bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

"Like Sugar in Milk": Reconstructing the genetic history of the Parsi population

Gyaneshwer Chaubey[1*], Qasim Ayub[2*], Niraj Rai[3*], Satya Prakash[3], Veena Mushrif-Tripathy[4], Massimo Mezzavilla[2], Ajai Kumar Pathak[1,5], Rakesh Tamang[6] , Sadaf Firasat[7], Maere Reidla [1,5], Monika Karmin[1,5,8] Deepa Selvi-Rani[3], Alla G. Reddy[3], Jüri Parik[1,5], Ene Metspalu[1,5], Siiri Rootsi[1], Kurush Dalal [9], , Shagufta Khaliq[10], Syed Qasim Mehdi[7], Lalji Singh[11], Mait Metspalu[1,5], Toomas Kivisild[1,12], Chris Tyler-Smith[2], Richard Villems[1,5†] and Kumarasamy Thangaraj[3†]

[1] Evolutionary Biology Group, Estonian Biocentre, Riia23b, Tartu, 51010, Estonia.
[2] The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK.
[3] CSIR-Centre for Cellular and Molecular Biology, Hyderabad, 500007, India.
[4] Department of Archaeology, Deccan College Post-Graduate and Research Institute, Pune, Maharashtra, 411006, India.
[5] Department of Evolutionary Biology, Institute of Molecular and Cell Biology, University of Tartu, Tartu, 51010, Estonia.
[6] Department of Zoology, University of Calcutta, Kolkata, 700073, India.
[7] Centre for Human Genetics and Molecular Medicine, Sindh Institute of Urology and Transplantation, Karachi, 74200, Pakistan.
[8] Department of Psychology, University of Auckland, Auckland, 1142, New Zealand.
[9] Centre for Archaeology (CfA), Centre for Extra Mural Studies (CEMS) University of Mumbai (Kalina Campus) Vidyanagri, Santacruz E Mumbai 400098, India.
[10] Department of Human Genetics & Molecular Biology, University of Health Sciences, Lahore, 54000, Pakistan.
[11] Genome Foundations Hyderabad, 500076, India
[12] Division of Biological Anthropology, University of Cambridge, Cambridge, CB2 3QG, UK.

*Equal contributions
†These authors share the senior authorship.

Correspondence to: Gyaneshwer Chaubey (gyanc@ebc.ee), Qasim Ayub (qa1@sanger.ac.uk) or Kumarasamy Thangaraj (thangs@ccmb.res.in)

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Background

The Parsis, one of the smallest religious community in the world, reside in South Asia. Previous genetic studies on them, although based on low resolution markers, reported both Iranian and Indian ancestries. To understand the population structure and demographic history of this group in more detail, we analyzed Indian and Pakistani Parsi populations using high-resolution autosomal and uniparental (Y-chromosomal and mitochondrial DNA) markers. Additionally, we also assayed 108 mitochondrial DNA markers among 21 ancient Parsi DNA samples excavated from Sanjan, in present day Gujarat, the place of their original settlement in India.

Results

Our extensive analyses indicated that among present-day populations, the Parsis are genetically closest to Middle Eastern (Iranian and the Caucasus) populations rather than their South Asian neighbors. They also share the highest number of haplotypes with present-day Iranians and we estimate that the admixture of the Parsis with Indian populations occurred ~1,200 years ago. Enriched homozygosity in the Parsi reflects their recent isolation and inbreeding. We also observed 48% South-Asian-specific mitochondrial lineages among the ancient samples, which might have resulted from the assimilation of local females during the initial settlement.

Conclusions

We show that the Parsis are genetically closest to the Neolithic Iranians, followed by present-day Middle Eastern populations rather than those in South Asia and provide evidence of sex-specific admixture from South Asians to the Parsis. Our results are consistent with the historically-recorded migration of the Parsi populations to South Asia in the 7thcentury and in agreement with their assimilation into the Indian sub-continent's population and cultural milieu "like sugar in milk". Moreover, in a wider context our results suggest a major demographic transition in West Asia due to Islamic-conquest.

Keywords: Parsi- Zoroastrian- Autosomes- mtDNA – Y chromosome- ancient DNA

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Background

The Parsi (or Parsee) community of the Indian sub-continent are a group of Indo-European speakers and adherents of the Zoroastrian faith, one of the earliest monotheisms that flourished in pre-Islamic Persia (present-day Iran)(1,2). Zoroastrianism was the official religion of Persia from 600 B.C. to 650 A.D.(3-5) and despite a long history of well-preserved culture, it now has a limited number of followers (6,7). The Heritage Institute estimates the total number of Zoroastrians to be around 137,000, with 69,000 living in India, roughly 20,000 in Iran (8) and 2,000-5,000 in Pakistan (8,9). This reduction in population is mainly due to strict marriage practices and low birth rates (7,10-12).

The Parsi trace their ancestry to the Zoroastrians of modern Iran who are followers of the Prophet Zoroster or Zarathushtra (3). In the seventh century the Zoroastrian Sassanian dynasty was threatened by Islamic conquest and a small group of Zoroastrians fled to Gujarat in present-day India, where they were called 'Parsi' (literally meaning 'people from Paras or Faras', the local term for Persia)(3,4,6,13). Several myths narrate their first arrival in the West Coast of India and settlement in Sanjan (Gujarat) (4,14-17). The most popular one mentioned in the *Qissa-e-Sanjaan* is that an Indian ruler called *Jadi Rana* sent a glass full of milk to the Parsi group seeking asylum (4,18). His message was that his kingdom was full with local people. The Zoroastrian immigrants put sugar (or a ring, in some versions of the story) into the milk to indicate an assimilation of their people into the local society, like "sugar in milk"(14,18) In contemporary India and Pakistan, we see their adoption of local languages (Gujarati, Sindhi) and economic integration while maintaining their ethnic identity and practicing strict endogamy (1,3,4,12,19-21).

Previous genetic analyses of the Parsis have focussed mainly on low-resolution uni-parental markers which have suggested their affinity with both West Eurasian and Indian populations(22-24). Autosomal analysis based on microsatellites or human leukocyte antigens (HLA) have revealed their intermediate position among the populations of South Asia and the Middle East/Europe(9,24). A study of mitochondrial DNA (mtDNA) variation reported 60% of South Asia lineages among the Pakistani Parsi population (23), whereas the

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

male lineages based on Y chromosome admixture estimates were almost exclusively Iranian (22). Based on these results, a male-mediated migration followed by assimilation of local South Asia females was concluded (23).

