

# The Genetic Legacy of Zoroastrianism in Iran and India: Insights into Population Structure, Gene Flow, and Selection

Saioa López,<sup>1,11,12,\*</sup> Mark G. Thomas,<sup>1,11</sup> Lucy van Dorp,<sup>1,2</sup> Naser Ansari-Pour,<sup>3</sup> Sarah Stewart,<sup>4</sup> Abigail L. Jones,<sup>5</sup> Erik Jelinek,<sup>1</sup> Lounès Chikhi,<sup>6,7</sup> Tudor Parfitt,<sup>8</sup> Neil Bradman,<sup>9</sup> Michael E. Weale,<sup>10</sup> and Garrett Hellenthal<sup>1,\*</sup>

Zoroastrianism is one of the oldest extant religions in the world, originating in Persia (present-day Iran) during the second millennium BCE. Historical records indicate that migrants from Persia brought Zoroastrianism to India, but there is debate over the timing of these migrations. Here we present genome-wide autosomal, Y chromosome, and mitochondrial DNA data from Iranian and Indian Zoroastrians and neighboring modern-day Indian and Iranian populations and conduct a comprehensive genome-wide genetic analysis in these groups. Using powerful haplotype-based techniques, we find that Zoroastrians in Iran and India have increased genetic homogeneity relative to other sampled groups in their respective countries, consistent with their current practices of endogamy. Despite this, we infer that Indian Zoroastrians (Parsis) intermixed with local groups sometime after their arrival in India, dating this mixture to 690–1390 CE and providing strong evidence that Iranian Zoroastrian ancestry was maintained primarily through the male line. By making use of the rich information in DNA from ancient human remains, we also highlight admixture in the ancestors of Iranian Zoroastrians dated to 570 BCE–746 CE, older than admixture seen in any other sampled Iranian group, consistent with a long-standing isolation of Zoroastrians from outside groups. Finally, we report results, and challenges, from a genome-wide scan to identify genomic regions showing signatures of positive selection in present-day Zoroastrians that might correlate to the prevalence of particular diseases among these communities.

## Introduction

Zoroastrianism developed from an ancient religion that was once shared by the ancestors of tribes that settled in Iran and northern India and is thought to have been founded by the prophet priest Zarathustra (Zoroaster in Greek). Since there is no context or documentation for the life of Zarathustra, his very existence is a matter for debate. Although some scholars have proposed that he lived in the 6<sup>th</sup> century BCE, i.e., during the Achaemenic period, most scholars now believe he lived around 1200 BCE, at a time when the ancient Iranians inhabited the areas of the Inner Asian Steppes (also a subject of great controversy<sup>1,2</sup>) prior to the great migrations south to modern Iran, Afghanistan, Northern Iraq, and parts of Central Asia.<sup>3</sup> Zoroastrianism became the state religion of three great Iranian empires: Achaemenid (559–330 BCE) founded by King Cyrus the Great and ended by the conquest of Alexander the Great, Parthian (c. 247 BCE–224 CE), and Sasanian (224–651 CE), during which time the religion as an imperial faith is best known. Zoroastrianism ceased to be the state religion of Iran after the Arab conquests (633–654 CE), although it is thought

that widespread conversion to Islam did not begin until about 767 CE.<sup>4</sup>

According to Parsi (i.e., Indian Zoroastrians) tradition, a group of Zoroastrians set sail from Iran to escape persecution by the Muslim majority. They landed on the coast of Gujarat (India) where they were permitted to stay and practice their religion. The date of the arrival has been the cause of speculation and varies between 785 CE<sup>5</sup> and 936 CE.<sup>6</sup> These dates, among others, are based on the *Qisseh-ye Sanjan*, a legendary account of the journey by sea from Iran and settlement in India.<sup>7</sup> However, maritime trade is known to have taken place between ethnic groups from Iran, including Zoroastrians, and peoples in India long before the arrival of Islam.<sup>8</sup> Down the subsequent centuries, the Indian Zoroastrians maintained contact with the Zoroastrians of Iran and later became an influential minority under British Colonial rule.

Zoroastrian communities today (2011 census) are concentrated in India (61,000 people), Southern Pakistan (1,675), and Iran—mainly in Tehran, Yazd, and Kerman—(14,000). In the last 200 years Zoroastrians, both Parsi and Irani, have formed diaspora communities in North America (14,306), Canada (6,422), Britain (5,000), Australasia (3,808), and the Middle East (2,030).

<sup>1</sup>Department Genetics, Evolution & Environment, University College London, London WC1E 6BT, UK; <sup>2</sup>Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London, London WC1E 6BT, UK; <sup>3</sup>Faculty of New Sciences and Technology, University of Tehran, Tehran 14395-1561, Iran; <sup>4</sup>SOAS, University of London, London WC1H 0XG, UK; <sup>5</sup>Liverpool Women's Hospital, Liverpool L8 7SS, UK; <sup>6</sup>CNRS, Université Paul Sabatier, Toulouse 31062, France; <sup>7</sup>Instituto Gulbenkian de Ciência, Oeiras 2780-156, Portugal; <sup>8</sup>Florida International University, Miami, FL 33199, USA; <sup>9</sup>Henry Stewart Group, London WC1A 2HN, UK; <sup>10</sup>Department Medical & Molecular Genetics, King's College London, London SE1 9RT, UK

<sup>11</sup>These authors contributed equally to this work

<sup>12</sup>Twitter: @Saioa\_l

\*Correspondence: saioa.lopez@ucl.ac.uk (S.L.), g.hellenthal@ucl.ac.uk (G.H.)

