



Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/156300/>

Version: Accepted Version

Article:

Ullah, I., Olofsson, J.K., Margaryan, A. et al. (2017) High Y-chromosomal differentiation among ethnic groups of Dir and Swat districts, Pakistan. *Annals of Human Genetics*, 81 (6). pp. 234-248. ISSN: 0003-4800

<https://doi.org/10.1111/ahg.12204>

This is the peer reviewed version of the following article: Ullah, I., Olofsson, J.K., Margaryan, A., Ilardo, M., Ahmad, H., Sikora, M., Hansen, A.J., Shahid Nadeem, M., Fazal, N., Ali, M., Buchard, A., Hemphill, B.E., Willerslev, E. and Allentoft, M.E. (2017), High Y-chromosomal Differentiation Among Ethnic Groups of Dir and Swat Districts, Pakistan. *Annals of Human Genetics*, 81: 234-248., which has been published in final form at <https://doi.org/10.1111/ahg.12204>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions.

Reuse

Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

High Y-chromosomal differentiation among ethnic groups in Dir and Swat districts, Pakistan

5 InamUllah^{1,2*}†, Jill K. Olofsson^{3*}, Ashot Margaryan², Melissa Ilardo², Habib Ahmad^{1,4},
Martin Sikora², Anders J. Hansen², Muhammad Shahid Nadeem⁵, Numan Fazal¹, Murad
Ali¹, Anders Buchard⁶, Brian E. Hemphill⁷, Eske Willerslev², Morten E. Allentoft²†

* Authors contributed equally to this work

† Authors for correspondence

- 10 1) Department of Genetics, Hazara University, Garden Campus, Mansehra 21300
Pakistan.
2) Centre for GeoGenetics, Natural History Museum, University of Copenhagen,
ØsterVoldgade 5-7, 1350 Copenhagen, Denmark.
3) Department of Animal and Plant Sciences, University of Sheffield, Western Bank,
15 Sheffield S10 2TN, United Kingdom.
4) Islamia University, Peshawar, Khyber Pakhtunkhwa, 25120 Pakistan
5) Department of Biochemistry, Faculty of Science, King Abdulaziz University, Jeddah
21589, Saudi Arabia.
6) Department of Forensic Medicine, University of Copenhagen, Frederik V's Vej 11,
20 2100 Copenhagen, Denmark
7) Department of Anthropology, University of Alaska, Fairbanks, Fairbanks, AK 99775,
United States.

Keywords: Y-STR, genetic drift, ethnicity, Hindu Raj, genetic differentiation

25 **Running title:** Y-chromosome diversity in Dir and Swat

Author for correspondence:

Morten E. Allentoft

Centre for GeoGenetics, Natural History Museum, University of Copenhagen,

30 ØsterVoldgade 5-7, 1350 Copenhagen, Denmark.

Phone: +4529824634. Email: meallentoft@snm.ku.dk

Summary

The ethnic groups that inhabit the mountainous Dir and Swat districts of northern Pakistan are marked by high levels of cultural and phenotypic diversity. To obtain
35 knowledge of the genetic diversity in this region, we investigated the Y-chromosomal diversity in five population samples representing the three main ethnic groups residing within these districts, including Gujar, Pashtun and Kohistani. A total of 27 Y-chromosomal short tandem repeats (Y-STRs) and 331 Y-chromosomal single nucleotide polymorphisms (Y-SNPs) were investigated. In the Y-STRs we observed very high and
40 significant levels of genetic differentiation in nine of the 10 pairwise between-group comparisons (R_{ST} 0.179 - 0.746) and the differences were mirrored in the Y-haplogroup frequency distribution. No genetic differences were found between the two Pashtun sub-ethnic groups Tarklani and Yusafzai ($R_{ST} = 0.000$). Utmankhels, also considered Pashtuns culturally, were not closely related to any of the other population samples
45 (R_{ST} 0.451 - 0.746). Thus, our findings provide examples of both associations and dissociations between cultural and genetic legacies. When analyzed within a larger continental-scale context, these five ethnic groups tribes fall mostly outside the previously characterized Y-chromosomal gene pools of the Indo-Pakistani sub-continent. Male founder effects, coupled with culturally and topographically based
50 constraints upon marriage and movement, is likely responsible for the high degree of genetic structure in this region.

Introduction

Pakistan is home to over 180 million people and at least 18 ethnic groups who speak more than 60 different local languages assigned to a wide array of linguistic stocks, including, but not limited to Indo-Iranian, Indo-Aryan, Tibeto-Burman, and Dravidian (Grimes & Grimes, 2000, Newcomb, 1986). Geographically, Pakistan is situated at the crossroad linking Western and Central Asia to South Asia. Historically, Pakistan was part of the British Indian Empire which, following the independence in 1947, was subdivided into the independent countries and kingdoms that today makes up the Indo-Pakistani sub-continent.

Despite being a country inhabited by a population of considerable ethnic diversity, the genetic legacy of many of the ethnic groups in Pakistan has remained largely unstudied. For example, the gene pools of the ethnic groups residing in the mountainous terrain of northern Pakistan and northeastern Afghanistan (Fig. 1) remain poorly understood. The ethnic and cultural diversity in this geographic region has been attributed to a dynamic history of repeated invasions by Aryans (Bernhard, 1983, Parpola, 1995, Parpola, 2009), Indo-Iranians (Jettmar, 1967, Jettmar, 1996), Macedonians (Birdwood, 1959), Arabs, and Mongols (Lapidus, 2002). It is also believed that the southern coast of the Persian Gulf, the Makran Coast of Pakistan, and the territory of present-day Afghanistan likely served as passages for human dispersal in prehistoric times (Derenko *et al.*, 2013), thereby providing a deep temporal dimension to the population dynamics within the region. Furthermore, the Hindu Kush, Hindu Raj, Karakoram and Himalayan highlands are believed to have served as physical barriers that channeled causeways of trade and communication along the Silk Route that linked the Mediterranean Basin and West Asia to Central Asia, South Asia and China for more than 16 centuries (Quintana-Murci *et al.*, 1999, Petraglia *et al.*, 2012, Vadime, 2001, Kuz'mina & Mair, 2008,

Hemphill & Mallory, 2004). It is therefore possible that the extant populations of the Hindu Kush and Hindu Raj highlands conserve traces of historic, and possibly even prehistoric, gene flow from geographically distant human populations (Hemphill, 2009, 80 Hemphill, 2013a, Hemphill *et al.*, 2013, Hemphill, 2013b).

Dir and Swat districts are located within the Khyber Pakhtunkhwa Province of northern Pakistan (Fig. 1). Both districts are divided into southern (or “lower”) and northern (or “upper”) regions, with the former including the foothills between the northern reaches of the Indus Valley to the south and the latter including the Hindu Raj 85 range of the greater Hindu Kush (Fig. 1). Altogether, Dir and Swat districts encompass a total of 5,284 and 6,226 km², respectively (Ali & Qaiser, 1986, Ahmad & Sirajuddin, 1996, Hazrat *et al.*, 2007). The major ethnic groups found in Dir and Swat districts are: (i) Pashtuns (also known as Pathans), a Pashto speaking (Eastern Iranian language) agriculturist ethnic group consisting of four widely recognized patrilineally-based social 90 groups (Bettani, Ghurghakhti, Karlani and Sarbani) which can be further subdivided into a number of sub-tribes known as *khels* or *zais* (Table S1) (Nüsser & Dickoré, 2002, Coningham & Young, 2015, Böhner & Lucarini, 2015, Caroe, 1992, Khan, 2008); (ii) Gujars, who speak Gojri (a lowland Indo-Aryan language) an agro- pastoral group with widespread clans residing in all parts of both districts who speak Gujari (a lowland 95 Indo-Aryan language), and (iii) Kohistanis, speakers of an array of Dardic languages, who practice a wide range of agricultural and transhumant herding subsistence strategies (Barth, 1956, Bangash, 2012). The Kohistanis are commonly thought to be descendants of ancient nomadic herders who were forced into the mountainous highlands from the low-lying fertile plains by Pashtun-speaking 100 agriculturalists from the west during the 16th century (Barth, 1956, Rome, 2008, Shah, 2013). According to Barth (1956), there is little reported intermarriage between

Pashtuns, Gujars and Kohistanis because they tend to live in isolation and discourage intermarriages with members of other ethnic groups. As a result, previous researchers have described the local populations of these ethnic groups as genetically isolated and
105 marked by high levels of inbreeding (Caroe, 1992, Glatzer, 2002, Mehdi *et al.*, 1999, Siddique, 2014). However, these studies have not studied the genetic relationships among the populations residing within and immediately adjacent to the Hindu Raj highlands in any detail.

Analyses of genetic variants of the human Y-chromosome are useful for inferring
110 patterns of current and past gene flow between human populations (reviewed in Oppenheimer, 2012). Due to the exclusively paternal non recombined inheritance pattern of the human Y-chromosome, the paternal line is easily traced using Y-chromosomal genetic variants, such as short tandem repeats (Y-STRs) and single nucleotide polymorphisms (Y-SNPs) (Oppenheimer, 2012). Y-STR analyses can be
115 used to resolve the genetic relationship and paternal gene flow between current human populations, whereas the slower mutating Y-SNPs can provide information on more ancient prehistoric or historic demographic events (Karafet *et al.*, 2008, Roewer, 2009, Larmuseau *et al.*, 2015).