These early studies of the Parsi populations relied mainly on low-resolution markers, limiting the power of the analyses (9,23-25), and the majority of Parsis (~98.8%), living in India, have been underrepresented in these studies. In the present study, we used a combination of high-resolution uni-parental and bi-parental genetic markers on modern samples, as well as mtDNA markers on ancient Parsi samples excavated from Sanjan, in present-day Gujarat, India (Fig. 1). The human remains from Sanjan *dokhama* (tower of silence) District Valsad, Gujarat, was excavated in 2004. The AMS (Accelerator Mass Spectrometry) dating of human remains suggest that the *dokhama* belongs to the 14th to 15th century A.D. (17).

We investigated whether or not the current Parsi people living in India and Pakistan are genetically related amongst themselves and with the present-day Iranian population, and if their genetic composition has been affected by the neighbouring Indian and Pakistani populations. We also examined runs of homozygosity (RoH) to study consanguinity. To address the extent to which the current Parsi populations assimilated local females during their long formation history, we compared their mtDNA haplogroup composition with ancient remains excavated in Sanjan, the initial settlement established by these migrants from Persia (17).

Results and Discussion

For autosomal analyses, we used Illumina HumanHap 650K genotyping chips on 19 Indian Parsi samples collected from Mumbai, and Illumina 2.5M genotyping chips for 24 Pakistani Parsi individuals from Karachi (Fig. 1 and Additional File 1). The combined Parsi dataset was merged with a global data from the published literature (26,27). Table S2 lists the populations and number of SNPs used for various analyses after quality control (Additional File 1). The mean allele frequency differentiation between the two Parsi (Indian and Pakistani) groups

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

was the lowest (F_{ST} Indian and Pakistani Parsi = 0.00033 $\pm$ 0.00025), followed by the differentiation of each from the Iranian population (0.011 $\pm$ 0.00021 and 0.012 $\pm$ 0.00025 for Pakistani and Indian Parsis, respectively), suggesting a common stock for both the Indian and Pakistani Parsis with the closest interpopulation affinity with populations from their putative homeland, Iran (Additional File 1: Fig. S1 and Additional File 1: Table S3). Collectively, in F_{ST}-based analysis, both of the Parsi groups showed significantly closer connection with West Eurasians than any of the Indian groups (two-tailed p value <0.0001).

We applied the default settings of the SmartPca programme implemented in the EIGENSOFT package (28) and performed principal-components analysis (PCA) with other Eurasian populations using autosomal SNP data (Additional File 1: Table S2 and Supplementary text). Our plot of the first and second principal components (PCs) clusters the Indian and Pakistani Parsis together, along the European-South Asian cline (Fig. 2). A plot of populationwise mean of eigenvalues showed their placement between the Pakistani and Iranian populations, indicating that the Parsis might have admixed from these two groups. Such an intermediate position of Parsis closer to Iranians than to their present geographic neighbours (Sindhi and Gujarati), suggests that the Parsis may have major ancestry from West Eurasians (Iranians) and minor ancestry with South Asians (Fig. 2 and Additional File: Fig. S1). In order to identify the ancestry components of Parsis more quantitatively, we applied the model-based clustering method ADMIXTURE (29). The best model (30,31) suggested eight major ancestral groups and identified three ancestral components within the Parsi (Fig. 3 and Additional File 1: Table S4). The distributions of these components among Indian and Pakistani Parsis were unique and similar to each other, but were distinct from their neighbouring South Asian populations (Additional File 1: Fig. S2a). The ANI (Ancestral North Indian) ancestry calculated from *f*4 ancestry estimate showed a substantially higher level of this ancestry among Parsis than any other South Asia population (Additional File 1: Fig. S2b). Supporting the F_{ST} and PCA results, the Middle-Eastern-specific (blue) ancestry component was significantly higher (two-tailed p value <0.0001), in the Parsis than in any other populations residing in South Asia that were examined. The present-day Iranian population exhibited a striking difference from the Parsis, mainly in carrying an additional European component (light blue) and

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

substantially lower South Asian ancestry (dark green) (Fig. 3 and Additional File 1: Fig. S2 and Table S4). It was suggested previously that the Islamic conquest had a major genomic impact on several Middle Eastern populations, including Iranians (32). Since Parsis diverged from Iranians just after this conquest, they may represent the genetic strata of Iran before Islamic-conquest. To test this scenario, we applied a formal test of admixture *f*3 statistics (Additional File 1: Table S5). For Iranians, a negative value with significant Z-scores indicated that they are descendants of a population formed by the admixture of Neolithic Iranians and populations from the Arabian -Peninsula, while for Parsis this test was positive with significant Z-scores. Therefore, it seems plausible that the additional light blue component may have been introduced to Iran after the exile of the Parsis, likely via recent gene flow from the Arabian Peninsula(32-34). The quantitative estimation of South Indian and Iranian ancestries among Parsis and their South Asian neighbours showed a significant level of differentiation in ancestry composition with an inclination of Parsis towards Iranian ancestry (two-tailed p value <0.0001) (Table 1 and Additional File 1: Supplementary text).

