

Unveiling the genetic tapestry of Kohistan: a population genetic analysis of autosomal STRs in a South Asian population

Received: 21 January 2025

Accepted: 6 February 2026

Published online: 13 February 2026

Cite this article as: Adnan A., Rakha A., Chen H.D. *et al.* Unveiling the genetic tapestry of Kohistan: a population genetic analysis of autosomal STRs in a South Asian population. *BMC Genomics* (2026). <https://doi.org/10.1186/s12864-026-12641-x>

Atif Adnan, Allah Rakha, Hao Dong Chen, Muhammad Ilyas, Muhammad Farhat Ullah, Shahid Nazir, Chuan Chao Wang & Hongbo Wang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Unveiling the Genetic Tapestry of Kohistan: A Population Genetic Analysis of Autosomal STRs in a South Asian Population

Atif Adnan^{1,2 *}, Allah Rakha³, Hao Dong Chen⁴, Muhammad Ilyas⁵, Muhammad Farhat Ullah⁶, Shahid Nazir³, Chuan Chao Wang^{4,7,8*} and Hongbo Wang^{9,10*}

1. School of Medical Science, Shandong Xiehe University, **Jinan, Shandong** Province 250109, **China**.
2. Department of Chemistry and Forensic Science, College of Science, Technology, Engineering, and Mathematics, Eastern Kentucky University, **Richmond, Kentucky**, 40475, **USA**.
3. Department of Forensic Sciences, University of Health Sciences, **Lahore**, 54600, **Punjab, Pakistan**.
4. Department of Anthropology and Ethnology, Institute of Anthropology, School of Sociology and Anthropology, Xiamen University, **Xiamen, Fujian** Province 361000, **China**
5. Center for OMIC Sciences, Islamia Collage University Peshawar, **Peshawar, KPK, Pakistan**
6. National Centre for Bioinformatics, Quaid-I-Azam University, **Islamabad** 44000, **Pakistan**.
7. Ministry of Education Key Laboratory of Contemporary Anthropology, Department of Anthropology and Human Genetics, School of Life Sciences, Fudan University, **Shanghai** 200438, **China**
8. State Key Laboratory of Genetic Engineering, Center for Evolutionary Biology, School of Life Sciences, Fudan University, **Shanghai** 200438, **China**
9. Department of Human Anatomy, School of Basic Medical Sciences, Shenyang Medical College. **Shenyang**, 110034, **Liaoning** Province **China**.
10. Liaoning Province Key Laboratory for Phenomics of Human Ethnic Specificity and Critical Illness, Shenyang Medical College, **Shenyang, Liaoning** Province, **China**.

*Correspondence

Atif Adnan (mirzaatifadnan@gmail.com)

Chuan Chao Wang (chuanchaowang@fudan.edu.cn)

Hongbo Wang (wanghongbo@symc.edu.cn)

ABSTRACT

[illegible]

structure, shaped by isolation and endogamy, and contribute to understanding South Asian genetic diversity.

Keywords: Kohistani population; Forensic Genetics; Population history; Autosomal STRs; South Asia

INTRODUCTION:

Northern Pakistan, characterized by its distinctive topography of high mountains and deep river gorges, along with its diverse ethnicities and languages, is an area of significant research interest due to its rich cultural heritage and ethnic diversity. This region serves as a crucial connection point between the Himalayas, the Hindukush, and the Karakoram Mountain ranges. Nestled within this rugged landscape is Kohistan, a hilly region geographically positioned in the northernmost part of Khyber Pakhtunkhwa (KPK) province, between Pakistan's western and eastern borders. The Indus River divides Kohistan into two parts: Hazara and Swat, which were unified in 1976 to form the Kohistan district (Figure 1), encompassing an area of 7,492 square kilometers. This rugged terrain has fostered the development of distinct ethnic groups and languages, isolated yet interconnected through centuries of migration and trade. Kohistan is home to various ethnic groups, the most prominent being the Kohistani people, who speak the Kohistani language. Classified as a Dardic language belonging to the Indo-European family [1], Kohistani plays a pivotal role in shaping community identity [2].

However, among the Dardic languages, Kohistani remains the least studied [3]. The origins and genealogy of the Kohistani people are largely unknown due to a scarcity of historical records and established traditions. Ancient mentions of the region appear in the Rig Veda, as well as in Greek and Chinese literature, reflecting its historical significance as a Silk Road crossroads. Throughout history, Buddhist, Hindu, and Muslim cultures have influenced Kohistan [4, 5], leaving an indelible mark on modern Kohistani society, which continues to reflect a blend of these traditions. This cultural diversity aligns with the complex genetic history of South Asia, shaped by ancient migrations and admixture events, as evidenced by genomic studies of the region [6, 7]. Despite its cultural richness, Kohistan has the lowest literacy rates in Pakistan, with 17% for males and 3% for females, and a population of 784,711 with an average annual growth rate of 2.70% (<https://www.pbs.gov.pk/>).

The Kohistani population, residing in a geographically isolated region of northern Pakistan, represents an understudied ethnic group with unique linguistic and cultural characteristics. Centuries of migration along the Silk Road have contributed to Kohistan's rich tapestry of ethnicities, making it a unique case study for exploring human genetic diversity. Social practices such as endogamy, prevalent within Kohistani society, likely influence the population's genetic structure, contributing to patterns of genetic drift and reduced diversity. Despite its historical significance as a crossroads between South and Central Asia, the genetic composition and population history of the Kohistani people remain poorly understood due to limited genetic studies.

To address these gaps, this study investigates the genetic structure, ancestry, and relationships of the Kohistani population using autosomal Short Tandem Repeat (STR) markers. STRs, known for their high polymorphism and forensic utility, enable robust comparisons with global datasets such as the 1000 Genomes Project [8], which provides a comprehensive reference for human genetic variation across 26 worldwide populations. These markers are DNA

sequences with repeating patterns of 2 to 7 base pairs found throughout the human genome, characterized by stable polymorphisms, short sequence lengths, and even chromosomal distribution, facilitating easy identification and analysis using PCR and sequencing [9, 10]. Due to their highly polymorphic and informative nature, STRs are considered the gold standard in forensic investigations, including paternity testing, sexual assault cases, kinship analysis, and missing person identification [11–15]. Their widespread use in forensic genetics is well-documented in foundational studies that established STRs as reliable markers for human identification and kinship analysis [14, 15]. In this study, we genotyped 200 unrelated individuals from the Kohistani ethnic group of Pakistan using a 6-dye GlobalFiler PCR Amplification kit (Thermo Fisher Scientific, Inc., Waltham, MA, USA). This kit includes 21 commonly used autosomal STR loci (D8S1179, D21S11, D7S820, CSF1PO, D3S1358, TH01, D13S317, D16S539, D2S1338, D19S433, VWA, TPOX, D18S51, D5S818, FGA, D12S391, D1S1656, D2S441, D10S1248, D22S1045, and SE33), along with three gender determination loci (Amelogenin, Yindel, and DYS391). By focusing on unrelated male individuals, we aim to minimize bias from familial relatedness while capturing the population's genetic diversity. This study employs an ethnogenetic approach, integrating genetic, linguistic, and geographical data to provide a comprehensive understanding of the Kohistani population's origins and evolutionary history. Such insights not only enrich our knowledge of human diversity but also contribute to forensic science and the preservation of cultural heritage in this historically significant region.

