scientific reports

Check for updates

OPEN

Contrasting maternal and paternal genetic histories among five ethnic groups from Khyber Pakhtunkhwa, Pakistan

Muhammad Tariq[1,2,5]✉, Habib Ahmad[2], Brian E. Hemphill[3], Umar Farooq[4] & Theodore G. Schurr[5]✉

Northwest Pakistan has served as a point of entry to South Asia for different populations since ancient times. However, relatively little is known about the population genetic history of the people residing within this region. To better understand human dispersal in the region within the broader history of the subcontinent, we analyzed mtDNA diversity in 659 and Y-chromosome diversity in 678 individuals, respectively, from five ethnic groups (Gujars, Jadoons, Syeds, Tanolis and Yousafzais), from Swabi and Buner Districts, Khyber Pakhtunkhwa Province, Pakistan. The mtDNAs of all individuals were subject to control region sequencing and SNP genotyping, while Y-chromosomes were analyzed using 54 SNPs and 19 STR loci. The majority of the mtDNAs belonged to West Eurasian haplogroups, with the rest belonging to either South or East Asian lineages. Four of the five Pakistani populations (Gujars, Jadoons, Syeds, Yousafzais) possessed strong maternal genetic affinities with other Pakistani and Central Asian populations, whereas one (Tanolis) did not. Four haplogroups (R1a, R1b, O3, L) among the 11 Y-chromosome lineages observed among these five ethnic groups contributed substantially to their paternal genetic makeup. Gujars, Syeds and Yousafzais showed strong paternal genetic affinities with other Pakistani and Central Asian populations, whereas Jadoons and Tanolis had close affinities with Turkmen populations from Central Asia and ethnic groups from northeast India. We evaluate these genetic data in the context of historical and archeological evidence to test different hypotheses concerning their origins and biological relationships.

South Asia is thought to be one of the first major geographic regions to be inhabited by anatomically modern humans (AMHs) as they dispersed out of Africa[1–4]. Archeological and anthropological evidence suggests the initial settlement of the region by AMH populations occurred 60–70 thousand years ago (kya)[3–13], presumably via a coastal route[1,9,14–16]. From this region, modern humans likely dispersed into East Asia, Southeast Asia, and Sahul[2,4,17]. Although Paleolithic and Mesolithic peoples left their mark in the area[10], major prehistoric and historic events with possible genetic consequences also occurred during the Neolithic period and later[18,19].

Today, South Asia is home to more than one billion people who belong to thousands of distinct socio-culturally, ethnically, and genetically diverse populations[17,20–27]. These include more than 2200 major population groups[28], over 450 tribal communities[29] and some 30 hunter-gatherer populations[30,31]. Ethnic groups claiming to have arrived and settled in South Asia include but are not limited to Afghans, Arabs, Armenians, Aryans, Chinese, Greeks, Huns, Iranians, Mongols, Persians, Scythians, Syrians, Tajiks, Turks, and Uzbeks[32]. The overwhelming majority of the current ethnic groups are reportedly endogamous[31,33], and speak an array of languages from various language families, including Indo-European, Dravidian, Tibeto-Burman, Austro-Asiatic, and Sino-Tibetan, each being differentially distributed throughout South Asia[34–36].

This differential distribution has been attributed to the impact of external influences on South Asian populations. The earliest evidence of farming-based economies in South Asia has been traced to the introduction of West Asian cultigens such as wheat and barley at Mehrgarh, Pakistan, dating to 8 kya[37–41]. From there, farming

[1]Centre for Omic Sciences, Islamia College, Peshawar 25120, Khyber Pakhtunkhwa, Pakistan. [2]Department of Genetics, Hazara University Mansehra, Mansehra 21120, Pakistan. [3]Department of Anthropology, University of Alaska, Fairbanks, AK 99775, USA. [4]College of Life Science, Northeast Forestry University, Harbin 150040, China. [5]Department of Anthropology, University of Pennsylvania, 3260 South Street, Philadelphia, PA 19104-6398, USA. ✉email: mtariq@icp.edu.pk; tgschurr@sas.upenn.edu

and sedentary lifeways spread further east, laying the foundation for the later Indus Valley (including the cities of Harappa and Mohenjo-Daro) and the Gangetic Valley civilizations, which arose between 4.6–3.9 and 3.5–2.5 kya, respectively[40,41]. Sometime around 3.5 kya, Indo-European-speaking nomadic pastoralists from the southern steppes, often called 'Aryans', crossed the Hindu Kush Mountains and expanded into the subcontinent[42–46]. Later, in the eighth century CE, an Arab-Muslim army invaded Sindh in the extreme western periphery and occupied the subcontinent for a brief period of time. At the beginning of the eleventh century CE, Turkic populations bearing Islamic culture entered South Asia from Afghanistan and began spreading Islamic culture from west to east[47–57]. Outside of northern Pakistan, this series of population expansions effectively generated the gene pool from which subsequent South Asia populations developed. Yet, in northern Pakistan, appreciable movement of Islamic populations whose ultimate origins are found in the Kandahar region of southern Afghanistan did not occur until the sixteenth century[58].

Recent genetic studies suggest that the major West Eurasian genetic contribution to South Asia derives from Neolithic Iranian and early Bronze Age steppe populations[59,60]. Other studies have further revealed contributions from Middle and Late Bronze Age steppe populations in South Asia, together with a Chalcolithic or Bronze Age Central Asian admixture scenario[61,62]. The various invasions and subsequent migrations are assumed to have resulted in major demographic expansions in the region, adding new languages and cultures to the mix of peoples already residing within the subcontinent. As a result of these processes, the majority of present-day Pakistani and Northwest Indian populations have relatively close affinities with West Eurasian populations[7,8,17,19,62–71].

Like South Asia as a whole, the population of Pakistan encompasses a diverse array of cultures with different communities distributed into different ethnic groups. Indo-European languages are spoken by more than 70% of Pakistani ethnic groups[31,33]. These languages have been connected to the so-called "Indo-Aryan invasion" from Central Asia that occurred approximately 3.5 kya and the subsequent establishment of the caste system. The actual extent of immigration by "Aryan" populations remains controversial[31], although at least some Indo-Iranian languages were likely introduced by immigrant Islamic groups from Afghanistan during the medieval period[72–76]. The speakers of these languages are further divided into different castes, sub-castes and tribes, reflecting the complex social organization of the region today[2,30,31,58,77–80]. Moreover, the region is home to followers of many religions, the major ones being Islam, Hinduism, Buddhism and Sikhism, with sizeable Christian and Jewish minorities also being present. All have likely contributed to the genetic and cultural diversity found in this region of South Asia.

While several studies have focused on Pakistan, genetic research on the ethnic groups in the Khyber Pakhtunkhwa Province (KPP) remains rather limited[81–87]. Thus, in this study, we conducted an extensive analysis of diversity in the mitochondrial DNA (mtDNA) and the non-recombining portion of the Y-chromosome (NRY) among members of the five major ethnic groups of Buner and Swabi Districts of KPP to elucidate their genetic history. The data generated were also compared with previously published information for geographically and ethnically diverse global populations to explore the maternal and paternal history of South Asian populations. The resulting data provide, for the first time, deep phylogeographic information about Pakistani population genetic diversity.

Results

Mitochondrial DNA diversity. *Genetic lineages.* The maternal genetic ancestry of 659 individuals from the five KPP ethnic groups was characterized through coding region single nucleotide polymorphism (SNP) genotyping and control region (CR) sequencing (Tables S1 and S2). A total of 54 different mtDNA haplogroups was detected among individuals of the five ethnic groups. Although sharing a number of haplogroups in common, the five populations differed significantly in the frequencies of many of these maternal lineages.

The majority (50.8%) of the identified haplogroups were of West Eurasian (WE) derivation. The most prevalent WE haplogroup was H, followed by U7, J1, W and HV. The relative proportion of these haplogroups was greatest among Gujar individuals (62.3%).

South Asian (SA) lineages were mainly represented by haplogroups deriving from macrohaplogroup M, which comprised approximately 39% of the individuals within the study populations. The most frequently occurring SA lineages were U2, followed by M3 and R5. The highest frequency of SA lineages occurred among Tanolis (47.8%).

All five ethnic groups also had a number of East Eurasian (EE) haplogroups, which together accounted for 10.2% of their mtDNAs. Haplogroup D was the most prevalent EE lineage, with A, C, F, M7, M9, M10 and Z also occurring at low frequencies. The frequency of such EE lineages was highest among Jadoons (15.2%), whereas they were far less common among Tanolis (5.2%) and Gujars (4.1%).

Certain haplogroups were confined to particular ethnic groups. For instance, J2 was present only among Gujars, while haplogroup M1 appeared only in Jadoons. Similarly, haplogroups M27, M76, R30 were only observed among Syeds, while M21, M33, M52, V and Z were nearly exclusive to Yousafzais. In addition, the more frequent haplogroups, such as D, HV, M5, M18, M30, N3, U2, U7 and W, differed considerably in their distribution among the five ethnic groups from KPP.

We further observed a high degree of CR sequence diversity among members of the five KPP ethnic groups (Table S2). Notably, between 43–74% of their mtDNAs presented unique HVS1 haplotypes. Each ethnic group also shared a modest number of mtDNA haplotypes with the other (avg = 12.8), with the Yousafzais sharing more than the other four ethnic groups. Only three mtDNA haplotypes were shared between all five ethnic groups, with these (H [#135], R5 [#196] and U7 [#284]) likely to be the founder haplotypes for the respective maternal lineages.

These patterns of diversity were mirrored in the median-joining (MJ) networks generated from HVS1 haplotypes present in the five ethnic groups (Supplemental Fig. 1). As evident from these networks, there is extraordinarily mtDNA diversity in all KPP populations. Overall, these ethnic groups shared a number of different M

haplogroups, the majority being of South Asian origin, as well as a number of haplogroup U mtDNAs, including those from both South Asian and West Eurasian lineages. Haplogroup W also appeared in multiple KPP groups, haplogroup H was seen in most groups, and J, K, T and R5 comprised many of the R-derived mtDNAs in these populations.

To further understand the maternal genetic background of KPP ethnic groups, we compared their mtDNA haplogroup frequency data with those from 77 Old World populations representing South Asia, Central Asia, East Asia, Middle East, Europe and the Caucasus (Table S3). We used these data and GPS coordinates associated with the populations to generate geospatial maps of mtDNA haplogroup frequencies across Eurasia. As seen in these maps (Fig. 1) and discussed above, KPP populations bear a set of maternal lineages that reflect the geographic regions from which they emerged and were dispersed over the past 40–50,000 years. Those lineages originating in East Asia (D and M_{EA}) and South Asia (M_{SA}) showed foci in those regions. Likewise, haplogroups having European (H, I) and Near East (HV, J, K, T, N1) origins were concentrated in those areas, although clearly having been spread into South Asia since evolving. The other lineages (U_{SA}, U_{WE}, W) were somewhat less concentrated in any one region. While providing insights into the distribution of these haplogroups across Eurasia, this analysis may have been affected by the sample sizes, hence, the relative haplogroup frequencies, used for the comparative populations.

Genetic variation and relatedness of KPP populations with other Asian groups. Summary statistics describing genetic diversity in the five KPP ethnic groups and other Pakistani populations are shown in Table 1. All five KPP groups exhibited great haplotypic diversity, with the highest occurring in Yousafzais (0.994) and the lowest among Tanolis (0.970). Nucleotide diversity estimates were also essentially the same for KPP and other Pakistani populations. Neutrality tests yielded significantly negative Tajima's D and Fu's F estimates for the KPP populations, suggesting that they experienced relatively recent expansions in population size.

To further elucidate the maternal genetic relationships between the five KPP ethnic groups and comparative populations, we generated pairwise F_{ST} values based on their HVS1 sequences (Table S4). The resulting estimates were then visualized in a Neighbor-joining (NJ) tree (Fig. 2). In the NJ tree, Jadoons, Syeds, Yousafzais and Gujars clustered together with the majority of Central Asian and KPP populations, but were clearly distinguished from Middle Eastern, Indian, East Asian, European and Caucasus populations. By contrast, the Tanolis were more closely positioned with Indian populations.

Analysis of molecular variation (AMOVA) was conducting using F_{ST} estimates based on HVS-1 haplotypes in KPP and comparative populations. In this analysis, 95.2% of the genetic variation occurred within all 82 Pakistani and comparative populations (Table 2). When grouping populations by country of origin or region, the genetic variation among countries or region accounted for only 2.6% of the variation, whereas 2.55% was explained by the differences between population samples within countries. A similar trend was observed for the AMOVA results based on ethnicity. Thus, KPP ethnic groups were not strongly differentiated from each other or regional populations in terms of their maternal lineage composition.

Y-Chromosome diversity. *Genetic lineages.* The analysis of NRY variation in 678 male individuals from the same ethnic groups yielded 11 distinct haplogroups defined by 54 SNP and 19 Y-STR markers (Tables S5 and S6). The majority of their Y-chromosomes fell into one of four haplogroups, namely, R1a1a-M17 (50%), R1b1a-M297 (17.4%), O3-M122 (13.9%) and L-M20 (7.1%) with another seven haplogroups comprising the remaining 11.6%. The haplogroup profiles of the five ethnic groups suggested that the genetic diversity in these groups was structured along ethnic boundaries.

