Received: 21 February 2018 | Revised: 3 December 2018 | Accepted: 4 December 2018

DOI: 10.1002/ajpa.23789

WILEY American Journal of PHYSICAL ANTHROPOLOGY

RESEARCH ARTICLE

The genetic legacy of the Yaghnobis: A witness of an ancient Eurasian ancestry in the historically reshuffled central Asian gene pool

Elisabetta Cilli[1†] | Stefania Sarno[2†] | Guido Alberto Gnecchi Ruscone[2†] | Patrizia Serventi[1,2] | Sara De Fanti[2] | Paolo Delaini[3] | Paolo Ognibene[3] | Gian Pietro Basello[4] | Gloria Ravegnini[5] | Sabrina Angelini[5] | Gianmarco Ferri[6] | Davide Gentilini[7] | Anna Maria Di Blasio[7] | Susi Pelotti[8] | Davide Pettener[2] | Marco Sazzini[2] | Antonio Panaino[3] | Donata Luiselli[1‡] | Giorgio Gruppioni[1‡]

[1]Laboratories of Physical Anthropology and Ancient DNA, Department of Cultural Heritage, University of Bologna, Ravenna, Italy

[2]Laboratory of Molecular Anthropology and Centre for Genome Biology, Department of Biological Geological and Environmental Sciences, University of Bologna, Bologna, Italy

[3]Department of Cultural Heritage, University of Bologna, Ravenna, Italy

[4]Department of Asian, African and Mediterranean Studies, University of Naples "L'Orientale", Naples, Italy

[5]Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[6]Department of Diagnostic and Clinical Medicine and Public Health, University of Modena and Reggio Emilia, Modena, Italy

[7]Centre for Biomedical Research and Technologies, Italian Auxologic Institute, IRCCS, Milan, Italy

[8]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy

Correspondence
Elisabetta Cilli, Department of Cultural Heritage, University of Bologna, via degli Ariani n 1, Ravenna 48121, Italy.
Email: elisabetta.cilli@unibo.it
and
Patrizia Serventi, Department of Cultural Heritage, University of Bologna, via degli Ariani n 1, Ravenna 48121, Italy.
Email: p.serventi87@gmail.com

Funding information
Fondazione del Monte di Bologna e Ravenna, Grant/Award Number: FdM/1200; FP7 Ideas:

Objectives: The Yaghnobis are an ethno-linguistic minority historically settled along the Yaghnob River in the Upper-Zarafshan Valley in Tajikistan. They speak a language of Old Sogdian origin, which is the only present-day witness of the Lingua Franca used along the Silk Road in Late Antiquity. The aim of this study was to reconstruct the genetic history of this community in order to shed light on its isolation and genetic ancestry within the Euro-Asiatic context.

Materials and Methods: A total of 100 DNA samples were collected in the Yaghnob and Matcha Valleys during several expeditions and their mitochondrial, Y-chromosome and autosomal genome-wide variation were compared with that from a large set of modern and ancient Euro-Asiatic samples.

Results: Findings from uniparental markers highlighted the long-term isolation of the Yaghnobis. Mitochondrial DNA ancestry traced an ancient link with Middle Eastern populations, whereas Y-chromosome legacy showed more tight relationships with Central Asians. Admixture, outgroup-*f3*, and D-statistics computed on autosomal variation corroborated Y-chromosome evidence, pointing respectively to low Anatolian Neolithic and high Steppe ancestry proportions in Yaghnobis, and to their closer affinity with Tajiks than to Iranians.

Discussion: Although the Yaghnobis do not show evident signs of recent admixture, they could be considered a modern proxy for the source of gene flow for many Central Asian and Middle Eastern groups. Accordingly, they seem to retain a peculiar genomic ancestry probably ascribable to an ancient gene pool originally wide spread across a vast area and subsequently reshuffled by distinct demographic events occurred in Middle East and Central Asia.

KEYWORDS

genome-wide SNPs, mitochondrial DNA, Tajikistan, Yaghnobis, Y-chromosome

[†]These authors equally contributed to the article.

[‡]Denotes co-senior authorship.

European Research Council, Grant/Award Number: ERC-2011-AdG295733; Università di Bologna, Grant/Award Number: RFO

1 | INTRODUCTION

Central Asia has represented for millennia a crossroads of individuals, genes, languages, cultures and goods, being in a key point of relevant human migration routes (Baskakov, 1960, 1966; CLI, 1989; Johanson & Csató, 1998; Kerimova, 1966; Menges, 1968; Poppe, 1970; Rastorgueva, 1966). Its history was deeply marked by considerable movements of different populations through time, which took place since the post Last Glacial Maximum, reached remarkable intensity in the Neolithic and the Bronze Age and progressively increased until Antiquity and Middle Ages (Beckwith, 2009; Litvinskij, 1998). In fact, this area was interested by the main political phenomena of the Eastern Ancient and Mediaeval World as an important satrapy of the Persian Empire, and as a trade center along the Silk Road, in particular with Samarkand. Its involvement in this trade route occurred particularly during the Sasanian Kingdom and after the Islamic Invasion of Central Asia (Le Strange, 1905; Wiesehöfer, 1996), when Central Asia was also affected by a number of migratory waves involving Turkic tribes (Barthold, 1962; PhTF, 2000) and the massive expansion of the Mongol Empire (Barthold, 1928; Heissig, 1978; HM, 1989; Spuler, 1961, 1985). In the context of these intense population movements, most of Central Asian history remains complex and contentious (Frye, 1998).

Several genetics and genomics studies have already focused on Central Asia populations based on modern samples (Chaix et al., 2007; Comas et al., 1998, 2004; Dulik et al., 2011; Heyer et al., 2009; Irwin et al., 2010; Marchi et al., 2017; Martínez-Cruz et al., 2011; Palstra et al., 2015; Quintana-Murci et al., 2004; Ségurel et al., 2008; Wells et al., 2001). Moreover, some ancient samples have been analyzed as well (Allentoft et al., 2015; Lazaridis et al., 2016), but including specimens mainly recovered from the northern boundaries of this territory (e.g., Northern Kazakhstan) and focusing on particular time frames, such as the Bronze Age. In fact, during this period, populations of the Eurasian steppe were supposed to have mediated the diffusion of the so-called steppe-ancestry westward, thus appreciably contributing to the present-day European and South Asian (SA) gene pools. However, in spite of these studies and in the light of the results recorded, still unanswered questions remain and some Central Asian populations deserve to be further investigated.

Among these populations, Yaghnobis are certainly an unusual group, a tile in the mosaic of the intriguing and complex history of Central Asia, very interesting from both an ethno-linguistic and a genetic perspective. They were proposed to be relic depositary of traces of a peculiar and ancient gene pool (Cilli et al., 2011; Cilli, De Fanti, Delaini, Panaino, & Gruppioni, 2013) and of an archaic language, which is a living Eastern variety of the ancient Middle Iranian language (Sogdian) spoken in Samarkand and Bukhara, that represented for centuries important pre-Islamic cultural crossroads along the Silk Road (Panaino, Gariboldi, & Ognibene, 2013; Panaino, 2008). In particular, Yaghnobi is the only living language that directly continues a particular variety of Sogdian, probably used in the mountains of that area and different in some respects from the Sogdian spoken in the cities of which we have many written documents (Benveniste, 1979; Negmatov, 1996; Panaino, 2008; Panaino et al., 2013; Sims-Williams, 1989; Vaissière, 2002, 2011).

Knowledge about the existence of their language dates from the second half of the nineteenth century (Iskanderkul expedition in 1870), when Yaghnobi was discovered and then written for the first time (Kuhn, 1881a, 1881b). For this reason, for centuries, the history of the Yaghnobis relied only on oral traditions passed down through generations (Andreev & Peščereva, 1957). These traditions bring back their origins to the fascinating history of Sogdians, the main ancient caravan merchants of the Silk Road. The Yaghnob Valley is a very interesting area, studied in historical, ethnographical, and geographical perspectives (Basello, Guizzo, & Ognibene, 2008; Baumer, 2012; Golden, 2011; Gunya, 2002; Panaino et al., 2013; Rose, 2010; Whitlock, 2003) but, until nowadays, it has not been deeply investigated from the archeological point of view and no relevant evidences about past settlements are highlighted. In fact, there are only some dated works concerning a large-scale archeological survey (Jakubov, 1979, 1988), while there are some relevant reports only about the Hermitage's expedition in the adjacent Matcha Valley (Lur'e, Semenov, & Stepanov, 2017).