2. MATERIALS AND METHODS:

2.1 Study Population and DNA extraction

This study aimed to characterize the genetic structure of the Kohistani population using 21 autosomal short tandem repeat (STR) loci amplified with the GlobalFiler™ PCR Amplification Kit. Autosomal STRs were selected for their high discriminatory power, compatibility with international forensic and

population databases, and proven utility in elucidating genetic diversity attributes established since their foundational use in human identification by Edwards et al. (1991) [16]. A total of 200 unrelated males were recruited from Upper and Lower Kohistan, districts within Khyber Pakhtunkhwa (KPK), Pakistan, geographically divided by the Indus River (Figure 1). To ensure ethnic homogeneity, participants were required to self-identify as Kohistani, confirm through a detailed questionnaire that both parents were of Kohistani descent, and demonstrate fluency in the Kohistani language, a key cultural identifier. Residency in either Upper or Lower Kohistan was verified with the assistance of local collaborators, including community elders. Individuals with known recent admixture (e.g., one non-Kohistani parent) were excluded. The exclusive inclusion of males was necessitated by logistical and cultural constraints in this remote region, where social norms and ethical considerations limited access to female participants. The sample size of 200 was determined based on statistical power calculations recommended for STR-based population studies [17, 18], offering a practical balance between representativeness and feasibility given Kohistan's estimated population of approximately 784,711. Selecting unrelated individuals minimized bias from familial clustering, and the cohort size aligns with those used in comparable studies of isolated populations [10, 11].

Sample collection was coordinated by the Research Registry of the Office of Research, Innovation, and Commercialization (ORIC) at Islamia College University, Peshawar, and carried out by a trained phlebotomy team to ensure standardized protocols and participant safety. De-identified samples, labeled only with sex and ethnic origin, were subsequently transferred to the research team to preserve confidentiality. The study was approved by the Bioethics Review Committee of Islamia College University (Approval No.: 866/ORIC/ICP, 23 April 2023) and conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013) and Pakistani national regulations. Written informed consent was obtained from all

participants. Genomic DNA was extracted from blood samples using the ReliaPrep™ Blood gDNA Miniprep System (Promega, Madison, USA) following the manufacturer's protocol. DNA sample concentrations were measured using a NanoDrop spectrophotometer (Thermo Scientific, Wilmington DE, USA). The final DNA concentration was adjusted to 1–2 ng/μl.

2.2 PCR amplification, and genotyping:

The extracted DNA samples were subjected to PCR amplification targeting 24 genetic markers, including 21 autosomal STR loci and three gender determination loci (Amelogenin, Yindel, and DYS391). Amplification was performed using the commercial kit on a GeneAmp PCR System 9700 (Thermo Fisher Scientific company), following the manufacturer's instructions. The resulting PCR products were analyzed on an 8-capillary ABI 3500 DNA Genetic Analyzer with POP-4™ polymer (Life Technologies). Genotypes were assigned using GeneMapper Software version 1.5 (Life Technologies). Allele nomenclature followed the allelic ladder provided with the kit [19], adhering to the guidelines of the International Society for Forensic Genetics (ISFG) [20, 21].

2.3 Quality control:

Negative (autoclaved deionized H₂O) and positive (AmpFlSTR® Control DNA 9947A) controls were included during DNA extraction, quantification, PCR amplification, and capillary electrophoresis. No amplification products were observed in the negative controls, while positive controls yielded genotypes consistent with known profiles.

2.4 Data analysis

Allele frequencies and key forensic statistical parameters (e.g., gene diversity, polymorphic information content, power of discrimination, and power of exclusion) were calculated using STRAF [17]. Additionally, principal

component analysis (PCA) was performed using STRAF to visualize the genetic structure and relationships of the Kohistani population with other populations. Hardy-Weinberg equilibrium (HWE), linkage disequilibrium (LD), and observed heterozygosity (Ho) were estimated using Arlequin 3.5 [22]. Autosomal STR data from the 1000 Genomes Project dataset were extracted using HipSTR (Haplotype inference and phasing for Short Tandem Repeats) [23]. Autosomal STR genotype data for 26 worldwide populations from the 1000 Genomes Project Phase 3, previously generated using HipSTR [17], were obtained from publicly available datasets [24]. These data were used for comparative analyses, including pairwise F_{st} genetic distances, multi-dimensional scaling (MDS), principal coordinates analysis (PCoA), and STRUCTURE, to assess the genetic relationships of the Kohistani population with global populations. Pairwise F_{st} genetic distances between the Kohistani population and 26 other global populations from the 1000 Genomes Project, representing five continents [8], were calculated using Arlequin 3.5 [22]. Nei's genetic distances between the Kohistani population and 22 Asian populations along with 26 populations from 1000 genome project were assessed using the Phylip 3.695 program [25]. The Phylogenetic tree [26] was visualized using Mega7 software [27]. Multi-Dimensional Scaling (MDS) and Principal Coordinates Analysis (PCoA) were used to visualize genetic relationships among populations based on Nei's genetic distances. MDS is a statistical method that projects pairwise genetic distances into a low-dimensional space (typically two dimensions) to minimize distortion of the original distances, producing two-dimensional plots to illustrate genetic affinities. PCoA, a specific form of metric MDS, applies eigenvalue decomposition to the distance matrix to maximize the variance explained by principal coordinates, providing a complementary visualization that quantifies the proportion of variance captured by each axis. In this study, MDS was used to compare the Kohistani population with global (1000 Genomes Project) and regional (Asian and South Asian) populations, while PCoA was also applied to regional populations to confirm clustering patterns,

Forensic parameters

[illegible]

homozygosity, reflecting a smaller effective population size and stricter endogamy. Bangladeshis [32], with a combined PD of 0.99991 for four STR loci including TPOX, exhibit higher diversity, likely due to less isolation. These forensic parameters demonstrate the high informativeness of most loci, underscoring the kit's utility for individual differentiation and paternity testing within the Kohistani population.

Hardy-Weinberg equilibrium (HWE)

None of the 21 loci were initially in HWE, with an excess of homozygotes at TPOX ($p < 0.05$). After sequential Bonferroni correction [33] (adjusted p -value = $0.05 / 21 \approx 0.00238$), all loci conformed to HWE (Supplementary Table 2). The pre-correction deviations, particularly at TPOX, suggest endogamy or population substructure, common in isolated groups. Similar patterns occur in Baloch populations, where HWE deviations at TPOX persist pre-correction due to consanguinity. Pashtuns show fewer deviations, with most loci in HWE post-correction, indicating less inbreeding than Kohistanis. Punjabis typically conform to HWE without correction, reflecting random mating in larger populations. Indian populations [34] with endogamous communities also show HWE deviations, aligning with Kohistani patterns. The Kohistani population's strong deviations underscore the impact of geographic isolation and cultural practices limiting gene flow.

Linkage equilibrium (LE)

Exact tests revealed significant linkage disequilibrium (LD) in 80 pairwise locus comparisons ($p < 0.05$). After Bonferroni correction [33] (adjusted p -value = $0.05 / 210 \approx 0.000238$), 16 pairs remained in LD (D18S51/TPOX, D1S1656/D13S317, D1S1656/D8S1179, FGA/D8S1179, D1S1656/FGA, FGA/D2S441, D13S317/TPOX, D10S1248/D19S433, D2S441/vWA, D1S1656/D21S11, D8S1179/vWA, D21S11/vWA, FGA/D21S11, D2S441/D8S1179, D2S441/D21S11 and D21S11/D8S1179). As many of these loci are located on different chromosomes, these significant results are

unlikely to reflect true linkage disequilibrium due to physical linkage. Instead, they likely arise from demographic and cultural factors, such as genetic drift, endogamy, and a small effective population size, which are prevalent in the isolated Kohistani population. This is consistent with the observed excess of homozygotes at loci like TPOX (Section 3, Hardy-Weinberg Equilibrium) and aligns with patterns seen in other endogamous South Asian populations, such as the Baloch, where significant LD across unlinked loci has been attributed to population substructure and inbreeding [25]. Compared to Pashtuns, who show fewer significant pairs (e.g., 10 pre-correction), the Kohistani population's higher number of significant associations reflects stronger genetic drift due to its smaller population size and stricter endogamy. Punjabis, with minimal significant LD, indicate random mating in a larger, less isolated population. These findings highlight the impact of Kohistani's geographic isolation and cultural practices, such as endogamy, on its genetic structure, mimicking LD patterns typically associated with linked loci [28, 29]. [35].