West Eurasian Lineages. The most common paternal haplogroup in the study populations was R1a1a-M17. It appeared at high frequencies among individuals of three of the five ethnic groups (Syeds, Yousafzais and Gujars). R1a1a-M17 is also one of the most common haplogroups in Eurasia, with high frequencies occurring in Eastern Europe, Central Asia, and South Asia[8,43,88–94]. In neighboring Afghanistan, R1a1a-M17 is frequent among Pashtuns (51.02%) and Tajiks (30.36%) but less so in Uzbeks (17.65%) and Hazaras (6.7%)[95]. In addition, it has been observed at high frequency (~80%) among Yousafzais of Swat Pakistan[82], a finding consistent with our data.

The second most common haplogroup was R1b1a-M297. It occurred in the Tanolis at a very high frequency but appeared at very low frequencies in the Jadoons, Yousafzais and Syeds, while being completely absent in the Gujars. Haplogroup R1b is the most frequent Y-chromosome lineage in Western Europe (>70%)[96–99], but also appears in South Asian populations at modest frequencies[44,100].

The other West Eurasian haplogroups appeared non-uniformly in the study populations. G2a-P15 and I2-P215 occurred at low frequency in only the Yousafzais. G2a is thought to have arisen in Anatolia and the Caucasus[101], and may be associated with the Neolithic expansion throughout the region[94]. G2a has also been observed throughout the Near East[102] and Mediterranean region[103], and occurs in South Asia in appreciable frequencies[44,95,104]. By contrast, I2-P215 may have arisen in the Balkans and central Europe[105], since it is commonly observed in Slavic speaking populations of southern Europe[106].

Gujars, Jadoons and Yousafzais exhibited haplogroup J2a-M410 at low frequencies, while J2b-M12 occurred at low frequency in the Gujars and Yousafzais. Previous studies have demonstrated that J2a-M410 and J2b-M12 are associated with the demic diffusion of Neolithic farmers in North Africa and Eurasia from Mesopotamia (Iraq and Syria)[107–109]. Both J2a-M410 and J2b-M12 (0–8%) also appear at low frequencies in populations inhabiting different parts of India[110]. In general, the presence of J2a-M410 and J2b-M12 in Pakistan and India has been considered indicative of gene flow from western Asia[43,44].

South Asian Lineages. Haplogroup H-M69, which is commonly observed in South Asia[60,111,112], occurred in all five ethnic groups at modest frequencies. The Gujars also had a moderate frequency of haplogroup L-M20, with this paternal lineage being present at low frequency among Syeds, Yousafzais, and Jadoons and completely

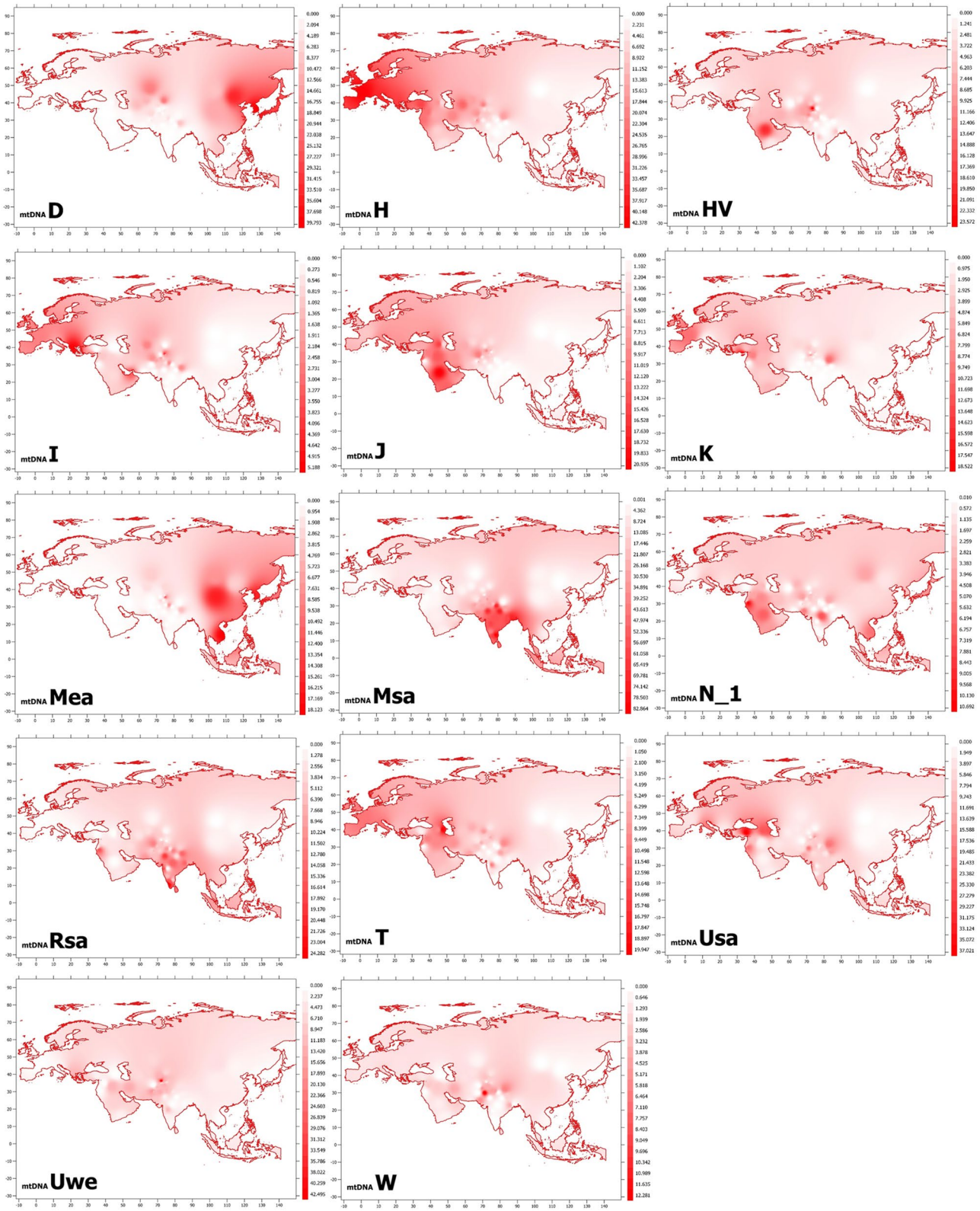

Figure 1. Geospatial map of mtDNA haplogroup frequencies in KPP ethnic groups and comparative populations. See the Methods section for a description of the mapping process, and Table S3 for the data on which the projection is based. With respect to the abbreviations in the different panels in the figure, "M_{EA}" indicates mtDNAs belonging to East Asian haplogroups deriving from macrohaplogroup M (C, D, G, M7-M9, Z), while "M_{SA}" denotes those belonging to South Asian haplogroups derived from M (M2-M10, M12, M18, etc.). Similarly, "R_{SA}" indicates mtDNAs derived from R haplogroups arising in South Asia (e.g., R2, R5, R6, R7, R9, R30, R31), "U_{WE}" denotes mtDNAs from U haplogroups common to West Eurasian populations (U2e, U3-U5, U7, U8), and "U_{SA}" mtDNAs from U haplogroups identified in South Asian populations (U*, U1, U2a-c).

Population	n	#	HD	ND	PD	RI	Tajima's D	Fu's FS
Gujars	122	56	0.975	0.022 ± 0.012	5.71 ± 2.76	0.016	−1.805	−25.194
Jadoons	99	66	0.988	0.022 ± 0.012	5.67 ± 2.74	0.011	−2.003	−25.268
Syeds	127	93	0.992	0.023 ± 0.012	5.89 ± 2.83	0.009	−2.094	−25.117
Tanolis	134	59	0.970	0.022 ± 0.012	5.76 ± 2.78	0.012	−1.664	−25.147
Yousafzais	177	131	0.994	0.022 ± 0.012	5.61 ± 2.71	0.011	−2.107	−25.055
Kashmiri	317	211	0.993	0.023 ± 0.012	5.89 ± 2.82	0.011	−2.117	−24.675
Makrani	100	60	0.984	0.027 ± 0.014	7.01 ± 3.32	0.006	−1.742	−24.936
Pathan	230	153	0.992	0.022 ± 0.012	5.61 ± 2.70	0.010	−2.225	−24.934
Saraiki	85	47	0.957	0.023 ± 0.012	5.98 ± 2.88	0.013	−1.637	−25.238
Sindhi	115	81	0.992	0.024 ± 0.013	6.16 ± 2.95	0.013	−1.579	−25.069
Balti	49	32	0.979	0.020 ± 0.019	5.12 ± 2.53	0.016	−1.772	−22.570
Bangash	25	17	0.973	0.012 ± 0.011	5.10 ± 2.56	0.045	−1.162	−7.036
Khattak	25	14	0.932	0.021 ± 0.011	5.34 ± 2.66	0.073	−0.413	−2.867
Mahsuds	25	10	0.917	0.019 ± 0.011	4.96 ± 2.50	0.028	−0.680	−0.142
Orakzai	25	18	0.967	0.022 ± 0.012	5.61 ± 2.79	0.045	−1.186	−7.811
Brahui	38	22	0.952	0.018 ± 0.010	4.63 ± 2.32	0.032	−1.619	−10.643
Hazara	23	21	0.992	0.022 ± 0.012	5.76 ± 2.86	0.018	−1.638	−15.567
Hunza	44	32	0.980	0.024 ± 0.013	6.13 ± 2.97	0.016	−1.912	−21.877
Kalash	44	11	0.830	0.015 ± 0.009	3.86 ± 1.98	0.064	−0.041	−0.249
Parsi	44	20	0.943	0.018 ± 0.010	4.53 ± 2.27	0.058	−1.501	−6.759

Table 1. Summary statistics for KPP ethnic groups and other Pakistani populations based on mtDNA HVS-1 haplotypes. n = number of samples; # = number of haplotypes; HD: Haplotype Diversity; ND: Nucleotide Diversity; PD: Pairwise differences; RI: Raggedness index; Citations and references for the comparative populations are provided in Table S9.

absent in Tanolis. Haplogroup L commonly occurs in populations from Pakistan, India and Afghanistan, and has spread into the Near East and Iran[44,84,94,113]. In addition, South Asian-specific haplogroup R2-M124 occurred at low frequencies among all KPP populations. Haplogroup R2-M124 has mainly been found in Indian, Iranian, Pakistani and Central Asian populations, and postulated to have a Central Asian origin[8,44,94,114–118].

East Eurasian Lineages. East Eurasian haplogroup O3-M122 occurred nearly exclusively in Jadoons, and otherwise appeared at very low frequency in Yousafzais, Gujars and Tanolis while being absent in Syeds. O3-M122 is the most common haplogroup among Han Chinese populations, occurring at frequencies of 50–60% in them[119–121]. It also occurs at very low frequencies in India and Pakistan, mostly likely due to the westward expansion of Tibeto-Burman speakers into South Asia[44].

By contrast, all five KPP populations possessed Q-MEH2 Y-chromosomes at low frequencies. The MEH2 marker occurs downstream of the M242 marker that helps to define this paternal lineage. Haplogroup Q-M242 probably originated in Central Asia and has been distributed widely in Northeast Asia, while also appearing at low frequencies in Europe and the Middle East, mostly likely to due to the influence of Mongolic and Turkic speaking populations[93,122]. Among the Pashtuns of Afghanistan, the frequency of haplogroup Q-M242 is about 18.4%[95].

These patterns of diversity were mirrored in the median-joining (MJ) networks generated from Y-STR haplotypes present in the five ethnic groups (Supplemental Fig. 2). Gujars, Syeds and Yousafzais all exhibited specific R1a lineages, although sharing some with other KPP ethnic groups in which other haplotypes from this paternal lineage also occurred. In addition, the Tanolis showed a wide range of R1b haplotypes, indicating their centrality to the paternal gene pool for this population. Similarly, Jadoons had mainly O3 haplotypes that appeared in a starlike cluster suggestive of an older founder event, whereas L haplotypes were dispersed among all ethnic groups with no specific pattern of clustering.

To further assess the paternal genetic background of KPP ethnic groups, we compared their NRY haplogroup frequencies with those from 82 Old World populations representing South Asia, Central Asia, East Asia, Middle East, Europe and the Caucasus (Table S7). We used these data and GPS coordinates associated with the respective populations to generate geospatial maps of NRY haplogroup frequencies across Eurasia. As seen in these maps (Fig. 3) and discussed above, KPP populations bear a set of paternal lineages that reflect the geographic regions from which they emerged and were dispersed into adjacent area at different points in time. Haplogroups H, L and R2 clearly have South Asian roots, with J2 arising in the Near East and O3 in South-East Asia, as previously above. Similarly, R1b appears to have arisen and spread into South Asia from Europe, while R1a shows a more complex pattern reflective of its dual origin in Eurasia. Thus, NRY diversity in KPP populations reveals these prehistoric expansions of paternal lineages into South-Central Asia, while also reflecting more recent population movements and ethnic group formation, as described below.