We computed *D* statistics (35) to determine the nature of gene flow and admixture of Parsis with their parental (Iranian) and neighbouring (Gujarati, Sindhi) populations (Table 2). Consistent with the previous analyses (Figs. 2 and 3), both of the Parsi populations shared a highly significant value of *D* with each other. On the other hand, the South Asian populations (Gujarati and Sindhi) had significant levels of gene flow with each other, as well as with both of the Parsi populations when evaluated with respect to the present-day Iranian population (Table 2). Two independent studies have recently reported the data from ancient Iranian samples (36,37). It was suggested that the early Neolithic Zagros sample showed closer affinity with the Iranian Zoroastrians (36). Here we estimated the *D* values of Parsis for Neolithic Iranians *vs* modern Iranians to compare the allele sharing. Our results demonstrated a significant level of genetic affinity between Parsis and Neolithic Iranians (Table 2 and Additional File 1: Supplementary text and Table S6). The outgroup *f*3 statistics of ancient Iranian samples supported the close affinity of the Parsis with the Neolithic Iranians (Additional File 1: Supplementary text, Fig. S3 and Table S6). Moreover, for modern populations, the outgroup *f*3 statistic test and identity-by-state (IBS) plots supported the

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

closer affinity of the Parsis with the West Eurasian populations than South Asians (Additional File 1: Figs. S4 and S5). In order to compare the shared drift with Iranian and Indian (South Munda) populations, we plotted the derived allele sharing values of Parsis and other Eurasian populations calculated with respect to the Iranian and South Munda (Indian) populations (Additional File 1: Supplementary text and Fig. S6). This analysis aligned Parsis closer to the Iranian axis between Pakistani and West Eurasian populations, confirming the historical interpretation of the most recent common ancestry of Parsis with the Iranians. TreeMix (38) analysis support these conclusions and shows the Parsis located between the South Asian and Iranians (Additional File 1: Fig. S7).

We computed a maximum likelihood (ML) tree and co-ancestry matrix based on the haplotype structure of the Parsi populations, applying the default settings of ChromoPainter and fineSTRUCTURE (version1) (39). The ML tree split South Asian and West Eurasian populations into two distinct clusters (Additional File 1: Fig. S8). All the Parsi individuals form a unique sub-cluster embedded within the major West Eurasian population trunk. The co-ancestry matrix plot clearly differentiated Parsis from their neighbours in sharing a large number of chunk counts with West Eurasian (mainly Iranian and Middle Eastern) populations (Additional File 1: Fig. S9 and Supplementary text). . Additionally, South Asian populations have donated a significantly higher number of chunks to Parsis than they received from them (two tailed p value <0.0001). However, the number of these chunks was significantly lower than the chunk counts shared between any pair of South Asian populations (two-tailed p value <0.0001) (Additional File 1: Fig. S9). . The fineSTRUCTURE and *D* statistic results thus largely suggest unidirectional minor gene flow from South Asians to Parsis (Table 2 and Additional File 1: Fig. S9). .

To further investigate the parental relatedness among Parsis (19), we analysed the Runs of Homozygosity (RoH) and Inbreeding coefficient in the population (Additional File 1: Fig. S10). For RoH calculations, we applied three window sizes (1,000 kb, 2,500 kb and 5,000 kb), requiring a minimum of 100 SNPs per window and allowing one heterozygous and five missing calls per window. Long RoH segments characterise consanguinity and also provide a distinctive record of the demographic history for a particular population (40,41). As expected, both of the Parsi populations carried a larger number of long segments relative to

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

their putative parental populations and present neighbours at the 1,000 kb window length, likely due to small population size and a high level of inbreeding. However, the Sindhi population from Pakistan also showed a higher level of inbreeding at the larger RoH window sizes, most likely due to an elevated level of cross-cousin marriages (Additional File 1: Fig. S10).

We used ALDER, a method based on linkage disequilibrium (LD), (42), to estimate the time of admixture between Parsis and their neighbouring South Asian populations. For this analysis we used the present-day Iranian *vs* Gujarati or Sindhi populations as their surrogate ancestors. We estimated the admixture time of Parsi groups to be around ~40 (95% CI 26-50) generations, which yields a time of 1160 years (assuming a generation time of 29 years), in good agreement with their historically-recorded migration to South Asia (Table 3). We also tested evidence for a more complex admixture history using MALDER (43) which can be used to infer multiple admixture events. The MALDER analyses confirmed the ALDER results, demonstrating only one admixture event 54 ± 8 generations ago. The ancestral sources with the highest amplitude in MALDER were Sardinian and Dai (Additional File 1: Table S7).

To investigate how drift shaped the functional genetic variation after admixture in the Parsis, we implemented the population branch statistic (44) using the Sindhi and Iranians as reference and outgroup. We analysed variants over the 99.9th percentile of the genomic distribution, focussing only on those that were annotated as missense, stop gain, stop loss, splice acceptor and splice donor using the Ensembl Variant Effect Predictor tool (45). This revealed a cluster of linked SNPs in the HLA region and a missense SNP in CD86 (rs1129055) with a high ancestral G allele frequency in the Parsi (0.87) (Additional File 1: Fig. S11 and Table S8). The frequency of this G allele is lower in the Iranians and Sindhi (0.60) and other South Asians (0.52) and East Asians (0.40). This polymorphism has been associated with the pathogenesis of pneumonia-induced sepsis and the G allele has been associated with a decreased risk of active brucellosis in Iranians (46). The G variant has also been associated as an eQTL for decreased expression of *IQBC1*, an IQ motif containing B1 gene that is highly expressed in EBV transformed B lymphocytes (45).

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To obtain a detailed understanding on the sex-specific South Asian and Iranian ancestries, we examined maternally-inherited mtDNA and paternally-inherited Y chromosome biallelic markers in a larger sample in both the Indian and Pakistani Parsi populations (Additional File 1: Fig. S12 and Tables S9 and S10). For the mtDNA analysis we were also able to assay 21 ancient samples from the Sanjan (21) region, for 108 diagnostic polymorphisms (Additional File 1: Table S11 and Supplementary text). Interestingly, we observed 48% South-Asian-specific lineages (haplogroups M2, M3, M5 and R5) among the ancient Parsi samples, which could potentially be explained in two ways. First, these haplogroups might have been carried by the migration of Zoroastrian refugees from Faras (Iran), a possibility which is supported by the presence of these clades in present-day Persian samples (9.9%) (34). Secondly, they might have resulted from the assimilation of local females during the initial settlement. The comparison of ancient and modern samples thus identified maternal lineages which can be considered as founding (surviving or lost), as well as those which were assimilated subsequently (Additional File 1: Fig. S12 and Tables S9-11). The Y chromosome profiles of Indian and Pakistani Parsi populations revealed a higher frequency of Middle-Eastern-specific lineages than South Asian ones in the Parsis (Additional File 1: Fig. S12 and Table S10). The PC analysis of both mtDNA and Y chromosome data placed all the Parsi groups close to each other and showed their contrasting clustering based on maternal or paternal ancestries (Fig. 4). For mtDNA, the Parsi cluster was closer to the Indian and Pakistani cluster (Fig. 4a), whereas for the Y chromosome it aligns between the Iranian and Pakistani populations (Fig. 4b). The Y-chromosomal PCA is similar to the autosomal PCA (Fig. 2 and Additional File: Fig. S1). The contrasting patterns of maternal and paternal ancestry support a largely female-biased admixture from the South Asian populations to the Parsis.