Kohistani population Ancestry content analysis with STRUCTURE

STRUCTURE analysis, incorporating data from the 1000 Genomes Project, provided insights into the ancestral composition of the Kohistani population alongside 26 global populations. The optimal number of ancestral populations was determined to be two ($K = 2$), with a high mean similarity score of 0.985 (Figure 2). At this level, two primary ancestry components (blue and orange) were identified. The Kohistani population showed a predominance of the blue component (95%), slightly less than in neighboring South Asian populations (99%), suggesting a strong genetic affinity within South Asia while also indicating a distinct genetic profile for the Kohistani. Further analysis at $K = 4$ (mean similarity score 0.972) revealed four ancestry components (green, blue, purple, and orange), with the Kohistani population almost exclusively characterized by the green component (99.5%), a trait shared only with other South Asian populations. This distinct genetic signature highlights the unique

genetic landscape of the Kohistani people, potentially reflecting historical migrations, genetic drift, or founder effects. These findings contribute to understanding genetic diversity within South Asia and emphasize the importance of studying underrepresented populations. The specific genetic components dominant in the Kohistani population suggest avenues for further research into the historical and evolutionary dynamics shaping South Asian genetic structure.

Kohistani and 1000 genome population comparison:

Analysis of F_{st} genetic distances between the Kohistani population and 26 worldwide populations from the 1000 Genomes Project revealed the closest association with the Bangladeshi population in Bangladesh (0.0085), followed by the Punjabi population from Lahore, Pakistan (0.0094). The most distant associations were with Peruvians from Lima, Peru (0.0233) and the Esan population from Nigeria (0.0178) (Supplementary Table 4). A multi-dimensional scaling (MDS) plot based on Nei's genetic distances placed the Kohistani population alongside other South Asian populations, including Punjabi from Lahore, Indian Gujarati from the USA, and Bengali from Bangladesh, indicating close genetic relationships (Figure 3). The Kohistani population appeared genetically distinct from European and African population clusters, suggesting relative genetic isolation. A neighbor-joining tree, based on F_{st} genetic distances, similarly positioned the Kohistani population as genetically distinct yet closely related to other South Asian populations (Figure 4). Principal component analysis (PCA), with PC1 explaining 49.3% and PC2 explaining 13.8% of the variance (Figure 5), also showed the Kohistani population clustering with South Asian populations, differentiating them from African, East Asian, and, to a lesser extent, European groups, highlighting regional genetic homogeneity. The interactivity test, based on allele frequency profiles of the 21 autosomal STR loci, was conducted to assess genetic interactions between the Kohistani population and the 26 worldwide populations from the 1000 Genomes Project

(Supplementary Figure 1). The test revealed significant genetic interactions ($p < 0.05$ after Bonferroni correction) between the Kohistani population and South Asian populations, particularly the Bangladeshi ($p = 0.0012$) and Punjabi from Lahore ($p = 0.0028$), indicating strong genetic affinity consistent with shared ancestry and regional gene flow. In contrast, weaker interactions were observed with European (e.g., CEU, $p = 0.124$), African (e.g., Esan, $p = 0.098$), East Asian (e.g., CHB, $p = 0.142$), and South American (e.g., Peruvian, $p = 0.115$) populations, reflecting greater genetic differentiation due to geographic and historical barriers to gene flow. These results, visualized as a heatmap in Supplementary Figure 1, complement the F_{st} genetic distances, MDS, and PCoA analyses (Figures 3–5), reinforcing the Kohistani population's close clustering with South Asian populations and its distinctiveness from other global groups. The strong interactions with South Asian populations align with linguistic and geographic ties, such as the Dardic language heritage and proximity to neighboring groups like Pashtuns and Punjabis.

Kohistani Population Comparison with Asian & South Asian Populations:

The multi-dimensional scaling (MDS) plot (Figure 6) and neighbor-joining (NJ) tree (Figure 7) consistently position Kohistani within a South Asian genetic continuum, showing strong affinity with South Asian populations like Punjabi_Pak, Saraki, Sindhi_Pak, Interior_of_Sindh, and Bangladeshi. The smallest Nei's genetic distance is with the Indian population (0.0716), followed by the Tadjik population (0.0724). Kohistani's proximity to Pashtuns and Bangladeshi reflects shared geographic and linguistic ties, such as their Dardic heritage and location in northern Pakistan. In contrast, greater distances from Central Asian (e.g., Uzbek: 0.111, Kazakh: 0.131) and East Asian (e.g., Uygur: 0.127, Malaysian: 0.127) populations highlight historical barriers to gene flow, such as geographic isolation and cultural differences. The Principal Coordinates Analysis (PCoA) (Figure 8) further supports this,

with PC1 (64.08%) separating South Asian populations from Central and East Asian groups, and PC2 (22.59%) showing Kohistani's tight clustering with South Asian groups like Pashtuns and Bangladeshi.

Addressing the ethnic diversity within the South Asian cluster, the Nei's genetic distance matrix reveals that Kohistani's smallest genetic distance among South Asian subgroups is with Sikh_Punjabi_Ind (0.0706), followed closely by India-Hindu (0.0716) and Indian_General_Pop (0.0732). These distances are smaller than those with other subgroups, such as Bangladeshi-1 (0.0726), Bangladeshi-2 (0.0836), Punjabi_Pak (0.1701), Saraki_Pak (0.1703), Interior Sindhi_Pak (0.1805), and Sindhi_Pak (0.2083). Within Indian subgroups, Kohistani shows greater distance from Muslim_Gujarat_Ind (0.1001), Shia_Muslim_AP_Ind (0.1048), and Brahmin_Punjab_Ind (0.0971). The MDS plot (Figure 6) visually confirms this, placing Kohistani closest to Sikh_Punjabi_Ind, India-Hindu, and Indian_General_Pop, while Punjabi_Pak, Saraki_Pak, and Sindhi_Pak are farther away. The NJ tree (Figure 7) supports this, with Kohistani branching most closely with Sikh_Punjabi_Ind and India-Hindu, alongside Bangladeshi populations, within a tight South Asian cluster. The small branch lengths between Kohistani and Sikh_Punjabi_Ind (0.0706) suggest recent gene flow or shared ancestry, likely influenced by geographic proximity in northern Pakistan and shared cultural practices. Larger distances to groups like Sindhi_Pak (0.2083) and Muslim_Gujarat_Ind (0.1001) reflect greater genetic differentiation due to historical endogamy or isolation.

Among Bangladeshi and Bengali subgroups, Kohistani's distance to Bangali (0.0726) is smaller than to Bangladeshi-2 (0.0836), indicating a closer affinity to the former, though both are farther than Sikh_Punjabi_Ind and India-Hindu. Among Punjabis, the distance to Sikh_Punjabi_Ind (0.0706) is much smaller than to Punjabi_Pak (0.1701) or Brahmin_Punjab_Ind (0.0971), pinpointing the Sikh Punjabi subgroup as the closest match. For Indian populations, India-Hindu (0.0716) and Indian_General_Pop (0.0732) are closer than

Muslim_Gujarat_Ind (0.1001) or Shia_Muslim_AP_Ind (0.1048), highlighting Kohistani's specific affinity with Hindu and general Indian populations over Muslim subgroups.