<http://dx.doi.org/10.1016/j.ajhg.2017.07.013>

© 2017 The Authors. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Zoroastrianism is a non-proselytizing religion, with a hereditary male priesthood of uncertain origins.<sup>9</sup> Among the Parsis, priestly families are distinguished from the laity. Priestly status is patrilineal, although there is also a strong matrilineal component with the daughters of priests encouraged to marry into priestly families. Remarkably, many priests preserve family genealogies that can be traced back to the purported time of arrival of Iranian Zoroastrians in India and beyond to an Iranian homeland.

Genetic data provide a means of examining the biological relationships of different populations and testing claims of common ancestry. Previous studies of Iranian Zoroastrians have suggested that they are genetically differentiated from their neighboring populations. For example, Farjadian et al.<sup>10</sup> analyzed mitochondrial DNA (mtDNA) variation in 14 different ethnic groups from Iran and observed that Zoroastrians and Jews were genetically distinct from other groups. In the same vein, Lashgary et al.<sup>11</sup> analyzed 14 bi-allelic loci from the non-recombining region of the Y chromosome (NRY) and observed a notable reduction in haplogroup diversity in Iranian Zoroastrians compared with all other groups. Furthermore, a recent study using genome-wide autosomal DNA found that haplotype patterns in Iranian Zoroastrians matched more than other modern Iranian groups to a high-coverage early Neolithic farmer genome from Iran.<sup>12</sup>

Less is known about the genetic landscape and the origins of Zoroastrianism in India, despite Parsis representing more than 80% of present-day Zoroastrians worldwide.<sup>13</sup> A study of four restriction fragment length polymorphisms (RFLPs) suggested a closer genetic affinity of Parsis to Southern Europeans than to non-Parsis from Bombay.<sup>14</sup> Furthermore, NRY haplotype analysis<sup>15</sup> and patterns of variation at the HLA locus<sup>16</sup> in the Parsis of Pakistan support a predominately Iranian origin.

Prompted by these observations, we explored the genetic legacy of Zoroastrianism in more detail by generating genome-wide autosomal and Y/mtDNA genotype data for Iranian and Indian Zoroastrian individuals and comparing them to other publicly available genetic data. With this study we aimed to address the following questions.

1. Which demographic processes (e.g., admixture, isolation) have contributed most to present-day genetic differences between Zoroastrian and non-Zoroastrian groups sampled from the same country?
2. Is there evidence of increased genetic homogeneity in Zoroastrian groups relative to non-Zoroastrian groups?
3. Is there genetic evidence for admixture events in Zoroastrian groups? Does this support historical records tracing the origin of Indian Zoroastrians (a.k.a. Parsis) to migrants from Persia?
4. Is there evidence that any inferred admixture events were sex biased?
5. Is there genetic evidence for the patrilineal descent of Parsi priests from a small number of founding individuals?

6. Using a genome-wide scan, can we reliably detect genomic signatures of positive selection in the Zoroastrian populations that may relate to the prevalence of diseases or other phenotypic traits in the community?

## Subjects and Methods

### Samples

Buccal swabs were collected from a total of 526 men from India, Iran, the United Arab Emirates, and the United Kingdom (see [Table S1](#)). Individuals sampled in the United Arab Emirates are mainly first-generation Parsis who left Aden following the communist coup in 1970, after which Asians were expelled (Aden was part of the Bombay Presidency until 1947 and the British left Aden in 1967–1968). Individuals sampled from the United Kingdom Zoroastrian population are mainly descendants of 19<sup>th</sup> century immigrants; the Zoroastrian Association was formed in 1861, at which time there were around 50 Zoroastrians living in the UK.<sup>17</sup> Swabs were stored in a DNA preservative solution containing 0.5% sodium doecyl sulfate and 0.05 M EDTA for transport purposes and DNA was purified by phenol-chloroform extraction/isopropanol precipitation. Work on the genetics of these samples is covered by the UK ethics committee London Bentham REC (formally the Joint UCL/UCLH Committees on the Ethics of Human Research: Committee A and Alpha, REC reference number 99/0196, Chief Investigator M.G.T.). The procedures followed were in accordance with the ethical standards of London Bentham REC and Zoroastrian Studies (Bombay). Informed consent was obtained from all individuals before samples were taken.