Today, the Yaghnobis represent an isolated ethno-linguistic population, historically settled between 2,200 and 2,600 m above sea level in the valley of the Yaghnob River, located in the Upper-Zarafshan area (North-West Tajikistan; Figure 1). The Upper Valley of the Yaghnob River is virtually inaccessible for several months of the year, due to considerable snowfall and to the complete absence of roads after the village of Nemetkon (Panaino, 2008). Geographical isolation helped the people of the Yaghnob Valley to preserve their language, culture, lifestyle, and, possibly, their gene pool for centuries. The Yaghnobis are Muslims even if they still preserve some peculiar traits that might link them to an early Iranian pre-Islamic past (Panaino, 2008).

In the remote past, the Yaghnobis were settled only along the eponymously named river but, in 1968–1969, most of them were relocated onto cotton plantations, in the plains surrounding Zafarobod, near to the Uzbekistan border. After 1978, some of them returned to the Yaghnob Valley, but later, during the Tajik Civil War (1992–1997), a more consistent part of the Yaghnobi population moved back to their homeland, where they still live today. The actual Yaghnobi population in the Upper Yaghnob Valley is estimated to be less than 500 individuals, which are settled in 21 villages and rely on an economy of subsistence funded on farming and agriculture (Panaino, 2008). Aside from the valley, considerable Yaghnobi communities are today located near Dushanbe and in Zafarobod, as a consequence of deportation, although Yaghnobis live also in Zumand as result of a voluntary migration in the second half of 19th century.

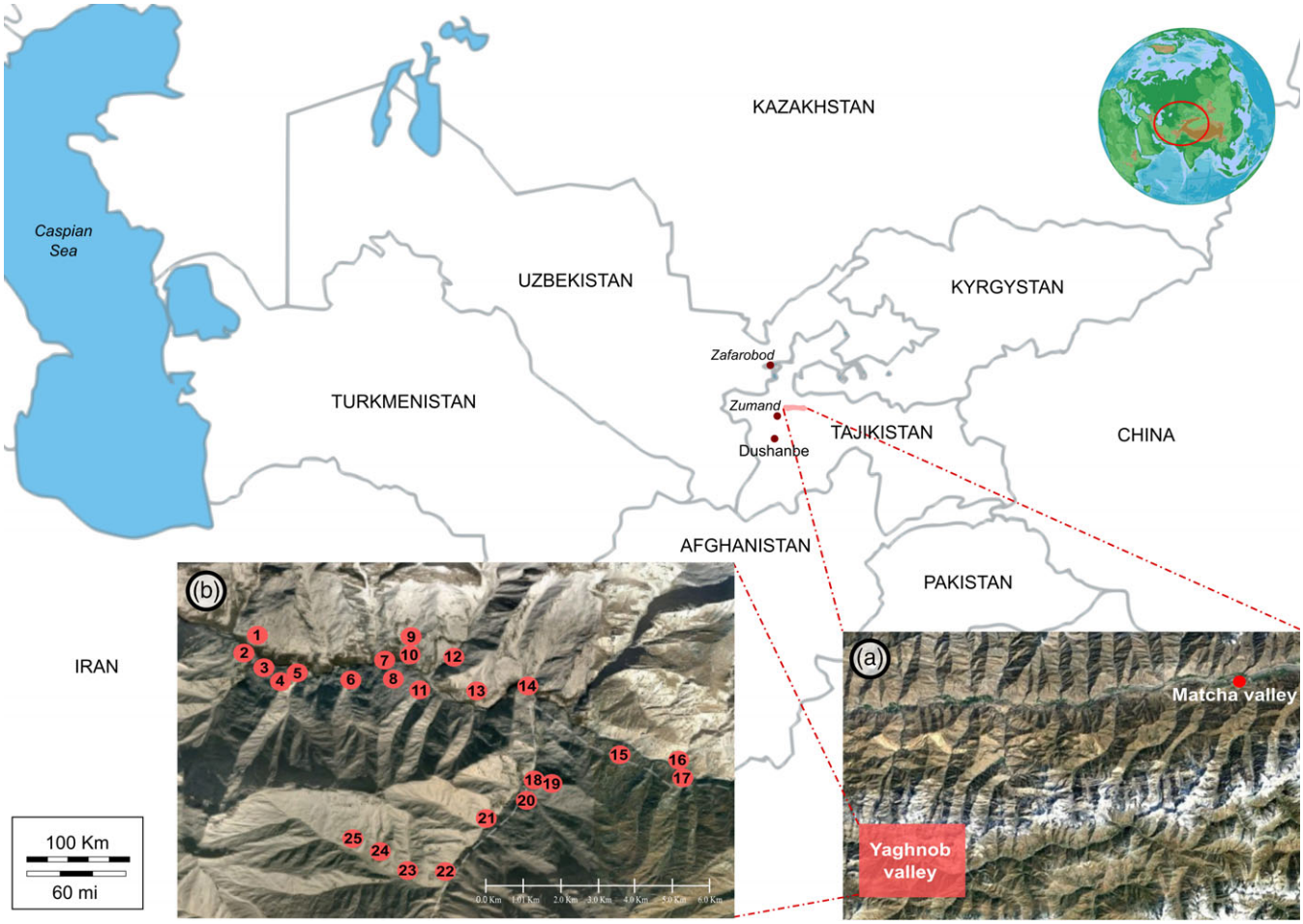

FIGURE 1 Map of Central Asia with (a) a detail of the locations of the Yaghnob and Matcha valleys and (b) the enlargement of the central part of the middle valley of Yaghnob River with the location of all the villages of the valley. Villages located on the Yaghnob Valley are represented by red circles: Bidev (1); Marghtimain (2); Dumzoi (3); upper Showeta (4); lower Showeta (5); upper Xisougudarv (6); Gelink (7); Nemetkon (8); Pullarovut (9); Kashe (10); Chokat (11); tag-i Chanor (12); Pitip (13); tang-i Dahana (14); Pskon (15); pad-i past (16); Dehibaland (17); lower Gharmen (18); upper Gharmen (19); Sokan (20); Simich (21); Dahana (22); Kul-i Mobayn (Bidun-i Kul) (23); lower Kul (Sar-i Kul) (24); upper Kul (Chentermen) (25). Some villages are no long inhabited, in fact they were not repopulated after the deportation of 1970s

To date, the genetic history of this ethnic group has not been fully investigated. Previous works on the Yaghnobis (Chaix et al., 2007; Cilli et al., 2011; Comas et al., 2004; Marchi et al., 2017; Martínez-Cruz et al., 2011; Ségurel et al., 2008; Wells et al., 2001; Yunusbayev et al., 2015) have only marginally described their genetic variability, mainly because of the limited number of samples or the low-resolution markers used. Moreover, in almost all these studies, the collected samples were from outside the Yaghnob Valley, in particular from multi-ethnic and cosmopolitan areas or from unspecified locations.

Contrary to what has been published so far, the present study makes use of a large spectrum of biological data collected by an ethno-linguistic and archeological mission of the University of Bologna, in the framework of a specific and larger multidisciplinary investigation (Panaino, 2008; Panaino et al., 2013). To shed light on their origins and to disentangle the genetic relationships between the Yaghnobis and other extant or ancient populations, a high percentage of the residual Yaghnobi population settled in the upper Yaghnob Valley (Figure 1) was sampled. Indeed, the relevance of this study lies in the large number of collected samples (N = 88), which constitutes approximately almost 20% of the whole population resident in the Yaghnob Valley, in addition 12 individuals from the adjacent Matcha Valley, not belonging to the Yaghnobi ethnic group, were collected. Characterization of their mitochondrial DNA (mtDNA), Y-chromosome and autosomal genome variation was used to set this ethno-linguistic isolate into the landscape of modern and ancient Euro-Asiatic genetic diversity, enabling us to infer the main demographic processes that have shaped the genetic history of the Yaghnobis.

2 | MATERIALS AND METHODS

2.1 | DNA sampling and extraction

Buccal swabs samples were obtained from 100 unrelated subjects residing in several villages (21) placed along the adjacent Yaghnob (N = 88) and Matcha (N = 12) Valleys of Tajikistan (Figure 1).

Appropriate written informed consents were obtained from all the individuals participating in the study. The bioethic committee of the University of Bologna released approval for the present study,

which was designed and conducted in accordance with relevant guidelines and regulations and according to the ethical principles for medical research involving human subjects stated by the WMA Declaration of Helsinki.

Genomic DNA was extracted from buccal swabs using a salting out protocol modified from Miller, Dykes, & Polesky (1988), and quantified with the Qubit 2.0 fluorometer (Invitrogen, Thermo Fisher Scientific Inc., Carlsbad, CA, USA). Details about samples, sampling strategy, and DNA extraction are available in Supporting Information.