Genetic variation and relatedness of KPP populations with other Asian groups. Molecular diversity estimates were calculated from Y-STR haplotypes in an effort to quantify the degree of paternal genetic variation in the five KPP ethnic groups and other Pakistani populations (Table 3). Gene diversity estimates showed that Gujars

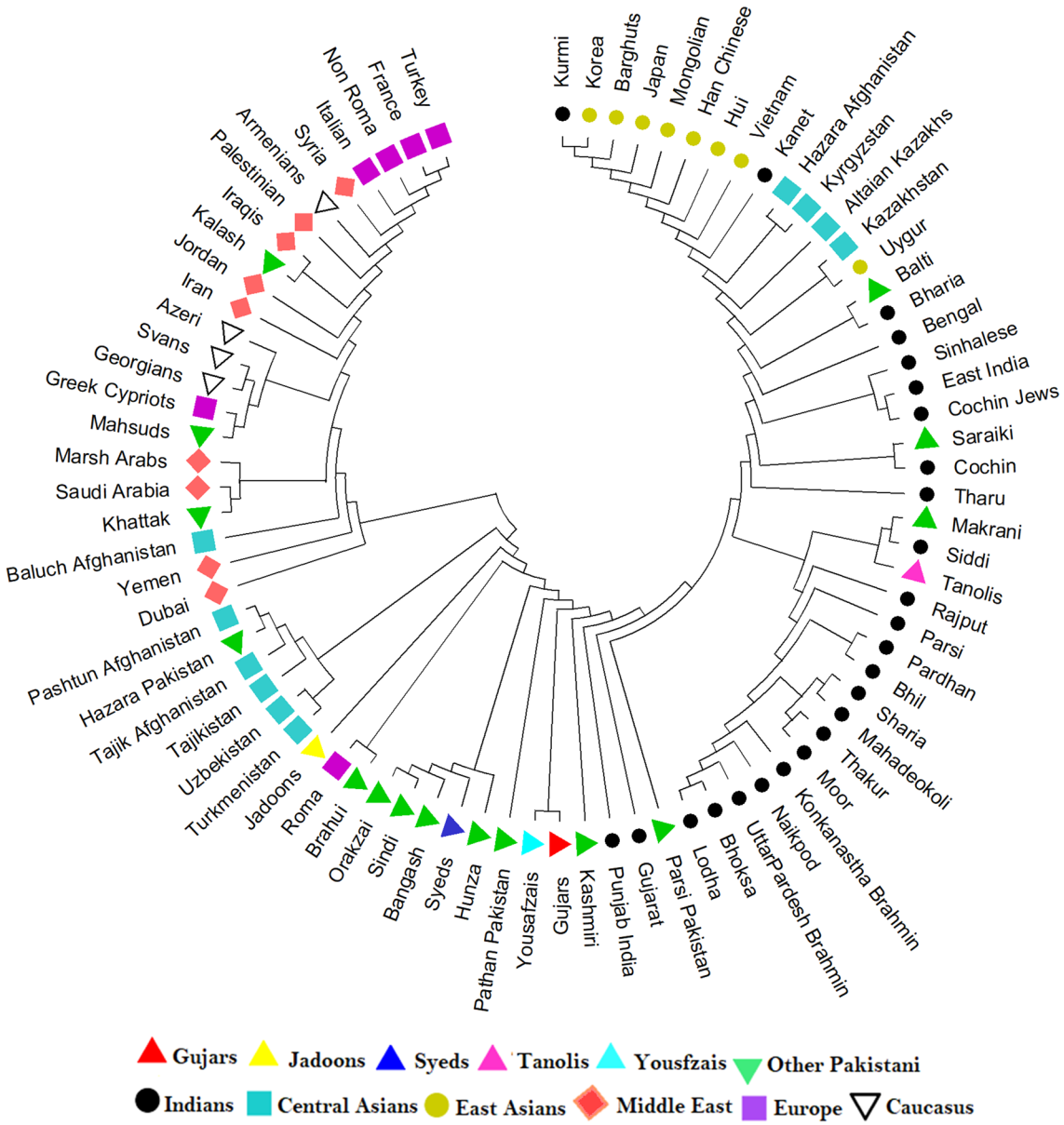

Figure 2. A Neighbor-Joining tree showing the genetic relationships between KPP ethnic groups and 77 world populations based on F_{ST} estimates from mtDNA HVS1 sequence data (Table S4).

were more diverse than the other four ethnic groups, with Syeds being the least diverse by far. Unlike the mtDNA data, NRY pairwise differences were greater among the Gujars than individuals of the other four ethnic groups.

We compared the Y-STR data from the five KPP ethnic groups with those from the comparative populations to place their genetic diversity in a broader context. Y-STR haplotypes for this analysis were reduced to 10-loci haplotypes in order to incorporate data from as many comparative populations as possible. Subsequently, we calculated pairwise R_{ST} values (Table S8) and visualized them in a NJ tree (Fig. 4).

The NJ tree revealed different genetic affinities of the KPP populations. Gujars from our study clustered tightly with a Gujars population from Punjab[123], Yousafzais from this study clustered closely with populations from South Afghanistan[104] and other Yousafzai groups[82,83]. In addition, the Syeds from our study were positioned close to a Tarklanis population from the Dir District of Pakistan[82]. By contrast, Jadoons and Tanolis generally clustered somewhat near each other and with Turkmen population from Central Asia and ethnic groups from northeast India.

We conducted an AMOVA using R_{ST} estimates generated from Y-STR haplotypes in KPP and comparative populations (Table 4). This analysis indicated that the great majority of Y-chromosome diversity occurs within populations (81.8%), with less than 20% occurring between groups across the 87 populations considered. When categorized by geography, genetic variance accounted for 3.8%, whereas 14.8% of total variance was explained by differences between population samples within geographic regions. These estimates were approximately the same for the AMOVA based on ethnicity. Thus, KPP and comparative populations were moderately genetically differentiated from each other based on their paternal lineages.

No Groups						
Source of variation	**d.f.**	**Sum of squares**	**Variance component**	**% of variance**	**Fixation Indices**	**P-value**
Among populations	72	1782.704	0.15213 Va	4.79		
Within populations	10,496	31,766.373	3.02652 Vb	95.21	F_{ST}: 0.04786	0.000 ± 0.000
Total	10,568	33,549.077	3.17865			
Grouped by Geography						
Among groups	6	865.080	0.08299 Va	2.60	F_{SC} (Va):0.026	0.000 ± 0.000
Among populations within groups	66	917.624	0.08129 Vb	2.55	F_{ST} (Vb):0.051	0.000 ± 0.000
Within populations	10,496	31,766.373	3.02652 Vc	94.85	F_{CT} (Vc):0.026	0.000 ± 0.000
Total	10,568	33,549.077	3.19080			
Grouped by Ethnicity						
Among groups	15	977.099	0.08197 Va	2.57	F_{SC} (Va): 0.025	0.000 ± 0.000
Among populations within groups	57	805.604	0.07913 Vb	2.48	F_{ST} (Vb): 0.051	0.000 ± 0.000
Within populations	10,496	31,766.373	3.02652 Vc	94.95	F_{CT} (Vc): 0.026	0.000 ± 0.000
Total	10,568	33,549.077	3.18762			

Table 2. AMOVA results for mtDNA HVS-1 sequences in KPP and comparative populations.

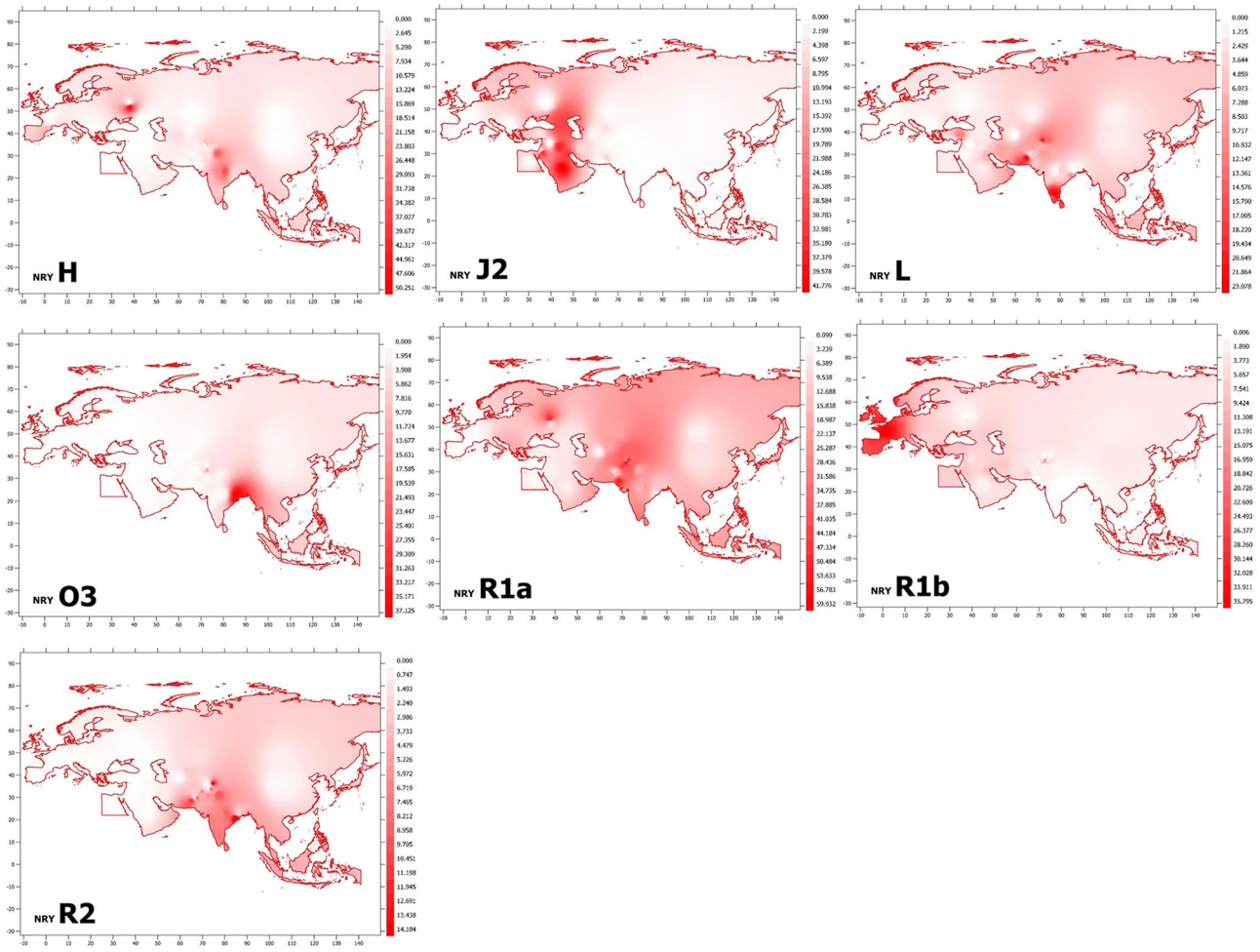

Figure 3. A geospatial map of NRY haplogroup frequencies in KPP ethnic groups and comparative populations. See the Methods section for a description of the mapping process, and Table S6 for the data on which the projection is based.

Populations	n	#	HD	PD
Gujars	124	37	0.910	5.47 ± 2.65
Jadoons	114	36	0.796	3.07 ± 1.61
Syeds	129	31	0.794	3.05 ± 1.60
Tanolis	134	32	0.795	3.21 ± 1.67
Yousafzais	177	85	0.905	4.32 ± 2.15
Yousafais_Old	146	90	0.966	5.07 ± 2.48
Gujars_Swat	20	10	0.758	5.29 ± 2.67
Kohestani	20	14	0.890	6.13 ± 3.04
Tarklani	20	9	0.837	2.65 ± 1.47
Utmankhail	20	6	0.684	2.27 ± 1.30
Yousafzais_Swat	20	10	0.800	3.43 ± 1.83
Pathan_Pakistan	270	152	0.973	5.61 ± 2.70
Kashmiri	101	68	0.981	6.38 ± 3.05
Hazara_Pakistan	153	73	0.910	4.15 ± 2.07
Punjabi	394	266	0.995	6.10 ± 2.91
Sheikh	180	100	0.984	6.15 ± 2.94
Gujars_Punjab	176	84	0.971	6.23 ± 2.97
Baluch	59	48	0.988	6.68 ± 3.20
Brahui	110	80	0.968	5.66 ± 2.73
Burusho	86	55	0.990	6.17 ± 2.96
Kalash	44	23	0.946	5.55 ± 2.72
Makrani	58	52	0.996	6.75 ± 3.23
Parsi_Pak	90	56	0.969	6.09 ± 2.93
Sindhi	122	97	0.990	5.97 ± 2.87

Table 3. Summary statistics for KPP ethnic groups and other Pakistani populations based on Y-STR haplotypes. n = number of samples; # = number of haplotypes; HD: Haplotype Diversity; PD: Pairwise differences; Citations and references for the comparative populations are provided in Table S10.