In conclusion, Kohistani's closest genetic relationships within the South Asian cluster are with the Sikh_Punjabi_Ind subgroup (0.0706), followed by India-Hindu (0.0716) and Indian_General_Pop (0.0732). These findings refine the analysis by specifying that Kohistani's strongest affinity is with Sikh Punjabis from India, likely due to shared ancestry and geographic proximity, rather than with Pakistani Punjabis (Punjabi_Pak: 0.1701) or other subgroups like Sindhi_Pak (0.2083) or Muslim_Gujarat_Ind (0.1001). Among Bangladeshi populations, Kohistani is closer to Bangladeshi-1 (0.0726) than Bangladeshi-2 (0.0836), but these are still more distant than the Sikh Punjabi and Hindu Indian subgroups. This precision accounts for ethnic diversity within South Asian populations and identifies Kohistani's closest genetic ties within the region.

CONCLUDING REMARKS:

This study provides a comprehensive analysis of the genetic structure of the Kohistani population from northern Pakistan using 21 autosomal STR loci, revealing insights into its forensic utility, genetic diversity, and population history. The high forensic efficiency of the STR panel, with a combined power of discrimination (CPD) of 0.999 999 999 999 999 999 999 999 999 999 378 and a combined power of exclusion (CPE) of 0.9999336, underscores its effectiveness for individual identification and kinship analysis in the Kohistani population. These values are comparable to those reported for other South Asian populations using the GlobalFiler™ kit, such as Pashtuns from Khyber Pakhtunkhwa (CPD: >0.999999, CPE: 0.9998–0.9999) [29] and Punjabis from Lahore (CPD: >0.999999, CPE: 0.9997–0.9999) [30]. However, the Kohistani population exhibits lower genetic diversity at certain loci, particularly TPOX (GD: 0.7389; Ho: 0.545; PD: 0.8852), compared to Pashtuns (GD: 0.73–0.92; Ho: 0.55–0.85) [29], Punjabis (GD: 0.75–0.95; Ho: 0.6–0.9) [30], and

Bangladeshis (CPD: 0.99991 for four STR loci including TPOX) [32]. This reduced diversity is likely attributable to the Kohistani population's small effective population size and strict endogamous practices, which amplify genetic drift and limit allelic variation, a pattern also observed in the Baloch population (GD: 0.74–0.90; Ho: 0.58–0.82) [31].

Initial deviations from Hardy-Weinberg equilibrium (HWE) across all loci, particularly an excess of homozygotes at TPOX ($p < 0.05$), were resolved after sequential Bonferroni correction, suggesting that these deviations stem from population substructure and inbreeding due to endogamy. Similar HWE deviations have been reported in other endogamous South Asian populations, such as certain Indian subgroups (e.g., Chandra et al., 2021 [34]), where consanguinity leads to increased homozygosity. In contrast, Pashtuns show fewer HWE deviations post-correction [29], and Punjabis often conform to HWE without correction [30], reflecting random mating in larger, less isolated populations. These findings highlight the Kohistani population's genetic distinctiveness, shaped by its geographic isolation in the mountainous Kohistan region and cultural practices limiting gene flow.

The significant non-random associations observed in 16 pairwise locus comparisons after Bonferroni correction, particularly for loci on different chromosomes (e.g., D18S51/TPOX, D1S1656/D8S1179), are unlikely to reflect true linkage disequilibrium due to physical linkage. Instead, these patterns likely result from demographic and cultural factors, such as genetic drift, endogamy, and a small effective population size, as seen in other isolated South Asian groups like the Baloch [31]. Compared to Pashtuns, who exhibit fewer significant LD pairs (10 pre-correction) [29], and Punjabis, who show minimal LD [30], the Kohistani population's elevated non-random associations underscore the stronger impact of isolation and inbreeding. These results align with studies of endogamous Indian populations, where increased LD is attributed to population substructure [34, 35], emphasizing the need to interpret such patterns cautiously in small, isolated populations.

Genetic distance analyses (Fst, Nei's distances), MDS, PCoA, and STRUCTURE revealed that the Kohistani population clusters closely with South Asian populations, particularly Bangladeshi (Fst: 0.0085), Punjabi from Lahore (Fst: 0.0094), and Pashtuns (Nei's distance: 0.0732), reflecting shared ancestry and regional gene flow. The STRUCTURE analysis at K=4 showed a nearly exclusive green ancestry component (99.5%) in Kohistani, shared with other South Asian populations, suggesting a distinct genetic signature shaped by historical migrations and founder effects. These findings are consistent with Khan et al. (2019) [35], who reported genetic links between Kohistani and West Eurasian populations, potentially via Bronze Age migrations along the Silk Road, but our study emphasizes stronger regional continuity with South Asian groups. The Kohistani population's greater genetic distance from Central Asian (e.g., Uzbek: 0.111) and East Asian (e.g., Uyghur: 0.127) populations highlights historical barriers to gene flow, such as the rugged terrain of the Himalayas and Hindukush.

The interactivity test (Supplementary Figure 1) further confirmed significant genetic interactions between Kohistani and South Asian populations (e.g., Bangladeshi, $p = 0.0012$; Punjabi, $p = 0.0028$), with weaker interactions with European, African, East Asian, and South American populations, reinforcing regional genetic homogeneity. These results align with the MDS and PCoA plots (Figures 3, 6, 8), which place Kohistani within a tight South Asian cluster, particularly close to Sikh Punjabi from India (Nei's distance: 0.0706) and Hindu Indian populations (0.0716). Compared to published data, the Kohistani population's genetic affinity with Sikh Punjabis and Hindu Indians is closer than with Pakistani Punjabis (0.1701) or Sindhi (0.2083) [34], likely due to shared Dardic linguistic heritage and geographic proximity in northern Pakistan.

This study contributes to forensic genetics by demonstrating the high discriminatory power of the GlobalFiler™ STR panel in the Kohistani population, supporting its use in forensic applications such as paternity

testing and individual identification in this understudied region. From a population genetics perspective, the findings elucidate the Kohistani population's unique genetic structure, shaped by endogamy and isolation, while confirming its place within the South Asian genetic continuum. The reduced diversity at loci like TPOX, combined with patterns of HWE deviations and non-random associations, mirrors findings in other endogamous groups (e.g., Baloch [31], Indian subgroups [34]), underscoring the impact of cultural practices on genetic variation. These findings are consistent with broader genomic studies of South Asia, which demonstrate a history of admixture between Ancestral South Indian (ASI) and Ancestral North Indian (ANI) populations, with Kohistani likely reflecting a predominance of ANI ancestry due to its northern geographic position [6, 7].

Future studies should incorporate mitochondrial DNA and Y-chromosomal markers to explore maternal and paternal lineages, providing a more comprehensive view of Kohistani's population history. Expanding the dataset to include other isolated South Asian populations, such as the Kalash or Burusho, could further clarify the role of endogamy and isolation in shaping regional genetic diversity. Additionally, integrating genomic data (e.g., whole-genome sequencing) with anthropological and linguistic data will enhance our understanding of Kohistani's historical connections to Silk Road migrations and its place within South Asia's complex genetic landscape.