2.2 | Mitochondrial DNA and Y-chromosome typing

A total of 79 samples were successfully sequenced for the first hypervariable segment (HVS-I) of the mtDNA control region and genotyped for 22 coding region single nucleotide polymorphisms (SNPs). MtDNA haplogroups were assigned using Haplogrep and Phylotree mtDNA phylogeny, Build 17 (van Oven & Kayser, 2009), based on both HVS-I sequence and on the analysis of the 22 coding region SNPs. The protocols and the list of markers used are presented in Supporting Information. HVS-I sequences will be made available upon request at the corresponding author or at the repository of the Molecular Anthropology Lab (http://www.bioanthropologybologna.eu).

All male samples (N = 49) were further genotyped for the 23 Y-chromosome short-tandem repeat (STR) loci implemented in the PowerPlex Y23 System (PPY23, Promega Corporation, Madison, WI), as described in the Supporting Information. Subsequently, three haplogroup-specific multiplexes (Ferri & Alù, 2012) followed by Minisequencing reaction were used to type 17 Y-chromosome SNPs according to the protocols described in Supporting Information.

2.3 | Statistical analyses on uniparental genetic markers

To set the Yaghnobi mtDNA and Y-chromosome variation into a wider genetic landscape, as well as to investigate potential sex-biased patterns of gene flow and admixture, we compared uniparental genetic profiles of the collected samples with those included in extended reference datasets extracted from the literature. Comparison population samples were selected to be representative of the following geographical areas: South-Eastern Europe, Middle East, Caucasus, Central Asia, South Asia and East Asia. Overall, ~7,600 mtDNA HVS-I sequences and ~9,700 Y-STRs profiles were retrieved from individuals belonging to 120 and 166 Eurasian populations, respectively (Supporting Information Tables S3 and S4).

Standard diversity indexes, as well as R_{st} and F_{st} pairwise genetic distances, were computed on STRs and HVS-I data using the Arlequin software v.3.5.1.2 (Excoffier & Lischer, 2010). Genetic relationships among population groups were explored through principal component analysis (PCA) based on haplotype data, and population samples resulting outliers in the PCA space were excluded from subsequent analyses. To further clarify spatial patterns of genetic variation, a spatial principal component analysis (sPCA) was also performed by integrating the genetic data with the information about the geographic distribution of examined populations by maximizing the product of variance and spatial autocorrelation. PCA and sPCA analyses were performed using the R *adegenet* package. Significance of Global and Local sPCA-detected structures (compared to random spatial distributions of genetic variability) were tested by means of a Monte Carlo procedure using 10,000 permutations, as implemented in the *global.rtest* and *local.rtest* functions of the *adegenet* package (Jombart, Devillard, Dufour, & Pontier, 2008; Jombart, 2008).

2.4 | Genome-wide SNPs genotyping and assembling of a reference dataset

Among the collected samples, those with DNA concentration higher than 30 ng/μL were selected for genome-wide SNPs genotyping. Thus 19 belonging to the Yaghnobi ethnic group were further genotyped for the 719,666 genome-wide SNPs implemented in the Illumina HumanOmniExpress-24 BeadChips v1.1 array (San Diego, CA) at the Laboratory of Molecular Biology of the Italian Auxologic Institute in Milan, on a HiScan SQ System (Illumina, San Diego, CA).

Quality control (QC) steps, along with computation of summary statistics, merging with reference datasets and analysis of runs of homozygosity, were performed using the PLINK software v.1.07 (Purcell et al., 2007). After the QC steps described in Supporting Information, the filtered dataset consisted of 15 Yaghnobi individuals successfully typed for 663,785 SNPs.

Such high quality data were then merged with several reference datasets overall made up of 43 Eurasian and 5 African modern populations for which comparable genome-wide information was available (Supporting Information Table S5), including also six Yaghnobi individuals previously typed by Yunusbayev et al. (2015) and here indicated as *Yaghnobi2* population. The same QC protocols described in Supporting Information were applied also to these reference datasets prior to merging procedures. The obtained dataset was finally merged also with publicly available genome-wide data generated for ancient samples from Europe, Central Asia and the Middle East (Allentoft et al., 2015; Lazaridis et al., 2016; Mathieson et al., 2015).

After QC, a final un-pruned dataset included 1,078 individuals typed for 260,409 SNPs. For most of the subsequent analyses we applying a Linkage Disequilibrium (LD) pruning procedure to exclude one SNP for each pair showing r^2 values higher than 0.2 within a 50-SNPs window, sliding five loci each time. Moreover, SNPs with a minor allele frequency lower than 0.01 were also removed, obtaining a pruned dataset of 105,798 SNPs, that was used to perform multivariate, admixture, ROH and population-based analyses. Genome-wide autosomal SNP-chip data generated during the current study are available upon request to the corresponding author or at the Molecular Anthropology Lab repository (http://www.bioanthropologybologna.eu).

2.5 | Inferences drawn from genome-wide data

PCA was performed on the pruned dataset including modern reference populations by using the *smartpca* function implemented in the EIGENSOFT package v6.0.1 (Patterson, Price, & Reich, 2006; Price et al., 2006). Ancient samples were subsequently projected on the obtained PCA space by means of the *lsqproject* = YES function. Admixture analyses were carried out using ADMIXTURE v1.21 to infer a maximum likelihood estimate of individual ancestry (Alexander

Novembre, & Lange, 2009). ADMIXTURE was run by testing two to nine hypothetical ancestral populations (*K*). Fifty replicates with a different random seed were run for each *K* in order to test for convergence, and a cross-validation error (CV) was calculated for each run to identify which *K* showed the best fit to the data. Results from these population structure analyses were also used to verify the homogeneity of the newly typed Yaghnobi samples (*N* = 15) with respect to those already available from the literature (*N* = 6), such that they could be considered as a single group (*N* = 21) when computing ROH and population-based tests.

ROH were estimated to calculate the number of genomic segments in homozygosity within each population of the pruned dataset with the aim of comparing the Yaghnobi inbreeding pattern with those of other human groups. Since this is an intra-individual-based analysis, only populations with more than 10 individuals were considered, to reduce results possibly biased by low sample sizes.

The three-population test (*f3*; Reich et al., 2009) and the D-statistics (Green et al., 2010) were calculated on the pruned dataset using the dedicated software implemented in the ADMIXTOOLS v3.0 package. *Z*-scores of the *f3* statistics were computed via a Block Jackknife approach to assess significance of the obtained results. Population trios showing significant *f3* values (i.e., *Z*-scores ≤ −2) were finally considered for admixture dating with the LD decay-based approach implemented in ALDER v1.03 (Loh et al., 2013) performed on the un-pruned dataset.

3 | RESULTS

3.1 | Yaghnobi mitochondrial affinity with Middle Eastern populations

MtDNA haplotypes of the control region and the corresponding haplogroup assignment for each individual are provided in Supporting Information Table S1.

Diversity indexes calculated from Yaghnobi mtDNA HVS-I data suggested appreciable genetic isolation of this ethnic group, by showing significantly lower values compared to those observed, on average, for the other population clusters included in the reference dataset (Supporting Information Table S2).

PCA computed on mtDNA HVS-I data generated a scatterplot where the first two principal components (PCs) attested the overall resemblance between genetic relationships and geographic distribution of the examined Eurasian populations (Supporting Information Figure S1). In particular, PC1 (12% of variance) described a longitudinal axis of mtDNA variation, where Central Asian populations occupied an intermediate position between East Asian (EA) and the bulk of Middle Eastern, Caucasian, and South-Eastern European samples. By contrast, PC2 (5.5% of variance) reflected a north–south cline of diversity, mainly differentiating SA (i.e., Indian) samples from the Central Eurasian ones.

The observed longitudinal and latitudinal gradients of mtDNA diversity were also confirmed by the significant Global structure (observed value = 0.0475, *p*-value = 0.0082) identified by sPCA. It depicted a mtDNA landscape characterized by two main global genetic components showing similar spatial autocorrelation indexes, with sPC1 being additionally described by remarkable higher values of variance (Figure 2 and Supporting Information Figure S2).

Both PCA and sPCA analyses pointed to a substantial difference in the Yaghnobi mtDNA genetic composition with respect to those of other Central Asian populations. In fact, the PCA plot highlighted that, while Tajiks from the Matcha Valley (TJM) clustered within the Central Asian group in the PCA plot, the Yaghnobis (YAG) appeared to be genetically closer to Middle Eastern populations (Supporting Information Figure S1). Similarly, sPC1 values revealed different clustering patterns for the Tajiks from the Matcha Valley and the Yaghnobis, along the East–West cline of mtDNA variation spanning from East Asia (large white squares) to South-Eastern Europe, the Caucasus and the Middle East (large black squares; Figure 2). In particular, the positive value obtained for the Yaghnobis (0.47, black square) contrasted with the negative values (white squares) observed for both the Tajiks from Matcha (−0.41) and, more generally, for the whole Central Asian cluster (on average −0.20; Supporting Information Table S2).