Discussion

Previous genetic research on Pakistani populations has largely been limited to studies of either Y-STR or mtDNA variation in a single ethnic group[83,85–87,123–131], or Y-chromosome and mtDNA analysis in many ethnic groups with limited sample sizes[81,82,84,132]. None of these studies broadly analyzed genetic diversity among the myriad ethnic groups residing in Pakistan. This study provides the first survey of mtDNA and NRY variation in members of five major ethnic groups inhabiting Buner and Swabi Districts of KPP.

The paternal and maternal gene pools of the KPP populations were found to consist of West Eurasian, South Asian and East Asian lineages. However, the patterns of mtDNA and NRY diversity amongst these populations are fairly different, suggesting contrasting paternal and maternal genetic histories for them. Based on mtDNA data, Syeds, Yousafzais and Jadoons have close affinities to one another, while NRY data reveal close affinities between Gujars, Syeds and Yousafzais. Overall, Yousafzais show greater genetic affinities with the Syeds than to any of the other three ethnic groups, whereas Tanolis, Jadoons and Gujars were outliers in the NJ trees relative to the other KPP populations, depending on the data set being analyzed. Such results are not entirely surprising, given that Yousafzais and Syeds claim to be Afghans, or at least in the latter case "Arabs." By contrast, Gujars are not considered Pathans at all, nor are Tanolis, while Jadoons are allegedly descendants of Pashtun leader Ghurghusht, the third and youngest son of Qais[58]. We elaborate on the histories of these ethnic groups below.

Yousafzais share more mtDNA haplotypes with all other ethnic groups than any of the other four and share the most with Syeds. This observation likely reflects both the larger sample size and the effective population size for Yousafzais. These data also suggest some degree of gene flow between these populations, or else the selection of marriage partners whose mtDNAs draws from the common maternal gene pool for South-Central Asia that developed over time[7,33,63,133,134]. The resulting diversity is now being resorted and reshuffled within extant ethnic populations.

The MJ networks of the major common NRY haplogroups show that flow of paternal lineages among the five ethnic groups is quite limited and consistent with high levels of male endogamy practiced by them. Similar Y-chromosome results have been previously reported for Central Asian and South Asian ethnic groups[44,82,91,100,135,136], but with less pronounced genetic differentiation in maternal lineages[135–138].

These findings are consistent with evidence from historical and ethnographic research involving populations from this region. According to Barth[80], there has been relatively little intermarriage between any of the members of these ethnic groups. They tend to live in isolation relative to other ethnic groups and discourage intermarriage between them. However, as Barth notes[80], "the Pathans allow marriages of equals, even when close relatives, and the giving of a daughter to a man of superior status but discourage the giving of a woman in marriage to inferior men. Landowners, as a group, thus, tend to marry endogamously but they also take some women in marriage

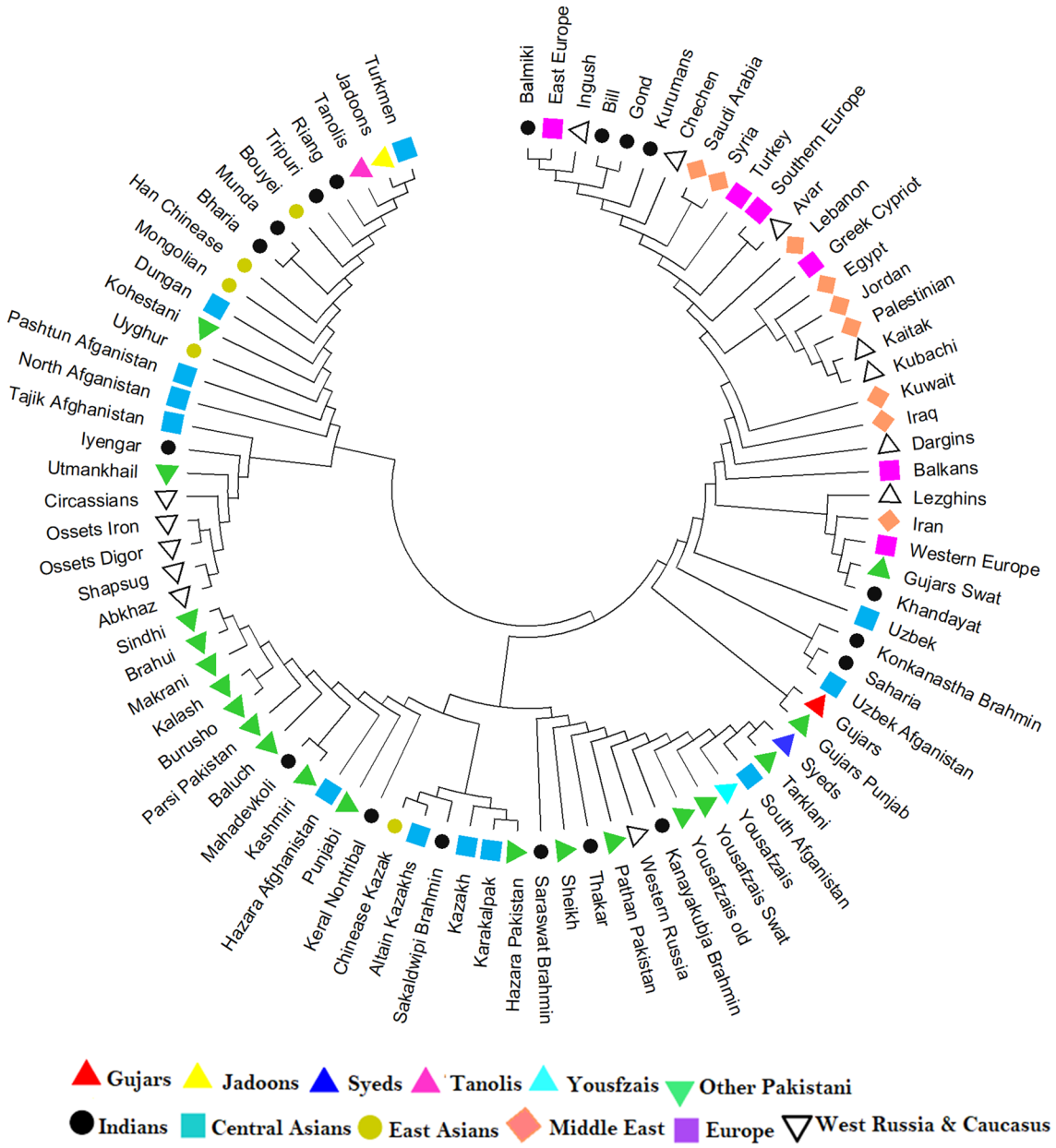

Figure 4. A Neighbor-Joining tree showing the genetic relationships between KPP ethnic groups and 82 world populations based on R_{ST} estimates from Y-STR haplotype data (Table S8).

from lower groups, whereas they will not give up their daughters in marriage to inferiors." These practices may possibly have shaped the population dynamics of KPP populations and led researchers to describe local populations as being genetically isolated and marked by high levels of inbreeding[139–141].

Such an explanation overlooks two important facts, though. First, as devout Muslims, Pathans believe that all individuals are equal in the eyes of the creator. Consequently, there is no absolute genealogical "litmus test" of worthiness equivalent to the Hindu notion of inborn purity and pollution[142,143]. Second, in the political sphere, Pathans are extremely competitive, and Pathan chiefs tend to spend far beyond the revenue generated by their landholdings[144]. Because of these two factors, economic advantage can outweigh inherited social status in arranging marital partners, especially when village leaders are seeking to consolidate their power in the political arena[80,142,145–149].

With respective to the different KPP ethnic groups, Gujars are characterized by having predominantly R1a1a-M17 Y-chromosomes, the frequency of which is the highest observed among the populations of the Indus Valley[89]. Otherwise, they are marked by 30% SA haplotypes and a low frequency of EA haplotypes. They also have a high frequency of South Asian haplogroup L-M20 compared to other KPP populations, supporting their historically documented affinities with various South Asian ethnic groups[150–153], especially those residing in the northwestern portion of South Asia[154].

No Groups						
Source of variation	**d.f.**	**Sum of squares**	**Variance component**	**% of variance**	**Fixation Indices**	**P-value**
Among populations	79	18,852.353	1.59547 Va	18.23	F_{ST}: 0.182	0.000 ± 0.000
Within populations	11,641	83,324.558	7.15785 Vb	81.77		
Total	11,720	102,176.911	8.75333			
Grouped by Geography (Ten Groups: Country or Region)						
Among groups	9	6701.154	0.33529 Va	3.81	F_{SC} (Va):0.153	0.000 ± 0.000
Among populations within groups	70	12,151.198	1.29605 Vb	14.75	F_{ST} (Vb):0.186	0.000 ± 0.000
Within populations	11,641	83,324.558	7.15785 Vc	81.44	F_{CT} (Vc):0.038	0.000 ± 0.000
Total	11,720	102,176.911	8.78919			
Grouped by Ethnicity						
Among groups	14	8879.462	0.44807 Va	5.10	FSC (Va):0.141	0.000 ± 0.000
Among populations within groups	65	9972.890	1.17945 Vb	13.43	FST (Vb):0.185	0.000 ± 0.000
Within populations	11,641	83,324.558	7.15785 Vc	81.47	FCT (Vc):0.051	0.000 ± 0.000
Total	11,720	102,176.911	8.78537			

Table 4. AMOVA results for Y-STR haplotypes in KPP and comparative populations.

Gujar maternal ancestry is largely congruent with their paternal genetic ancestry. Their mtDNAs are largely of WE origin, although some derive from SA and EA[81]. Gujars also possess haplogroups linking them with West Asia (HV, U7, W) while having relatively few EA mtDNAs. Based on this pattern of diversity, they show strong genetic affinities with the Yousafzais, Kashmiri and other Pakistani populations.

With respect to their origins, one hypothesis proposes that Gujars expanded into India from Central Asia, while another suggests they came from Georgia via Afghanistan in the fifth century CE. Based on our data, Gujars generally resemble Iranians who mixed with local populations rather than populations from the Caucasus[155]. By contrast, Gujars of Jammu and Kashmir show mtDNA affinities with populations from Uttar Pradesh and Arunachal Pradesh to the east[156]. Thus, our data suggest a more complex origin for KPP Gujars. One scenario could involve an indigenous population with genetic affinities to geographically proximate Jats and Rajputs mixing with Indo-Iranian or Turkic speaking Muslim populations, and then migrating into the region from the steppes of Central Asia[157].

Unlike other KPP ethnic groups, the Jadoons exhibit a strongly East Asian paternal ancestry, with NRY haplogroups O3-M122 and Q-MEH2 representing 82.5% of their Y-chromosomes. Although O3-M122 is very rare among South Asian populations[44], Q-M242 appears at modest frequencies in them. In the NRY NJ tree, the Jadoons occupy an outlier position relative to other KPP populations, but exhibit affinities with Turkmen from Central Asia. In this regard, Mongol expansions into Central-South Asia probably brought NRY haplogroups C3, O3, and Q to Pakistan during the medieval period, and NRY diversity in Kazakh populations from Central Asia was probably similarly influenced during this time[94,158,159]. The rest of their Y-chromosomes belong to either WE or SA haplogroups, and appear similar to types present in other KPP populations, suggesting some degree of gene flow between them.

Jadoons mtDNA shows greatest similarity to groups from WE followed by SA with less affinity to EA groups. As such, Jadoons exhibit a pattern of extra-regional affinities that are generally like those observed among Gujars, Syeds and Yousafzais. The neighbor-joining tree also identifies affinities to European Roma populations and to other South-Central Asian groups. As such, these results corroborate previous studies that identify a genetic affinity of Roma populations to South Asian groups, especially those residing in the northwestern region of the subcontinent[160,162]. Viewed as a whole, genetic diversity among Jadoons appears to reflect a scenario in which male-mediated gene flow into the region was followed by these immigrant males subsequently marrying indigenous females thereby yielding a maternal gene pool similar to those possessed by members of other Pakistani and Central Asian ethnic groups.

Most Syeds possess NRY haplogroup R1a1a-M17, along with a unique array of Y-STR haplotypes of this patrilineage that is coupled with low prevalence of Y-chromosomal variations common to South and East Asians. This combination aligns Syeds with other Pakistani and Central Asian ethnic groups while distancing them from ethnic groups of the rest of the subcontinent and East Asia. Matrilineal genetics yield a similar pattern. Syeds are marked by high frequencies of WE lineages coupled with low frequencies of lineages common to South and East Asians. While this pattern aligns Syeds with other Pakistani and some Indian Samples[7,63,85,87,118,125–127,132], they are distinct through their hgh frequencies of haplogroup U2, T2, M9, X and R30 mtDNAs.

Although Syeds are hypothesized to have come from the Near East and entered South and Central Asia during the Mongol invasions of the thirteenth and fourteenth centuries, mtDNA and NRY data instead support a scenario in which Syeds have an ultimate origin in Afghanistan coupled with long-standing gene flow with Central Asian populations[163]. Moreover, the topography of northern Pakistan with its formidable mountains and narrow, steep-sided valleys may have fostered the establishment of localized endogamous social groups that, over time, developed into largely reproductively isolated distinct ethnic groups.