In conclusion, this study highlights the Kohistani population's genetic distinctiveness within the South Asian context, driven by endogamy and geographic isolation, while affirming its close genetic ties to neighboring populations like Bangladeshi, Punjabi, and Pashtun. These findings enrich our understanding of South Asian genetic diversity, provide valuable forensic data, and underscore the importance of studying underrepresented populations to unravel the intricate patterns of human genetic variation.

DECLARATIONS

Ethics Approval and Consent to Participate

This study was reviewed and approved by the Ethics Committee of Islamia College Peshawar, Khyber Pakhtunkhwa, Pakistan (Approval No. 866/ORIC/ICP, dated April 3, 2023). Written informed consent was obtained from all participants prior to their inclusion in the study. All experimental procedures were conducted in accordance with the ethical principles outlined in the Declaration of Helsinki (1964), as amended.

Consent for Publication

All participants provided informed consent for the publication of aggregated and anonymized genetic data derived from their samples, with all direct identifiers removed to protect privacy.

Availability of Data and Materials

The full genotype dataset is not publicly shared due to the potential risk of re-identification inherent in forensic-grade genetic markers. However, anonymized data may be made available upon submission of a scientifically justified request to the authors (AA: mirzaatifadnan@gmail.com or MI: milyas@icp.edu.pk), subject to review and approval by the relevant institutional ethics and data governance authorities.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Department of Science and Technology of Liaoning Province, China (grant no. 2024021179-JH2/1019).

Author Contributions:

A. Adnan and C.C. Wang conceived the study. M. Ilyas, M.F. Ullah, and S. Nazir collected the samples. A. Adnan and S. Nazir performed the experiments. A. Adnan, A. Rakha, H.D. Chen, H. Wang, and C.C. Wang analyzed the data. A. Adnan drafted the manuscript. A. Adnan and C.C. Wang revised the manuscript. All authors reviewed and approved the final manuscript.

Acknowledgements

We sincerely thank all the participants from the Kohistani community for their generous contribution, trust, and engagement in this research. We are especially grateful to the community elders and local collaborators in Upper and Lower Kohistan for their guidance, support, and invaluable assistance in facilitating culturally respectful and ethically sound fieldwork.

This work is dedicated to my newborn nephew, **Muhammad Adan Ahmad**, whose arrival has brought immense joy and inspiration to our family.

Legends

Table 1: Forensic efficiency and statistical parameters on 21 autosomal STR loci in the Kohistani population from KPK, Pakistan (n=200)

Figure 1: Location of the sampling collection area on map.

Figure 2: Structure analysis results between Kohistani population and 1000 genome populations (K=2 to K=15).

Figure 3: Two-dimensional plot from multi-dimensional scaling analysis of Nei's values based on 21 autosomal STRs between Kohistani population, 1000 genome populations and European populations from STRidER.

Figure 4: NJ Tree between Kohistani population and 1000 genome populations based on Nei's formula.

Figure 5: Principal component analysis (PCA) between Kohistani population and 1000 genome populations.

Figure 6: Two-dimensional plot from multi-dimensional scaling analysis of Nei's values based on 15 overlapping autosomal STRs between Kohistani population, and 22 regional populations.

Figure 7: NJ Tree between Kohistani population, and 22 regional populations based on Nei's formula.

Figure 8: Principal component analysis (PCA) between Kohistani population, and 22 regional populations.

Electronic Supplementary materials (ESM)

Supplementary Figure 1: Interactivity test results showing genetic relationships between the Kohistani population (KHS) and reference populations from the 1000 Genomes Project. Lower values represent stronger genetic interactions, indicating closer affinity.

Supplementary Table 1: The distribution of allele frequencies on 21 STRs loci in Kohistani population from KPK, Pakistan

Supplementary Table 2: Sequential Bonferroni corrections for p values from HWE exact tests on 21STRs loci in Kohistani population from KPK, Pakistan.

Supplementary Table 3: Sequential Bonferroni corrections for p values from exact tests for LE on 21 STRs loci in Kohistani population from KPK Pakistan.

Supplementary Table 4: Pairwise Fst genetic distances between Kohistani population from Pakistan and 26 worldwide reference populations from 1000 genome project phase 3

REFERENCES:

1. Hallberg DG. The languages of Indus Kohistan. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics; 1992.

2. Baart J. Sustainable development and the maintenance of Pakistan's indigenous languages. In: Conference on the State of the Social Sciences and Humanities: Current Scenario and Emerging Trends. Islamabad: HEC & UNESCO, Islamabad, Pakistan; 2004. p. 202-13.
3. Dani AH. History of Northern Areas of Pakistan: upto 2000 A.D. Rev. 1st ed. [i.e. 3rd ed.]. Lahore: Sang-e-Meel Publications; 2001.
4. Bennett LA, editor. Encyclopedia of world cultures. 4: Europe: Central, Western, and Southeastern Europe / Linda A. Bennett, vol. ed. 1. publ. Boston, Mass: Hall; 1992.
5. Ali N, Ilyas M, Ahmad H, Khan S, Shafique W, Rehman G ur, et al. Extraction and Amplification of mtDNA HVR1 from Old and Degraded Samples of Human Skeletal Collected from Khyber Pakhtunkhwa province of Pakistan. *Adv life sci.* 2021;8:246-50.
6. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh P-R, Govindaraj P, et al. Genetic Evidence for Recent Population Mixture in India. *The American Journal of Human Genetics.* 2013;93:422-38.
7. Reich D, Thangaraj K, Patterson N, Price AL, Singh L. Reconstructing Indian population history. *Nature.* 2009;461:489-94.
8. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015;526:68-74.
9. Hammond HA, Jin L, Zhong Y, Caskey CT, Chakraborty R. Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet.* 1994;55:175-89.
10. Sánchez-Diz P, Acosta MA, Fonseca D, Fernández M, Gómez Y, Jay M, et al. Population data on 15 autosomal STRs in a sample from Colombia. *Forensic Sci Int Genet.* 2009;3:e81-82.
11. Adnan A, Rakha A, Noor A, van Oven M, Ralf A, Kayser M. Population data of 17 Y-STRs (Yfiler) from Punjabis and Kashmiris of Pakistan. *International Journal of Legal Medicine.* 2017. <https://doi.org/10.1007/s00414-017-1611-9>.
12. Adnan A, Rakha A, Lao O, Kayser M. Mutation analysis at 17 Y-STR loci (Yfiler) in father-son pairs of male pedigrees from Pakistan. *Forensic Science International: Genetics.* 2018. <https://doi.org/10.1016/j.fsigen.2018.07.001>.

13. Adnan A, Zhan X, Kasim K, Rakha A, Xin XJ. Population data and phylogenetic structure of Han population from Jiangsu province of China on GlobalFiler STR loci. *International Journal of Legal Medicine*. 2018. <https://doi.org/10.1007/s00414-018-1815-7>.
14. Butler JM. Short tandem repeat typing technologies used in human identity testing. *BioTechniques*. 2007;43:ii-v.
15. Jobling MA, Gill P. Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics*. 2004;5:739-51.
16. Edwards A, Civitello A, Hammond HA, Caskey CT. DNA typing and genetic mapping with trimeric and tetrameric tandem repeats. *Am J Hum Genet*. 1991;49:746-56.
17. Gouy A, Zieger M. STRAF-A convenient online tool for STR data evaluation in forensic genetics. *Forensic Sci Int Genet*. 2017;30:148-51.
18. Butler JM. *Advanced topics in forensic DNA typing: methodology*. San Diego, CA: Elsevier Academic Press; 2012.
19. Ludeman MJ, Zhong C, Mulero JJ, Lagacé RE, Hennessy LK, Short ML, et al. Developmental validation of GlobalFiler™ PCR amplification kit: a 6-dye multiplex assay designed for amplification of casework samples. *Int J Legal Med*. 2018;132:1555-73.
20. Bodner M, Bastisch I, Butler JM, Fimmers R, Gill P, Gusmão L, et al. Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER). *Forensic Science International: Genetics*. 2016;24:97-102.
21. Tillmar AO, Kling D, Butler JM, Parson W, Prinz M, Schneider PM, et al. DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. *Forensic Science International: Genetics*. 2017;29:269-75.
22. Excoffier L, Lischer HEL. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*. 2010;10:564-7.
23. Willems T, Zielinski D, Yuan J, Gordon A, Gymrek M, Erlich Y. Genome-wide profiling of heritable and de novo STR variations. *Nat Methods*. 2017;14:590-2.