The described differences between the Yaghnobis and the Tajiks from Matcha become even more evident when we directly compared their (F_{st}) genetic distances with those computed for all comparison populations. While the TJM showed the lowest genetic distances when contrasted with the other Central Asian groups, the YAG revealed an increased genetic affinity (i.e., lower distances) with geographically distant groups from the Middle East (Supporting Information Figure S3).

Consistent with these findings, the Yaghnobi mtDNA haplogroup composition (Supporting Information Table S1) turned out to be mainly characterized by typical West Eurasian H (~19.1%) and HV (~38.2%) lineages, and the absence of the Asian M haplogroup.

3.2 | Yaghnobi Y-chromosome relatedness with Central Asian populations

Y-STR haplotypes of each individual and the corresponding haplogroup assignment are provided in Supporting Information Table S3.

The Yaghnobi Y-chromosome haplotype diversity (*h*) was found to lay within the range of variability observed for the other examined populations (Supporting Information Table S4). However, their mean number of pairwise differences (MNPD) and nucleotide diversity (π) appeared to be significantly lower with respect to the mean calculated for reference population groups (Supporting Information Table S6). This is concordant with results obtained from the analyses of mtDNA data indicating appreciable reduction in Yaghnobi genetic variability, possibly due to some degrees of isolation.

PCA performed on Y-STR data, after the exclusion of outlying reference samples, set the examined populations along a triangular distribution of Y-chromosome variation whose extremes were represented respectively by EA, Central Asian, and Middle Eastern geographic clusters (Supporting Information Figure S4). Consistently with their geographic location, Tajiks from Matcha (TJM) showed their affinities with the group of Central Asian populations, whereas the Yaghnobis (YAG) were distributed along the cline of Central Asian to Middle Eastern genetic variation, occupying an intermediate, but decentralized position between these two extremes.

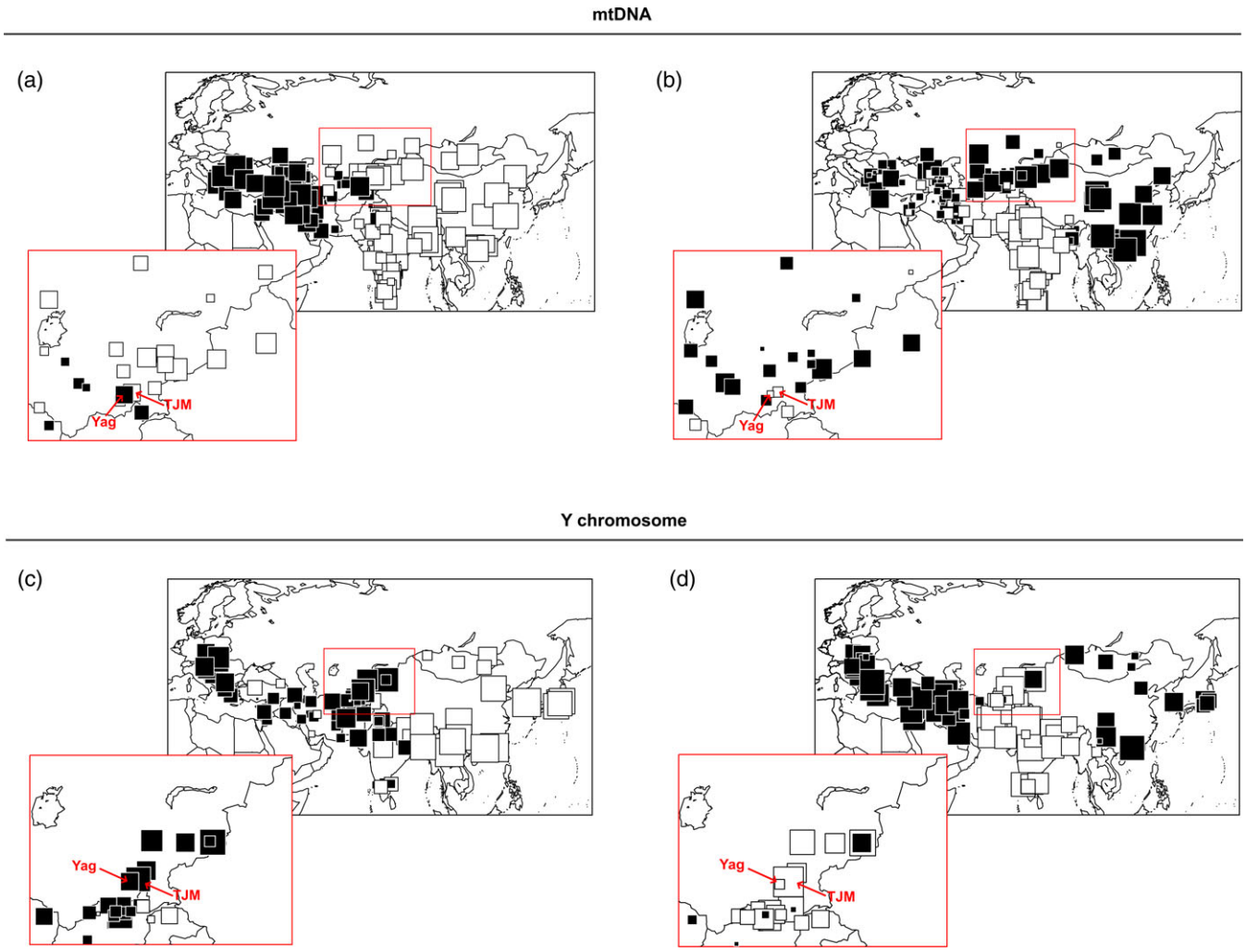

FIGURE 2 sPCA performed on mtDNA HVS-I and Y-chromosome extended datasets. The list of the populations included in mtDNA and Y-chromosome comparison datasets are provided in Supporting Information Tables S3 and S4, respectively. The first two global components are depicted in a multi-panel figure for both mtDNA––sPC1 (a); sPC2 (b)––and Y-chromosome––sPC1 (c); sPC2 (d). Positive values are represented by black squares, while negative values are represented by white squares. The size of the square is proportional to the absolute value of sPC scores. The enlarged box within each panel details the position of the newly-collected populations of Yaghnobis (YAG) and Tajiks from the valley of Matcha (TJM)

Such an intermediate position between the Middle Eastern and Central Asian population clusters was also reflected by the Yaghnobi Y-chromosome haplogroup composition (Supporting Information Table S3). In fact, equal representation (at frequencies of ~30% each) of both the typical Middle-Eastern J2-M172 and the Eastern European/Central Asian R1a-M17 lineages was observed, along with the R1b-M269* and K-M9 haplogroups which are present at 21% and 12%, respectively. Other lineages (such as E1b1b-M35, G2a-U5, and R1b-M269) are present at minor frequencies being observed in single individuals.

This genetic pattern was further explored by means of sPCA, highlighting the presence of a significant Global structure (observed value = 0.0599, *p*-value = 0.0028), with the first two global components (sPC1 and sPC2) showing the highest spatial autocorrelation indexes (Figure 2). In particular, sPC1 pointed to an East–West cline of Y-chromosome genetic variation that extended from the large negative values (on average −0.66) of EA groups (white squares) to the large positive ones (on average 0.41) of Southern European populations (black square; Figure 2). In this gradient, Middle Eastern and Central Asian samples occupied an intermediate position, showing on average low positive values (0.10 and 0.28, respectively). In this context, the Yaghnobis clustered with the other Central Asian groups (black squares), showing a positive sPC1 value (0.38) that was comparable with the average calculated for all Central Asian populations (Supporting Information Table S4). sPC2 instead differentiated a SA-like negative component (white squares), which was significantly less present in the Yaghnobis (−0.07) as compared to the mean value obtained for the other Central Asian populations (−0.21; Figure 2).

To directly test Y-chromosome differences between Yaghnobis and Tajiks from Matcha, their (R_{st}) genetic distances with respect to other Eurasian population groups were compared against one another. Accordingly, YAG and TJM were found to be similarly distant from all the considered reference populations, with the sole partial exception of a slightly higher similarity of TJM with some Central Asian groups (Supporting Information Figure S5).

3.3 | Setting the Yaghnobi population into the Eurasian genomic landscape

From an autosomal SNP perspective, the Yaghnobis showed high levels of intra-individual homozygosity, especially in terms of average number of homozygous segments (14.8), which was the second highest value among all reference populations (Supporting Information Table S5), following that observed in the Balochi sample (15.7; Supporting Information Figure S6). This finding, in line with the reduced diversity observed for uniparental markers, further corroborates the hypothesis of strong genetic drift having acted on this population, plausibly as a consequence of long-term isolation.