In this study we present information on 27 Y-STR and 331 Y-SNP loci for five
120 ethnically distinct groups from Dir and Swat districts in Pakistan. We apply a series of genetic analyses in order to investigate the genetic relationships among these groups. The ethnic groups included in this study are characterized by having different lifestyles; low elevation valley agriculturists (Pashtuns), mountainous nomadic herders (Gujars), and transhumant herders (Kohistanis). Gujar and Kohistani individuals, as well as
125 members of the three patrilineally-based Pashtun subethnic groups (Tarklanis, Utmankhels and Yusafzais), were sampled (see Table S1 for ethnic divisions among

Pashtuns/Pathans). We characterize Y-STR genetic diversity within and among these ethnically distinct groups, thereby uncovering a relatively unexplored part of the modern human gene pool. Although some of the major population groups from this region are included in the Human Genome Diversity Panel (HGDP) and have therefore been included in worldwide genetic studies (e.g., Shi *et al.*, 2010, Cann *et al.*, 2002), few studies have looked at the micro-geographic patterns of genetic diversity within population groups of northwestern Pakistan. We investigate whether current information about common history, culture, and language is reflected in the genetic relationships among the populations residing within or adjacent to the Hindu Raj highlands. As such, our data offers an excellent opportunity to test the nature and extent of the relationship between genetic and cultural affinity. We hypothesize that ethnicity has exerted a greater effect on the genetic associations present among the human populations residing within Dir and Swat districts than simple geographic propinquity.

140

Materials and Methods

Sampling and DNA extraction

A total of 100 saliva samples from males of five ethnically distinct population samples (Tarklanis, Yusafzais, Kohistanis, Gujars, and Utmankhels) were collected from individuals residing in Swat and Dir districts of northern Pakistan (Fig. 1). Members of three of these population samples (Tarklanis, Yusafzais, and Utmankhels) are commonly recognized as patrilineally-based sub-groups within Pashtuns (Pathans) ethnic group. Ethnicity was self-declared and all participants gave their informed written consent after the aims and procedures of the study were explained to them. Great care was taken to avoid sampling related individuals. First and foremost, 4-5 visits to the communities were initially undertaken to carefully select individuals and

150

record their names and ethnic relationships. At the day of sampling all the volunteers were informed to meet at their *hujra* (meeting place) under the guidance of a *malak* (a local counsellor among the elders). Before sampling, the elders and the volunteers were again interviewed to exclude closely related individuals, especially first degree paternal relatives.

Genomic DNA was isolated using a modified phenol:chloroform method as previously described (Ralser *et al.*, 2006) and DNA concentrations were determined on a Qubit flourometer (Invitrogen, life technology, cat. Number Q32857) using the Qubit dsDNA HR Assay Kit (Invitrogen, cat. Number Q32854).

Y-STR and Y-SNP datasets

A total of 27 Y-STR loci were amplified with the Yfiler[®] Plus PCR Amplification kit (ThermoFisher Scientific, Cat. No. 4484678) and PCR products were separated and evaluated according to manufacturer's protocols with the modifications described by Olofsson *et al.* (2015a). All samples were genotyped in duplicates in the ISO17025-certified forensic genetics laboratory at the Department of Forensic Medicine, Section of Forensic Genetics, University of Copenhagen, Denmark, and concordant results were obtained between the first and second typing of all the samples. All haplotypes were reported to the Y-chromosomal haplotype reference database (YHRD) (Willuweit & Roewer, 2015) under the accession numbers YA004265 to YA004269 and are presented in Table S2.

Initial assignment of Y-chromosomal haplogroups was carried out using genotypes of Y-SNPs included on the Infinium[®]OmniExpressExome-8 v.1.3 BeadChip array. A total of 1,641 Y-SNPs are included on the array, of which 1,226 passed genotyping filters

(call rate $\geq 90\%$) among the individuals included in the study. The Y-SNPs that passed the genotyping filters were intersected with the ISOGG Y-DNA SNP index (http://isogg.org/tree/index.html, version 10.103), resulting in a final set of 331 haplogroup-defining Y-SNPs. Individual haplogroups were assigned as the most derived haplogroup where the individual's genotype matched the derived allele. The shorthand version of the ISOGG nomenclature was used, where the main haplogroup, or sub-haplogroup, is followed by the most derived Y-SNP to which the Y-chromosome could be typed (Table S3). Markers in parenthesis followed by an "x" indicate downstream markers for which the samples were typed but were found to be in the ancestral state.

Analyses

Population genetic parameters were estimated for the five ethnically distinct Pakistani population samples and for the meta-population of Dir and Swat districts, combining all of the individuals included in the study, using a framework previously described (Olofsson *et al.*, 2015a). Genetic distances between population samples were evaluated as pairwise R_{ST} distances calculated in Arlequin v. 3.5.1.2 (10,000 permutations; Excoffier & Lischer, 2010) and visualized through nonmetric multidimensional scaling (MDS) in the statistical software R v. 3.2.1 using the isoMDS function of the MASS package. Median joining networks of haplotypes were constructed in the program Network v. 5.0.0.0 (http://www.fluxus-engineering.com) and weights (1-5) were given to the included loci based on the inverted diversities (1: DYS449, DYS458, DYS481, DYS518, DYS576, DYS627; 2: DYS19, DYS389B, DYS390, DYS392, DYS393, DYS437, DYS448, DYS533, DYS570, DYS635; 3: DYS438, DYS439, DYS456; 4: DYS389I; 5: DYS391, DYS460, YGATAH4) (Olofsson *et al.*, 2015a). The multi-copy loci in this kit, DYS385 and DYS387S1 were excluded for estimations of genetic

distances (R_{ST}) and construction of median joining networks as is common practice,
205 resulting in 23 Y-STRs for these analyses. Furthermore, individuals with haplotypes
displaying duplication events, null or intermediate alleles were excluded in the network
analyses but for genetic distances null and intermediate alleles were counted as missing
data. As is standard for Y-STR analyses, the alleles of the DYS389II locus were
converted to the DYS389B nomenclature by subtracting the repeat number of DYS389I
210 from that of DYS389II.

To place the diversity observed for the patrilineal gene pool of Swat and Dir districts in
a greater geographic and ethnic context we constructed two datasets based upon
previously published Y-STR data. One dataset encompassed 27 population samples
(including the five from this study) from the Indian sub-continent and Southwest Asia
215 (Roewer *et al.*, 2009, Haber *et al.*, 2012, Perveen *et al.*, 2014, Lee *et al.*, 2014, Qamar *et al.*,
et al., 2002, Tabassum *et al.*, 2017) (Table S4). The other dataset encompassed 53
worldwide population samples (including the five from this study and the Yusafzais
population from Tabassum *et al.*, 2017) from the HGDP panel (Cann *et al.*, 2002,
Rosenberg, 2006, Vermeulen *et al.*, 2009), with the criterion that the populatio was
220 represented by at least five males genotyped for both Y-STRs and Y-SNPs (Table S4).
To be able to merge the different data sets, the data were limited to 15 (DYS19,
DYS389I, DYS389B, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438,
DYS439, DYS448, DYS456, DYS458, DYS635, and YGATAH4) of the 23 Y-
STRs loci. Only studies and loci typed with the commercial kits
225 AmpFLSTR® Yfiler® PCR Amplification kit or Yfiler® Plus PCR Amplification kit
(both ThermoFisher Scientific) were included in the comparisons (Table S4). This
conservative approach ensures that there is no difference in nomenclature of the alleles
between the different studies. R_{ST} values between all groups were calculated in Arlequin

v. 3.5.1.2. The same package was used for the analyses of molecular variance
230 (AMOVA) between all groups and between groupings based on country of origin and
reported ethnicity (Table S5, S6). The text associated with Table S5 outlines the
rationale behind these ethnicity aggregates. MDS plots were constructed based on the
pairwise R_{ST} genetic distance matrix with R software package v. 3.2.1 as described
above.

235

Results

Genetic diversity

The 100 individuals in this study self-identify as members of one of three major ethnic
groups: Pashtuns (Pathans), Kohistanis, or Gujars. Pashtuns are further represented by
240 individuals from three widely recognized patrilineally-based divisions, Tarklanis,
Utmankhels, and Yusafzais (Table 1). Analyses of the 27 Y-STRloci resulted in the
identification of a total of 82 haplotypes of which 75 were unique (Table 1). The
percentage of unique haplotypes within each of the five population samples varied from
100% (20 out of 20) among Kohistanis to 45% (9 out of 20) among Utmankhels (Table
245 1). Seven haplotypes were shared between two to six individuals within the meta-
population of Dir and Swat district and all but two haplotypes were population-specific
(Table 1). The non-population-specific haplotypes were shared between four and five
individuals within the meta-population of Dir and Swat districts, respectively. These
include a haplotype shared by three Yusafzai individuals and one Tarklani individual
250 and a haplotype shared by four Gujars and one Kohistani individual. As a result of the
differences in unique haplotype frequencies, haplotype diversity also varied between
population samples, ranging from 1.00 among Kohistanis to 0.93 among Utmankhels
(Table 1). The overall power of discrimination was relatively high (0.82) for the

combined dataset but varied widely low (0.60) in Utmankhels to high (1.00) in
255 Kohistanis, when the five ethnic groups were considered separate populations (Table 1).

Information on Y-SNPs was used to assign a Y-chromosomal haplogroup (Karafet *et al.*, 2008, Larmuseau *et al.*, 2015) to each individual. A relatively large number of haplogroups was observed (Table 1, Table S2, S3) and the spectrum of these
260 haplogroups was consistent with previous studies (Qamar *et al.*, 2002, Karafet *et al.*, 2008, Sengupta *et al.*, 2006, Lee *et al.*, 2014, Kivisild *et al.*, 2003, Zhao *et al.*, 2009, Chennakrishnaiah *et al.*, 2013). However, 85% of the studied individuals carry one of four haplogroups (H1-M69, G2b-M283, L1-M22(xM274), and R1a-M417, Page 7) and there are large differences in the frequencies of these four haplogroups between the five
265 population samples (Table 1, Table S2, S3). For example, haplogroup G2b-M283 occurs with very high frequency (0.80; 0.77-0.83) among Utmankhels, but is completely absent among members of three out of the other four population samples (Table 1, Table S2, S3). In contrast, haplogroup R1a-M417, Page 7 occurs among members of all five population samples but frequencies range from high (0.80; 0.77-0.83) among
270 Yusafzais and Tarklanis to low (0.10; 0.07-0.13) among Utmankhels (Table 1). Due to small sample sizes the 90% confidence intervals are relatively large and overlap for some haplogroups (Table 1).