Conclusions

In conclusion, our investigation has not only contributed substantial new data, but has also provided a more comprehensive insight into the population structure of Parsis and their genetic links to Iranians and South Asians. We show that the Parsis are genetically closer to Middle Eastern (Iranian and Caucasian) populations than those in South Asia and provide evidence of sex-specific admixture with prevailing female gene flow from South Asians to the

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Parsis. Our results are consistent with the historically-recorded migration of the Parsi populations to South Asia in the 7th century and in agreement with their assimilation into the Indian sub-continent's population and cultural milieu "like sugar in milk".

Methods

A detailed description of material and methods can be found in the supplementary text (Additional File 1). The Ethics Committee of the participating institutions (EBC, Estonia, CCMB, India and Human Materials and Data Management Committee, Wellcome Trust Sanger Institute) approved the study. The modern Parsi samples were pooled from three independent collections: two from Mumbai, India, and one from Karachi, Pakistan (Fig. 1). Illumina 650K and 2.5M chips were used to genotype 19 Indian and 24 Pakistani Parsi individuals, respectively, following the manufacturers' specifications. We merged our newly-generated data of 43 samples with the relevant reference datasets of 829 samples published elsewhere (Additional File 1: Supplementary text and Table S2). For mtDNA control and coding region markers, we genotyped 117 Indian and 50 Pakistani Parsi samples (Additional File 1: Table S9). We followed phylotree (build 17) to classify them into haplogroups. For Y chromosome genotyping, 90 Pakistani samples were genotyped either by sequencing or by PCR-RFLP for the relevant Y chromosome markers, whereas 84 Indian samples were assayed for 80 Y chromosomal SNPs by using Sequenom mass array technology (Additional File 1: Table S10).

Ancient DNA samples were excavated from Sanjan, Gujarat in 2001 (Additional File 1: Supplementary Text). Archaeological analysis and AMS dating were consistent with these remains belonging most likely to the migrant Parsis during 8-13th Century (Supplementary text). The teeth obtained from 21 of these specimens were brought to the ancient DNA (aDNA) laboratory of the CSIR-Centre for Cellular and Molecular Biology, Hyderabad, India for the extraction and genotyping. We followed our standard published protocol to isolate DNA from the teeth (47) (Additional File 1: Supplementary text).

For autosomal analyses, after data curation and merging (Additional File 1: Supplementary text), we first used method of Cockerham and Weir (48) to estimate mean pairwise *F*st. Further, we performed PC analysis on pruned data using the smartpca v. 7521 (28)

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

program (with default settings). We also used the F_{ST}:Yes method of smartpca to calculate the F_{ST} with standard errors. We ran unsupervised ADMIXTURE v.1.23, 25 times each for K=2 to K=12, and used the method described previously to choose the best K value (31). The F statistics were performed by the ADMIXTOOL package v.3 (35) and the haplotype-based analysis was performed by Chromopainter and fineSTRUCTURE v.1 (39). The ML tree of world populations was constructed using TreeMix v.1.12 (38) and the RoH were calculated using PLINK 1.9 (49). ALDER v 1.03 (42) and MALDER v.1.0 (43) were used to calculate the time and number of admixture events. The population branch statistic method (PBS) (44) was used to identify genomic regions under selection in the Parsi population.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Declarations

Acknowledgements and funding

We are thankful to Dr. Shernaz Cama, Director Parzor Foundation, India for her help and critical comments. Tuuli Reisberg for technical assistance. Support was provided by: Estonian Personal grants PUT-766 (GC, MK, AKP); EU European Regional Development Fund through the Centre of Excellence in Genomics to Estonian Biocentre and Project No. 2014-2020.4.01.15-0012, and Estonian Institutional Research grants IUT24-1 (RV, MM, SR, EM, MR, TK); The Council of Scientific and Industrial Research (CSIR), Government of India (KT); Wellcome Trust grant 098051 (QA and CTS); WZCF (World Zarathushti Cultural Foundation), Parsi foundation and IAS, Indian Archaeological Society (VMT); PIRSES-GA-2012-318979 grant (MK). ERC Starting Investigator grant (FP7 - 261213) (T.K.). AKP was supported by the European Social Fund's Doctoral Studies and Internationalisation Programme DoRa. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The analyses were performed in the High Performance Computer Center of the University of Tartu, Estonia and the Wellcome Trust Sanger Institute, Hinxton, UK.

Data availability

The data are available at NCBI GEO Accession no GSE97086 and the data repository of the Estonian Biocentre: www.ebc.ee/free_data.

Author contributions

GC, QA, RV, KT and NR conceived the project and designed the experiments. GC, QA, NR, SP, AKP, RT, SF, MR, MK, DSR, AGR, EM, SR, and SK carried out haploid DNA genotyping and analysis. GC, QA, MaM and TK analysed autosomal data. VMT, KD and SRW performed archaeological study. VMT and NR collected ancient samples. NR isolated DNA from ancient remains and performed genotyping. SQM, LS, MM, CTS, RV and KT provided samples and reagents. GC, QA, NR, VMT and MaM wrote the manuscript with the help of TK, CTS, RV, MM and KT.

Competing interests

None.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References

1. Palsetia JS. The Parsis of India. BRILL; 2001. 368 p.

2. Axelrod P. myth and identity in the indian Zoroastrian community. Journal of Mithraic Studies 1980;3(1-2):150-165.

3. Mirza HDK. Outlines of Parsi history. Mirza; Mumbai, 1974.

4. Hinnells J, Williams A. Parsis in India and the Diaspora. Routledge; 2007. 304 p.

5. Hinnells, J.R, The modern Zoroastrian diaspora In: Migration: the Asian experience, edited by Judith M. Brown and Rosemary Foot. New York, New York, St. Martin's Press, 1994. 56-82.

6. Roy, T.K.; Unisa, S.; Bhatt, M.,Growth of Parsi Population in India. Mumbai: National Commission for Minorities, 2004.