When PCA was applied to the pruned dataset excluding ancient samples, PC1 (3.6% of variance) described a rough gradient of genomic variation extending from Europe to East Asia, while PC2 (0.8% of variance) mainly distinguished the European populations from the SA ones (Figure 3a). In such a context, both the newly typed and the previously studied Yaghnobi individuals clustered together and occupied an intermediate position between the Iranian and the Tajik/Turkmen population samples, with some individuals slightly overlapping with Tajiks.

Ancient samples (Supporting Information Table S5) were then projected onto the PCA space described above and found to cluster within the area of the plot occupied by modern European and Middle Eastern populations (Supporting Information Figure S7), as previously observed in other studies (Lazaridis et al., 2016; Mathieson et al., 2015).

Inclusion of ancient samples turned out to be highly informative for the reconstruction of the Yaghnobi genomic ancestry when ADMIXTURE analysis was applied. The lowest CV-errors for all the performed ADMIXTURE replicates were observed by testing $K = 5$ and $K = 6$ ancestral populations (Supporting Information Figure S8), leading to the identification of an African-like (*green*), a SA-like (*yellow*), and an EA-like (*blue*) modern genetic component. At $K = 6$ (Figure 3b), the three main components that describe the genetic

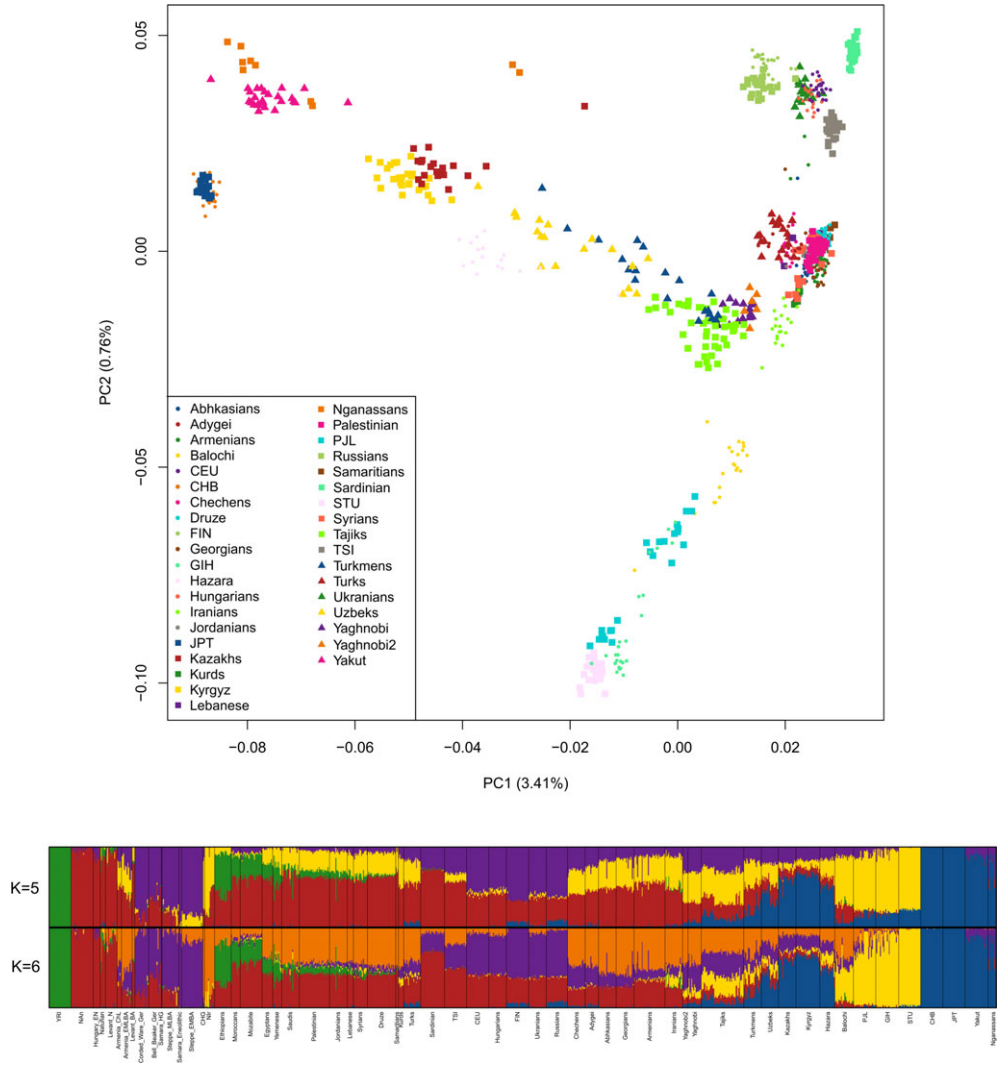

FIGURE 3 Principal component analysis performed on 37 modern Eurasian populations included in the pruned dataset (a). ADMIXTURE analyses testing $K = 5$ (top) and $K = 6$ (bottom) ancestral populations (b). For each K only runs with the highest log-likelihood values were plotted. Ancient samples were grouped according to the following labels: NAn (Neolithic from Anatolia), Hungary_EN (early Neolithic from Hungary), Levant_N (Neolithic Levant), Armenia_ChL (chalcolithic from Armenia), Armenia_EMLBA (early/middle/Late bronze age from Armenia), Levant_BA (Levant bronze age), Samara_HG (Samara hunter-gatherer), steppe_MLBA (steppe Middle/late bronze age), Steppe_EMBA (Steppe early/middle bronze age), CHG (Caucasus hunter-gatherer), NIr (Neolithic and chalcolithic from Iran). A complete list of modern and ancient populations included in the tested dataset are provided in Supporting Information Table S5

structure of ancient samples were also detected, in accordance with previous studies (Lazaridis et al., 2016). They were refereed according to the groups deriving most of their ancestry from them (i.e., the Neolithic Anatolian component, NAn, *red*; the Early/Middle Bronze Age Steppe component, EMBA_Steppe, *purple*; the Neolithic Iranian component, NIr, *orange*). Interestingly, on average 44% of the Yaghnobi ancestry was accounted for by the NIr component, similarly to what observed for modern Iranians (46%). Nevertheless, Yaghnobis differed from present-day Iranian samples due to a higher proportion of the EMBA_Steppe component (32% vs. 11%) and a lower percentage of the NAn one (8% vs. 25%). The Tajiks presented a genomic background more similar to that of Yaghnobis than modern Iranians, although they showed appreciable internal variability. In fact, some Tajik individuals presented higher levels of EA ancestry, while others had instead lower EA proportions, similarly to Yaghnobis. The main difference between Yaghnobi and Tajik populations was that the former showed a lower proportion (10% vs. 21%) of SA ancestry, coupled with a higher contribution of the NIr component (44% vs. 30%).

3.4 | Disentangling genomic relationships between the Yaghnobis and several Eurasian populations

To better understand the genomic relationships between Yaghnobis, Iranians and Tajiks suggested by results of ADMIXTURE analyses, more refined demographic hypotheses about the Yaghnobi genetic history were tested. For this purpose, an outgroup *f3* statistic in the form of *f3* (*Yoruba; test*, *X)* was used to compare patterns of shared genetic drift between each of the tested group (i.e., Yaghnobis, Iranians and Tajiks) with respect to all the other populations included in the reference dataset (X). Accordingly, Yaghnobis and Tajiks were found to be similarly distant to all Eurasian populations more than what resulted by comparing Yaghnobis and Iranians (Supporting Information Figure S9), thus suggesting a closer genetic affinity of Yaghnobis to Tajiks than to modern Iranians.

To corroborate this finding, an ad-hoc four populations *D*-statistics was calculated by testing a demographic model in the form of a tree *(Tajiks*, *Iranians; Yaghnobis*, *Yoruba)*. In fact, such an analysis can assess whether Yaghnobis show substantial recent gene flow with Tajiks (i.e., positive *D*) or instead if they are more closely related to Iranians (i.e., negative *D*). In accordance with outgroup *f3* results, the obtained *D* values suggested the presence of recent gene flow between Yaghnobis and Tajiks ($D = 0.0138 \pm 0.0007$; *Z*-score = 17.84).

The *f3* statistic was also used to formally assess the occurrence of admixture events by testing each population as admixed between a combination of every other two populations of the pruned dataset considered as parental proxies. Interestingly, Yaghnobis did not show any negative *f3* values when tested as target of admixture. On the contrary, they presented several significant negative values when considered as parental source of admixture (Supporting Information Table S7). In accordance with ADMIXTURE analyses, the main pattern that emerged from these *f3*-tests is that Yaghnobis showed a good fit as a parental proxy of admixtures for modern Middle Eastern and Caucasian populations, also involving an African- or a NAn-related ancestry group as the second parental source. They further showed a good fit as a parental proxy for admixture between Central Asian populations (e.g., Uzbeks, Kazakhs, Kyrgyzs, Hazaras, Tajiks, and Turkmens) and EAs as the second source group. Moreover, *f3* tests detected admixture between Tajiks and a source of SA ancestry (as suggested by ADMIXTURE results), involving Sri Lankan Tamil and Yaghnobis as parental proxies (Supporting Information Table S7).