Genetic differentiation

275 The genetic distances between the five groups, as estimated from the Y-STR markers using pairwise R_{ST} , are mostly very large and highly significant, ranging from 0.179 to 0.746, except for the pairwise comparison between Tarklanis and Yusafzais (Table 2, Fig. S1). Despite being considered different ethnic subgroups of Pashtuns, members of

these two groups are not significantly different from each other genetically ($R_{ST} = 0$, $p =$
280 0.604).

The genetic structure is also evident in the median joining network of Y-STR
haplotypes (Fig. 2, Table 1). Members of the Tarklani and Yusafzai subgroups of
Pashtuns are mostly found together, being separated by only a few mutational steps
285 (Fig. 2). This is in contrast to the Utmankhels and Gujars, who, with the exception of
some outliers, form distinct groups separated by a large number of mutational steps
from the other groups. There are no shared haplotypes within the Kohistani group;
hence they appear more scattered in the network. Nevertheless, the majority of
haplotypes are still found close together in relative proximity to the Tarklani/Yusafzai
290 aggregate(Fig. 2).

Genetics, ethnicity and geography

To examine the genetic variation in a broader context we included population samples
from a wider geographic range. We used 15 Y-STR loci for a worldwide data set and for
295 a data set representing the Indian sub-continent and Southwest Asia (Table S4). The
results are summarized in: (i) matrices of pairwise R_{ST} values for both datasets (Table
S7), (ii) a pair of AMOVA analyses for the 27 Indo-Pakistani sub-continent and
Southwest Asia population samples (Tables S5, S6), and (iii) MDS plots for both
datasets (Fig. 3 and Fig. S2). In the AMOVA analysis, *c.* 92% of the genetic variation
300 occurs within the 27 population samples from the Indo-Pakistani sub-continent and
Southwest Asia. When grouping these population samples by country of origin (Table
S5A), the genetic variation among the countries accounts for only 2.2%, whereas 5.6%
of the total variation is explained by difference between population samples within
countries. However, when the 27 samples are instead grouped by ethnic relationships,

305 differences between the ethnic groups account for 4.1% of the total variation, while the
variation between population samples within the ethnic groups accounts for 3.4% of the
total variation (Table S5B). For the grouping based on ethnicity, the Utmankhels were
treated as a separate group due to its profound genetic differences to all other population
groups included in the AMOVA analyses. When the Utmankhels were instead
310 considered part of the Pashtun/Pathan ethnic group, the variation among populations
within ethnic groups increased to close to 5% (Table S6) indicating that this particular
population accounts for a large amount of the between population variation.

Despite the inclusion of 27 population samples from the Indo-Pakistani sub-continent
315 and Southwest Asia, most of the genetic variation in the MDS is still defined by the five
population samples from Dir and Swat districts (Fig. 3). In this data set, limited to 15
STR loci, there are still large genetic differences between the samples from Dir and
Swat districts (Table S7, Fig. S1, S2, and S3), but the reduction resolution implies that
the differentiations between Gujars and Kohistanis and between Yusafzais and
320 Kohistanis become non-significant (Table S7).

Several specific observations can be made. The Gujar sample and the Baluch (Balochi)
ethnic groups from Afghanistan (Haber *et al.*, 2012) are both outliers and occupy the
same area in the MDS plot (Fig. 3), whereas the Baluch (Balochi) sample from Pakistan
325 (Cann *et al.*, 2002, Rosenberg, 2006, Vermeulen *et al.*, 2009) occupies a more central
position. However, the genetic distances between these samples are non-significant after
correction for multiple testing. The Kohistanis occupy a more central position within the
MDS plot adjacent to a large number of other sampled ethnic groups from the Indo-
Pakistani subcontinent and Southwest Asia. Noticeably, the Utmankhel sample is
330 separated by very large and highly significant genetic distances from all other groups

(Table S7), and within the MDS plots (Fig. 3 and S2) this sample occupies an isolated position. The Tarklanis and Yusufzais are marked by very similar genetic distances to the other groups included in this analysis (Table S7, Fig. 3). These results are generally mirrored when the MDS is constructed from the worldwide data set (Fig. S2, Table 335 S7). Surprisingly, the three sub-ethnic groups of the Pashtuns sampled from Dir and Swat (Tarklanis, Utmankhels, Yusufzais) still represent outliers, observed far outside most of the known Y-STR genetic diversity in Indo-Pakistani sub-continent and Southwest Asia (Fig. 3, S2, and Table S7).

340 *Detailed analysis of two Y-chromosomal haplogroups*

To get a more detailed picture of the relationship between the five population samples from Dir and Swat districts we constructed haplotype (15 Y-STR loci) networks for individuals assigned to Y-SNP haplogroups (i) G-Page94 [(G2a-L30(xL14, L13, M278) and G2b-M283)], (ii) H1-M69, and (iii) L1-M22(xM274), and included previously 345 published datasets from Pakistan and Afghanistan (Haber *et al.*, 2012, Vermeulen *et al.*, 2009)(Fig. 4). Most of the Utmankhels possess haplogroup G-Page94 (G2b-M283, more specifically) and they all cluster closely together (owing to highly similar Y-STR profiles) and with a couple of individuals from both Afghanistan and Pakistan (Fig. 4A). Only one Kohistani and one Gujar individual have a Y-SNP profile assigned to the G- 350 Page94 haplogroup, and these two individuals share the same Y-STR haplotype, which is clearly separated from the haplotypes observed among the sampled Utmankhel individuals (Fig. 4A).

The Y-STR network with individuals assigned to SNP-haplogroup H1-M69 is more diffuse and many individuals are separated by a larger number of mutational steps. 355 However, most Kohistanis are found within this network, and many cluster together, sharing the same Y-STR haplotype (Fig. 4B). The network of STR-haplotypes of

individuals assigned to SNP-haplogroup L1-M22(xM274) shows at least two defined groups (Fig. 4C). All but one Gujar individual in this network share the same Y-STR haplotype, which is also shared by a single Kohistani individual (even when extended to the full 27 Y-STR loci haplotype; Table 1 and Fig. 2). Only a single Gujar individual is found in the other sub-group within the network.

Discussion

Genetic diversity and differentiation in Dir and Swat

Our analyses of patrilineal genetic diversity among males of the five ethnic groups from Dir and Swat districts of Pakistan have yielded several insights. First, the level of Y-STR haplotype diversity within each ethnic group is generally high and comparable to average global values (Purps *et al.*, 2014), except for the Utmankhel sample, which displays less diversity and fewer unique haplotypes (Table 1). Second, the five groups display an extreme level of genetic differentiation, both among themselves (Table 2, Fig. S1) and in relation to other groups from this geographic region (Fig. 3, Table S7). Based on the 23 single-copy Y-STR loci, the average R_{ST} between these five ethnic groups is very high (0.38, Table 2), with an extreme R_{ST} of 0.75 observed between Tarklanis and Utmankhels (Table 2). The middle range R_{ST} values (e.g., 0.1-0.2) found between some of the ethnic groups (Gujar – Kohistani, Tarklani – Kohistani, Yusafzai – Kohistani) are comparable to genetic distances reported previously between population groups from the Indo-Pakistani sub-continent (Alam *et al.*, 2010, Seema Nair *et al.*, 2011, Perveen *et al.*, 2014) and the Middle East (Triki-Fendri *et al.*, 2015). It is intriguing that Kohistanis represent the common denominator in these middle range values, for they likely represent the indigenous population of the region with the other likely representing more recent immigrants (Barth, 1956); Tarklanis and Yusafzais

occupying the low-lying regions of southern Dir and Swat and Gujars the rugged
higher-altitude Upper Swat. The extreme genetic distances we observe ($R_{ST} > 0.4$) in
385 several of the pairwise comparisons (Gujars –Tarklanis, and the Utmankhels compared
to any of the other population samples) are unusual and higher than observed between
most human populations - even when occupying different continents (e.g., Purps *et al.*,
2014). The very large genetic distances result from a number of non-overlapping, fixed,
or almost fixed alleles in the five population samples (Table S2). For example in the
390 three Pashtun population samples allele 11 is almost completely fixed for the DYS392
locus whereas a large number of alleles are found in the Gujars and the Kohistanis.
Similarly the Utmankhels have allele distributions that are skewed from the mean of the
whole dataset, for example showing an almost complete fixation alleles in DYS448,
DYS458 and DYS635. Small sample sizes can inflate the genetic distances and with
395 just 20 sampled individuals from each group, the R_{ST} values should be interpreted with
caution. However, we note that such extreme genetic distances have been observed
previously between other ethnic groups living in relative geographic proximity (Zeng *et al.*,
2014), when they have experienced prolonged and severe genetic isolation coupled
with long-standing endogamy (Zeng *et al.*, 2014, Roewer *et al.*, 2013, Gaikwad *et al.*,
400 2006). As such, it is perhaps not unexpected to observe large genetic distances between
the ethnic groups of Swat and Dir districts given their isolated residential localities,
their cultural preferences for endogamous marriages, as well as their differences in
subsistence practices, lifestyles, and language (Barth, 1956). The high differentiation
could be an effect of male founder effects (see below) and might not be mirrored in
405 genome-wide autosomal data, but further studies are needed to clarify this.
Nevertheless, our results indicate that isolated lifestyles and cultural preferences can
have a very large impact on genetic distances between geographically proximate
populations.