This explanation corroborates results obtained in other mtDNA studies from South Asia[63,85,132]. In which populations residing west of the Indus Valley possess mtDNA lineages largely of West Eurasian derivation with limited contributions from South Asia and East Eurasia, while those found to the south and to the east are characterized by mtDNA profiles that feature higher frequencies of deep-rooted lineages indigenous to South Asia[63]. Likewise ethnic groups from KPP are marked by generally show close affinities to one another, but share only distant affinities to populations from Iran, Uzbekistan and Kazakhstan[81,85]. Sharma et al.[134] further noted the mtDNA divergence between ethnic groups of Jammu and Kashmir in northern India to be greater than within Pakistani groups or populations from Europe and the Caucasus. Such results not only document the limited impact of the medieval incursion of different Pashtun ethnic groups from Afghanistan and the Iranian Plateau into the northwestern periphery of South Asia, but also suggest that such introgressive events involved non-local individuals of both sexes, rather than being limited to males[134,164].

Tanolis, whose communities are restricted to hilly area of Swabi District along the border with Buner and Haripur Districts of KPP, have predominantly R1b1a-P297 Y-chromosomes, along with a low frequency of SA and EA haplotypes. Based on several studies, haplogroup R1b is thought to have spread with pastoralism and Indo-European speakers into South Asia[165–168]. For this reason, the Tanolis are relatively dissimilar to other KPP and comparative populations. From a mtDNA perspective, Tanolis have a high frequency of haplogroup N3, which arose in Western Eurasia, as well as higher frequencies of SA haplogroups such as M2-M6 than other populations. As a result, Tanolis show genetic similarities with Siddi[169] and other populations from India.

Given this genetic profile, Tanolis may be an outlier within the Indo-Pakistani subcontinent. While suggested to have Turkic roots, and also claiming Pashtun ancestry tracing to the fifteenth century CE, the Tanolis appear to have a different genetic origin than the other four KPP ethnic groups. It is possible that they have experienced significant genetic drift, perhaps due to founder effects, which would affect the frequencies of their paternal lineages. Yet, the latter scenario is not consistent with the high proportion (47.8%) of South Asian mtDNA lineages observed in Tanolis relative to other KPP ethnic groups.

Yousafzais are the most genetic diverse KPP population in this study. While exhibiting a high frequency of R1a1a Y-chromosomes, they also have a mixture of other NRY haplogroups, including West Eurasian G2, I2, J1 J2, South Asian H, L, R2 and East Asian O3, Q. Four of these haplogroups (G2a, R1a, J2a1b, I2a) are likely associated with male-mediated migrations related to Neolithic farming[45,98,101,170,171]. They also exhibit genetic affinities with other Yousafzais populations[82,83]. The Yousafzais also exhibit considerable mtDNA diversity. Their maternal lineages are largely of WE derivation, with moderate frequencies of SA and low frequency of EA haplotypes also being present. Based on these data, the Yousafzais show genetic similarities to Central Asian and other Pakistani populations.

Overall, Yousafzais are marked by affinities to local non-Pathan groups both paternally and maternally. Such findings suggest that Yousafzais absorbed a number of local males, perhaps through religious conversion of the most successful landholders[80], and also integrated local females into their population, either as spouses and daughters of local non-Pathan converts or through hypergamous unions with Yousafzai men[80]. Both avenues of gene flow are well-documented throughout South Asia[55,145,146,172–175], including regions of northern Pakistan such as Gilgit-Baltistan[147], most likely reflect endogamous practices that involved the assimilation of foreign females into the populations.

Conclusions

As described above, the patterns of mtDNA and NRY diversity amongst the KPP ethnic groups are fairly different, suggesting contrasting paternal and maternal genetic histories for them. We have attempted to situate these data in the context of archaeological, ethnographic, historical, genetic, and linguistic evidence to better explain the complex pattern of ethnic diversity in Pakistan and the KPP region. Yet, as shown in this genetic analysis, there are many uncertainties in the population histories of these ethnic groups. Future analysis of mitogenome and whole genome sequences will greatly facilitate the testing of the hypothesized origins and biological relationships of KPP populations outlined in this study.

Materials and methods

Sample and data collection. Ethnographic fieldwork and sample collection were undertaken in 13 villages located within Buner and Swabi Districts of KPP in 2014–15 (Fig. 5). Within the Buner District, the villages included Bajkata, Channar Swari, Dewana Baba, Kingargalai, Sonigram, Swari Bazar, Takhtaband, while in Swabi District they included Dalori Gadoon, Dobyan, Gani Chatra, Kabgany Gadoon, Utla and Yar Hussain. A total of 700 unrelated male volunteers from five endogamous ethnic groups were the subjects of this research effort. Prior to starting this study, the Institutional Biomedical Ethics Committee of Hazara University Mansehra reviewed the project details and approved the protocol for obtaining informed written consent from study participants (ref # 73/HU/ORIC/IBC/2013). All experimental procedures were carried out in accordance with the approved guidelines of the Research Ethics Committee of Hazara University Mansehra. This research was also approved by the Institutional Biomedical Ethics Committee, Islamia College Peshawar (ref #530/ORIC/ICP) and the University of Pennsylvania IRB #8.

DNA analysis. *Genomic DNA preparation.* Saliva samples were collected from all participants with informed consent written in English and Urdu. Genomic DNAs were isolated from the buccal cell samples using a modified protocol of Aidar and Line[176]. DNA concentrations were measured with a NanoDrop ND-1000 spectrophotometer and normalized to 1 ng/µl.

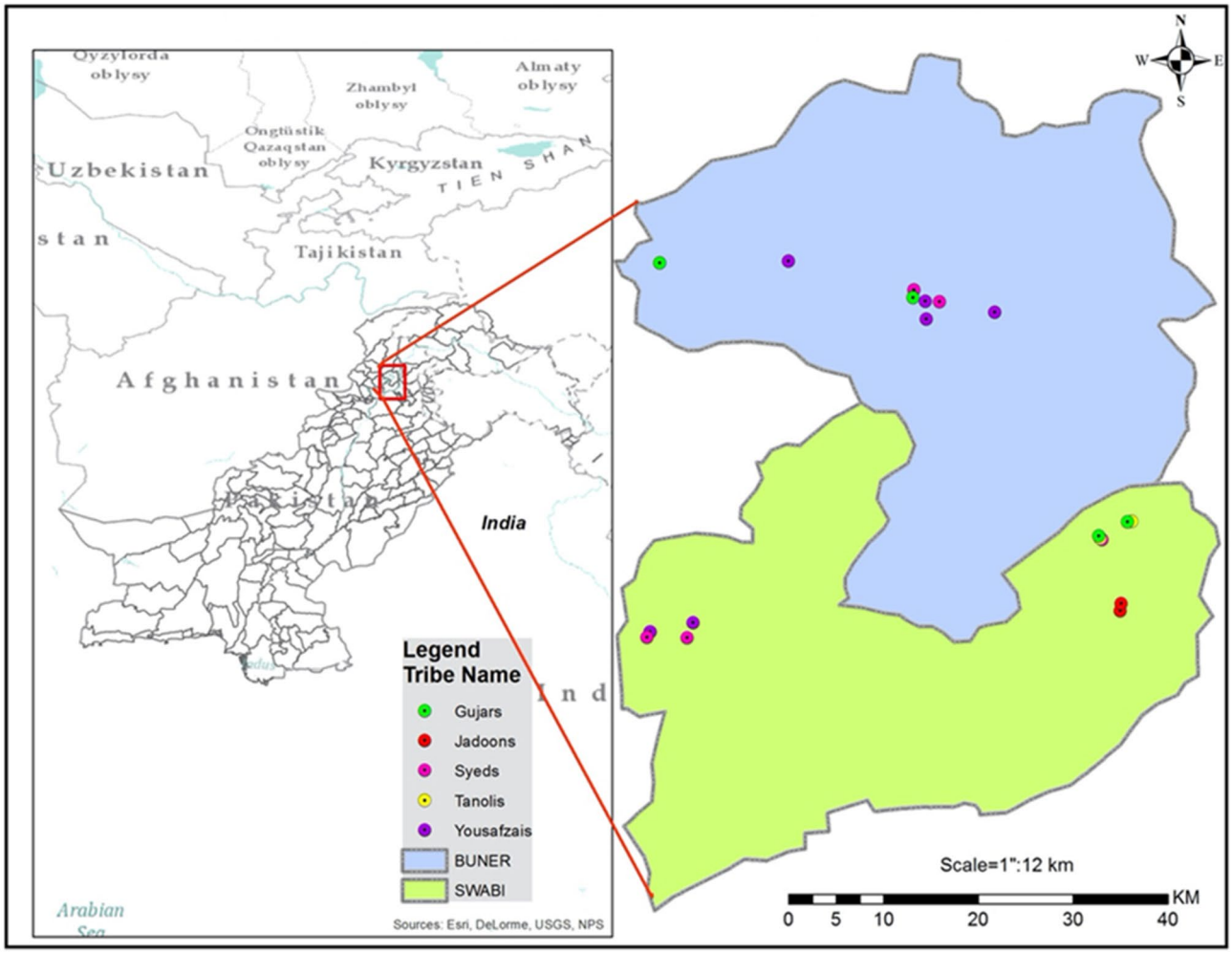

Figure 5. A map of Pakistan showing the locations of fieldwork in the Khyber Pakhtunkhwa Province. DNA samples were collected from areas in which each ethnic group was highly concentrated. Gujars, Syeds and Yousafzais were sampled from both Buner and Swabi Districts, while the Jadoons and Tanolis were only sampled in the Swabi District. The map was created with the ArcGIS software, v10.3.1., based on source map from ESRI https://www.esri.com/en-us/home.

Mitochondrial DNA analysis. Maternal genetic ancestry was elucidated through analysis of mtDNA variation. A total of 659 individuals from five ethnic groups were surveyed for mtDNA variation. All samples were screened for phylogenetically informative single nucleotide polymorphisms (SNPs) defining the major branches (haplogroups) of the human mtDNA phylogeny with custom TaqMan assays[177–179]. All TaqMan assays were read on an ABI Prism 7900 HT Fast Real-Time PCR System. SDS v2.3 was used to analyze all runs, and the resulting allelic calls checked through visual inspection.

Samples were then surveyed for sequence variation through control region (CR) sequencing using published methods[177–179]. All variable positions were determined relative to the revised Cambridge Reference Sequence (rCRS)[180,181]. The CR sequence data defined maternal haplotypes in these individuals, and all haplogroups were ascertained relative to existing mtDNA databases, such as Phylotree version 17[182,183]. A single PCR–RFLP test was also used to screen mtDNA samples for the 14,766 SNP that characterizes haplogroup HV.

Y-chromosome analysis. Paternal genetic ancestry was elucidated through analysis of NRY variation in male participants. The DNAs of 678 individuals were screened for phylogenetically informative SNPs in a hierarchical fashion according to published information and previously published methods[93,184–186]. The SNP genotyping involved 47 custom TaqMan assays that were read using the ABI 7900HT Fast Real-Time PCR System[177–179]. They were then additionally surveyed for variation at 17 Y-STR loci using the ABI Y-filer Kit, as previously described[177–179]. Two other STR loci, along with six insertion–deletion (indel) SNPs, were genotyped in a separate custom multiplex assay[177–179]. The multiplex PCRs were read on an ABI 3130xl Genetic Analyzer with POP-4 polymer using GeneScan™-500 LIZ™ size standards. The resulting data were analyzed using GeneMapper1 ID Software v3.2. STR allele sizes were identified based on previous recommendations[187]. Quality control procedures included checking SNP genotypes for phylogenetic consistency and comparing the data with haplogroups

predicted from STR profiles (http://www.hprg.com/hapest5/index.html). The paternal haplotype for each sample was designated by its full 19-STR locus profile.

Y chromosome lineages (haplogroups) were defined as the unique combinations of SNP and STR data present in the samples. DYS389b was calculated by subtracting DYS389I from DYS389II, which was used for all statistical network analyses. Each male sample was assigned a SNP haplogroup following the conventions outlined by the Y-chromosome Consortium[93,184] and detailed in PhylotreeY[188]. All of the Y-STR haplotypes were further checked for their haplogroup status using Athey's (http://www.hprg.com/hapest5/) and the Nevgen Y-DNA (http://www.nevgen.org/) haplogroup predictors. The SNPs and STR alleles defined the haplogroups and haplotypes, respectively, for each male individual.

Comparative populations. We compared the mtDNA and NRY data obtained from the five Pakistani ethnic groups to those from populations in South Asia, Central Asia, East Asia, Middle East and Europe in an effort to place the genetic histories of these five Pakistani ethnic groups within a broader framework. For the mtDNA analysis, we examined a total of 11,411 mtDNA HVS1 sequences, including 659 from this study and the rest from comparative populations in South, Central and East Asia, Europe, Caucasus, and the Near East (Table S9). In addition, we compared 12,519 Y-STR haplotypes including 678 from this study and the rest from comparative populations in South, Central and East Asia, Europe, Caucasus, and the Near East (Table S10). All Y-STR haplotypes were reduced to ten loci (DYS19, DYS389I, DYS398b, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439) to allow for the broadest comparison possible.