All significant negative *f3* results that involved Yaghnobis as a parental population were further tested with ALDER to infer the dates of the observed admixture events (Supporting Information Table S8). For this purpose, we used the un-pruned dataset by excluding ancient samples since their high levels of missing data and differences in time ranges could bias estimations of admixture dating. As expected, not all of the best candidate admixture events pointed out by *f3* tests were significant when tested with ALDER, probably due to the high sensitivity of this method to close relatedness between parental populations and/or parental and target populations (i.e., remarkable LD sharing between them). Nevertheless, significant ALDER results detected admixture in Central Asians (e.g., Hazaras, Kazakhs, Turkmens, and Uzbeks) involving both Yaghnobis and EA populations, suggesting a range of 14–25 generations since admixture. Conversely, Tajiks did not present significant LD decay curves because of their close genetic distance and gene flow with Yaghnobis, as identified in the previously described analyses, which resulted in them sharing too long range LD. A second significant admixture detected by ALDER is the one involving the Yaghnobis and Yemenis as parental proxy for the Iranians as admixed group which present a wide range of 33 ± 18 generations since admixture. This wide interval is due to the dispersion of LD decay data points (Supporting Information Figure S10) as consequence of the close genetic proximity between the groups involved (i.e., Yemenis, Iranians, and Yaghnobis). Finally, Yaghnobis and Yoruba or Moroccans were suggested as parental groups in admixture events occurred in Palestinians or Mozabites (20–35 generations since admixture).

4 | DISCUSSION

The aim of the present study was to shed light on the genetic history and variability of the Yaghnobis, an enigmatic and interesting population, whose language and oral records link them to Sogdians, the ancient merchants of the Silk Route, were speaking the Lingua Franca of that area. Accordingly, we sought to disentangle their genetic history and elucidate their relationships with other Euro-Asiatic populations by taking advantage of data available in literature from both modern and ancient samples. For these purposes, we first applied an in-depth and accurate sampling strategy, in terms of number of collected samples and surveyed geographical locations, and we performed an exhaustive characterization of the Yaghnobi gene pool by using three complementary sets of genetic markers: mtDNA, Y-chromosome and genome wide autosomal SNPs. Finally, we integrated the obtained genetic evidence with historical and linguistic records available for the investigated population.

Our analyses showed that the Yaghnobi matrilineal gene pool comprises a restricted panorama of haplogroups, mainly represented by West Eurasian lineages, such as H and HV. Accordingly, the mitochondrial genetic ancestry of the Yaghnobis seems to reveal an ancient link with populations from the Middle East. This connection

could be explained by the preservation of an ancestral mtDNA background due to long-term isolation, and/or by the fact that such a population was scarcely involved in relatively recent events of gene-flow occurred with neighboring Central Asian groups, events that are indeed generally male-biased. In fact, the mitochondrial DNA variability emerging from these analyses corroborates specific peculiarity shared between the Yaghnobi language (strongly related to a variety of the ancient Middle Iranian language) and those spoken on the Iranian plateau, representing an example of close connection between linguistic (Bielmeier, 1989; Payne, 1989; Sims-Williams, 1989; Windfuhr, 1989) and genetic history (Cavalli-Sforza et al., 1994).

Moreover, in the light of the possible effect of past male-biased migration events, the Yaghnobi Y-chromosome legacy turned out to be characterized by a more tight relationship with other Central Asian populations than with the Middle Eastern ones. This supports the hypothesis that admixture with Central Asians was mainly mediated by male-driven migrations, as already supposed or, in some instances, proved for other particular migration events, like those occurred during the Bronze Age, from Pontic-Caspian Steppe to Europe (Goldberg, Günther, Rosenberg, & Jakobsson, 2017; Haak et al., 2015) or from Central Asia/Pontic Steppe to India (Silva et al., 2017). Nevertheless, Y-chromosome data alone may not clearly predict the direction of this gene flow (i.e., from the Yaghnobis to Central Asians or vice versa). The picture is further complicated as the Yaghnobi paternal lineages presented some peculiar signatures that still differentiate this population from the other Central Asian ones. For instance, the Yaghnobis showed a low proportion of typical SA paternal ancestry coupled with a reduced diversity. In addition, within the Y-chromosome genetic landscape, they shifted in an outlying position occupied between the Central Asian and Middle Eastern population clusters. These latter features could, once again, indicate that possible isolation effects may have occurred in the Yaghnobi population.

Since the patterns observed for uniparentally-inherited markers could be, to some extent, distorted by the small effective population sizes of such genetic systems, we attempted to disentangle the open questions regarding the Yaghnobi genetic history by investigating their autosomal genome-wide profiles. The results of these analyses turned out to be more concordant with inferences drawn from the Y-chromosome data compared to mtDNA ones. In fact, although the Yaghnobis, according to PCA (Figure 3a), seem to occupy an intermediate position in the Eurasian genomic landscape (i.e., between Central Asian and Middle Eastern population clusters) ADMIXTURE analyses further suggested that they are genetically closer to Tajiks than to Iranians (Figure 3b). However, in the interpretation of such admixture patterns, it is important to note that the Iranian population (Behar et al., 2010) included in our extended dataset was not clearly characterized and we do not know if these samples were evenly collected from the multitude of ethnic groups settled in the Iranian territory. Certainly, they appeared to be located in the geographical cline of Eurasian genetic variability and they seemed quite homogenous according to their distribution in the PCA plot (Figure 3a). On the other hand, in the datasets utilized for uniparental markers analyses (Supporting Information Tables S1 and S2), the sampling was definitely more detailed and accurate for the populations which were historically or linguistically connected with the Yaghnobis and consequently useful to elucidate the related pattern of genetic variation, as in the case of groups from the Iranian plateau (Di Cristofaro et al., 2013; Farjadian et al., 2011; Metspalu et al., 2004). Overall, the pattern highlighted by ADMIXTURE analysis (i.e., higher similarity of the Yaghnobis to the Tajiks than to Iranians) was mainly accounted for by the low Anatolian Neolithic ancestry fraction observable in the Yaghnobis (8%) as compared to modern Iranians (25%). Moreover, high proportions of the Steppe ancestry were detected in both the Yaghnobi (32%) and Tajik samples, but not in the Iranian ones (11%).

The performed outgroup *f3* statistics confirmed the relatedness between the Yaghnobis and the Tajiks and the computed *D*-statistics also showed evidence of recent allele sharing between these two groups, which could be instead a sign of a prolonged but low proportion gene flow. Despite that, the Yaghnobi and the Tajik populations are still well-distinguishable from a genomic perspective, especially due to the lower SA (10% vs. 21%) and higher Neolithic Iranian (44% vs. 30%) ancestry proportions observed in the Yaghnobis. This could be explained by the preservation of an ancient genomic background in the Yaghnobis due to long-term isolation. In fact, in accordance with reduced diversity of uniparental markers, also ROH analyses suggested that genetic drift has strongly shaped the gene pool of present-day Yaghnobis. This could be the result of cultural practices (e.g., consanguinity and endogamy) and/or of prolonged geographical isolation (Pemberton et al., 2012) or demographic events such as a recent bottleneck, as the deportation suffered by the Yaghnobis in the recent past (Panaino, 2008).

The possibility of this long-term isolation is further supported by results from three-population tests which indicated how the Yaghnobis did not present any signature of recent admixture. On the other hand, according to both *f3* and ALDER analyses, they seem to have played the role of source parental population in admixture events occurred in Central Asian or Middle Eastern groups. In particular, most of the obtained evidence pointed to some Central Asian groups as the result of admixture between EA populations and a genomic background for which the Yaghnobis could be considered as a reliable present-day proxy. These demographic processes are set in a timeframe ranging from 650 to 350 years ago considering 25 years per generation (or 750–420 considering 30 years/generation) and they probably constitute a signature ascribable to the several waves of westward Turkic and Mongol nomadic migrations (Barthold, 1962; Barthold, 1928; Spuler, 1985). A different event involved admixture between the Arab speaking Yemenis and a Yaghnobi-like gene pool resulting in the formation of Iranians. This event while showing a wide time interval (1,275–375 years ago) it post-dates and could be tentatively associated to a number of historical events emerging as consequence of the Arab-Islamic invasion of Central Asia, to which the disappearance of the Sogdian language progressively followed (Yoshida, 2016).

Other events dated by ALDER involved admixture between groups of Sub-Saharan African ancestry and a Yaghnobi-like gene pool resulting in the formation of Middle Eastern populations. The timings of these events (900–500 years ago) and the populations involved correspond with the results obtained from other genetic studies (Hellenthal et al., 2014; Moorjani et al., 2011). These studies explained these admixtures as consequence of the Arab slave trade (Gordon,

1998) that brought Sub-Saharan African ancestries into Middle East and Central Asia post-dating the Arab-Islamic expansion.