### Genome-wide Genotyping with the Human Origins Array

71 of these individuals (29 Iranian Zoroastrians, 17 Iranian Farsi, 13 Indian Zoroastrians, and 12 Indian Hindu), all belonging to the lay (i.e., non-priest) population, were genotyped using the Affymetrix Human Origins array. This array targets 627,421 SNPs with well-documented ascertainment, although we note that our inference techniques use haplotype information that has been shown to be less affected by ascertainment bias.<sup>18,19</sup> SNPs and individuals were pruned to have genotyping rate greater than 0.95 using PLINK v.1.9.<sup>20</sup> Genotypes for the Indian Zoroastrian and Indian Hindu individuals are available in plink format (see [Web Resources](#)). Genotypes for the Iranian Zoroastrians and the Iranian Farsi were made publicly available by Broushaki et al.<sup>12</sup>

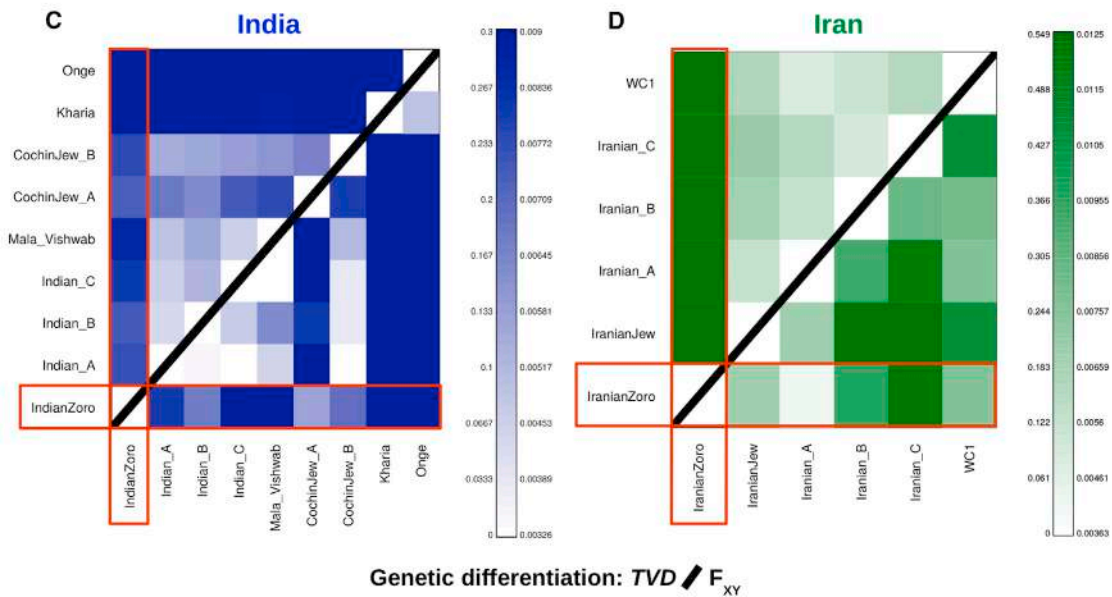
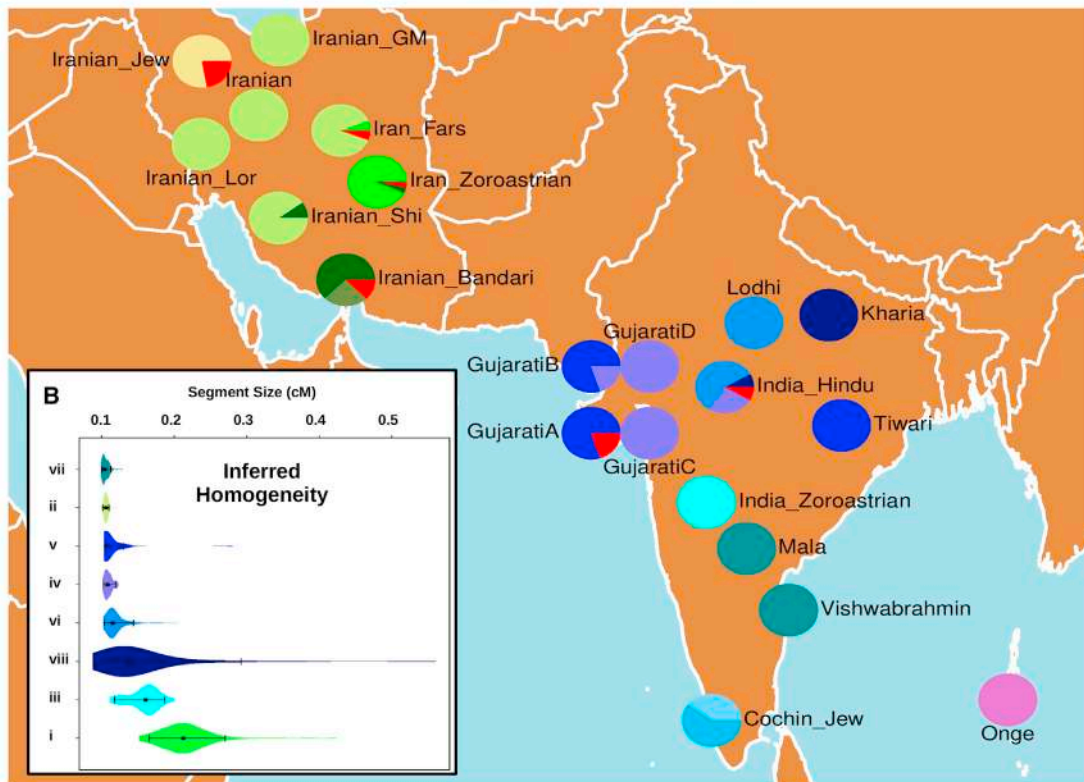
The above dataset was merged with modern populations in the Human Origins dataset of Lazaridis et al.,<sup>21</sup> which includes 17 labeled populations from India and Iran (here we use their same labels throughout). We also included other high-coverage ancient samples: an Early Neolithic individual from Iran (WC1),<sup>12</sup> a Mesolithic hunter-gatherer from Luxembourg (Loschbour), Neolithic individuals from Germany (LBK), Anatolia (Bar8<sup>22</sup>), Georgia (KK1<sup>23</sup>), and Hungary (NE1<sup>24</sup>), a 4,500-year-old genome from Ethiopia (Mota<sup>25</sup>), and a 45,000-year-old genome from western Siberia Ust-Ishim.<sup>26</sup> In total, the merge contained 2,553 individuals and 525,796 overlapping SNPs.