Statistical analyses. Haplotype diversity (h), nucleotide diversity (p) and pairwise differences were calculated for mtDNA HVS-1 sequences (np 16,024–16,400) and Y-STR haplotypes using Arlequin v. 3.5.2.1[189]. Descriptive statistical indices, as well as Tajima's D[190] and Fu's FS[191] neutrality tests, were calculated using the same software. Pairwise F_{ST} and R_{ST} distances values between populations were calculated from HVS-1 sequences and Y-STR haplotypes using the same software. F_{ST} values were estimated with the Tamura and Nei (1993) model of sequence evolution. The resulting matrices were visualized in Neighbor-joining trees using MEGA version X[192]. For both the mtDNA and NRY data sets, the genetic structure of Pakistani and comparative populations was examined through analysis of molecular variance (AMOVA) in Arlequin v. 3.5.2.1[189].

Phylogenetic analysis. Median-joining networks[193] were constructed for both mtDNA HVS-1 sequences and Y-STR haplotypes using Network version 5.0.1.1[193,194] to explore the phylogenetic history of the genetic lineages encompassed within the five sampled ethnic groups. For mtDNA HVS-1 sequences, the mutation-weighting scheme was based on that described by Bandelt and coworkers[195], in which fast-evolving sites were given lower weights relative to less mutable sites. All variants known to result from homopolymeric C expansions (e.g., A16182C, A16183C) or to occur at mutational hotspots in the mtDNA CR (e.g., T16519C) were excluded from the haplotypes used in this analysis.

The NRY haplotypes used to generate the networks for specific haplogroups consisted of 17 Y-STR loci. Y-STR loci were weighted according to their individual mutation rates[196] by applying a fivefold weighting scheme with higher weights given to slowly evolving markers and lower weights to faster evolving markers. The multicopy marker DYS385 was not used in the analysis because the differentiation between its alleles was not possible to ascertain using the Y-Filer kit[187].

Geospatial frequency maps. The GPS coordinates for the KPP and comparative populations were determined and used to create geospatial maps of haplogroup frequencies with QGIS Desktop v.3.20.0 using the EPSG 4326 coordinate system. The resulting maps were exported at a scale of 1:48,000,000. Continent and country boundary vector data were procured from free-use, publicly available World Health Organization assets. The data points were organized by latitude/longitude coordinates from Tables S3 and S7, with the geospatial coordinates being used as the sample points in an Inverse Distant Weighted (IDW) interpolation calculation. A weight of 3 was used in these calculations to help clarify the produced raster visualization's color ramp and decrease the known disadvantage of IDWs with irregular sample point distributions that produces visual peaks and pits around sample points. The resulting rasters were then clipped by the vector land boundaries and their color ramps clipped to 20 values between the min and max before being exported in PDF format[197,198]. This interpolation does not take into account natural land boundaries, water boundaries, or cultural boundaries that would affect the falloff of influence from neighboring sample points, since it is solely based on geographic distance.

Data availability

The majority of the data discussed in this paper are provided in the Supplementary Tables, including mtDNA control region sequences and Y-SNP and Y-STR data. The mtDNA HVS1 sequences have also been deposited in the NCBI GenBank at https://www.ncbi.nlm.nih.gov/genbank/ under Accession numbers XXXXX-XXXXX.

Received: 1 April 2021; Accepted: 10 December 2021
Published online: 19 January 2022

References

1. Macaulay, V. *et al.* Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308**, 1034–1036. https://doi.org/10.1126/science.1109792 (2005).
2. Majumder, P. P. The human genetic history of South Asia. *Curr. Biol.* **20**, R184-187. https://doi.org/10.1016/j.cub.2009.11.053 (2010).

3. Mellars, P., Gori, K. C., Carr, M., Soares, P. A. & Richards, M. B. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 10699–10704. https://doi.org/10.1073/pnas.1306043110 (2013).
4. Yelmen, B. *et al.* Ancestry-specific analyses reveal differential demographic histories and opposite selective pressures in modern South Asian Populations. *Mol. Biol. Evol.* **36**, 1628–1642. https://doi.org/10.1093/molbev/msz037 (2019).
5. Templeton, A. Out of Africa again and again. *Nature* **416**, 45–51. https://doi.org/10.1038/416045a (2002).
6. Bae, C. J., Douka, K. & Petraglia, M. D. On the origin of modern humans: Asian perspectives. *Science* **358**, 9067. https://doi.org/10.1126/science.aai9067 (2017).
7. Metspalu, M. *et al.* 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. https://doi.org/10.1186/1471-2156-5-26 (2004).
8. Kivisild, T. *et al.* The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* **72**, 313–332. https://doi.org/10.1086/346068 (2003).
9. Thangaraj, K. *et al.* Reconstructing the origin of Andaman Islanders. *Science* **308**, 996. https://doi.org/10.1126/science.1109987 (2005).
10. James, H. A. & Petraglia, M. Modern human origins and the evolution of behavior in the later Pleistocene record of South Asia. *Curr. Anthropol.* **46**, S3–S27. https://doi.org/10.1086/444365 (2005).
11. Field, J. S., Petraglia, M. D. & Lahr, M. M. The southern dispersal hypothesis and the South Asian archaeological record: Examination of dispersal routes through GIS analysis. *J. Anthropol. Archaeol.* **26**, 88–108 (2007).
12. Reyes-Centeno, H. *et al.* Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 7248–7253. https://doi.org/10.1073/pnas.1323666111 (2014).
13. Dennell, R. & Petraglia, M. D. The dispersal of Homo sapiens across southern Asia: how early, how often, how complex?. *Quatern. Sci. Rev.* **47**, 15–22 (2012).
14. Forster, P. & Matsumura, S. Evolution. Did early humans go north or south?. *Science* **308**, 965–966. https://doi.org/10.1126/science.1113261 (2005).
15. Field, J. S. & Lahr, M. M. Assessment of the southern dispersal: GIS-based analyses of potential routes at oxygen isotopic stage 4. *J. World Prehist.* **19**, 1–45 (2005).
16. Underhill, P. A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43–62. https://doi.org/10.1046/j.1469-1809.2001.6510043.x (2001).
17. Xing, J. *et al.* Genetic diversity in India and the inference of Eurasian population expansion. *Genome Biol* **11**, R113. https://doi.org/10.1186/gb-2010-11-11-r113 (2010).
18. Kivisild, T. *et al.* The Place of the Indian Mitochondrial DNA Variants in the Global Network of Maternal Lineages and the Peopling of the Old World. In *Genomic Diversity* (eds Papiha, S. S. *et al.*) 135–152 (Springer, Boston, MA, 1999).
19. Moorjani, P. *et al.* Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93**, 422–438. https://doi.org/10.1016/j.ajhg.2013.07.006 (2013).
20. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206. https://doi.org/10.1038/nature18964 (2016).
21. Pagani, L. *et al.* Genomic analyses inform on migration events during the peopling of Eurasia. *Nature* **538**, 238–242. https://doi.org/10.1038/nature19792 (2016).
22. Thapar, B. K. A Harappan Metropolis beyond the Indus Valley. In *Ancient Cities of the Indus* (ed. Possehl, G. L.) 196–202 (Carolina Academic Press, 1979).
23. Possehl, G. L. *Harappan civilization: A contemporary perspective* (Aris & Phillips, 1982).
24. Shaffer, J. G. & Lichtenstein, D. A. Ethnicity and change in the Indus valley cultural tradition. Old Probl. New Perspect. *Archaeol. South Asia* **2**, 117–126 (1989).
25. Mughal, M. R. The Decline of the Indus Civilization and the Late Harappan Period in the Indus Valley. *Lahore Museum J.* **3**, 1–22 (1990).
26. Possehl, G. L. Revolution in the urban revolution: The Emergence of the Indus urbanization. *Annu. Rev. Anthropol.* **19**, 261–282 (1990).
27. Nath, A. Rakhigarhi: A Harappan metropolis in the Saraswati-Drishadvati divide. *Puratattva* **28**, 39–45 (1998).
28. Singh, K. S. *India's Communities: A-G. People of India National Series Delhi* Vol. 4 (Oxford University Press, 1998).
29. Singh, K. S. *People of India: An Introduction* (Anthropological Survey of India, Kolkata, 1992).
30. Majumder, P. P. People of India: Biological diversity and affinities. *Evol. Anthropol.* **6**, 100–110 (1998).
31. Silva, M. *et al.* A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC Evol. Biol.* **17**, 88 (2017).
32. Ahloowalia, B. S. *Invasion of the Genes: Genetic Heritage of India* (Eloquent Books, 2009).
33. Ayub, Q. & Tyler-Smith, C. Genetic variation in South Asia: assessing the influences of geography, language and ethnicity for understanding history and disease risk. *Brief. Funct. Genomic. Proteomic.* **8**, 395–404. https://doi.org/10.1093/bfgp/elp015 (2009).
34. Southworth, F. *Linguistic Archaeology of South Asia* (Routledge, 2004).
35. Kachru, B. B., Kachru, Y. & Sridhar, S. N. *Language in South Asia* 122–127 (Cambridge University Press, 2008).
36. Kosambi, D. D. *The Culture and Civilization of Ancient India in Historical Outline* (Vikas Publishing House, 1991).
37. Jarrige, J. F. Economy and society in the Early Chalcolithic/ Bronze Age of Baluchistan: New perspectives from recent excavations at Mehrgarh. *South Asian Archaeol.* **1979**, 93–114 (1981).
38. Jarrige, J. F. Mehrgarh Neolithic. *Pragdhara* **18**, 135–154 (2008).
39. Costantini, L. The Beginning of Agriculture in the Kachi Plain: The Evidence of Mehrgarh. In *South Asian Archaeology* (ed. Allchin, B.) 29–33 (Cambridge University Press, 1984).
40. Kenoyer, J. M. The Indus Valley Tradition of Pakistan and western India. *J. World Prehist.* **5**, 331–385. https://doi.org/10.1007/BF00978474 (1991).
41. Kenoyer, J. M. *Ancient Cities of the Indus Valley Civilization* (Oxford University Press, 1998).
42. Hemphill, B. E. & Mallory, J. P. Horse-mounted invaders from the Russo-Kazakh steppe or agricultural colonists from western Central Asia? A craniometric investigation of the Bronze Age settlement of Xinjiang. *Am J Phys. Anthropol.* **124**, 199–222. https://doi.org/10.1002/ajpa.10354 (2004).
43. Sahoo, S. *et al.* A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proc. Natl. Acad. Sci. U.S. A.* **103**, 843–848. https://doi.org/10.1073/pnas.0507714103 (2006).
44. 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. *Am. J. Hum. Genet.* **78**, 202–221. https://doi.org/10.1086/499411 (2006).
45. de Barros Damgaard, P. *et al.* The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, eaar7711 (2018).
46. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, eaat7487 (2019).
47. Wolpert, S. *A New History of India* (Oxford University Press, 2000).
48. Anthony, D. W. & Ringe, D. The Indo-European Homeland from Linguistic and Archaeological Perspectives. *Annu. Rev. Linguist.* **1**, 199–219 (2015).