410 The genetic distinction between members of these ethnic groups is further underscored
by differential haplogroup frequencies (Table 1). The only haplogroup shared by
members of all five population samples is R1a-M417,Page7, which is not surprising as
this haplogroup occurs widely throughout the Eurasian continent, especially among
populations found in Central Asia and the Indo-Pakistani sub-continent (Karafet *et al.*,
415 2008, Novelletto, 2007, Rosser *et al.*, 2000, Semino *et al.*, 2000, Sengupta *et al.*, 2006,
Underhill *et al.*, 2010, Underhill *et al.*, 2015, Pamjav *et al.*, 2012).

Genetics and ethnicity

It is widely recognized that cultural factors such as language and group associations,
420 can sometimes play a role in forming the genetic structure among human populations,
especially those found in remote areas where populations are small and isolated
(Gaikwad *et al.*, 2006, Ayub & Tyler-Smith, 2009). Our AMOVA analyses confirm that
this is also the case for the Indo-Pakistani sub-continent, where 4.1% of the genetic
variation is explained by ethnicity whereas only 2.3% is explained by country of origin
425 (Table S5). Hence, members of the studied ethnic groups were found to be more similar
genetically to population samples assigned to their respective ethnicity than to their
country of origin (Fig. 3, Table S5).

Unlike Gujars, Kohistanis, and especially Utmankhels, the Tarklanis and Yusafzais
430 samples cannot be differentiated from each other genetically with the 23 analyzed Y-
STR markers ($R_{ST}=0$, Table 2), and the Y-SNP data show that the majority of these
individuals carry variants of haplogroup R1a-M417,Page7, that are intermingled in a
loosely defined group in the network (Fig. 2). Recent studies have dissected the R1a-
M417,Page7 haplogroup in greater detail (Pamjav *et al.*, 2012, Underhill *et al.*, 2015)

435 and it is reasonable to hypothesize that the Pakistani individuals from this study
assigned to haplogroup R1a-M417,Page7 belong to one of the sub-haplogroups of R1a-
Z95, such as R1a-Z2125, R1a-M560, or R1a-M780 (Underhill *et al.*, 2015). Although
Tarklanis and Yusafzais consider themselves to be distinct subgroups of Pashtuns
(Table S1), several studies have suggested that they share many cultural and linguistic
440 characteristics (Caroe, 1992, Khan, 2008), which is clearly mirrored in our genetic data.
In this particular case, our results suggest that both historic and current gene flow
between members of these sub-groups (i.e., patrilineal clans) prevails despite their
residence within remote areas of the Hindu Kush-Hindu Raj highlands. However, a
large pool of shared common Y-chromosomal ancestry could also explain the close
445 genetic affinity between these two subgroups. With the exception of the genetic distance
to the Utmankhels, neither of these two populations was significantly different from
other Pashtun (Pathan) groups from Afghanistan and Pakistan after Bonferroni
correction for multiple comparisons(Fig. 3 and Table S7).

450 Utmankhels also consider themselves to be Pashtuns (Table S1), but with R_{ST} distances
of 0.45-0.75 (23 loci) to the other four population samples from Dir and Swat districts
(Table 2) and 0.24-0.67 (15 loci) to populations from the Indo-Pakistani sub-continent
and Southwest Asia (Table S7), they are genetically different from any other sample
from this geographic region included in this study (see also Fig. 3). This is also reflected
455 in the haplogroup networks where most Utmankhels form a very distinct cluster within
haplogroup G-Page94 (Figs. 2, 4, Table 1). This haplogroup is common among ethnic
groups residing in the Caucasus but it is also found in medium to low frequencies
among ethnic groups residing in the Middle East and southern Europe (Rootsi *et al.*,
2012, Kivisild *et al.*, 2003). As such, the Utmankhels may be considered a genetic
460 outlier within the Indo-Pakistani sub-continent (Fig. S2), at least with regard to the Y-

chromosome. Such results suggest that they either have a different genetic origin than members of the other Pashtun sub-groups included in this analysis or that the Utmankhel male lineage has been subjected to severe genetic drift, perhaps due to a male founder effect or genetic bottleneck followed by isolation. The latter scenario is consistent with the lower genetic diversity observed among Utmankhels relative to that seen among members of the other sampled groups from Dir and Swat districts (Table 1). These results are intriguing given the oral tradition that members of the current Utmankhel clan are all descendants of a single adopted son of unknown origin (Barfield, 2010, Caroe, 1992). This could explain the apparent genetic isolation of the Utmankhel male lineage, although the presence of other Y-SNP haplogroups in the population sample (Table 1) indicates that least some male-mediated geneflow must have occurred in either ancient or recent times or that the bottleneck was not quite as dramatic as proposed (i.e. only one male). We note that our findings do not question the ethnic descriptions of the Utmankhels as a sub-ethnic group of the Pashtuns, but rather underline the fact that close cultural associations may arise without a closely shared genetic history. Interestingly the Utmankhels are not significantly different from a number of other populations from Eurasia, in particular many of the European populations included in the HGDP (Table S7), suggesting a closer affinity to population groups of Europe than to populations from the Indo-Pakistani sub-continent. Ancient connections with an European-derived gene pool could possibly explain why Utmankhels appear as a genetic outlier in the Indo-Pakistani sub-continent.

The Gujar population sample is also much differentiated genetically from the other populations residing in Swat and Dir districts but shares relatively close affinities to other populations from Pakistan and Afghanistan, in particular to the Baluch population samples from the region (Fig. 3, Table S7). This observation could support previously

suggested cultural connections, such as a shared transhumant lifestyle (Nijjar, 2008, Barth, 1956, Adamec, 2011) between Gujars and Baluchis despite linguistic differences (Grierson, 1903-1928, Morgenstierne, 1932, Strand, 1973). The high proportion of
490 individuals sharing haplotype L1-M22(xM274) could again be the result of strong genetic drift. This haplogroup is today found in West Asia and the Indo-Pakistani sub-continent (Jobling & Tyler-Smith, 2003, Kivisild *et al.*, 2003). The data could also indicate recent gene flow between Gujars and Kohistanis, since these share haplotypes within haplogroup H1-M69, G2a-L30(xL14, L13,M278), and L1-M22(xM274) (Table 1
495 and Fig. 4B). Haplogroup L1-M22(xM274) is found in low frequency among Kohistanis but is the most frequent haplogroup among Gujars and thus recent paternal gene flow from Gujars to Kohistanis can be speculated. The Gujars are more recent immigrants to Upper Swat and the opportunity for gene flow is therefore in place, but more data are needed to test this hypothesis.

500 In contrast to the other four ethnic groups included in this study, Kohistanis are more genetically diverse and not significantly different from any of the other population samples from the Indo-Pakistani sub-continent, with the exception of the Utmankhels when the data-set is restricted to 15 Y-STRs (Figs.3, S2, Table S7). However, when all 23 single-copy Y-STR loci are considered they are indeed significantly different from
505 all other population from Dir and Swat districts reflecting the ability of the rapidly mutating Y-STRs included in the YfilerPlus kit to differentiate between individuals to a higher degree. The exact relationships within haplogroup H1-M69 (the most frequent haplogroup within Kohistanis) between Kohistanis and members of other ethnic groups of Pakistan and Afghanistan are unclear (Fig. 4B). Our results could suggest that
510 Kohistanis are more genetically admixed and have perhaps experienced less isolation than the other four ethnic groups from Dir and Swat districts included in the study.

Conclusions

We have characterized the genetic diversity in paternal lineages of five ethnic groups residing in the mountainous Dir and Swat districts of the Khyber Pakhtunkhwa Province, in northern Pakistan. With the exception of Tarklanis and Yusafzais, we have documented very high levels of genetic differentiation of the male lineages between the groups. Such differences suggest either a lack of shared ancestry, perhaps due to several distinct ancient or historic migrations into this region, and/or bottlenecks and isolation events resulting in severe genetic drift in the local male gene pools. The Y-STR and Y-SNP data we present here do not offer sufficient resolution to investigate these scenarios further but the results provide a strong impetus to resolve the demographic history of this region with genome-scale analyses. Also, investigations of the maternal lineages via mitochondrial genomes should be highly informative as they may depict a different genetic history if dispersal and gene flow differ between males and females. Such a pattern has been described in geographic areas largely influenced by European settlers such as South America (Roewer *et al.*, 2013, Fridman *et al.*, 2014) and Greenland (Pereira *et al.*, 2015, Helgason *et al.*, 2006, Olofsson *et al.*, 2015b) and it is very likely that the same pattern would be observed among the ethnic groups of the Dir and Swat districts given a common preference for patrilineal first cousin marriages coupled with post-marital virilocality (Donnan, 1988, Hussain & Bittles, 1998, Saadat & Tajbakhsh, 2013, Saify & Saadat, 2012, Wahab & Ahmad, 1996).

In concurrence with previous studies, we find that ethnicity provides a more accurate predictor of genetic associations than simple geographic propinquity. However, our data also illustrates a clear exception in that Utmankhels are not related to the other Pashtun groups genetically. Thus, their cultural association could either be a more recent phenomenon not explained by shared ancestry, or alternatively, that a founder event

such as a putative adoption among the Utmankhels, followed by strong genetic drift,
540 have simply erased the genetic links but not the cultural connections.

Acknowledgements

The authors would like thank the Ethnogenetic Project (No. 20-1409) titled
“Ethnogenetic elaboration of KP through dental morphology and DNA analysis” at
545 Hazara University, Mansehra, Pakistan for assisting in sample collection. We thank
Anders Holmer from Section of Forensic Genetics, University of Copenhagen for
technical assistance. The research was funded by the Indigenous 5000 Ph.D Fellowship
Program of the Higher Education Commission of Pakistan. JKO is funded by ERC grant
ERC-2014-STG-638333 and NERC grant NE/M00208X/1. Centre for GeoGenetics is
550 funded by the Danish National Research Foundation and the Lundbeck Foundation.
MEA is funded by the Villum Foundation (Young Investigator Programme, Grant No.
10120). Lastly, we express our gratitude to the individuals we have sampled in this
study for their voluntary participation, without which our work would not have been
possible. IU, MA, NF collected the samples. IU and MSN performed the lab work.
555 JKO, AM, IU, MS, AB and MI analysed the data. MEA, HA, BEH, AJH and EW
designed the study, supervised, and provided technical guidance. JKO, MEA, IU, and
BEH wrote the manuscript with input from all co-authors.