7. Katrak SK. Who are the Parsees? printed at Pakistan Herald Press; 1965.

8. Zoroastrian Demographics (Internet). (updated 2005; cited 04.04.2016). Available from: http://heritageinstitute.com/zoroastrianism/demographics/index.htm

9. Mohyuddin A, Mehdi S. HLA analysis of the Parsi (Zoroastrian) population in Pakistan. Tissue Antigens 2005;66(6):691-695.

10. Pocha JS. Parsis: The Vanishing Breed. Archaeological and Cultural News of the Iranian World, CAIS London, 2007.

11. Modī B. Parsi Gujarati: Vanishing Dialect: Vanishing Culture. Lincom Europa; 2011.

12. Shroff Z, and Castro MC. The potential impact of intermarriage on the population decline of the Parsis of Mumbai, India. DEMOGRAPHIC RESERACH 2011;25:545-564.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

13. Sekar CC. Some aspects of Parsi demography. Hum Biol 1948;20(2):47-89.

14. Paymaster RB. Early History of the Parsees in India from their landing in Sanjan to 1700 AD. Zartoshti Dharam Sambandhi Kelavni Apnari ane Dnyan Felavnari Mandli; 1954.

15. Gupta S, Dalal KE, Dandekar A, Nanji R, Aravazhi P, Bomble S, et al. On the Footsteps of Zoroasition Parsis in lndia Excavations of Sanjan on the West Coast 2004.

16. Gupta S, Dalal K, Dandekar A, Nanji R, Mitra R, Pandey R, et al. A Preliminary Report on the Excavations at Sanjan. Puratattva 2002;32:182-198.

17. Mushrif-Tripathy V, Walimbe SR. Human Skeletal Remains from the Medieval Site of Sanjan: Osteobiographic Analysis. Archaeopress; 2012.

18. Singh Nancy. The sugar in the milk: the Parsis in India. India: Institute for Development Education (Madras, India)., Madras Parsi Zarthosti Anjuman; 1986. 149 p.

19. Karkal M. Marriage among Parsis. Demography India 1975;4(1):128-145.

20. Karkal M. Survey of Parsi Population of Greater Bombay. Bombay: Bombay Parsi Panchayat 1982.

21. Jonnalagadda M, Ozarkar S, Mushrif-Tripathy V. Population affinities of Parsis in the Indian subcontinent. Int. J. of Osteoarchaeology 2011;21(1):103-110.

22. Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, et al. Y-chromosomal DNA variation in Pakistan. Am J Hum Genet 2002;70(5):1107-1124.

23. Quintana-Murci L, Chaix R, Wells RS, Behar DM, Sayar H, Scozzari R, et al. Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. Am J Hum Genet 2004;74(5):827-845.

24. Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MG, Nino-Rosales L, Ninis V, et al. Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet 2006;2(12):e215.

25. Undevia JV. Population Genetics of the Parsis. Field Museum , Florida, USA; 1972.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

26. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science 2008;319

27. Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature 2014;505(7481):87-91.

28. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2(12):e190.

29. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res 2009;19(9):1655-1664.

30. Chaubey G, Metspalu M, Choi Y, Mägi R, Romero IG, Soares P, et al. Population Genetic Structure in Indian Austroasiatic speakers: The Role of Landscape Barriers and Sex-specific Admixture. Mol Biol Evol 2011;28(2):1013-1024.

31. Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet 2011;89(6):731-744.

32. Haber M, Gauguier D, Youhanna S, Patterson N, Moorjani P, Botigué LR, et al. Genome-Wide Diversity in the Levant Reveals Recent Structuring by Culture. PLOS Genetics 2013;9(2):e1003316-e1003316.

33. Grugni V, Battaglia V, Hooshiar Kashani B, Parolo S, Al-Zahery N, Achilli A, et al. Ancient migratory events in the middle East: new clues from the y-chromosome variation of modern iranians. PloS one 2012;7(7):e41252.

34. Derenko M, Malyarchuk B, Bahmanimehr A, Denisova G, Perkova M, Farjadian S, et al. Complete mitochondrial DNA diversity in Iranians. PloS one 2013;8(11):e80673.

35. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics 2012;192(3):1065-1093.

36. Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, et al. Early Neolithic genomes from the eastern Fertile Crescent. Science 2016.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

37. Lazaridis I, Nadel D, Rollefson G, Merrett DC, Rohland N, Mallick S, et al. Genomic insights into the origin of farming in the ancient Near East. Nature 2016.

38. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 2012;8(11):e1002967.

39. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet 2012;8(1):e1002453.

40. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, et al. Runs of homozygosity in European populations. Am J Hum Genet 2008;83(3):359-372.

41. Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF, et al. Genomic runs of homozygosity record population history and consanguinity. PloS one 2010;5(11):e13996.

42. Loh PR, Lipson M, Patterson N, Moorjani P, Pickrell JK, Reich D, et al. Inferring admixture histories of human populations using linkage disequilibrium. Genetics 2013;193(4):1233-1254.

43. Pickrell JK, Patterson N, Loh PR, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci U S A 2014;111(7):2632-2637.

44. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science 2010;329(5987):75-78.

45. . The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 2015;348(6235):648.

46. Eskandari-Nasab E, Moghadampour M, Najibi H, Hadadi-Fishani M. Investigation of CTLA-4 and CD86 gene polymorphisms in Iranian patients with brucellosis infection. Microbiol Immunol 2014;58(2):135-141.

47. Rai N, Taher N, Singh M, Chaubey G, Jha AN, Singh L, et al. Relic excavated in western India is probably of Georgian Queen Ketevan. Mitochondrion 2014;14(1):1-6.