### Principal Component Analysis (PCA)

We performed PCA on all the South Asian and West European populations included in the merge using PLINK 1.9

**A** *fineSTRUCTURE* clusters:

- (i) **Iranian Zoroastrians (28)**
- (ii) **Iranian\_A (54)**
- Iranian\_B (7)**
- Iranian\_C (2)**
- Iranian Jews (7)**
- (iii) **Indian Zoroastrians (13)**
- (iv) **Indian\_A (16)**
- (v) **Indian\_B (24)**
- (vi) **Indian\_C (25)**
- (vii) **Mala\_Vishwabrahmin (26)**
- (viii) **Kharia (13)**
- Onge (11)**
- Cochin Jews\_A (2)**
- Cochin Jews\_B (3)**
- Others**



**Figure 1. Clustering, Homogeneity, and Genetic Differentiation of the Iranian and Indian Populations**

(A) Each color inside the pies represents the proportion of individuals from each population label that is assigned to each *fineSTRUCTURE* cluster (“Others” include all clusters containing primarily individuals outside Iran and India), with the total number of individuals included in each cluster shown inside brackets in the legend.

(legend continued on next page)

after linkage disequilibrium (LD) pruning using—indep-pairwise 50 5 0.5.

## Phasing

We jointly phased the autosomal chromosomes for all individuals in the merge using SHAPEIT<sup>27</sup> with default parameters and the linkage disequilibrium-based genetic map build 37.

## Chromosome Painting and fineSTRUCTURE

We classified our 2,545 modern individuals into 230 groups, with the majority of these groups based on population labels.<sup>21</sup> The exceptions to this are the individuals from Iran and India and neighboring populations (see Figure S1) who were re-classified into new, genetics-based groups using results from the clustering algorithm fineSTRUCTURE that groups individuals into genetically homogeneous clusters based entirely on patterns of shared ancestry identified by CHROMOPAINTER.<sup>28</sup> Briefly, CHROMOPAINTER uses a “chromosome painting” approach that compares patterns of haplotype sharing between each recipient chromosome and a set of donor chromosomes.<sup>28</sup> For the CHROMOPAINTER analysis used for our fineSTRUCTURE analysis, which is the first painting protocol described in this paper and referred to throughout as the “fineSTRUCTURE painting,” we painted each of the 696 individuals from the populations described above using all other 695 individuals as donors. Here we initially estimated the mutation/emission (Mut, “-M”) and switch rate (Ne, “-n”) parameters using 10 steps of the Expectation-Maximization (E-M) algorithm, for chromosomes 1, 4, 15, and 22, and for every 10 individuals (following van Dorp et al.<sup>29</sup> and Broushaki et al.<sup>12</sup>), which gave estimated Mut and Ne of 0.00091 and 320.9197, respectively. These values were then fixed before running CHROMOPAINTER across all chromosomes to produce a “painting profile” giving the proportion of genome-wide DNA each individual shares with each other donor individual in this analysis. All chromosomes were then combined to estimate the fineSTRUCTURE normalization parameter “c,” which was 0.28. Following Leslie et al.,<sup>30</sup> we then ran fineSTRUCTURE using this c value and performing 2,000,000 iterations of Markov-Chain-Monte-Carlo (MCMC), sampling an inferred clustering every 10,000 iterations. Following the recommended approach described by Lawson et al.,<sup>28</sup> we next used fineSTRUCTURE to find the single MCMC sampled clustering with highest posterior probability and performed 100,000 additional hill-climbing steps to find a nearby state with even higher posterior probability. This hill-climbing approach grouped these 695 individuals into 207 clusters, which we then merged into a tree using fineSTRUCTURE’s greedy algorithm that merges pairs of clusters, one step at a time, until only two super-clusters remain.

Based on this tree, exploration of genetic similarity among individuals (see below), and visual inspection of haplotype sharing patterns among our 207 clusters, we classified these clusters into genetically homogeneous groups, choosing a level of the tree where there were 50 final total clusters, among which Iranian Zoroastrians and Parsis each formed their own distinct groups. At this level of the tree, we note that the 10,000-year-old Neolithic Ira-

nian WC1 clustered with other modern Iranians, but nonetheless we re-classified WC1 as its own cluster, so that we ended up with 51 final total clusters we use throughout this paper (see Table S2, Figure S1).