References

- 560 Adamec, L.W. (2011) *Historical dictionary of Afghanistan*. Scarecrow Press.
Ahmad, M. & Sirajuddin, A. (1996) Ethnobotanical profile of Swat. In: *Proceeding of first
training workshop on Ethnobotany and its application to conservation, Islamabad,
Pakistan*) *Proceeding of first training workshop on Ethnobotany and its application to
conservation, Islamabad, Pakistan*.

- 565 Alam, S., Ali, M.E., Ferdous, A., Hossain, T., Hasan, M.M. & Akhteruzzaman, S. (2010) Haplotype diversity of 17 Y-chromosomal STR loci in the Bangladeshi population. *Forensic science international: genetics*, 4, e59-e60.
- Ali, S.I. & Qaiser, M. (1986) A phytogeographical analysis of the phanerogams of Pakistan and Kashmir. *Proceedings of the Royal Society of Edinburgh. Section B. Biological Sciences*, 89, 89-101.
- 570 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. *Briefings in functional genomics & proteomics*, 8, 395-404.
- Bangash, S. (2012) Socio-economic conditions of post-conflict Swat: a critical appraisal. *J. Peace Dev. II. FATA Research Centre, Islamabad*.
- 575 Barfield, T. (2010) *Afghanistan: A cultural and political history*. Princeton University Press.
- Barth, F. (1956) Ecologic relationships of ethnic groups in Swat, North Pakistan. *American Anthropologist*, 58, 1079-1089.
- Bernhard, W. (1983) Ethnogenesis of South Asia with special reference to India. *Anthropologischer Anzeiger*, 93-110.
- 580 Birdwood (1959) A History of the Pathans: Review. *The Geographical Journal*, 125, 414-416.
- Böhner, J. & Lucarini, V. (2015) Prevailing climatic trends and runoff response from Hindukush-Karakoram-Himalaya, upper Indus basin. *arXiv preprint arXiv:1503.06708*.
- Cann, H.M., De Toma, C., Cazes, L., Legrand, M.-F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B. & Cambon-Thomsen, A. (2002) A human genome diversity cell line panel. *Science*, 296, 261-262.
- 585 Caroe, O. (1992) *The Pathans (1958)*. Karachi: Oxford University Press.
- Chennakrishnaiah, S., Perez, D., Gayden, T., Rivera, L., Regueiro, M. & Herrera, R.J. (2013) Indigenous and foreign Y-chromosomes characterize the Lingayat and Vokkaliga populations of Southwest India. *Gene*, 526, 96-106.
- 590 Coningham, R. & Young, R. (2015) *The archaeology of South Asia: from the Indus to Asoka, c. 6500 BCE–200 CE*. Cambridge University Press.
- Derenko, M., Malyarchuk, B., Bahmanimehr, A., Denisova, G., Perkova, M., Farjadian, S. & Yepiskoposyan, L. (2013) Complete mitochondrial DNA diversity in Iranians. *PLoS one*, 8, e80673.
- 595 Donnan, H. (1988) *Marriage among Muslims: preference and choice in northern Pakistan*. Brill.
- Excoffier, L. & Lischer, H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, 10, 564-567.
- 600 Fridman, C., Gonzalez, R., Pereira, A. & Cardena, M. (2014) Haplotype diversity in mitochondrial DNA hypervariable region in a population of southeastern Brazil. *International journal of legal medicine*, 128, 589-593.
- Gaikwad, S., Vasulu, T. & Kashyap, V. (2006) Microsatellite diversity reveals the interplay of language and geography in shaping genetic differentiation of diverse Proto-Australoid populations of west-central India. *American journal of physical anthropology*, 129, 260-267.
- 605 Glatzer, B. (2002) The Pashtun tribal system. *Concept of tribal society*, 5, 265-282.
- Grierson, G. (1903-1928) *Linguistic Survey of India (11 volumes)*. Calcutta: Office of the Superintendent of Government Printing, India.
- 610 Grimes, B.F. & Grimes, J.E. (2000) *Ethnologue: Languages of the world*. SIL International.
- Haber, M., Platt, D.E., Bonab, M.A., Youhanna, S.C., Soria-Hernanz, D.F., Martínez-Cruz, B., Douaihy, B., Ghassibe-Sabbagh, M., Rafatpanah, H. & Ghanbari, M. (2012) Afghanistan's ethnic groups share a Y-chromosomal heritage structured by historical events. *PLoS one*, 7, e34288.
- 615 Hazrat, A., Shah, J., Ali, M. & Iqbal, I. (2007) Medicinal value of Ranunculaceae of Dir valley. *PAKISTAN JOURNAL OF BOTANY*, 39, 1037.

- Helgason, A., Pálsson, G., Pedersen, H.S., Angulalik, E., Gunnarsdóttir, E.D., Yngvadóttir, B. & Stefánsson, K. (2006) mtDNA variation in Inuit populations of Greenland and Canada: migration history and population structure. *American Journal of Physical Anthropology*, 130, 123-134.
- 620 Hemphill, B.E. (2009) Bioanthropology of the Hindu Kush High Lands: A Dental Morphology Investigation. *Pakistan Heritage*, 1, 19-36.
- Hemphill, B.E. (2013a) Grades, gradients, and geography: a dental morphometric approach to the population history of South Asia. In: *Anthropological Perspectives on Tooth Morphology: Genetics, Evolution, Variation* G.R. Scott & J.D. Irish (eds.) *Anthropological Perspectives on Tooth Morphology: Genetics, Evolution, Variation*. Cambridge: Cambridge University Press.
- 625 Hemphill, B.E. (2013b) A View to the North: Biological Interactions across the Intermontane Borderlands during the Last Two Millennia B.C. In: *South Asian Archaeology 2007, Volume I* T. M. & F. D (eds.) *South Asian Archaeology 2007, Volume I*. Oxford: Archaeopress-BAR.
- 630 Hemphill, B.E., Ali, I., Blaylock, S. & Willits, N. (2013) Are the Kho an indigenous population of the Hindu Kush?: A dental morphometric approach. In: *South Asian Archaeology 2007* M. Tosi & D. Frenez (eds.) *South Asian Archaeology 2007*. Oxford: Archaeopress-BAR.
- 635 Hemphill, B.E. & Mallory, J. (2004) 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. *American Journal of Physical Anthropology*, 124, 199-222.
- Hussain, R. & Bittles, A. (1998) The prevalence and demographic characteristics of consanguineous marriages in Pakistan. *Journal of biosocial science*, 30, 261-275.
- 640 Jettmar, K. (1967) The Middle Asiatic Heritage of Dardistan.(Islamic Collective Tombs in Punjab and Their Background). *East and West*, 17, 59-82.
- Jettmar, K. (1996) Approaches to the History of North Pakistan. In: *Proceedings of the Second International Hindukush Cultural Conference* *Proceedings of the Second International Hindukush Cultural Conference*. Oxford University Press, USA.
- 645 Jobling, M.A. & Tyler-Smith, C. (2003) The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics*, 4, 598-612.
- Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L. & Hammer, M.F. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome research*, 18, 830-838.
- 650 Khan, T.M. (2008) *The Tribal Areas of Pakistan, a Contemporary Profile*. Sang-e-Meel Publications.
- Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.-V. & Stepanov, V. (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *The American Journal of Human Genetics*, 72, 313-332.
- 655 Kuz'mina, E.E. & Mair, V.H. (2008) *The prehistory of the Silk Road*. University of Pennsylvania Press.
- Lapidus, I.M. (2002) *A history of Islamic societies*. Cambridge University Press.
- 660 Larmuseau, M.H., Van Geystelen, A., Kayser, M., Van Oven, M. & Decorte, R. (2015) Towards a consensus Y-chromosomal phylogeny and Y-SNP set in forensics in the next-generation sequencing era. *Forensic Science International: Genetics*, 15, 39-42.
- Lee, E.Y., Shin, K.-J., Rakha, A., Sim, J.E., Park, M.J., Kim, N.Y., Yang, W.I. & Lee, H.Y. (2014) Analysis of 22 Y chromosomal STR haplotypes and Y haplogroup distribution in Pathans of Pakistan. *Forensic Science International: Genetics*, 11, 111-116.
- 665 Mehdi, S., Qamar, R., Ayub, Q., Khaliq, S., Mansoor, A., Ismail, M., Hammer, M., Underhill, P. & Cavalli-Sforza, L. (1999) The Origins of Pakistani Populations. In: *Genomic Diversity* *Genomic Diversity*. Springer.