48. Cockerham CC, Weir BS. Covariances of relatives stemming from a population undergoing mixed self and random mating. Biometrics 1984;40(1):157-164.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

49. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. BMC Biol 2015;4(1):1-16.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Table 1. The South Indian (SIND), and Iranian ancestry among Parsis and neighbouring populations

Population (X)	SIND (SE)	Iranian (SE)
Pathan	31 (2.4)	56.5 (1.8)
Sindhi	22.9 (2.7)	63.1 (1.9)
Parsi_Pakistan	6.4 (2.4)	76.6 (1.5)
Parsi_India	8.8 (2.5)	74.6 (1.7)
Gujarati	58.1 (2)	34.3 (1.6)

South Indian ancestry= Yoruba,Papua;X,French/Yoruba,Papua;South_India,French
Iranian ancestry= Yoruba,Papua;X,South_India/Yoruba,Papua;Iranian,South_India

Table 2. The test of geneflow (*D* statistics) between Parsis, modern Iranian, Neolithic Iranian, Sindhi and Gujarati

Gp1	Gp2	Gp3	D value	Z score
Parsi_India	Parsi_Pakistan	Sindhi	0.0309	35.229
Parsi_India	Parsi_Pakistan	Iranians	0.0318	42.175
Parsi_India	Gujaratis	Sindhi	-0.0015	-1.89
Parsi_Pakistan	Parsi_India	Sindhi	0.0309	34.184
Parsi_Pakistan	Parsi_India	Iranians	0.0313	39.061
Parsi_Pakistan	Gujaratis	Sindhi	-0.002	-2.482
Sindhi	Parsi_India	Parsi_Pakistan	0	-0.036
Sindhi	Parsi_India	Gujaratis	-0.0038	-4.609
Sindhi	Parsi_Pakistan	Gujaratis	-0.0038	-4.688
Sindhi	Iranians	Parsi_Pakistan	-0.0124	-16.085
Sindhi	Iranians	Parsi_India	-0.0124	-15.308
Sindhi	Iranians	Gujaratis	-0.016	-17.608
Gujaratis	Parsi_Pakistan	Parsi_India	-0.0005	-0.898
Gujaratis	Iranians	Parsi_Pakistan	-0.0154	-21.245
Gujaratis	Iranians	Sindhi	-0.0211	-22.669
Gujaratis	Parsi_India	Sindhi	-0.0054	-5.982
Gujaratis	Parsi_Pakistan	Sindhi	-0.0058	-6.7
Iran_Neolithic	Sindhi	Iranians	-0.0002	-0.141
Iran_Neolithic	Gujaratis	Iranians	-0.008	-5.075
Iran_Neolithic	Parsi_India	Iranians	0.005	3.2
Iran_Neolithic	Parsi_Pakistan	Iranians	0.0058	4.083

D= (Gp1,Yoruba;Gp2,Gp3)

Table 3. The formal text of admixture using ALDER method

Reference 1	Reference 2	Admixed	Gen. Time	p value	Z score
Iranian	Gujarati	Parsi_India	38.26 ± 12.16	**0.0017**	**3.15**
Iranian	Sindhi	Parsi_India	32.96 ± 9.42	0.013	2.48
Iranian	Sindhi	Parsi_Pakistan	41.32 ± 8.93	**1.7×10^{-5}**	**4.3**
Iranian	Gujarati	Parsi_Pakistan	30.74 ± 14.04	0.029	2.19

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure Legends:

Fig. 1. The geographical distribution and sampling locations of modern and ancient Parsi samples. The population data is obtained from Parzor foundation, New Delhi, India.

Fig. 2. Principal component analysis of the combined autosomal genotypic data of individuals from Eurasia. The inset shows a plot of mean eigenvalues of Parsi and other closeby populations.

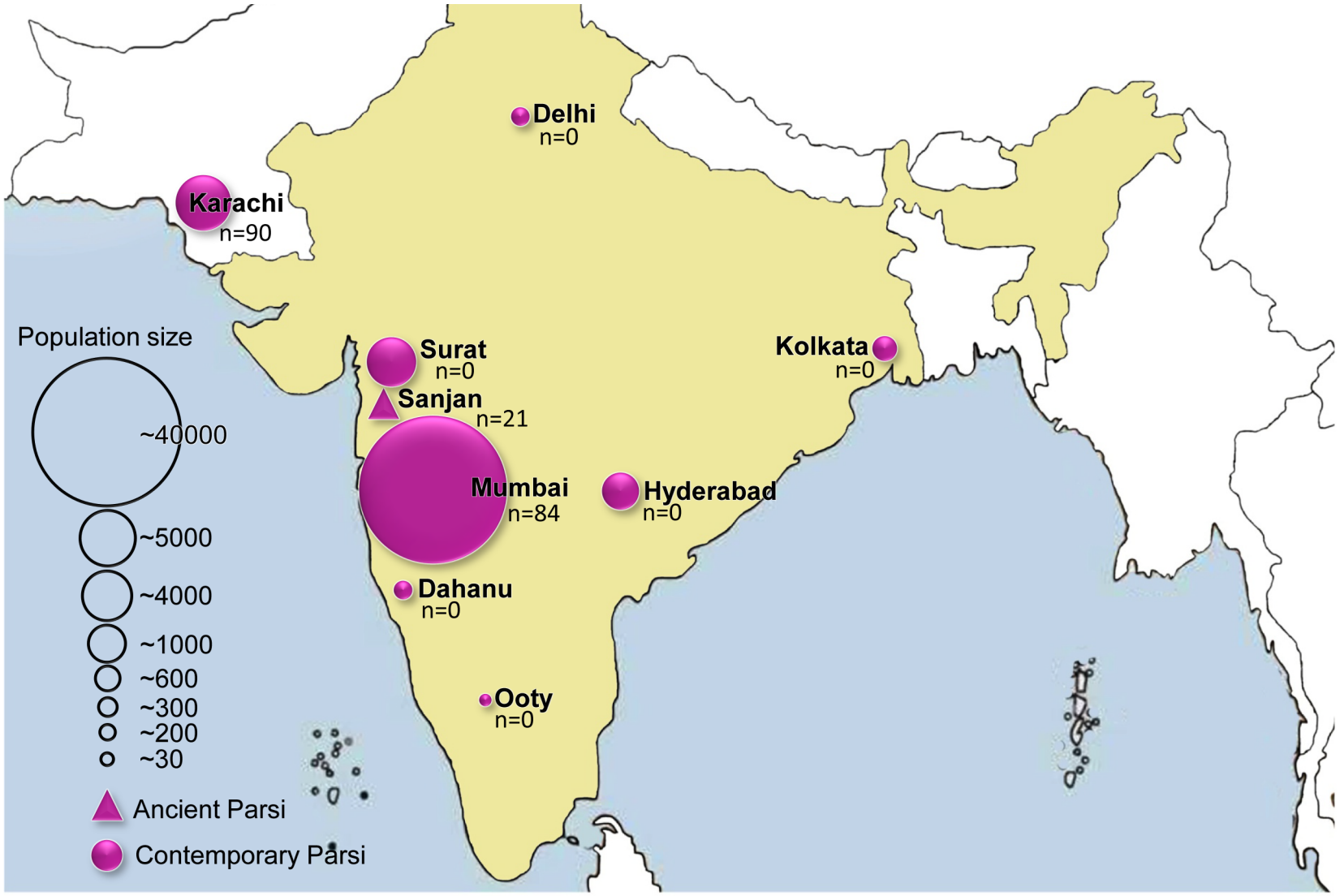
Fig. 3. Results of ADMIXTURE analysis (k = 8) of world populations with a zoom-in on Iranian, Parsis and other South Asian populations.