We then painted all 230 modern and 8 ancient samples using all 230 modern groups as donors, following the “leave-one-out” approach, as described by Hellenthal et al.,<sup>31</sup> which is designed to make painting profiles comparable. In particular, if each donor group  $\{1, \dots, K\}$  contains  $\{n_1, \dots, n_K\}$  individuals, respectively, the set of donors is fixed to contain  $n_k - 1$  individuals from each of the  $K$  groups. This is to account for the fact that individuals cannot be painted using themselves as a donor, so that individuals within each of these  $K$  donor groups can only ever be painted using  $n_k - 1$  individuals from their own group label. We refer to this second painting protocol where  $K = 230$  as the “all donors painting” throughout. Note that a primary difference between this painting and the fineSTRUCTURE painting described above is that we now use group labels, based in part on clustering results, which are required for our leave-one-out approach. When using haplotype information for this painting, we initially estimated the mutation/emission and switch rate parameters as described above, giving estimated Mut and Ne of 0.000704 and 223.5674, respectively.

We also performed a slightly different version of this painting where Iranian and Indian populations were excluded as donors, using the leave-one-out approach described above for all 216 remaining groups, a third painting protocol with  $K = 216$  that we refer to throughout as the “non-Indian/Iranian donors painting.” We did this to mitigate genetic effects due to recent isolation in our Indian and Iranian groups when inferring the best surrogates to sources of admixture in our groups. Mut and Ne parameters (0.00069 and 225.32, respectively) were re-estimated for this new scenario as described above. To help determine whether admixture from outside groups or independent drift effects due to genetic isolation are driving genetic differences among sampled groups within Iran and India<sup>29,30</sup> (i.e., lower right triangles of Figures 1C, 1D, S2, and S3), we further included all non-Indian groups as donors when painting Indian groups (“non-Indian painting”) and all non-Iranian, non-Parsi groups when painting Iranian groups (“non-Iranian/Parsi painting”). This is because in this paper we infer Iranians as important contributors to the DNA of Parsis, making them vital to include when evaluating whether genetic differences among Indian groups are due to admixture.

## TVD, $F_{XY}$ , and $F_{ST}$ between Iranian and Indian Groups

We quantified differences in the painting profiles between all Iranian and Indian groups by applying the metric total variation distance (TVD) as described in Leslie et al.<sup>30</sup> using the formula:

$$TVD_{XY} = 0.5 \sum_{k=1}^K (f_k^X - f_k^Y).$$

where  $f_k^X$  and  $f_k^Y$  are the average genome-wide proportion of DNA that individuals from the recipient groups  $X$  and  $Y$ , respectively, match to donor group  $k \in [1, \dots, K]$  as inferred by CHROMOPAINTER. We also inferred pairwise TVD among

---

(B) Distribution of CHROMOPAINTER’s inferred lengths of haplotype segments (in cM) copied intact from a single donor, when allowing 13 randomly sampled individuals from each group (roman numerals in A) to copy from the other 12 individuals with the same label. (Black dot = median values, bars = 95% empirical quantiles across individuals.)

(C and D) Comparison of pairwise TVD based on the all donors painting (upper triangle) and  $F_{XY}$  mitigating recent drift effects (lower triangle) based on the (C) non-Indian donors painting for India and (D) non-Iranian/Parsi donors painting for Iranian groups.



individuals assigned to the same group as defined by our final 51 fineSTRUCTURE-inferred clusters, results of which are discussed in the previous section.

Independent drift effects in groups X and Y since their split can generate genetic differences between them without requiring any outside introgression since this split. To elucidate whether inferred genetic differences are more attributable to ancestry from outside sources, we followed the approach in van Dorp et al.<sup>29</sup> designed to mitigate these drift effects. In particular we calculated:

$$F_{XY} = \frac{TVD_{XY}}{0.5(\widetilde{TVD}_X + \widetilde{TVD}_Y)}$$

where  $\widetilde{TVD}_X$  equals

$$\widetilde{TVD}_X = 0.5 \sum_{i=1}^{22} \frac{L_i}{L} \left[ \sum_{k=1}^K |f_{ik}^X - f_k^Y| \right]$$

where  $L_i$  is the number of SNPs in chromosome  $i \in [1, \dots, 22]$ ,  $L$  the total number of SNPs across all the 22 chromosomes, and  $f_{ik}^X$  is the average proportion of DNA that individuals from X match to donor group  $k$  when painting only chromosome  $i$ . This approach scales genetic differences between the two groups by differences across chromosomes within each group, exploiting how each chromosome should be subjected to independent drift.<sup>32</sup> For this analysis we used (1) the non-Iranian/Parsi donors painting for Iranian populations and (2) the non-Indian donors painting for Indian populations, which similarly attempt to attenuate drift effects within each Iran and Indian group by matching their DNA to only groups outside of their countries (thus disallowing “self-copying” in Iranian/Indian groups).<sup>29</sup> For comparison purposes,  $F_{XY}$  was also calculated for the Iranian and Indian groups using the all donors painting.