- Morgenstierne, G. (1932) *Report on a linguistic mission to north-western India*. Indus Publication.
- 670 Newcomb, L. (1986) The Islamic Republic of Pakistan: country profile. *International demographics*, 5, 1-8.
- Nijjar, B.S. (2008) *Origins and History of Jats and Other Allied Nomadic Tribes of India: 900 BC-1947 AD*. Atlantic Publishers & Dist.
- 675 Novelletto, A. (2007) Y chromosome variation in Europe: Continental and local processes in the formation of the extant gene pool. *Annals of human biology*, 34, 139-172.
- Nüsser, M. & Dickoré, W.B. (2002) A tangle in the triangle: vegetation map of the eastern Hindukush (Chitral, northern Pakistan). *Erdkunde*, 37-59.
- Olofsson, J.K., Mogensen, H.S., Buchard, A., Børsting, C. & Morling, N. (2015a) Forensic and population genetic analyses of Danes, Greenlanders and Somalis typed with the Yfiler® Plus PCR amplification kit. *Forensic Science International: Genetics*, 16, 232-236.
- 680 Olofsson, J.K., Pereira, V., Børsting, C. & Morling, N. (2015b) Peopling of the North Circumpolar Region—insights from Y chromosome STR and SNP typing of Greenlanders. *PloS one*, 10, e0116573.
- 685 Oppenheimer, S. (2012) Out-of-Africa, the peopling of continents and islands: tracing uniparental gene trees across the map. *Phil. Trans. R. Soc. B*, 367, 770-784.
- Pamjav, H., Fehér, T., Németh, E. & Pádár, Z. (2012) Brief communication: New Y-chromosome binary markers improve phylogenetic resolution within haplogroup R1a1. *American journal of physical anthropology*, 149, 611-615.
- 690 Parpola, A. (1995) 15. The problem of the Aryans and the Soma: Textual-linguistic and archaeological evidence. *The Indo-Aryans of Ancient South Asia: Language, Material Culture and Ethnicity*, 1, 353.
- Parpola, A. (2009) *Deciphering the Indus script*. Cambridge University Press.
- Pereira, V., Tomas, C., Sanchez, J.J., Syndercombe-Court, D., Amorim, A., Gusmão, L., Prata, M.J. & Morling, N. (2015) The peopling of Greenland: further insights from the analysis of genetic diversity using autosomal and X-chromosomal markers. *European Journal of Human Genetics*, 23, 245-251.
- 695 Perveen, R., Rahman, Z., Shahzad, M.S., Israr, M., Shafique, M., Shan, M.A., Zar, M.S., Iqbal, M. & Husnain, T. (2014) Y-STR haplotype diversity in Punjabi population of Pakistan. *Forensic science international. Genetics*, 9, e20.
- 700 Petraglia, M.D., Alsharekh, A., Breeze, P., Clarkson, C., Crassard, R., Drake, N.A., Groucutt, H.S., Jennings, R., Parker, A.G. & Parton, A. (2012) Hominin dispersal into the Nefud desert and Middle Palaeolithic settlement along the Jubbah palaeolake, northern Arabia. *PLoS One*, 7, e49840.
- 705 Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S.M., Santos, L.H., Anslinger, K. & Bayer, B. (2014) A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, 12, 12-23.
- Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C. & Mehdi, S.Q. (2002) Y-chromosomal DNA variation in Pakistan. *The American Journal of Human Genetics*, 70, 1107-1124.
- 710 Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., Mcelreavey, K. & Santachiara-Benerecetti, A.S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet*, 23, 437-441.
- Ralsler, M., Querfurth, R., Warnatz, H.-J., Lehrach, H., Yaspo, M.-L. & Krobitsch, S. (2006) An efficient and economic enhancer mix for PCR. *Biochemical and biophysical research communications*, 347, 747-751.
- 715 Roewer, L. (2009) Y chromosome STR typing in crime casework. *Forensic science, medicine, and pathology*, 5, 77-84.
- Roewer, L., Nothnagel, M., Gusmão, L., Gomes, V., González, M., Corach, D., Sala, A., Alechine, E., Palha, T. & Santos, N. (2013) Continent-wide decoupling of Y-chromosomal genetic
- 720

- variation from language and geography in native South Americans. *PLoS Genet*, 9, e1003460.
- 725 Roewer, L., Willuweit, S., Stoneking, M. & Nasidze, I. (2009) A Y-STR database of Iranian and Azerbaijani minority populations. *Forensic Science International: Genetics*, 4, e53-e55.
- Rome, S.-I. (2008) *Swat State (1915-1969) From Genesis to Merger: An Analysis of Political, Administrative, Socio-Political, and Economic Developments.* Karachi: Oxford University Press.
- 730 Rootsi, S., Myres, N.M., Lin, A.A., Järve, M., King, R.J., Kutuev, I., Cabrera, V.M., Khusnutdinova, E.K., Varendi, K. & Sahakyan, H. (2012) Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. *European journal of human genetics*, 20, 1275-1282.
- Rosenberg, N.A. (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of human genetics*, 70, 841-847.
- 735 Rosser, Z.H., Zerjal, T., Hurler, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E. & Barbujani, G. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *The American Journal of Human Genetics*, 67, 1526-1543.
- 740 Saadat, M. & Tajbakhsh, K. (2013) Prevalence of consanguineous marriages in west and south of Afghanistan. *J Biosoc Sci*, 45, 799-805.
- Saify, K. & Saadat, M. (2012) Consanguineous marriages in Afghanistan. *Journal of biosocial science*, 44, 73-81.
- 745 Seema Nair, P., Geetha, A. & Jagannath, C. (2011) Y-short tandem repeat haplotype and paternal lineage of the Ezhava population of Kerala, south India. *Croat Med J*, 52, 344-350.
- Semino, O., Passarino, G., Oefner, P.J., Lin, A.A., Arbuzova, S., Beckman, L.E., De Benedictis, G., Francalacci, P., Kouvatsi, A. & Limborska, S. (2000) The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: AY chromosome perspective. *science*, 290, 1155-1159.
- 750 Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S., Edmonds, C.A., Chow, C.-E.T., Lin, A.A., Mitra, M., Sil, S.K. & Ramesh, A. (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *The American Journal of Human Genetics*, 78, 202-221.
- 755 Shah, G.A. (2013) Administration of Dir under Nawab Shah Jehan. *Pakistan Annual Research Journal*, 49.
- Shi, W., Ayub, Q., Vermeulen, M., Shao, R.-G., Zuniga, S., Van Der Gaag, K., De Knijff, P., Kayser, M., Xue, Y. & Tyler-Smith, C. (2010) A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Molecular biology and evolution*, 27, 385-393.
- 760 Siddique, A. (2014) *The Pashtun Question: The Unresolved Key to the Future of Pakistan and Afghanistan.* Hurst & Company Limited.
- Strand, R.F. (1973) Notes on the Nūristāni and Dardic Languages. *Journal of the American Oriental Society*, 297-305.
- 765 Tabassum, S., Ilyas, M., Ullah, I., Israr, M. & Ahmad, H. (2017) A comprehensive Y-STR portrait of Yousafzai's population. *International Journal of Legal Medicine*, 1-2.
- Triki-Fendri, S., Sánchez-Diz, P., Rey-González, D., Ayadi, I., Carracedo, Á. & Rebai, A. (2015) Paternal lineages in Libya inferred from Y-chromosome haplogroups. *American journal of physical anthropology*, 157, 242-251.
- 770 Underhill, P.A., Myres, N.M., Rootsi, S., Metspalu, M., Zhivotovsky, L.A., King, R.J., Lin, A.A., Chow, C.-E.T., Semino, O. & Battaglia, V. (2010) Separating the post-Glacial coancestry

- of European and Asian Y chromosomes within haplogroup R1a. *European Journal of Human Genetics*, 18, 479-484.
- 775 Underhill, P.A., Poznik, G.D., Rootsi, S., Järve, M., Lin, A.A., Wang, J., Passarelli, B., Kanbar, J., Myres, N.M. & King, R.J. (2015) The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *European Journal of Human Genetics*, 23, 124-131.
- Vadime, E. (2001) *The Silk Roads: Highways of Culture and Commerce.* UNESCO publishing/Berghahn Books.
- 780 Vermeulen, M., Wollstein, A., Van Der Gaag, K., Lao, O., Xue, Y., Wang, Q., Roewer, L., Knoblauch, H., Tyler-Smith, C. & De Knijff, P. (2009) Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms. *Forensic Science International: Genetics*, 3, 205-213.
- Wahab, A. & Ahmad, M. (1996) Biosocial perspective of consanguineous marriages in rural and urban Swat, Pakistan. *Journal of biosocial science*, 28, 305-313.
- 785 Willuweit, S. & Roewer, L. (2015) The new Y chromosome haplotype reference database. *Forensic Science International: Genetics*, 15, 43-48.
- Zeng, Z., Garcia-Bertrand, R., Calderon, S., Li, L., Zhong, M. & Herrera, R.J. (2014) Extreme genetic heterogeneity among the nine major tribal Taiwanese island populations detected with a new generation Y23 STR system. *Forensic Science International: Genetics*, 12, 100-106.
- 790 Zhao, Z., Khan, F., Borkar, M., Herrera, R. & Agrawal, S. (2009) Presence of three different paternal lineages among North Indians: a study of 560 Y chromosomes. *Annals of human biology*, 36, 46-59.

795

Figure legends

800 **Figure 1.** *Map of study area*

Map of Pakistan with focus on Dir and Swat districts. Sampling localities for each of the five ethnic groups are indicated. Upper and lower dashed lines indicate the Hindu Kush and Hindu Raj ranges respectively.

805 **Figure 2.** *Network analysis*

Median joining network based on the Y-STR haplotypes (23 loci) of the five population samples. The circle sizes indicate the number of individuals with shared Y-STR haplotypes (smallest circles = one individual). The lengths of the connecting branches

indicate the number of mutational steps separating the haplotypes (shortest branch
810 lengths = one mutational step).

Figure 3. *Multidimensional scaling plot of regional populations*

Multidimensional scaling (MDS) analysis of pairwise genetic distances, estimated as
 R_{ST} (15 Y-STR loci), for 27 selected populations from the Indo-Pakistani sub-continent
815 and neighbouring countries (stress = 0.1544333). See Table S4 for a detailed list of the
included populations, number of individuals, and references.

Figure 4. *Y-chromosome haplogroup-specific networks*

Median joining network based on Y-STR haplotypes (15 loci) with individuals assigned
820 to (A) Y-SNP haplogroups G-Page94, (B) H1-M69, and (C) Y-SNP haplogroup L1-
M22(xM274). The circle sizes indicate the number of individuals that share the same Y-
STR profile for these 15 loci. The smallest circles represent one individual. The lengths
of the connecting branches indicate the number of mutational steps. The shortest
branches represent one mutational step.