Nevertheless, we must observe that the chronological frames based on modern genetic data and their dating still have a timespan of uncertainty, and that in reality their chronological starting point might be much older. This would imply that the potential connections with the suggested migratory phenomena would be significantly fitting. When more data will be available from ancient genomes, the timing of these admixture events would be better resolved and precisely traced back to historical events.

5 | CONCLUSIONS

In conclusion, our study has not only provided new genomic data on accurately characterized and previously unsurveyed Yaghnobi communities, but it has also contributed to a more comprehensive knowledge of the genetic history of this peculiar population. In fact, we provided a fine dissection of the genomic ancestry of Yaghnobis, suggesting that they could represent a relic of an ancient gene pool shared by several Middle Eastern groups, which has more recently contributed to shape the genomic landscape of present-day Central Asian populations. Our analyses highlighted that the Yaghnobi genetic diversity could be thus considered as a cross-section of a genetic background widespread in a large area of Eurasia before that a series of historical demographic events substantially reshuffled the genetic structure of human populations residing in this region. A private history of isolation or at least of long-term reduced admixture with Central Asian populations, with the sole exception of some degree of recent gene flow from neighboring Tajik groups, might explain the preservation of such a peculiar genomic profile in the Yaghnobis. This is highly consistent with the location of this ethnic group in a remote mountainous area and is conceivably with oral traditions and linguistic connections that could have retained the Yaghnobi genetic and linguistic identities over a long period of time. Certainly, in the light of the new advances in the field of paleogenomics, a fundamental and advocated contribute to the understanding of the genetic history of Yaghnobis will be represented by the possibility to analyze ancient specimens from that area and in particular from Sogdian settlements.

ACKNOWLEDGMENTS

The authors would like to thank all the donors who kindly provided their biological samples making this study possible. We are indebted to the team of the scientific missions undertaken in the Yaghnob Valley, under the supervision of Professor Antonio Panaino. This work was supported by Fondazione del Monte di Bologna e Ravenna, by the European Research Council ERC-2011-AdG295733 and by the RFO of the University of Bologna.

ORCID

Elisabetta Cilli https://orcid.org/0000-0003-0407-267X

REFERENCES

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664.

Allentoft, M. E., Sikora, M., Sjogren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*(7555), 167–172.

Andreev M. S., & Peščereva, E. M. (1957). *Yagnobskie teksty*. Moskva – Leningrad: AN SSSR.

Barthold, V. V. (1962). *Four studies on the history of Central Asia*. Leiden: Brill.

Barthold, W. (1928). *Turkestan. Down to the Mongol invasion*. London: Oxford University Press.

Basello, G. P., Guizzo, D., & Ognibene, P. (2008). *Sulla punta di uno spillo. Rapporto scientifico sulla prima missione etnolinguistica nella valle dello Yaghnob*. Milano: Mimesis.

Baskakov, N. A. (1960). *The Turkic languages of Central Asia: Problems of planned culture contact*. London: Central Asian Research Centre.

Baskakov, N.A. (1966). Tjurskie jazyki. Obščie svedenija i tipologičeskaja charakteristika. *Jazyki narodov SSSR*. II: Tjurskie jazyki. Moskva: Nauka, 7–65.

Baumer, C. (2012). *The history of Central Asia: The age of the steppe warriors*. London: I.B. Tauris.

Beckwith, C. I. (2009). *Empires of the silk road: A history of Central Eurasia from the bronze age to the present*. Princeton: Princeton University Press.

Behar, D. M., Yunusbayev, B., Metspalu, M., Metspalu, E., Rosset, S., Parik, J., … Villems, R. (2010). The genome-wide structure of the Jewish people. *Nature*, *466*(7303), 238–242.

Benveniste, É. (1979). *Études sogdiennes*. Weisbaden: Reichert.

Bielmeier, R. (1989). Yaghnōbī. *Compendium Linguarum Iranicarum. Herausgegeben von R. Schmitt* (pp. 480–488). Wiesbaden: Reichert.

Cavalli-Sforza, L. L., Menozzi, P., & Piazza, A. (1994). *The history and geography of human genes*. Princeton: Princeton University Press.

Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M. F., Mobasher, Z., Austerlitz, F., & Heyer, E. (2007). From social to genetic structures in Central Asia. *Current Biology*, *17*(1), 43–48.

Cilli, E., De Fanti, S., Delaini, P., Panaino, A. C. D., & Gruppioni, G. (2013). A contribution to the genetic history of Yaghnobis of Tajikistan through mitochondrial DNA analysis. In A. C. D. Panaino, A. Gariboldi, & P. Ognibene (Eds.), *Yagnobi studies I. Papers from the Italian missions in Tajikistan* (pp. 141–150). Milan: Mimesis.

Cilli, E., Delaini, P., Costazza, B., Giacomello, L., Panaino, A., & Gruppioni, G. (2011). Ethno-anthropological and genetic study of the Yaghnobis; an isolated community in Central Asia. A preliminary study. *Journal of Anthropological Sciences*, *89*, 189–194.

CLI. (1989). *Compendium Linguarum Iranicarum*. Hrsg von R. Schmitt. Wiesbaden: Reichert.

Comas, D., Calafell, F., Mateu, E., Pérez-Lezaun, A., Bosch, E., Martínez-Arias, R., … Bertranpetit, J. (1998). Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *American Journal of Human Genetics*, *63*(6), 1824–1838.

Comas, D., Plaza, S., Wells, R. S., Yuldaseva, N., Lao, O., Calafell, F., & Bertranpetit, J. (2004). Admixture, migrations, and dispersals in Central Asia: Evidence from maternal DNA lineages. *European Journal of Human Genetics: EJHG*, *12*(6), 495–504.

de la Vaissière, É. (2002). *Histoire des merchands sogdiens*. Paris: College de France, Institut des Hautes Etudes Chinoises.

Vaissière, É. de la (2011). Sogdiana in history and archeology. *Encyclopædia Iranica*, online edition: http://www.iranicaonline.org/articles/sogdiana-iii-history-and-archeology.

Di Cristofaro, J., Pennarun, E., Mazières, S., Myres, N. M., Lin, A. A., Temori, S. A., … Chiaroni, J. (2013). Afghan Hindu Kush: Where Eurasian sub-continent gene flows converge. *PLoS One*, *8*(10), e76748.

Dulik, M. C., Osipova, L. P., & Schurr, T. G. (2011). Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS One*, *6*(3), e17548.

Excoffier, L., & Lischer, H. (2010). Arlequin suite ver 3.5: A new series of programs to perform population genetics analyses under Linux and windows. *Molecular Ecology Resources*, *10*(3), 564–567.

Farjadian, S., Sazzini, M., Tofanelli, S., Castrì, L., Taglioli, L., Pettener, D., … Luiselli, D. (2011). Discordant patterns of mtDNA and ethno-linguistic variation in 14 Iranian ethnic groups. *Human Heredity*, *72*(2), 73–84.

Ferri, G., & Alù, M. (2012). Development of six-Y-SNPs assay for forensic analysis in European population. In *DNA in Forensics 2012*, 5th International EMPOP Meeting––8th International Forensic Y-User Workshop. Innsbruck.

Frye, R. N. (1998). *The heritage of Central Asia. From antiquity to the Turkish expansion*. Princeton: Markus Wiener.

Goldberg, A., Günther, T., Rosenberg, N. A., & Jakobsson, M. (2017). Ancient X chromosomes reveal contrasting sex bias in Neolithic and bronze age Eurasian migrations. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(10), 2657–2662.

Golden, P. B. (2011). *Central Asia in world history*. Oxford: Oxford University Press.

Gordon, M. (1998). *Slavery in the Arab world* (New ed.). New York, NY: New Amsterdam Books.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., … Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*(5979), 710–722.

Gunya, A. N. (2002). Yaghnob Valley. In *Nature, history and changes of a mountain community development in Tajikistan*. Moscow: Institute of Geography, Russian Academy of Sciences.

Haak, W., Lazaridis, I., Patterson, N., Rohland, N., Mallick, S., Llamas, B., … Reich, D. (2015). Massive migration from the steppe was a source for indo-European languages in Europe. *Nature*, *522*(7555), 207–211.

Heissig, W. (1978). *Die Mongolen*. München: DTV Verlag.

Hellenthal, G., Busby, G. B. J., Band, G., Wilson, J. F., Capelli, C., Falush, D., & Myers, S. (2014). A genetic atlas of human admixture history. *Science*, *343*(6172), 747–751.

Heyer, E., Balaresque, P., Jobling, M. A., Quintana-Murci, L., Chaix, R., Segurel, L., … Hegay, T. (2009). Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC Genetics*, *10*, 49.