The weighted  $F_{ST}$  for these groups was also calculated based on independent SNPs using PLINK 1.9, which implements the method introduced by Weir and Cockerham.<sup>33</sup>

### Exploring Relative Amounts of Genetic Diversity within Groups

For a comparison of techniques, we used the following three distinct approaches to quantify the relative amounts of genetic diversity within groups.

#### CHROMOPAINTER Analyses to Infer Relative Amounts of Genetic Diversity within Groups

We performed a fourth analysis using CHROMOPAINTER that is analogous to that in van Dorp et al.<sup>29</sup> to assess the relative genetic diversity within the 8 clusters inferred by fineSTRUCTURE that contained Iranian or Indian individuals and had sample size greater than or equal to 13 (which is the number of Parsi individuals): Indian\_A, Indian\_B, Indian\_C, Parsis, Iranian\_A, Iranian\_Zoroastrian, Kharia, and Mala\_Vishwabrahmin (each defined in Table S2). For each of these 8 clusters, we randomly subsampled 13 individuals and painted each individual using only the other 12 individuals from their respective cluster as donors, using 50 steps of CHROMOPAINTER E-M algorithm inferring the switch and emission rates (i.e., “-i 50 -in -iM”). We refer to this fourth painting protocol throughout as the “within-group-diversity painting.” For each individual, we calculated average segment size by dividing the total centimorgan lengths of genome-wide DNA copied from all donors by the total expected number of haplotype segments copied from all donors. This average segment size can be thought of as capturing the relative amount of genome-

wide haplotype diversity in each group, with a larger average segment size reflecting relatively less genome-wide diversity.

#### PLINK IBD Analysis to Infer Relative Amounts of Genetic Diversity within Groups

We also inferred within-group genetic diversity across all pairwise combinations of the same 13 individuals within each of the above genetic clusters using the IBD coefficient  $PI\_HAT$  implemented in PLINK v.1.9, on a dataset where SNPs were first pruned to remove those in high linkage disequilibrium ( $r^2 > 0.2$ ) in a sliding window of 250 SNPs.

#### fastIBD Analysis to Infer Relative Amounts of Genetic Diversity within Groups

In order to explore within-group genetic diversity using a third approach, which allows SNPs to be in LD as in our CHROMOPAINTER-based estimates of segment size, we applied fastIBD using the software BEAGLE v.3.3.2.<sup>34</sup> For each cluster, we used fastIBD to infer the pairwise IBD fraction between each pairing of the 13 sub-sampled individuals. For each chromosome of each cluster, fastIBD was run for 10 independent runs and an IBD threshold of  $10^{-10}$  for every pairwise comparison of individuals as recommended by Browning and Browning,<sup>34</sup> though we note results were similar using an IBD threshold of 0.0001.

### Inferring Admixture Events via a Mixture Modeling Approach, GLOBETROTTER, f3-Statistics, and TreeMix

As noted previously,<sup>30,31</sup> the inferred CHROMOPAINTER painting profiles are often not an ideal summary of shared ancestry patterns, as for example donor groups with larger sample sizes may be disproportionately represented in these paintings. In order to account for this, we performed additional analyses to “clean” the raw CHROMOPAINTER output. In particular, we applied the Bayesian mixture modeling approach described in Broushaki et al.<sup>12</sup> to infer proportions of ancestry for all recipient groups (which we term “targets”) in relation to other included groups that represent potential “surrogates” to sources of ancestry. Here we performed two analyses: (1) including all 229 modern groups excluding the target as potential surrogates and (2) using all 229 modern and 8 ancient groups as potential surrogates (i.e., 237 surrogate groups in total). The aim of this mixture modeling approach is to identify which subset of these 229–237 potential surrogates best reflect the sources of ancestry in the target group. We then use this subset of surrogates in our admixture analysis described below. However, we note that any inferred proportions from this mixture model analysis cannot necessarily be interpreted as reflecting proportions of admixture from distinct source groups. Instead, this mixture modeling step is primarily used to summarize the clearest patterns of shared ancestry between the target and surrogate groups and to restrict the set of surrogates used in our subsequent admixture analysis to help increase power and precision.

We applied GLOBETROTTER,<sup>31</sup> a haplotype-based approach to identify, describe, and date recent admixture events, to test for evidence of admixture separately in each of 24 “target” groups from Iran, India, Pakistan, and Armenia. Roughly speaking, GLOBETROTTER infers admixture in a target group using two (interlocking) steps. The first infers the genetic make-up of the putative admixing source groups, and the second infers the date of admixture. For the first step we used the all donors painting from CHROMOPAINTER for each target group, as this GLOBETROTTER inference step requires each surrogate and target group to be painted using the same (or a very similar) set