49. Hemphill, B. E. Biological affinities and adaptations of Bronze Age Bactrians: III. An initial craniometric assessment. *Am. J. Phys. Anthropol.* **106**, 329–348. https://doi.org/10.1002/(sici)1096-8644(199807)106:3%3c329::Aid-ajpa6%3e3.0.Co;2-h (1998).
50. Hemphill, B. E. Biological affinities and adaptations of Bronze Age Bactrians: IV. A craniometric investigation of Bactrian origins. *Am. J. Phys. Anthropol.* **108**, 173–192. https://doi.org/10.1002/(sici)1096-8644(199902)108:2%3c173::Aid-ajpa4%3e3.0.Co;2-3 (1999).
51. Hemphill, B. E. Grades, gradients, and geography: A dental morphometric approach to the population history of South Asia. *Anthropological Perspectives on Tooth Morphology: Genetics, Evolution, Variation* (2013).
52. Hemphill, B. E. A View to the North: Biological Interactions across the Intermontane Borderlands during the Last Two Millennia B.C. In *South Asian Archaeology 2008* Vol. 1 (eds Tosi, M. & Frenez, D.) (Archaeopress-BAR, Oxford, 2013).
53. Hemphill, B. E. Assessing Odontometric Variation among Populations. In *A Companion to Dental Anthropology* (eds Irish, J. D. & Scott, G. R.) (Wiley, Malden, MA, 2016).
54. Farah, C. E. *Islam* (Baron's Educational Series, New York, 2003).
55. Titus, M. T. *Islam in India and Pakistan* (Munshiram Manoharlal, New Delhi, 2005).
56. Keay, J. *India* (Grove Press, 2000).
57. Eaaswarkhanth, M. *et al.* Diverse genetic origin of Indian Muslims: evidence from autosomal STR loci. *J. Hum. Genet.* **54**, 340–348 (2009).
58. Caroe, O. *The Pathans* (Oxford University Press, 1958).
59. Broushaki, F. *et al.* Early Neolithic genomes from the eastern Fertile Crescent. *Science* **353**, 499–503. https://doi.org/10.1126/science.aaf7943 (2016).
60. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536**, 419–424. https://doi.org/10.1038/nature19310 (2016).
61. de Barros Damgaard, P. *et al.* The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, 7711. https://doi.org/10.1126/science.aar7711 (2018).
62. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, 7487. https://doi.org/10.1126/science.aat7487 (2019).
63. Quintana-Murci, L. *et al.* Where west meets east: The complex mtDNA landscape of the southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845. https://doi.org/10.1086/383236 (2004).
64. McElreavey, K. & Quintana-Murci, L. A population genetics perspective of the Indus Valley through uniparentally-inherited markers. *Ann. Hum. Biol.* **32**, 154–162. https://doi.org/10.1080/03014460500076223 (2005).
65. Thangaraj, K. *et al.* The influence of natural barriers in shaping the genetic structure of Maharashtra populations. *PLoS ONE* **5**, e15283. https://doi.org/10.1371/journal.pone.0015283 (2010).
66. Metspalu, M. *et al.* Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am. J. Hum. Genet.* **89**, 731–744. https://doi.org/10.1016/j.ajhg.2011.11.010 (2011).
67. Sharma, G. *et al.* Genetic affinities of the central Indian tribal populations. *PLoS ONE* **7**, e32546. https://doi.org/10.1371/journal.pone.0032546 (2012).
68. Gazi, N. N. *et al.* Genetic structure of Tibeto-Burman populations of Bangladesh: evaluating the gene flow along the sides of Bay-of-Bengal. *PLoS ONE* **8**, e75064. https://doi.org/10.1371/journal.pone.0075064 (2013).
69. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494. https://doi.org/10.1038/nature08365 (2009).
70. Chaubey, G. *et al.* "Like sugar in milk": Reconstructing the genetic history of the Parsi population. *Genome Biol.* **18**, 110. https://doi.org/10.1186/s13059-017-1244-9 (2017).
71. Chaubey, G., Kadian, A., Bala, S. & Rao, V. R. Genetic Affinity of the Bhil, Kol and Gond Mentioned in Epic Ramayana. *PLoS ONE* **10**, e0127655. https://doi.org/10.1371/journal.pone.0127655 (2015).
72. Green, N. Tribe, diaspora, and sainthood in Afghan history. *J. Asian Stud.* **67**, 171–211 (2008).
73. Rahim, M. A. *History of the Afghans in India* (Pakistan Publishing House, 1961).
74. Strand, R. F. Notes on the Nuristani dardic languages. *J. Am. Orient. Soc.* **93**, 297–305 (1973).
75. Morgenstierne, G. Iranian elements in Khowar. *Bull. Schl. Orient. Stud.* **8**, 657–671 (1936).
76. Morgenstierne, G. *Report on a Linguistic Mission to North-Western India* (Instituttet for sammenlignende Kulturforskning, 1932).
77. Menges, K. H. Sociolinguistics and South-Asia. *Central Asiatic J.* **29**, 25–34 (1985).
78. Basu, A. *et al.* Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13**, 2277–2290. https://doi.org/10.1101/gr.1413403 (2003).
79. Wangkumhang, P. *et al.* Insight into the peopling of Mainland Southeast Asia from Thai population genetic structure. *PLoS ONE* **8**, e79522. https://doi.org/10.1371/journal.pone.0079522 (2013).
80. Barth, F. *Political Leadership among Swat Pathans* (The Athlone Press, London, 1959).
81. Bhatti, S. *et al.* Genetic analysis of mitochondrial DNA control region variations in four tribes of Khyber Pakhtunkhwa, Pakistan. *Mitochondrial DNA A* **28**, 687–697. https://doi.org/10.3109/24701394.2016.1174222 (2017).
82. Ullah, I. *et al.* High Y-chromosomal differentiation among ethnic groups of Dir and Swat Districts, Pakistan. *Ann. Hum. Genet.* **81**, 234–248. https://doi.org/10.1111/ahg.12204 (2017).
83. Tabassum, S., Ilyas, M., Ullah, I., Israr, M. & Ahmad, H. A comprehensive Y-STR portrait of Yousafzai's population. *Int. J. Legal. Med.* **131**, 1241–1242. https://doi.org/10.1007/s00414-017-1550-5 (2017).
84. Qamar, R. *et al.* Y-chromosomal DNA variation in Pakistan. *Am. J. Hum. Genet.* **70**, 1107–1124. https://doi.org/10.1086/339929 (2002).
85. Rakha, A. *et al.* Forensic and genetic characterization of mtDNA from Pathans of Pakistan. *Int. J. Legal Med.* **125**, 841–848. https://doi.org/10.1007/s00414-010-0540-7 (2011).
86. Lee, E. Y. *et al.* Analysis of 22 Y chromosomal STR haplotypes and Y haplogroup distribution in Pathans of Pakistan. *Forens. Sci. Int. Genet.* **11**, 111–116. https://doi.org/10.1016/j.fsigen.2014.03.004 (2014).
87. Aziz, S., Nawaz, M., Afridi, S. G. & Khan, A. Genetic structure of Kho population from north-western Pakistan based on mtDNA control region sequences. *Genetica* **147**, 177–183. https://doi.org/10.1007/s10709-019-00060-8 (2019).
88. Underhill, P. A. *et al.* The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *Eur. J. Hum. Genet.* **23**, 124–131. https://doi.org/10.1038/ejhg.2014.50 (2015).
89. Underhill, P. A. *et al.* Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *Eur. J. Hum. Genet.* **18**, 479–484. https://doi.org/10.1038/ejhg.2009.194 (2010).
90. Mirabal, S. *et al.* Y-chromosome distribution within the geo-linguistic landscape of northwestern Russia. *Eur. J. Hum. Genet.* **17**, 1260–1273. https://doi.org/10.1038/ejhg.2009.6 (2009).
91. Singh, M., Sarkar, A. & Nandineni, M. R. A comprehensive portrait of Y-STR diversity of Indian populations and comparison with 129 worldwide populations. *Sci. Rep.* **8**, 15421. https://doi.org/10.1038/s41598-018-33714-2 (2018).
92. Trivedi, R. *et al.* Genetic Imprints of Pleistocene Origin of Indian Populations: A Comprehensive Phylogeographic Sketch of Indian Y-Chromosomes. *Int. J. Hum. Genet.* **8**, 97–118. https://doi.org/10.1080/09723757.2008.11886023 (2008).
93. Karafet, T. M. *et al.* New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res.* **18**, 830–838. https://doi.org/10.1101/gr.7172008 (2008).

94. Wells, R. S. *et al.* The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc. Natl. Acad. Sci. U.S.A* **98**, 10244–10249. https://doi.org/10.1073/pnas.171305098 (2001).
95. Haber, M. *et al.* 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 (2012).
96. Myres, N. M. *et al.* A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *Eur. J. Hum. Genet.* **19**, 95–101. https://doi.org/10.1038/ejhg.2010.146 (2011).
97. Sole-Morata, N. *et al.* Analysis of the R1b-DF27 haplogroup shows that a large fraction of Iberian Y-chromosome lineages originated recently in situ. *Sci. Rep.* **7**, 7341. https://doi.org/10.1038/s41598-017-07710-x (2017).
98. Balaresque, P. *et al.* A predominantly neolithic origin for European paternal lineages. *PLoS Biol.* **8**, e1000285. https://doi.org/10.1371/journal.pbio.1000285 (2010).
99. Kayser, M. *et al.* Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Hum. Genet.* **117**, 428–443. https://doi.org/10.1007/s00439-005-1333-9 (2005).
100. Singh, M., Sarkar, A., Kumar, D. & Nandineni, M. R. The genetic affinities of Gujjar and Ladakhi populations of India. *Sci. Rep.* **10**, 2055. https://doi.org/10.1038/s41598-020-59061-9 (2020).
101. Lacan, M. *et al.* Ancient DNA reveals male diffusion through the Neolithic Mediterranean route. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 9788–9791. https://doi.org/10.1073/pnas.1100723108 (2011).
102. Hammer, M. F. *et al.* Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 6769–6774. https://doi.org/10.1073/pnas.100115997 (2000).
103. Di Giacomo, F. *et al.* Clinal patterns of human Y chromosomal diversity in continental Italy and Greece are dominated by drift and founder effects. *Mol. Phylogenet. Evol.* **28**, 387–395. https://doi.org/10.1016/s1055-7903(03)00016-2 (2003).
104. Lacau, H. *et al.* Afghanistan from a Y-chromosome perspective. *Eur. J. Hum. Genet.* **20**, 1063–1070. https://doi.org/10.1038/ejhg.2012.59 (2012).
105. Rootsi, S. *et al.* Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in europe. *Am. J. Hum. Genet.* **75**, 128–137. https://doi.org/10.1086/422196 (2004).
106. Peričić, M. *et al.* High-resolution phylogenetic analysis of Southeastern Europe traces major episodes of paternal gene flow among Slavic populations. *Mol. Biol. Evol.* **22**, 1964–1975. https://doi.org/10.1093/molbev/msi185 (2005).
107. Al-Zahery, N. *et al.* Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Mol. Phylogenet. Evol.* **28**, 458–472. https://doi.org/10.1016/s1055-7903(03)00039-3 (2003).
108. Semino, O. *et al.* Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am. J. Hum. Genet.* **74**, 1023–1034. https://doi.org/10.1086/386295 (2004).
109. Di Giacomo, F. *et al.* Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Hum. Genet.* **115**, 357–371. https://doi.org/10.1007/s00439-004-1168-9 (2004).
110. Singh, S. *et al.* Dissecting the influence of Neolithic demic diffusion on Indian Y-chromosome pool through J2–M172 haplogroup. *Sci. Rep.* **6**, 19157. https://doi.org/10.1038/srep19157 (2016).
111. Gunther, T. *et al.* Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 11917–11922. https://doi.org/10.1073/pnas.1509851112 (2015).
112. Kivisild, T. The study of human Y chromosome variation through ancient DNA. *Hum. Genet.* **136**, 529–546. https://doi.org/10.1007/s00439-017-1773-z (2017).
113. Terreros, M. C., Rowold, D. J., Mirabal, S. & Herrera, R. J. Mitochondrial DNA and Y-chromosomal stratification in Iran: relationship between Iran and the Arabian Peninsula. *J. Hum. Genet.* **56**, 235–246. https://doi.org/10.1038/jhg.2010.174 (2011).
114. Pathak, A. K. *et al.* The genetic ancestry of modern Indus Valley populations from Northwest India. *Am. J. Hum. Genet.* **103**, 918–929. https://doi.org/10.1016/j.ajhg.2018.10.022 (2018).
115. Khurana, P. *et al.* Y chromosome haplogroup distribution in Indo-European speaking tribes of Gujarat, western India. *PLoS ONE* **9**, e90414. https://doi.org/10.1371/journal.pone.0090414 (2014).
116. Grugni, V. *et al.* Ancient migratory events in the Middle East: new clues from the Y-chromosome variation of modern Iranians. *PLoS ONE* **7**, e41252. https://doi.org/10.1371/journal.pone.0041252 (2012).
117. Zerjal, T. *et al.* Y-chromosomal insights into the genetic impact of the caste system in India. *Hum. Genet.* **121**, 137–144 (2007).
118. Thanseem, I. *et al.* Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* **7**, 42. https://doi.org/10.1186/1471-2156-7-42 (2006).
119. Yan, S., Wang, C. C., Li, H., Li, S. L. & Jin, L. An updated tree of Y-chromosome Haplogroup O and revised phylogenetic positions of mutations P164 and PK4. *Eur. J. Hum. Genet.* **19**, 1013–1015. https://doi.org/10.1038/ejhg.2011.64 (2011).
120. Shi, H. *et al.* Y-chromosome evidence of southern origin of the East Asian-specific haplogroup O3–M122. *Am. J. Hum. Genet.* **77**, 408–419. https://doi.org/10.1086/444436 (2005).
121. Xue, Y. *et al.* Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics* **172**, 2431–2439. https://doi.org/10.1534/genetics.105.054270 (2006).
122. Zhong, H. *et al.* Extended Y chromosome investigation suggests postglacial migrations of modern humans into East Asia via the northern route. *Mol. Biol. Evol.* **28**, 717–727. https://doi.org/10.1093/molbev/msq247 (2011).
123. Zahra, F. T. *et al.* Genetic polymorphism of Y-chromosomal STRs in Gujjar population of Punjab. *Int. J. Legal Med.* **134**, 1333–1334. https://doi.org/10.1007/s00414-019-02227-6 (2020).
124. Bhatti, S., Aslamkhan, M., Attimonelli, M., Abbas, S. & Aydin, H. H. Mitochondrial DNA variation in the Sindh population of Pakistan. *Aust. J. Forensic Sci.* **49**, 201–216 (2017).
125. Rakha, A. *et al.* EMPOP-quality mtDNA control region sequences from Kashmiri of Azad Jammu & Kashmir, Pakistan. *Forensic Sci. Int. Genet.* **25**, 125–131. https://doi.org/10.1016/j.fsigen.2016.08.009 (2016).
126. Siddiqi, M. H. *et al.* Genetic characterization of the Makrani people of Pakistan from mitochondrial DNA control-region data. *Leg. Med.* **17**, 134–139. https://doi.org/10.1016/j.legalmed.2014.09.007 (2015).
127. Hayat, S. *et al.* Mitochondrial DNA control region sequences study in Saraiki population from Pakistan. *Leg. Med.* **17**, 140–144. https://doi.org/10.1016/j.legalmed.2014.10.010 (2015).
128. Perveen, R. *et al.* Y-STR haplotype diversity in Punjabi population of Pakistan. *Forensic Sci. Int. Genet.* **9**, e20-21. https://doi.org/10.1016/j.fsigen.2013.12.004 (2014).
129. Khan, S. *et al.* Ethnogenetic analysis reveals that Kohistanis of Pakistan were genetically linked to west Eurasians by a probable ancestral genepool from Eurasian steppe in the bronze age. *Mitochondrion* **47**, 82–93. https://doi.org/10.1016/j.mito.2019.05.004 (2019).
130. Bhatti, S. *et al.* Genetic perspective of uniparental mitochondrial DNA landscape on the Punjabi population, Pakistan. *Mitochondrial DNA Part A* **29**, 714–726 (2018).
131. Firasat, S. *et al.* Y-chromosomal evidence for a limited Greek contribution to the Pathan population of Pakistan. *Eur. J. Hum. Genet.* **15**, 121–126. https://doi.org/10.1038/sj.ejhg.5201726 (2007).
132. Zubair, M. *et al.* Mitochondrial DNA diversity in the Khattak and Kheshgi of the Peshawar Valley, Pakistan. *Genetica* **148**, 195–206. https://doi.org/10.1007/s10709-020-00095-2 (2020).
133. Kivisild, T. *et al.* Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. *Curr. Biol.* **9**, 1331–1334. https://doi.org/10.1016/s0960-9822(00)80057-3 (1999).