Fig. 4. Principal component analysis using haplogroup frequencies for a) mtDNA, and b) Y-chromosome, **b)** in Parsis, Iranian, Central Asian, Pakistani and Indian populations. The details of populations is given in supplementary text.

Legends Additional files

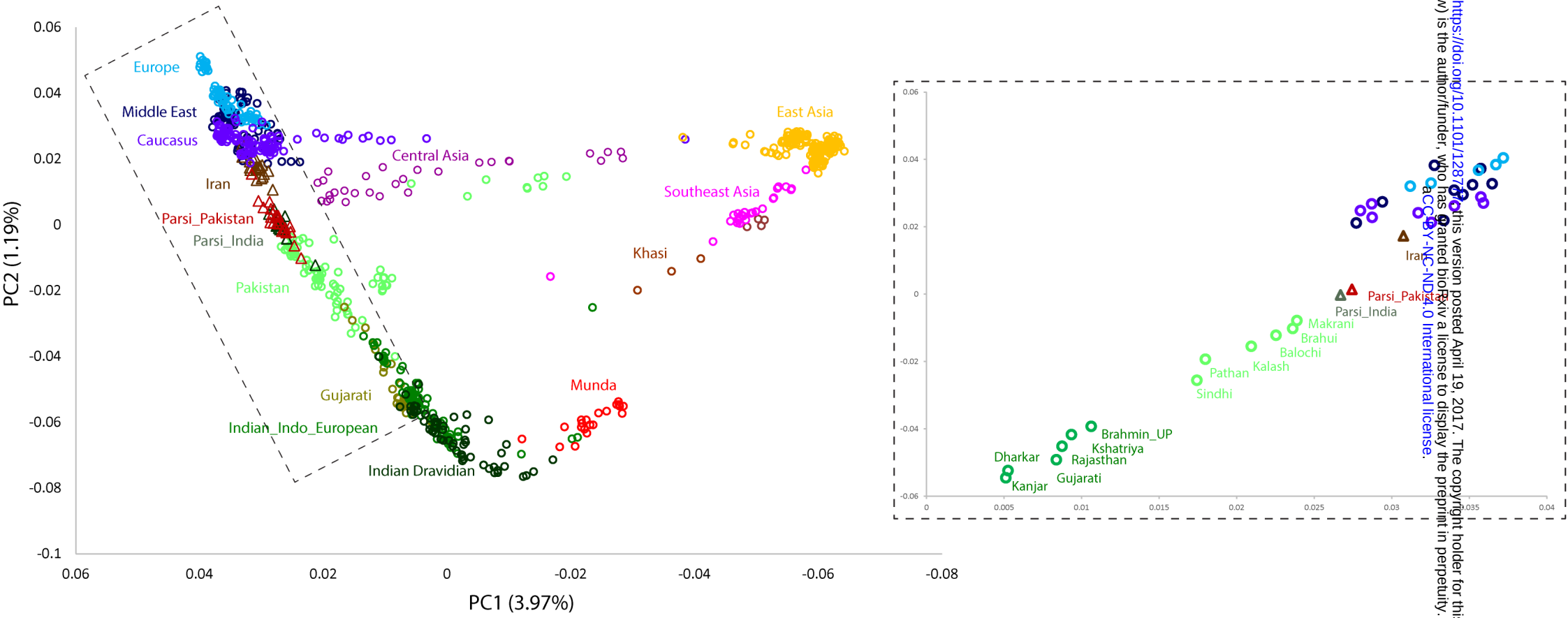
Additional File 1: Supplementary text explaining the archaeological details of ancient samaples; isolation of ancient DNA, genotyping and statistical analyses. 12 figures and 11 tables are also incorporated in this file.