of donors.<sup>31</sup> For the second step, we used CHROMOPAINTER to generate ten painting samples per haploid genome for each Iranian, Indian, Pakistani, and Armenian individual, under a different painting where each of these individuals is painted excluding any individuals from their assigned group as donors. We refer to this fifth painting protocol as “GLOBETROTTER painting.” We do this GLOBETROTTER painting to follow the suggested protocol in Hellenthal et al.,<sup>31</sup> as including individuals from your own group as donors when painting often substantially masks signals of admixture, particularly when generating the linkage disequilibrium (LD) decay curves critical to dating admixture. This is because individuals (unsurprisingly, but unhelpfully) match large segments of their genome to other individuals from their own group. While we could also use this GLOBETROTTER painting for the first step that infers the genetic make-up of the admixing source groups, for each target group we would then have had to re-paint every surrogate group similarly excluding that target group’s individuals as donors. For computational economy we instead used the same all donors painting for each target group, which previous work suggests makes little difference in practice for these sample sizes and which we explore further below.<sup>31</sup> For each target population, we included only the surrogate groups that contributed to our mixture modeling approach described above, separately under the two mixture modeling scenarios using as surrogates (1) modern groups only and (2) modern and ancient groups. We inferred admixture dates using the default LD decay curve range of 1–50 cM and bin size of 0.1 cM when considering the distance between genome segments. An exception to this is cases where the inferred admixture date was >60 generations ago using this default curve range and bin size, in which case we re-estimated dates using a curve range of 1–10 cM and a bin size of 0.05 cM, as using a similar range and bin size has been shown previously to more reliably estimate older dates of admixture.<sup>31</sup> In each analysis we used 5 iterations of GLOBETROTTER’s alternating source composition and admixture date inference (num.mixing.iterations: 5) and 100 bootstrap re-samples to infer confidence intervals around the point estimates of the date of admixture. Furthermore, in each case analyses were run twice, once using the option null.ind:0 and once with null.ind:1 to assess the effect of standardizing against a pseudo (null) individual, an approach designed to account for spurious signals of linkage disequilibrium that are not attributable to admixture.<sup>31</sup> Only results for null.ind = 1 are shown, as results for null.ind = 0 were very consistent. For comparison, we also performed an additional GLOBETROTTER analysis using the surrogates inferred under (2) when using CHROMOPAINTER results from the non-Indian/Iranian donors painting, this time using the same CHROMOPAINTER painting for both the first and second steps of GLOBETROTTER described above.

As a very different means of inferring admixture, we also used ADMIXTOOLS<sup>32</sup> to calculate  $f_3$  statistics,  $f_3(X; A, B)$ , a commonly used test to detect admixture in a target population X presumed to have received DNA from two ancestral source populations represented by surrogate groups A and B. We inferred admixture separately in the Indian and Iranian Zoroastrians, using all pairwise combinations of the other populations in the dataset, plus the ancient samples, as possible admixture sources A and B.

Additionally, we used TreeMix<sup>35</sup> to infer a bifurcating tree that merges four groups: our Indian and Iranian Zoroastrian groups, and the groups with largest sample size from each of Iran and India. We also included the Yoruba as the outgroup (root) popula-

tion, assessed results when further including a single migration event among populations in the tree, and accounted for linkage disequilibrium between SNPs in windows of 500 SNPs (-k 500).

### Positive Selection Tests

We used the XP-EHH (cross population extended haplotype homozygosity) statistic<sup>36</sup> to detect signatures of recent positive selection by comparing populations with similar demographic histories. In particular we inferred putative regions of positive selection in Zoroastrians of Iran and India, using as reference populations the clusters Iranian\_A and Indian\_A (for the latter only the individuals labeled as India\_Hindu and Gujarati were used due to usage restriction of the other samples for selection tests<sup>21</sup>), respectively. We note that we are applying XP-EHH to populations we infer here to be admixed (see Results). While XP-EHH has been applied to admixed populations before,<sup>37</sup> we caution that this presumably may lead to spurious findings, as proportions of DNA inherited from an introgressing group (which may have more or less linkage disequilibrium than the ancestral group) will vary randomly across genetic regions. This difficulty notwithstanding, normalized XP-EHH scores were calculated using SELSCAN v.1.1.0,<sup>38</sup> with positive XP-EHH values indicating potential selection in the Zoroastrian populations and negative XP-EHH values indicating potential selection in the non-Zoroastrian populations. SNP annotations were obtained using ANNOVAR.<sup>39</sup>

We performed 100 permutation tests to establish the empirical distributions of XP-EHH values across the genome for both the Indian and Iranian populations, both to determine significance thresholds and to assess the reliability of our selection inference. For each permutation, we randomly partitioned our Zoroastrians and non-Zoroastrians into two different groups, and then calculated XP-EHH comparing these two groups. The threshold values at significance level of 0.01% (quantiles 0.0001 and 0.9999) from the empirical distribution combining all 100 permutations were used to determine the significance of the XP-EHH test. These values were -4.46 and 4.46 for the Iran and -4.37 and 4.37 for India.

### Y Chromosome Typing and Mitochondrial DNA Sequencing

In order to explore sex-biased admixture and to evaluate claims of patrilineal inheritance among the Parsi priests, all the 526 samples collected for this study were typed for Y chromosome (490 successful samples) and mitochondrial DNA (518 successful sequencing) (see Table S1). Y chromosomes were typed for 6 STRs (DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393) and at 11 biallelic loci (92R7, M9, M13, M17, M20, SRY1465, SRY4064, SRY10831, sY81, Tat, and YAP) as described by Thomas et al.<sup>40</sup> and for the biallelic marker 12f2 as described by Rosser et al.<sup>41</sup> Microsatellite repeat numbers were assigned according to the nomenclature of Kayser et al.<sup>42</sup> For a subset of the samples (Parsi priests), four additional Y chromosome microsatellites (DYS389I, DYS389II, DYS425, and DYS426) were typed as described by Thomas et al.<sup>40</sup> Y chromosome haplogroups (Yhg) were defined by the 12 biallelic markers according to the Y-Chromosome consortium and the most recent nomenclature of ISOGG.