134. Sharma, I. *et al.* Ancient Human Migrations to and through Jammu Kashmir-India were not of Males Exclusively. *Sci. Rep.* **8**, 851. https://doi.org/10.1038/s41598-017-18893-8 (2018).
135. Heyer, E. *et al.* Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC Genet.* **10**, 49. https://doi.org/10.1186/1471-2156-10-49 (2009).
136. Fornarino, S. *et al.* Mitochondrial and Y-chromosome diversity of the Tharus (Nepal): a reservoir of genetic variation. *BMC Evol. Biol.* **9**, 154–154. https://doi.org/10.1186/1471-2148-9-154 (2009).
137. Chandrasekar, A. *et al.* Updating phylogeny of mitochondrial DNA macrohaplogroup M in India: dispersal of modern human in South Asian Corridor. *PLoS ONE* **4**, e7447. https://doi.org/10.1371/journal.pone.0007447 (2009).
138. Di Cristofaro, J. *et al.* Afghan Hindu Kush: where Eurasian sub-continent gene flows converge. *PLoS ONE* **8**, e76748. https://doi.org/10.1371/journal.pone.0076748 (2013).
139. Berrenberg, J. Beyond kinship algebra. Values and the riddle of Pashtun marriage structure. *Z. Ethnol.* **2003**, 269–292 (2003).
140. Mehdi, S. Q. *et al.* The Origins of Pakistani Populations. In *Genomic Diversity: Applications in Human Population Genetics* (eds Singh, S. S. *et al.*) (Springer, New York, NY, 1999).
141. Siddique, A. *The Pashtun Question: The Unresolved Key to the Future of Pakistan and Afghanistan* (Hurst & Company Ltd., 2014).
142. Mines, M. Muslim social stratification in India: The basis for variation. *Southwest. J. Anthropol.* **28**, 333–349 (1972).
143. Momin, A. The Indo-Islamic tradition. *Sociol. Bull.* **26**, 242–258 (1977).
144. Ferrier, J. P. *History of the Afghans* (J. Murray, 1858).
145. Ahmad, Z. Muslim caste in Uttar Pradesh. *Econ. Wkly* **14**, 325–336 (1962).
146. Wright, J. T. P. Can there be a melting pot in Pakistan? Interprovincial marriage and national integration. *Contemp. South Asia* **3**, 131–144 (1994).
147. Sökefeld, M. Stereotypes and boundaries: Pathan in Gilgit, northern Pakistan. In *Countries and Peoples of the Hindu Kush* (eds Kushwe, W. W. *et al.*) 280–299 (Wiley, St. Petersburg, 1998).
148. Kılıc, R. Sayyids and Shar? fs in the Ottoman State: On the Borders of the True and the False. *Muslim World* **96**, 21–35 (2006).
149. Lone, M. A. Institutional and Structural Changes in Pakhtoon Family and Marriage Systems in Kashmir Valley of Jammu and Kashmir, India. *Res. J. Fam. Commun. Consum. Sci.* **1**, 1–6 (2013).
150. Singh, D. E. *Islamization in Modern South Asia: Deobandi Reform and the Gujjar Response* (Walter de Gruyter, 2012).
151. Khatra, P. S. & Sharma, V. Socio-economic issues in the development of nomadic Gujjars. *Indian J. Agric. Econ.* **47**, 448–449 (1992).
152. Ibbetson, D. *Landmarks in Indian Anthropology: Punjab Castes, Races, Castes, and Tribes of the People of Punjab* (Cosmo Publications, 1987).
153. Bingley, A. *Caste, Tribes & Culture of Rajputs* (Ess Ess Publications, London, 1978).
154. Saini, J., Kumar, A., Matharoo, K., Sokhi, J. & Bhanwer, A. Genomic diversity and affinities in population groups of North West India: an analysis of Alu insertion and a single nucleotide polymorphism. *Gene* **511**, 293–299 (2012).
155. Yardumian, A. *et al.* Genetic diversity in Svaneti and its implications for the human settlement of the Highland Caucasus. *Am. J. Phys. Anthropol.* **164**, 837–852. https://doi.org/10.1002/ajpa.23324 (2017).
156. Singh, M., Sarkar, A., Kumar, D. & Nandineni, M. R. The genetic affinities of Gujjar and Ladakhi populations of India. *Sci. Rep.* **10**, 1–12 (2020).
157. Balgir, R. & Sharma, J. Genetic markers in the Hindu and Muslim Gujjars of northwestern India. *Am. J. Phys. Anthropol.* **75**, 391–403 (1988).
158. Dulik, M. C., Osipova, L. P. & Schurr, T. G. Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS ONE* **6**, e17548. https://doi.org/10.1371/journal.pone.0017548 (2011).
159. Zerjal, T., Wells, R. S., Yuldasheva, N., Ruzibakiev, R. & Tyler-Smith, C. A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am. J. Hum. Genet.* **71**, 466–482. https://doi.org/10.1086/342096 (2002).
160. Marushiakova, E. & Popov, V. *Gypsies (Roma) in Bulgaria* Vol. 18 (Peter Lang Pub Incorporated, 1997).
161. Fraser, A. M. *The Gypsies* 374 (Wiley-Blackwell, 1995).
162. Nagy, M. *et al.* Searching for the origin of Romanies: Slovakian Romani, Jats of Haryana and Jat Sikhs Y-STR data in comparison with different Romani populations. *Forensic Sci. Int.* **169**, 19–26. https://doi.org/10.1016/j.forsciint.2006.07.020 (2007).
163. Laruelle, M., Peyrouse, S. & Axyonova, V. *The Afghanistan-Central Asia Relationship: What Role for the EU?* (Universitäts-und Landesbibliothek Sachsen-Anhalt, 2013).
164. ArunKumar, G. *et al.* Genome-wide signatures of male-mediated migration shaping the Indian gene pool. *J. Hum. Genet.* **60**, 493 (2015).
165. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211. https://doi.org/10.1038/nature14317 (2015).
166. Poznik, G. D. *et al.* Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat. Genet.* **48**, 593–599. https://doi.org/10.1038/ng.3559 (2016).
167. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172. https://doi.org/10.1038/nature14507 (2015).
168. Batini, C. *et al.* Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat. Commun.* **6**, 7152. https://doi.org/10.1038/ncomms8152 (2015).
169. Shah, A. M. *et al.* Indian Siddis: African descendants with Indian admixture. *Am. J. Hum. Genet.* **89**, 154–161. https://doi.org/10.1016/j.ajhg.2011.05.030 (2011).
170. Chikhi, L., Nichols, R. A., Barbujani, G. & Beaumont, M. A. Y genetic data support the Neolithic demic diffusion model. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11008–11013 (2002).
171. Haak, W. *et al.* Ancient DNA from European early neolithic farmers reveals their near eastern affinities. *PLoS Biol.* **8**, e1000536. https://doi.org/10.1371/journal.pbio.1000536 (2010).
172. Gutala, R. *et al.* A shared Y-chromosomal heritage between Muslims and Hindus in India. *Hum. Genet.* **120**, 543–551 (2006).
173. Lal, K. S. *Indian Muslims: Who are They?* (Voice of India New Delhi, 1990).
174. Papiha, S. Genetic variation in India. *Hum. Biol.* **1996**, 607–628 (1996).
175. Aarzoo, S. S. & Afzal, M. Gene diversity in some Muslim populations of North India. *Hum. Biol.* **77**, 343–353 (2005).
176. Aidar, M. & Line, S. R. A simple and cost-effective protocol for DNA isolation from buccal epithelial cells. *Braz. Dent. Dent. J.* **18**, 148–152 (2007).
177. Gaieski, J. B. *et al.* Genetic ancestry and indigenous heritage in a Native American descendant community in Bermuda. *Am. J. Phys. Anthropol.* **146**, 392–405. https://doi.org/10.1002/ajpa.21588 (2011).
178. Schurr, T. G. *et al.* Clan, language, and migration history has shaped genetic diversity in Haida and Tlingit populations from Southeast Alaska. *Am. J. Phys. Anthropol.* **148**, 422–435. https://doi.org/10.1002/ajpa.22068 (2012).
179. Zhadanov, S. I. *et al.* Genetic heritage and native identity of the Seaconke Wampanoag tribe of Massachusetts. *Am. J. Phys. Anthropol.* **142**, 579–589. https://doi.org/10.1002/ajpa.21281 (2010).
180. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465. https://doi.org/10.1038/290457a0 (1981).

181. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147. https://doi.org/10.1038/13779 (1999).
182. Kloss-Brandstatter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**, 25–32. https://doi.org/10.1002/humu.21382 (2011).
183. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386-394. https://doi.org/10.1002/humu.20921 (2009).
184. Consortium & Y. C,. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12**, 339–348. https://doi.org/10.1101/gr.217602 (2002).
185. Cruciani, F. *et al.* A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am. J. Hum. Genet.* **88**, 814–818. https://doi.org/10.1016/j.ajhg.2011.05.002 (2011).
186. Francalacci, P. *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341**, 565–569. https://doi.org/10.1126/science.1237947 (2013).
187. Gusmao, L. *et al.* DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. *Forensic Sci. Int.* **157**, 187–197. https://doi.org/10.1016/j.forsciint.2005.04.002 (2006).
188. van Oven, M., Van Geystelen, A., Kayser, M., Decorte, R. & Larmuseau, M. H. Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum. Mutat.* **35**, 187–191. https://doi.org/10.1002/humu.22468 (2014).
189. Excoffier, L. & Lischer, H. E. Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and Windows. *Mol. Ecol. Resour.* **10**, 564–567. https://doi.org/10.1111/j.1755-0998.2010.02847.x (2010).
190. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
191. Fu, Y. X. Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**, 915–925 (1997).
192. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549. https://doi.org/10.1093/molbev/msy096 (2018).
193. Bandelt, H. J., Forster, P. & Rohl, A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* **16**, 37–48. https://doi.org/10.1093/oxfordjournals.molbev.a026036 (1999).
194. Bandelt, H. J., Forster, P., Sykes, B. C. & Richards, M. B. Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743–753 (1995).
195. Bandelt, H. J., Quintana-Murci, L., Salas, A. & Macaulay, V. The fingerprint of phantom mutations in mitochondrial DNA data. *Am. J. Hum. Genet.* **71**, 1150–1160. https://doi.org/10.1086/344397 (2002).
196. Goedbloed, M. *et al.* Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR® Yfiler® PCR amplification kit. *Int. J. Legal Med.* **123**, 471. https://doi.org/10.1007/s00414-009-0342-y (2009).
197. QGIS. https://docs.qgis.org/3.16/en/docs/gentle_gis_introduction/spatial_analysis_interpolation.html.
198. Wu, Y.-H. & Hung, M.-C. in *Applications of Spatial Statistics* (2016).

Acknowledgements

We are grateful to all the volunteers who participated in this study and donated their DNA samples for analysis. We also thank Robert Bryant from the Department of Anthropology at the University of Pennsylvania for conducting the geospatial mapping work. M.T. was supported by a HEC IRSIP Fellowship that allowed him to conduct genetic analyses at the University of Pennsylvania. This study was also supported by Faculty Research Funds awarded to T.G.S. by the University of Pennsylvania.

Author contributions

M.T., H.A.: Conceived and designed the study; M.T., U.F.: collected saliva samples for DNA analysis; T.G.S.: contributed reagents, materials, and analytical tools; M.T.: performed the experiments, and collected and analyzed the data; M.T., T.G.S., B.E.H.: wrote the manuscript incorporating contributions from other authors. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-05076-3.

Correspondence and requests for materials should be addressed to M.T. or T.G.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.