987) Molecular evolutionary genetics. Columbia University Press, New York
18. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595

19. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50
20. Eaaswarkhanth M, Haque I, Ravesh Z, Romero IG, Meganathan PR, Dubey B, Khan FA, Chaubey G, Kivisild T, Tyler-Smith C, Singh L, Thangaraj K (2009) Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. Eur J Hum Genet 18:354–363
21. Irwin JA, Ikramov A, Saunier J, Bodner M, Amory S, Rock A, O'Callaghan J, Nuritdinov A, Atakhodjaev S, Mukhamedov R, Parson W, Parsons TJ (2010) The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns. Int J Leg Med 124:195–204
22. Palanichamy MG, Sun C, Agrawal S, Bandelt HJ, Kong QP, Khan F, Wang CY, Chaudhuri TK, Palla V, Zhang YP (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. Am J Hum Genet 75:966–978
23. Rajkumar R, Banerjee J, Gunturi HB, Trivedi R, Kashyap VK (2005) Phylogeny and antiquity of M macrohaplogroup inferred from complete mt DNA sequence of Indian specific lineages. BMC Evol Biol 5:26
24. Thangaraj K, Chaubey G, Singh VK, Vanniarajan A, Thanseem I, Reddy AG, Singh L (2006) In situ origin of deep rooting lineages of mitochondrial macrohaplogroup 'M' in India. BMC Genomics 7:151
25. Fornarino S, Pala M, Battaglia V, Maranta R, Achilli A, Modiano G, Torroni A, Semino O, Santachiara-Benerecetti SA (2009) Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. BMC Evol Biol 9:154
26. Maji S, Krithika S, Vasulu TS (2009) Phylogeographic distribution of mitochondrial DNA macrohaplogroup M in India. J Genet 88:127–139
27. Metspalu M, Kivisild T, Metspalu E, Parik J, Hudjashov G, Kaldma K, Serk P, Karmin M, Behar DM, Gilbert MT, Endicott P, Mastana S, Papiha SS, Skorecki K, Torroni A, Villems R (2004) Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. BMC Genet 5:26
28. Zimmermann B, Brandstatter A, Duftner N, Niederwieser D, Spiroski M, Arsov T, Parson W (2007) Mitochondrial DNA control region population data from Macedonia. Forensic Sci Int Genet 1:e4–e9
29. McElreavey K, Quintana-Murci L (2005) A population genetics perspective of the Indus Valley through uniparentally-inherited markers. Ann Hum Biol 32:154–162
30. Chandrasekar A, Kumar S, Sreenath J, Sarkar BN, Urade BP, Mallick S, Bandopadhyay SS, Barua P, Barik SS, Basu D, Kiran U, Gangopadhyay P, Sahani R, Prasad BVR, Gangopadhyay S, Lakshmi GR, Ravuri RR, Padmaja K, Venugopal PN, Sharma MB, Rao VR (2009) Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in South Asian corridor. PLoS ONE 4:e7447
31. 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
32. Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V, Das P, Hegde M, Molinari L, Zapata G, Weber JL, Belmont JW, Patel PI (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet 2:e215
33. Bär W, Brinkmann B, Budowle B, Carracedo A, Gill P, Holland M, Lincoln PJ, Mayr W, Morling N, Olaisen B, Schneider PM, Tully G, Wilson M (2000) DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. Int J Leg Med 113:193–196
34. Parson W, Roewer L (2010) Publication of population data of linearly inherited DNA markers in the International Journal of Legal Medicine. Int J Leg Med 124:505–509