825

830

Supplementary files, legends

Figure S1. MDS plot of Dir and Swat

Multi-dimensional scaling (MDS) analysis of pairwise genetic distances, estimated as
835 R_{ST} (23 Y-STR loci), for the five population samples in this study (stress = 1.32472e-
16).

Figure S2: Worldwide MDS plot

Multi-dimensional scaling (MDS) analysis of pairwise genetic distances, estimated as
840 R_{ST} (15 Y-STR loci) for a) 53 population samples (from HGDP), including the five
population samples from Dir and Swat as well as the Yusafzai population from
Tabassum et al 2017; b) 48 populations samples from the HGDP and the Yusafzai
population from Tabassum et al 2017; c-g) 49 population samples including the samples
from the HGDP, the Yusafzai sample from Tabassum et al.(2017), and one of the five
845 population samples (as indicated in sub-figures) analyzed in this study. See Table S4 for
a detailed list of the included populations, number of individuals, and references. Stress
values as indicated in the separate sub-figures.

850 Table S1: Ethnic divisions of Pashtuns

Major ethnic and sub-ethnic groups of Pashtuns/Pathans residing in the Khyber
Pakhtunkhwa province of Pakistan.

Table S2: Genotype data

855 Haplotypes of 27 Y-STRs amplified with the Yfiler®Plus PCR amplification kit and Y-
SNP haplogroups for 100 individuals from five ethnically distinct populations of the Dir
and Swat district of northern Pakistan. For the Y-STR loci intermediate alleles and

duplication events are highlighted. Y-SNP haplogroup names according to International Society of Genetic Genealogy (ISOGG).

860

Table S3: Y-SNP calls and haplogroups

An overview of the Y-SNP derived mutations and haplogroup assignment for each individual.

865 **Table S4: Population samples**

An overview of the 68 population samples included in the larger comparative analyses. Sample sizes and references to the original studies are shown. Groups marked with ^a were used both for the regional analyses (MDS, AMOVA) and the world wide analysis (MDS). Groups marked with ^b were used only for the regional analyses (MDS, AMOVA). Groups marked with ^c were used only for the world wide MDS analysis.

870

Table S5. AMOVA test + description of rationale behind ethnicity aggregates

Analyses of molecular variance (AMOVA) when population samples are grouped based on country of origin and ethnicity, respectively.

875

Table S6. Alternative AMOVA test

Analyses of molecular variance (AMOVA) when population samples are grouped based ethnicity including Utmankhels in the Pashtun ethnic group.

880 **Table S7A+B. Genetic distances, R_{ST}**

3A) Regional R_{ST} analysis of population samples from the Indian subcontinent. The genetic distances, pairwise R_{ST} values, below the diagonal and the corresponding p-

values above the diagonal(15 Y-STR loci), between all populations.3B) Worldwide R_{ST} analysis. The genetic distances, pairwise R_{ST} values, below the diagonal and the
885 corresponding p-values above the diagonal based on the Y-chromosomal haplotype frequencies (15 Y-STR loci), between all populations.

Table 1: Genetic diversity

Number of individuals sharing a Y-STR haplotype	Sub-population					Meta-population of Dir and Swat District
	Kohistanis	Gujars	Yusafzais	Tarklanis	Utmankhels	
1 (unique)	20 ^a	16	15	17 ^d	9	75
2			1		1	2
3			1 ^c	1	1	2
4		1 ^b				1 ^e
5						1 ^f
6					1	1
Number of haplotypes	20	17	17	18	12	82
Sample size	20	20	20	20	20	100
Frequency of unique haplotypes	1.00	0.80	0.75	0.85	0.45	0.75
Haplotype diversity	1.00	0.98	0.99	0.99	0.93	0.99
Power of discrimination	1.00	0.85	0.85	0.90	0.60	0.82

Y-SNP haplogroup	Kohistanis	Gujars	Yusafzais	Tarklanis	Utmankhels	Combined
G2a-L30(xL14, L13,M278)	1 (0.05; 0.03-0.07)	1 (0.05; 0.03-0.07)				2 (0.02; 0.01-0.03)
G2b-M283				2 (0.10; 0.07-0.13)	16 (0.80; 0.77-0.83)	18 (0.18; 0.17-0.19)
H1-M69	10 (0.50; 0.46-0.54)	1 (0.05; 0.03-0.07)				11 (0.11; 0.10-0.12)
J2a-L25			2 (0.10; 0.07-0.13)			2 (0.02; 0.01-0.03)
J2b-M241			1 (0.05; 0.03-0.07)	1 (0.05; 0.03-0.07)		2 (0.02; 0.01-0.03)
L1-M22(xM274)	1 (0.05; 0.03-0.07)	11 (0.55; 0.51-0.59)	1 (0.05; 0.03-0.07)			13 (0.13; 0.12-0.14)
O2-IMS-JST0213554(xP164)		1 (0.05; 0.03-0.07)				1 (0.01; 0.006-0.014)
Q-M242(xL56, L57, L214)	2 (0.10; 0.07-0.13)					2 (0.02; 0.01-0.03)
Q-L56,L57(xL54)					2 (0.10; 0.07-0.13)	2 (0.02; 0.01-0.03)
R-M207,M734,P224,P280(xM173)	1 (0.05; 0.03-0.07)	2 (0.10; 0.07-0.13))		1 (0.05; 0.03-0.07)		4 (0.04; 0.03-0.05)
R-M734,P224,P280(xM173)		1 (0.05; 0.03-0.07)				1 (0.01; 0.006-0.014)
R1a-M417,Page7	5 (0.25; 0.21-0.29)	3 (0.15; 0.12-0.18)	16 (0.80; 0.77-0.83)	16 (0.80; 0.77-0.83)	2 (0.10; 0.07-0.13)	42 (0.42; 0.40-0.44)

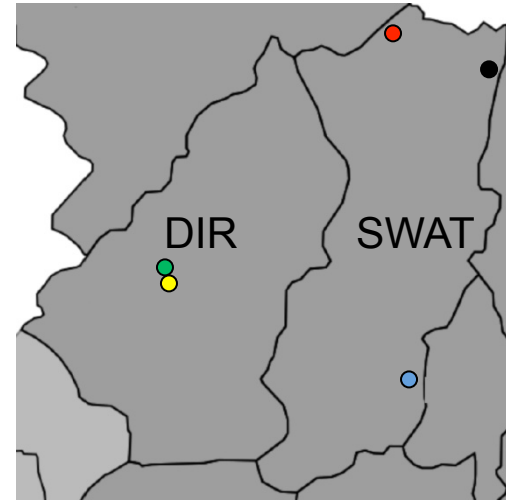
Genetic diversity in the 27 Y-STR loci and frequencies of Y-SNP haplogroups within five ethnic groups from Dir and Swat Districts and the meta-population of Dir and Swat Districts, combining all the 100 analysed individuals in this study. The values reported for the Y-SNP haplogroups represent the observed number of individuals followed by (in brackets) the frequency, and the 90% confidence interval.

^a One haplotype shared with four Gujar individuals; ^b Shared with one Kohistani individual; ^c Shared with one Tarklani individual; ^d One haplotype shared with three Yusafzai individuals; ^e Shared between three Yusafzai and one Tarklani individuals; ^f Shared between four Gujar and one Kohistani individuals.

Table 2. Genetic differentiation

	Gujar	Kohistani	Tarklani	Utmankhel	Yusafzai
Gujar	-	0.003±0.0005*	0.000±0.0005*	0.000±0.0005*	0.000±0.0005*
Kohistani	0.179	-	0.000±0.0005*	0.000±0.0005*	0.001±0.0002*
Tarklani	0.465	0.197	-	0.000±0.0005*	0.604±0.0048
Utmankhel	0.451	0.517	0.746	-	0.000±0.0005*
Yusafzai	0.395	0.154	0	0.702	-

The genetic distances calculated as pairwise R_{ST} values based on 23 of the 27 Y-STR loci. R_{ST} values below the diagonal and the corresponding P-values above the diagonal. * Significant at 0.05 significant level with correction for multiple testing ($0.05/10 = 0.005$).