Fig. 1



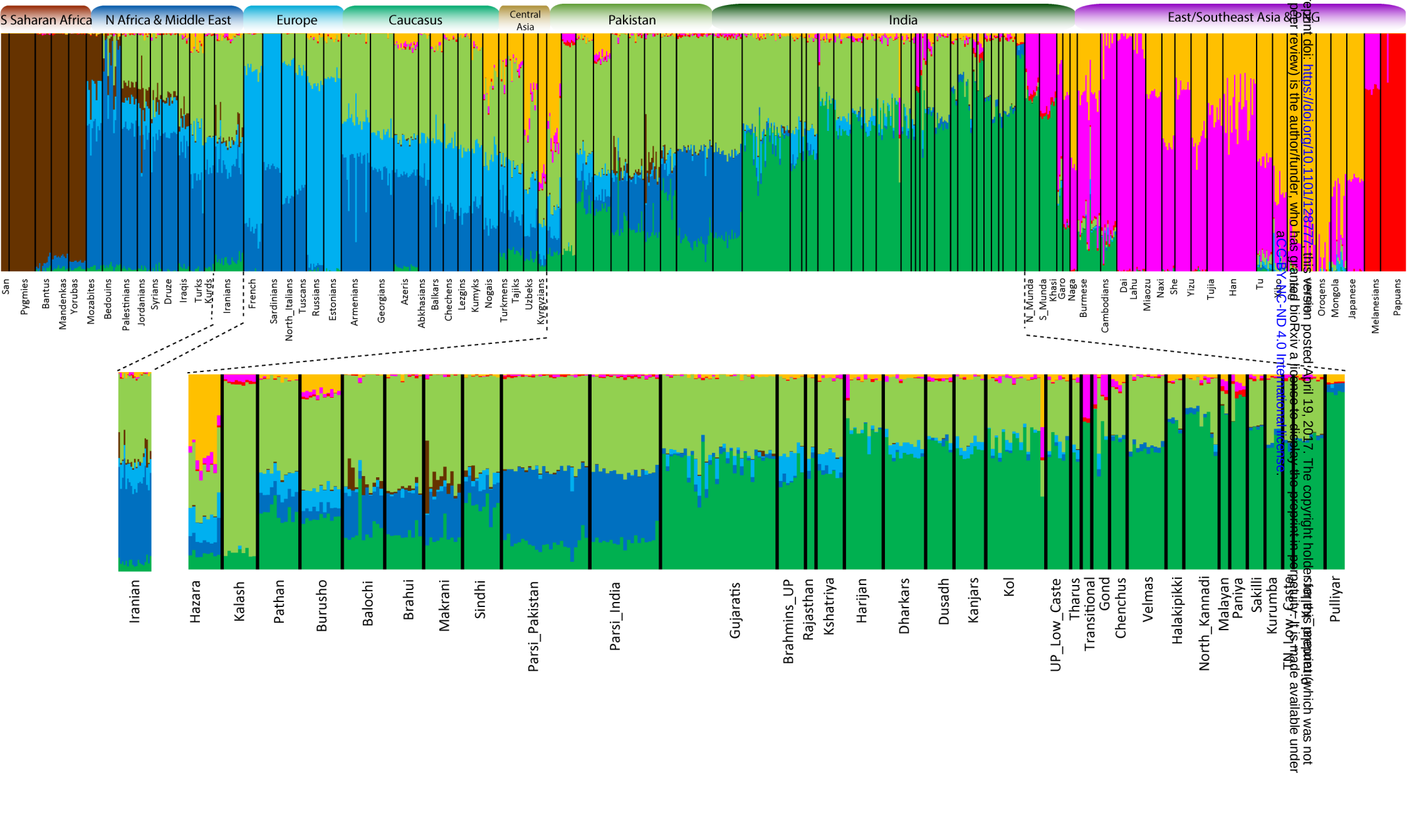
bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 2



bioRxiv preprint doi: https://doi.org/10.1101/128702 this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Fig. 3



bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

bioRxiv preprint doi: https://doi.org/10.1101/128777; this version posted April 19, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

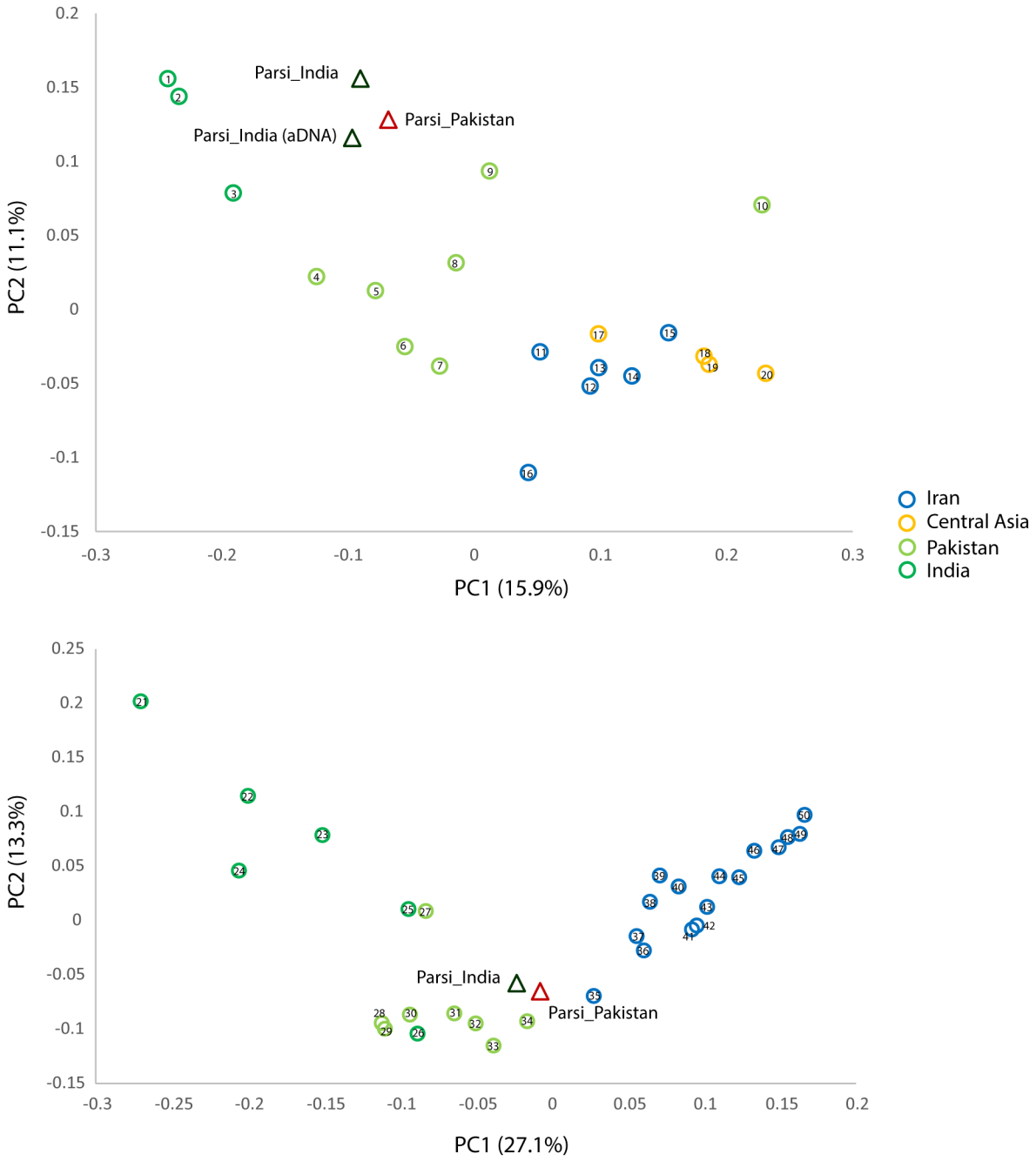



Fig. 4