24. Willems T, Gymrek M, Highnam G, The 1000 Genomes Project Consortium, Mittelman D, Erlich Y. The landscape of human STR variation. *Genome Res.* 2014;24:1894–904.
25. Felsenstein J. PHYLIP (Phylogeny Inference Package) Version 3.69. 2009.
26. Saif R, Nadeem S, Khaliq A, Zia S, Iftekhhar A. Mathematical Understanding of Sequence Alignment and Phylogenetic Algorithms: A Comprehensive Review of Computation of Different Methods. *Advancements in Life Sciences.* 2022;9:401–11.
27. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution.* 2016;33:1870–4.
28. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics.* 2000;155:945–59.
29. Batool Z, Perveen R, Sheikh N, Khan AA, Ilyas M, Shahzad M, et al. Genetic analysis of 15 autosomal STRs in Yousafzai population of Pakistan. *Int J Legal Med.* 2018;132:1635–6.
30. Javed F, Shafique M, Rani N, Rubab A, Shahid AA. Allele frequency data of 15 autosomal STRs in Arain population of Pakistan. *Int J Legal Med.* 2022;136:557–8.
31. Khan AA, Perveen R, Sheikh N, Abbasi BHA, Batool Z, Shahzad M, et al. Genetic polymorphism of 15 autosomal short tandem repeats in Baloch population of Pakistan. *Int J Legal Med.* 2019;133:775–6.
32. Kido A, Susukida R, Oya M, Fujitani N, Kimura H, Hara M. Allele frequency distributions of four STR loci vWA, TH01, TPOX and F13A01 in three Asian populations (Japanese, Bangladeshis and Indonesians). *International Congress Series.* 2003;1239:105–8.
33. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society.* 1995;Vol. 57 No.1:289–300.
34. Chandra D, Mishra VC, Anthwal A, Gupta N, Bhardwaj AK, Raina A, et al. Genetic polymorphism of 21 STR markers in the representative sample of Indian population. *Gene Reports.* 2021;25:101411.
35. Khan S, Ahmad H, Khan MS, Mehmood SA, Nisar A, Feng X. Ethnogenetic analysis reveals that Kohistanis of Pakistan were genetically

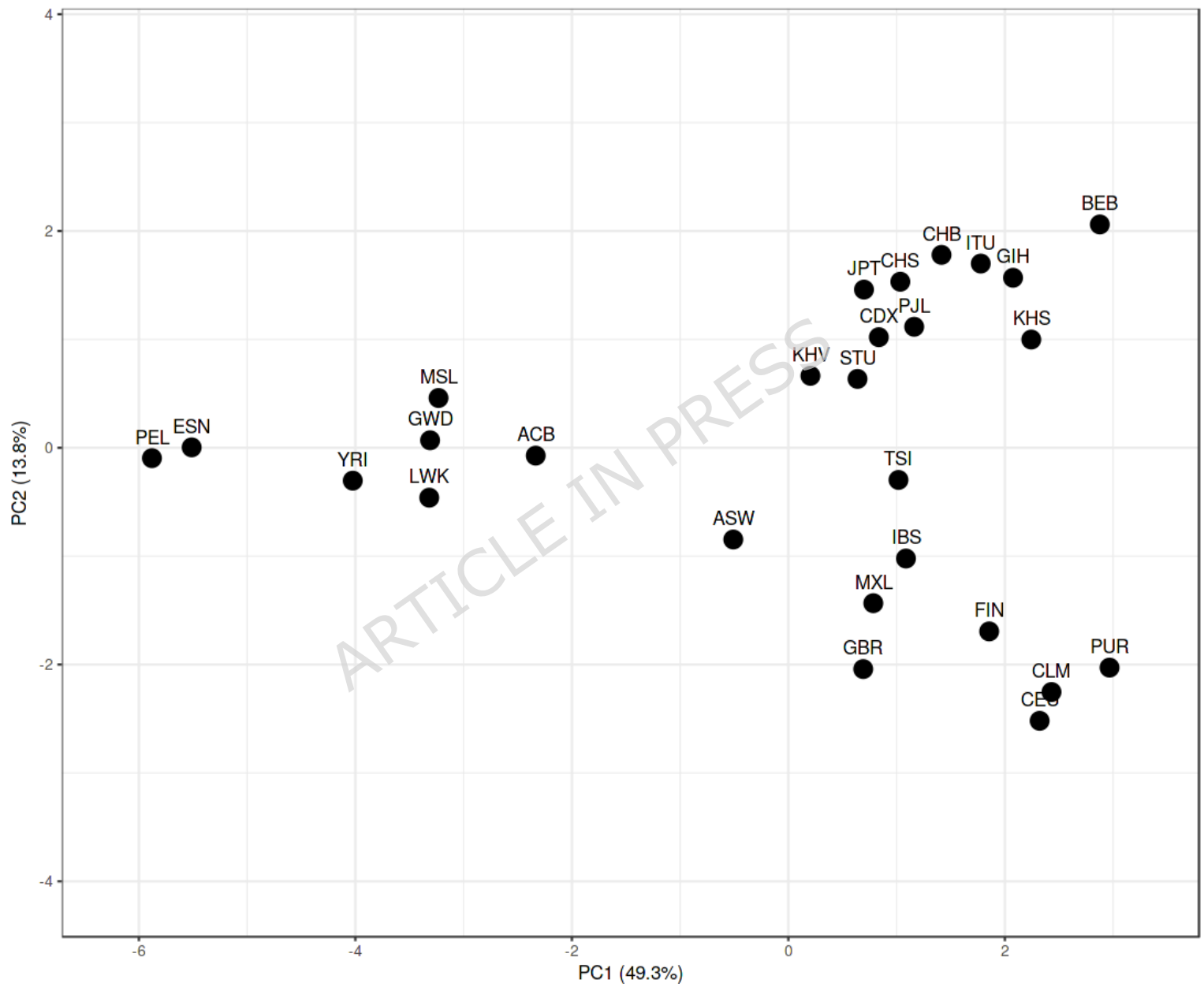
linked to west Eurasians by a probable ancestral genepool from Eurasian steppe in the bronze age. Mitochondrion. 2019;47:82–93.