HM (1989). Giovanni di Pian di Carpine. *Storia dei Mongoli*. E. Menestò (Ed.). Spoleto: CISAM.

Irwin, J. A., Ikramov, A., Saunier, J., Bodner, M., Amory, S., Röck, A., … Parsons, T. J. (2010). The mtDNA composition of Uzbekistan: A microcosm of central Asian patterns. *International Journal of Legal Medicine*, *124*(3), 195–204.

Jakubov, J. (1979). *Pargar v VII-VIII vv. n.ė. (Verchnij Zeravšan v ėpochu rannego srednevekov'ja)*. Dušanbe: Doniš.

Jakubov, J. (1988). *Rannesrednevekovye sel'skie poselenija gornogo Sogda*. Dušanbe: Doniš.

Johanson, L., & Csató, É. A. (Eds.). (1998). *The Turkic languages*. London: Routledge.

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405.

Jombart, T., Devillard, S., Dufour, A.-B., & Pontier, D. (2008). Revealing cryptic spatial patterns in genetic variability by a new multivariate method. *Heredity*, *101*(1), 92–103.

Kerimova, A.A. (1966). Tadžikskij jazyk. *Jazyki narodov SSSR*. I: Indoevropejskie jazyki. Moskva: Nauka, 212–236.

Kuhn, A. (1881a). Svedenija o Jagnaubskom narode. *Turkestanskie Vedomosti*, *3*, 11–12.

Kuhn, A. (1881b). Svedenija o Jagnaubskom narode. *Turkestanskie Vedomosti*, *4*, 14–15.

Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D. C., Rohland, N., Mallick, S., … Reich, D. (2016). Genomic insights into the origin of farming in the ancient near east. *Nature*, *536*(7617), 419–424.

Le Strange, G. (1905). *The lands of the eastern caliphate: Mesopotamia, Persia and Central Asia from the Moslem conquest to the time of Timur*. Cambridge: Cambridge University Press.

Litvinskij, B. A. (1998). *La civilisation de l'Asie centrale antique*. Leidorf: Rahden/Westf.

Loh, P.-R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J. K., Reich, D., & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics*, *193*(4), 1233–1254.

Lur'e, P., Semenov, N., & Stepanov, A. (2017). Stennaja rospis' iz Chisoraka. *Bulletin of Miho Museum*, *17*, 61–74.

Marchi, N., Hegay, T., Mennecier, P., Georges, M., Laurent, R., Whitten, M., … Heyer, E. (2017). Sex-specific genetic diversity is shaped by cultural factors in inner Asian human populations. *American Journal of Physical Anthropology*, *162*(4), 627–640.

Martínez-Cruz, B., Vitalis, R., Ségurel, L., Austerlitz, F., Georges, M., Théry, S., … Heyer, E. (2011). In the heartland of Eurasia: The multilocus genetic landscape of central Asian populations. *European Journal of Human Genetics: EJHG*, *19*(2), 216–223.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., … Reich, D. (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, *528*(7583), 499–503.

Menges, K. H. (1968). *The Turkic languages and peoples: An Introduction to Turkic studies*. Wiesbaden: Harrassowitz.

Metspalu, M., Kivisild, T., Metspalu, E., Parik, J., Hudjashov, G., Kaldma, K., … Villems, R. (2004). Most of the extant mtDNA boundaries in south and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genetics*, *5*(1), 26.

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research*, *16*(3), 1215.

Moorjan, i. P., Patterson, N., Hirschhorn, J. N., Keinan, A., Hao, L., Atzmon, G., … Reich, D. (2011). The history of African gene flow into southern Europeans, Levantines, and Jews. *PLoS Genetics*, *7*, e1001373.

Negmatov, N. N. (1996). Sogdiana. Part two. Ustrushana, Ferghana, Chach and Ilak. In B. A. Litvinsky (Ed.), *History of civilization of Central Asia. III: The crossroad of civilizations: A.D. 250 to 750* (pp. 259–280). Paris: UNESCO.

Palstra, F. P., Heyer, E., & Austerlitz, F. (2015). Statistical inference on genetic data reveals the complex demographic history of human populations in Central Asia. *Molecular Biology and Evolution*, *32*(6), 1411–1424.

Panaino, A. (2008). Attraversando la Valle dello Yaghnob: passato, presente e (possibile) futuro di una minoranza etno-linguistica. In G. P. Basello, D. Guizzo, & P. Ognibene (Eds.), *Sulla punta di uno spillo. Rapporto preliminare della spedizione etnografica italiana nella valle dello Yaghnob* (pp. 1–77). Milano: Mimesis.

Panaino, A., Gariboldi, A., & Ognibene, P. (2013). *Yaghnobi studies I. Papers from the Italian missions in Tajikistan*. Milano: Mimesis.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and Eigenanalysis. *PLoS Genetics*, *2*(12), e190.

Payne, J. (1989) Pāmir Languages). *Compendium Linguarum Iranicarum. Herausgegeben von R. Schmitt* (p. 417). Wiesbaden: Reichert.

Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of Homozygosity in worldwide human populations. *The American Journal of Human Genetics*, *91*(2), 275–292.

PhTF (2000). *Philologiae Turcicae fundamenta. Iussu et auctoritate Unionis Universae Studiosorum Rerum Orientalium, auxilio et opera Unitarum Nationum Educationis Scientiae Culturae Ordinis, una cum praestantibus Turcologis ediderunt Jean Deny, Kaare Gronbech, Helmut Scheel et al.* III: *History of the Turkic Peoples in the pre-Islamic Period*. In: H. R. Roemer (Ed.) with the assistance of W.E. Scharlipp, Wiesbaden: Steiner.

Poppe, N. (1970). *Mongolian language handbook*. Washington: Center for applied linguistics.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), 559–575.

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., … McElreavey, K. (2004). Where west meets east: The complex mtDNA landscape of the southwest and central Asian corridor. *American Journal of Human Genetics*, *74*(5), 827–845.

Rastorgueva, V.S. (1966). Iranskie jazyki. *Jazyki narodov SSSR*. I: Indoevropejskie jazyki. Moskva: Nauka, 194–211.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature*, *461*(7263), 489–494.

Rose, J. (2010). The Sogdians: Prime movers between boundaries. *Comparative Studies of South Asia, Africa and the Middle East*, *30*(3), 410–419.

Ségurel, L., Martínez-Cruz, B., Quintana-Murci, L., Balaresque, P., Georges, M., Hegay, T., … Vitalis, R. (2008). Sex-specific genetic structure and social organization in Central Asia: Insights from a multi-locus study. *PLoS Genetics*, *4*(9), e1000200.

Silva, M., Oliveira, M., Vieira, D., Brando, A., Rito, T., Pereira, J. B., … Soares, P. (2017). A genetic chronology for the Indian subcontinent points to heavily sex-biased dispersals. *BMC Evolutionary Biology*, *17*(1), 88.

Sims-Williams, N. (1989). Sogdian). *Compendium Linguarum Iranicarum. Herausgegeben von R. Schmitt* (pp. 173–192). Wiesbaden: Reichert.

Spuler, B. (1961). *Les Mongols dans l'histoire*. Paris: Payot.

Spuler, B. (1985). *Die Mongolen in Iran: Politik, Verwaltung und Kultur der Ilchanzeit 1220–1350*. Leiden: Brill.

van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, *30*(2), E386–E394.

Wells, R. S., Yuldasheva, N., Ruzibakiev, R., Underhill, P. A., Evseeva, I., Blue-Smith, J., … Bodmer, W. F. (2001). The Eurasian heartland: A continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences*, *98*(18), 10244–10249.

Whitlock, M. (2003). *Land beyond the river: The untold story of Central Asia*. London: St. Martin's Press.

Wiesehöfer, J. (1996). *Ancient Persia: From 550 BC to 650 AD*. London: I.B. Tauris.

Windfuhr, G. L. (1989). New Iranian languages). *Compendium Linguarum Iranicarum. Herausgegeben von R. Schmitt* (pp. 246–250). Wiesbaden: Reichert.

Yoshida, Y. (2016). SOGDIAN LANGUAGE i. Description. In *Encyclopædia Iranica*. http://www.iranicaonline.org/articles/sogdian-language-01.

Yunusbayev, B., Metspalu, M., Metspalu, E., Valeev, A., Litvinov, S., Valiev, R., … Villems, R. (2015). The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genetics*, *11*(4), e1005068.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Cilli E, Sarno S, Gnecchi Ruscone GA, et al. The genetic legacy of the Yaghnobis: A witness of an ancient Eurasian ancestry in the historically reshuffled central Asian gene pool. *Am J Phys Anthropol*. 2019;1–12. https://doi.org/10.1002/ajpa.23789