- Utmankheils
- Tarklanis
- Yusafzais
- Kohistanis
- Gujars

- GUJARS
- KOHISTANIS
- TARKLANIS
- UTMANKHELIS
- YUSAFZAIS

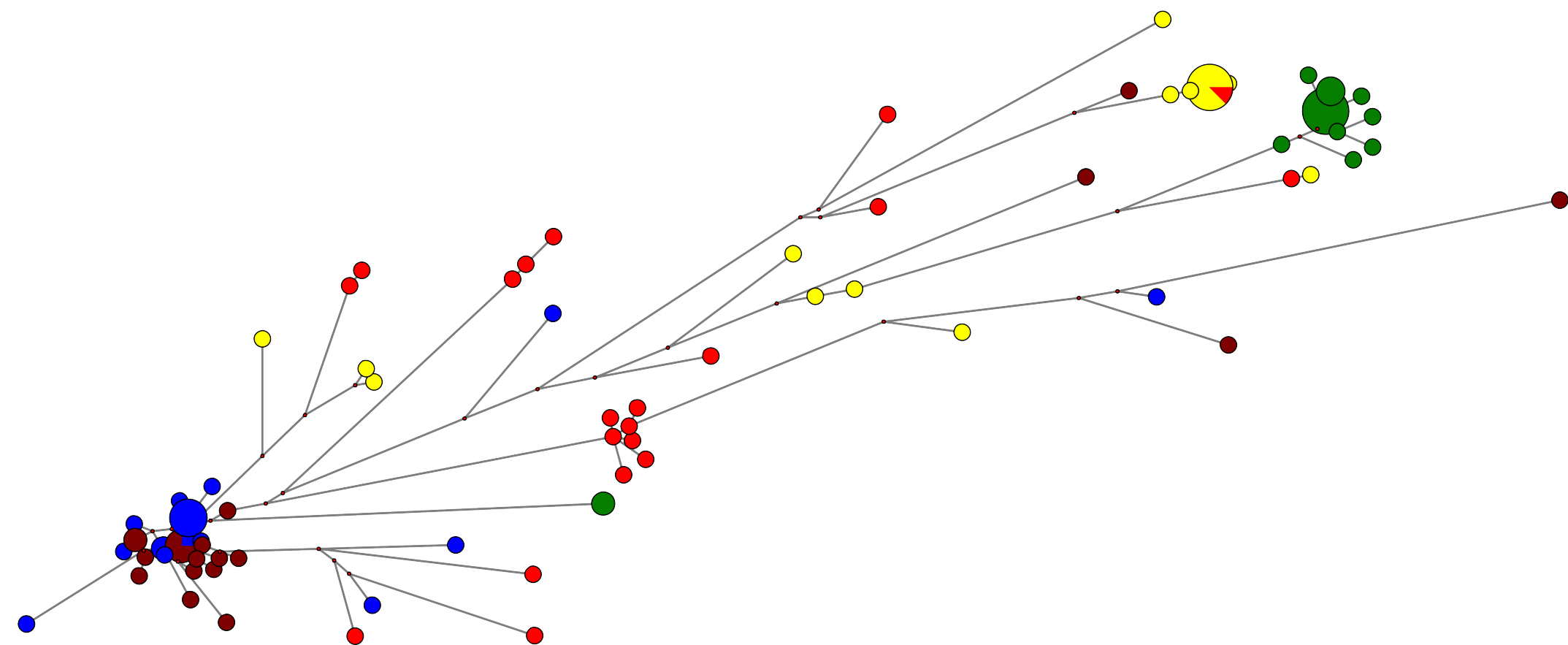


Fig. 3

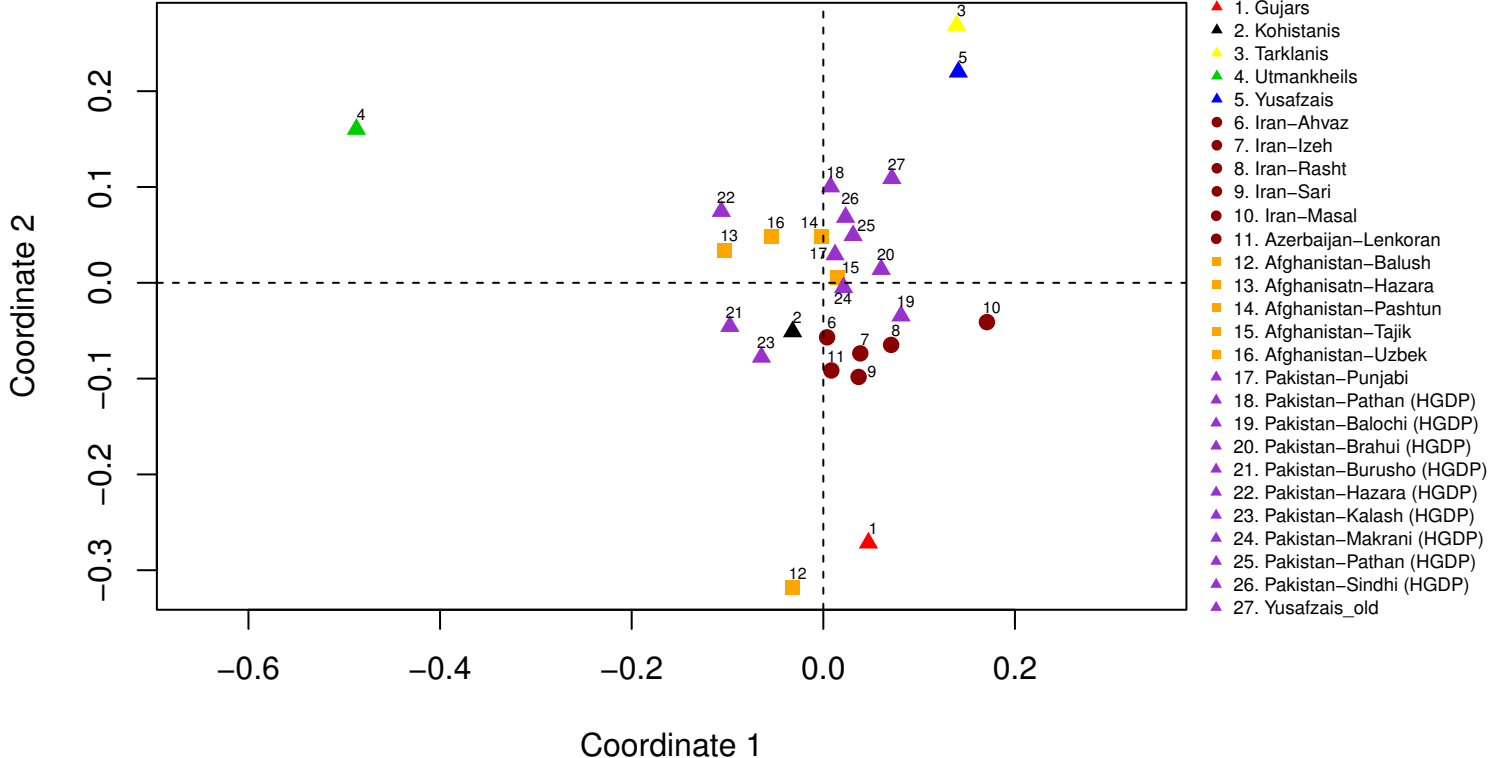
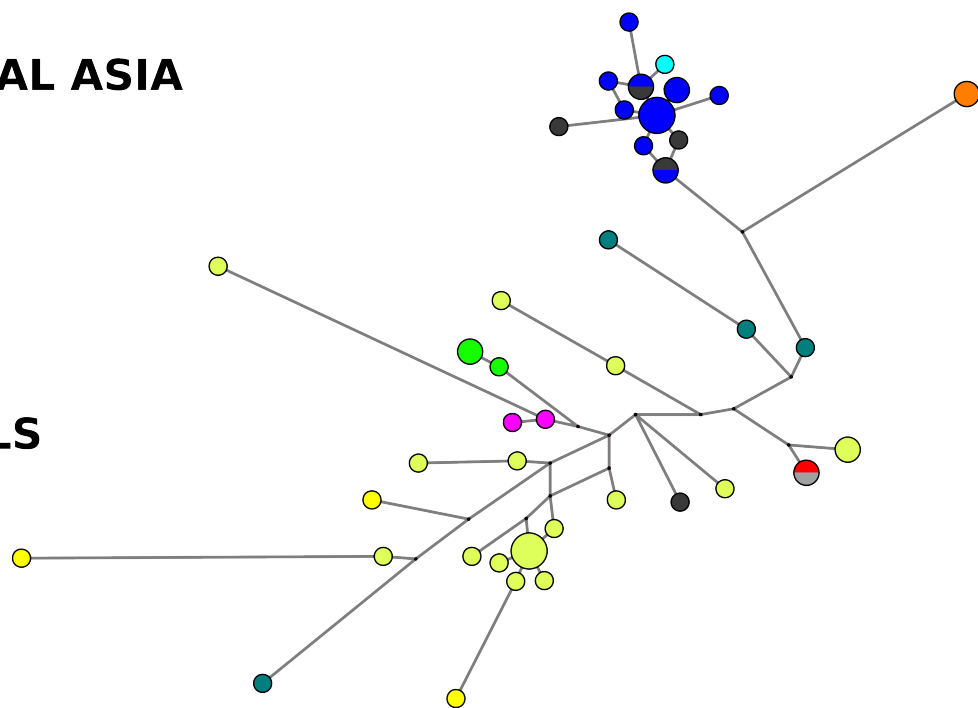
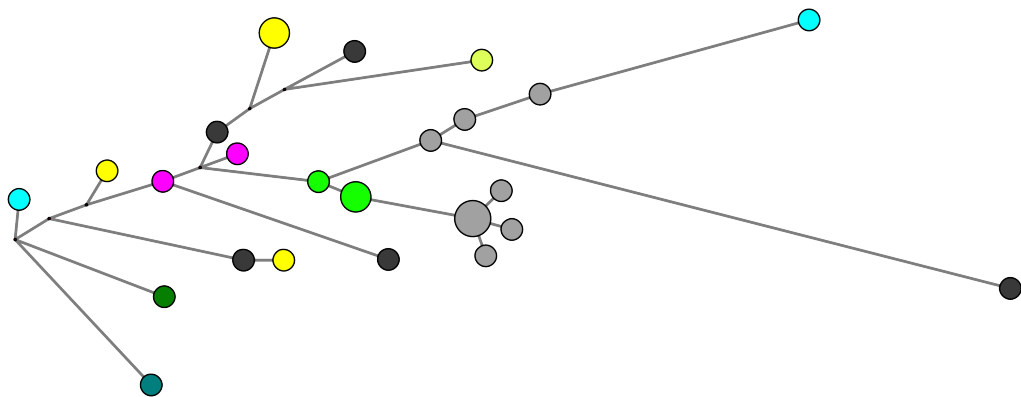


Fig. 4

- A)
- GUJARS
 - NON-CENTRAL ASIA
 - HAZARA
 - BRAHUI
 - BURUSHO
 - KALASH
 - PASHTUNS
 - KOHISTANIS
 - TARKLANIS
 - TAJIK
 - UTMANKHELIS



- B)
- NON-CENTRAL ASIA
 - KOHISTANIS
 - PASHTUNS
 - BALUCH
 - BRAHUI
 - BURUSHO
 - HAZARA
 - KALASH
 - TAJIK



- C)
- BALUCH
 - BRAHUI
 - BURUSHO
 - GUJARS
 - HAZARA
 - KALASH
 - KOHISTANIS
 - MAKRANI
 - NON-CENTRAL ASIA
 - PASHTUNS
 - TAJIK
 - UZBEK
 - YUSAFZAIS