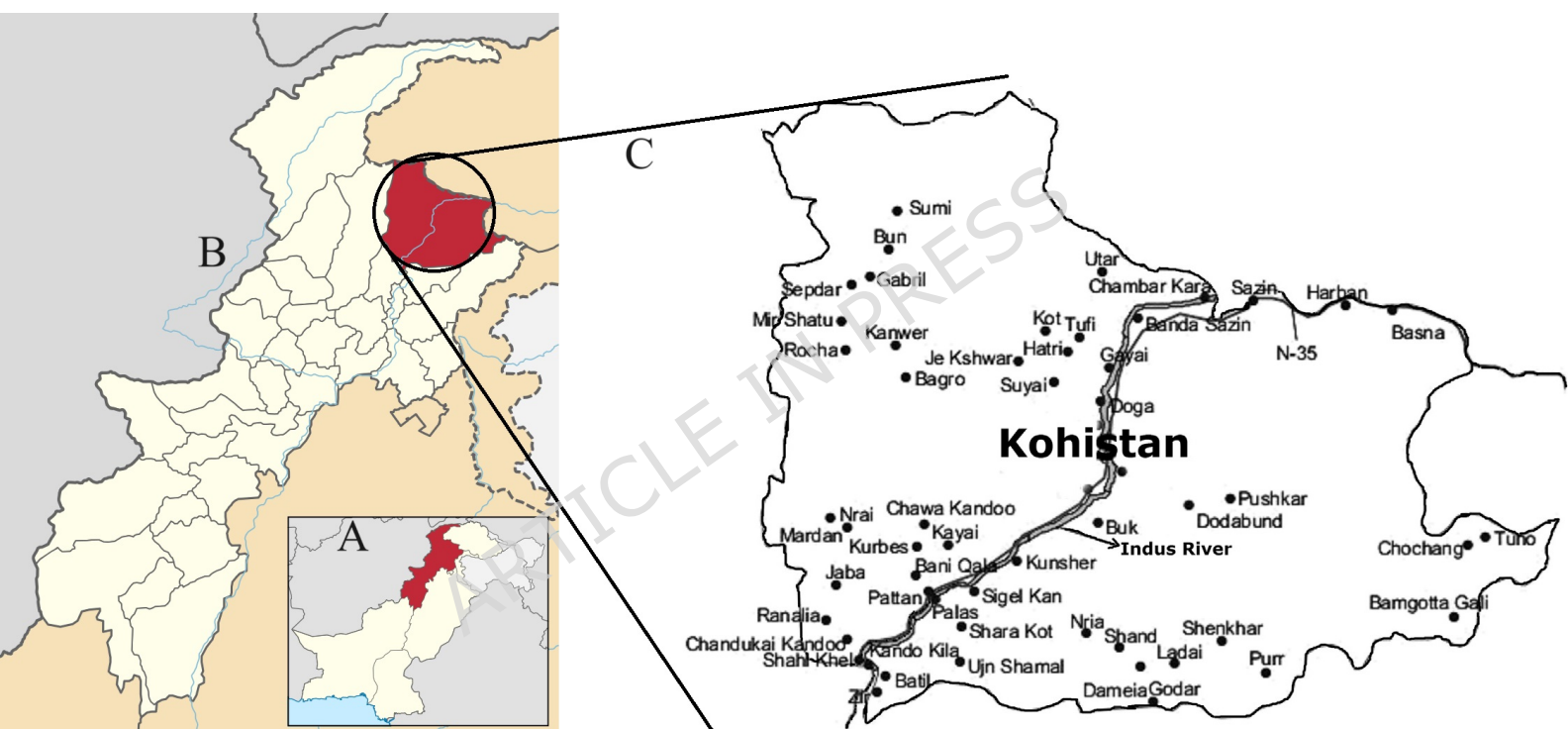
Table 1: Forensic efficiency and statistical parameters on 21 autosomal STR loci in the Kohistani population from KPK, Pakistan (n=200).

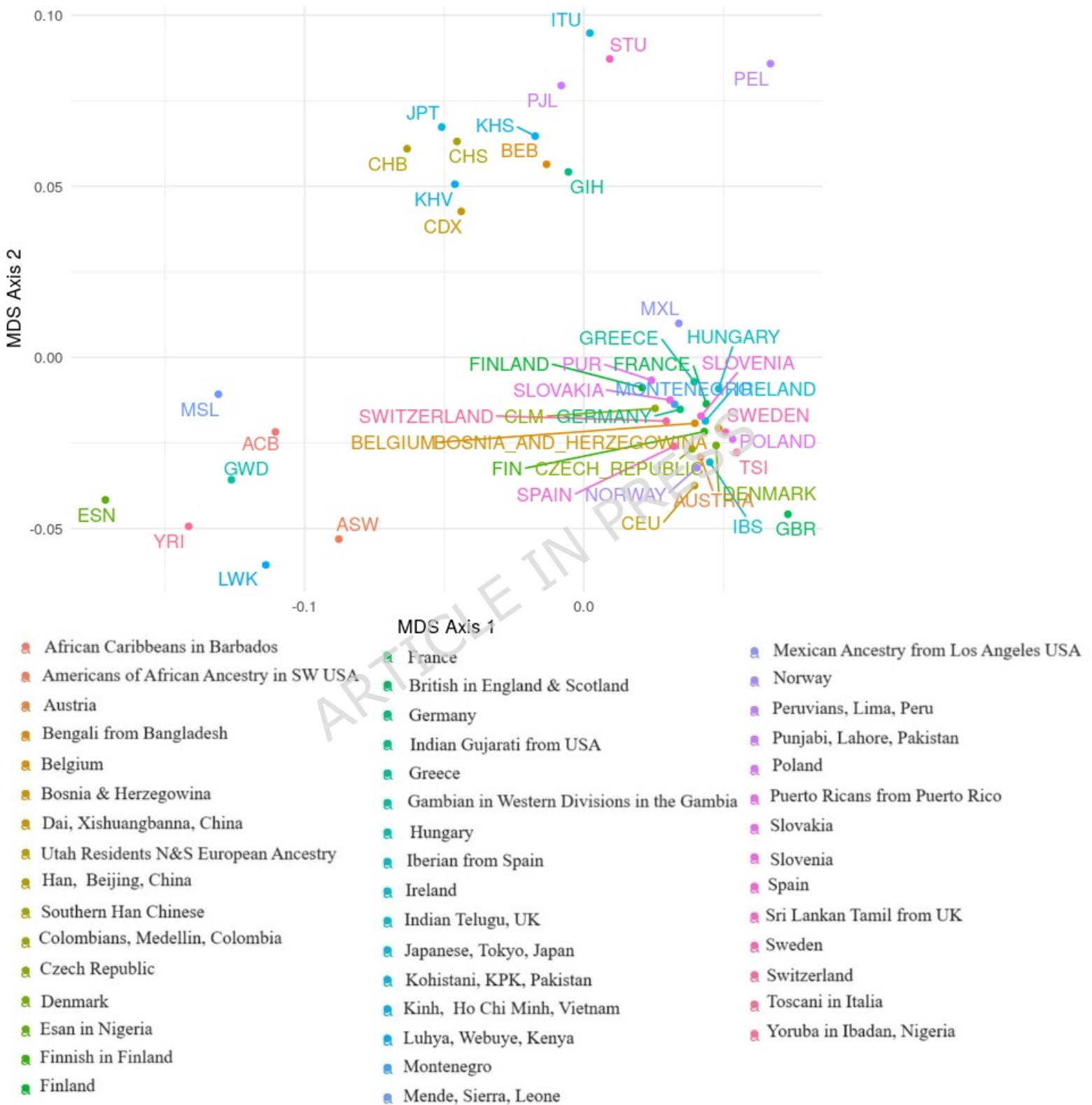
locus	Nall	GD	PIC	PM	PD	Hobs	PE	TPI
CSF1PO	9	0.7998 25	0.7679 96	0.0731	0.9269	0.645	0.3483 91	1.41
D10S12 48	12	0.8327 94	0.8087 88	0.0675 5	0.9324 5	0.9	0.7954 2	5
D12S39 1	12	0.8650 13	0.8477 48	0.0474	0.9526	0.955	0.9084 98	11.11
D13S31 7	10	0.8361 65	0.8122 69	0.0607	0.9393	0.865	0.7246 34	3.7
D16S53 9	9	0.8176 82	0.7917 77	0.0649 5	0.9350 5	0.69	0.4129 61	1.61
D18S51	15	0.8612 66	0.8439 98	0.0602 5	0.9397 5	0.92	0.8364 33	6.25

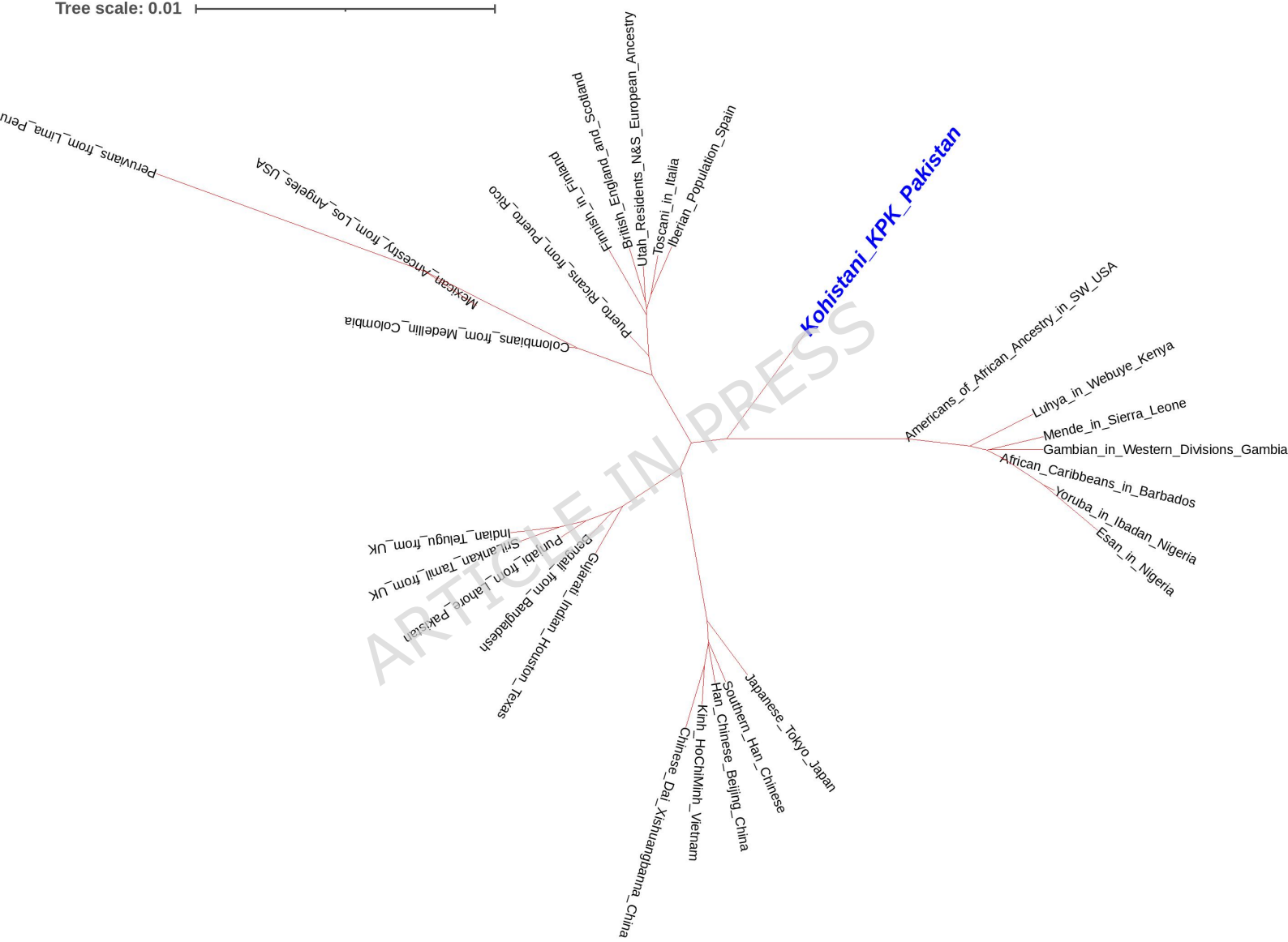
D19S43		0.8969	0.8860				0.8776	8.33
3	26	17	67	0.033	0.967	0.94	2	
D1S165		0.8758	0.8603	0.0408	0.9591		0.8570	7.14
6	13	9	22	5	5	0.93	17	
D21S11		0.8870	0.8740				0.4761	1.85
	21	55	22	0.0313	0.9687	0.73	81	
D22S10		0.8456	0.8264				0.9899	99.5
45	14	39	34	0.0695	0.9305	0.995	76	
D2S133		0.8892	0.8761	0.0371	0.9628		0.8467	6.67
8	14	23	62	5	5	0.925	21	
D2S441		0.8227	0.7968	0.0746	0.9253		0.7446	4
	12	57	31	5	5	0.875	9	
D3S135		0.7976	0.7657				0.3691	1.47
8	12	94	79	0.0724	0.9276	0.66	31	
D5S818		0.8131	0.7875	0.0822	0.9177		0.9290	14.29
	12	33	33	5	5	0.965	23	
D7S820		0.8363	0.8129				0.7648	4.35
	10	53	15	0.061	0.939	0.885	91	
D8S117		0.8764	0.8611				0.2619	1.18
9	16	29	91	0.039	0.961	0.575	48	
FGA		0.9106	0.9011	0.0258	0.9741		0.9799	49.5
	26	64	61	5	5	0.99	06	
SE33		0.9638	0.9600	0.0100	0.9899			100
	50	6	74	5	5	1	1	
TH01		0.8310	0.8068	0.0657	0.9342		0.6177	2.63
	10	78	5	5	5	0.81	3	
TPOX		0.7388	0.7000				0.2299	1.1
	12	97	04	0.1148	0.8852	0.545	99	
vWA		0.8374	0.8155				0.4928	1.92
	14	56	57	0.0495	0.9505	0.74	14	

Nall: Number of alleles; **GD:** gene diversity; **PIC:** polymorphism information content; **Hobs:** observed heterozygosity; **PD:** power of discrimination; **PM:** matching probability; **PE:** power of exclusion; **TPI:** typical paternity index;









K=2

K=3

K=4

K=5

K=6

K=7

K=8

K=9

K=10

K=11

K=12

K=13

K=14

K=15

Kohistani, KPK, Pakistan

Esan in Nigeria

Gambian in Western
Divisions in the GambiaLuhya in Webuye,
Kenya

Mende in Sierra Leone

Yoruba in Ibadan, Nigeria

African Caribbeans
in BarbadosAmericans of African
Ancestry in SW USAColombians from
Medellin, ColombiaMexican Ancestry from
Los Angeles USA

Peruvians from Lima, Peru

Puerto Ricans, Puerto Rico

Dai, Xishuangbanna, China

Han, Beijing, China

Southern Han Chinese

Japanese, Tokyo, Japan

Kinh, Ho Chi Minh, Vietnam

Utah Residents with N&S
European Ancestry

Finnish in Finland

British in England &
Scotland

Iberian in Spain

Toscani in Italia

Bengali from Bangladesh

Indian Gujarati, USA

Indian Telugu from UK

Punjabi, Lahore, Pakistan

Sri Lankan Tamil from UK