The mitochondrial DNA hyper variable segment 1 (HVS-1) was sequenced as described by Thomas et al.<sup>43</sup> Sequences were obtained for all samples between positions 16,027 and 16,400 according to the numbering scheme of Anderson et al.<sup>44</sup>

mtDNA haplotypes were assigned to haplogroups (iMhg) using HaploGrep.<sup>45</sup>

Unbiased genetic diversity,  $h$ , and its standard error were calculated using the formula given by Nei<sup>46</sup> and significant differences in calculated values were found using a standard two-tailed  $z$  test. Populations were compared using  $F_{ST}$  based on haplotype or haplogroup frequencies, estimated from analysis of molecular variance (AMOVA)  $\Phi_{ST}$  values,<sup>47,48</sup> and using the exact test for population differentiation.<sup>49</sup> Assessment of the significance of pairwise  $F_{ST}$  values was based on 10,000 permutations of the data and 10,000 Markov steps were used in the exact test. Patterns of genetic differentiation were visualized using principal coordinates (PCO) analysis performed on a similarity matrix calculated as  $1 - F_{ST}$ , based on Yhg and iMhg frequencies. Values along the main diagonal of the similarity matrix, representing the similarity of each population sample to itself, were calculated from the estimated genetic distance between two copies of the same population sample (for  $\Phi_{ST}$ -based  $F_{ST}$ , the resulting self-similarity values simplify to  $n/(n - 1)$ , where  $n$  is the sample size).

Y chromosome and mtDNA admixture proportions were estimated using the likelihood-based method LEA,<sup>50</sup> based on Yhg and inferred iMhg frequencies. We ran 5,000,000 Monte Carlo iterations of the coalescent simulation and discarded the first 10,000 iterations as burn-in.

The coalescence time of clusters of Y chromosomes belonging to the same UEP-defined haplogroup was estimated by finding the average square difference (ASD) between the inferred ancestral haplotype (in this case the modal haplotype) and all observed chromosomes.<sup>51,52</sup> The 95% confidence interval for this estimate was calculated as described in Thomas et al.<sup>53</sup> using 50,000 iterations. The microsatellite mutation rate was set to 15/7,856, based on data from three published studies.<sup>54–56</sup> This analysis was restricted to haplogroups containing a high-frequency modal haplotype (>50%) where the ancestral state could be inferred with confidence.

## Results

### Zoroastrians Are Genetically Differentiated from Non-Zoroastrians, with Different Historical Ancestry in Parsis Relative to Non-Zoroastrian Indians

Most of the Iranian Zoroastrians (see [Subjects and Methods](#) and [Table S1](#) for a description of the samples used in this work) are positioned within the autosomal genetic variation of other sampled Iranian samples in a PCA of West Eurasian individuals ([Figure S4](#)). Interestingly, 2 of the 29 Iranian Zoroastrians (YZ020 and YZ024) look genetically different from the others and were inferred by fineSTRUCTURE to cluster with other non-Zoroastrian Iranians ([Figures S1](#) and [S5](#)). In order to study the common ancestry of the genetically homogeneous majority of our sampled Iranian Zoroastrians, these two individuals were excluded from further analysis. The Parsis (i.e., Indian Zoroastrians) form a more wide-ranging cluster along PC1, falling inside Iranian, Pakistani, and Indian groups ([Figure S4](#)).

We clustered some of our sampled individuals, including all Indians and Iranians, into 51 genetically homogeneous groups that exhibited good correlation be-

tween genetic similarity and population label ([Figures 1A](#) and [S1](#), [Table S2](#); see [Subjects and Methods](#) for explanation of clustering approach). One of these final 51 clusters contained all 13 Parsis, forming the “Parsis” (or equivalently “Indian Zoroastrians”) group we use throughout the remainder of this study. A separate cluster contained 27 of 29 Iranian Zoroastrians plus a single Farsi individual that was very genetically similar to self-identified Zoroastrians ([Figure S5](#), [Table S2](#)), and these 28 individuals form the “Iranian Zoroastrian” group we use throughout the remainder of this study. The remaining 2 Iranian Zoroastrians were excluded as outliers as noted above, exhibiting genetic patterns more similar to the Lebanese, Turkish, or Bandari individuals they clustered with than to Zoroastrians ([Figure S5](#)). Though there are further splits of our final Iranian and Indian Zoroastrian clusters if we increase the number of clusters beyond 51 (see [Figure S1](#)), we note that individuals within each of our final assigned Zoroastrian clusters are very genetically similar to one another. For example, our haplotype-based genetic distance measure TVD<sup>30</sup> (see [Subjects and Methods](#)) ranged from 0.0289 to 0.3284 among pairwise comparisons of our 28 Iranian Zoroastrians and from 0.04 to 0.1772 among our 13 Parsis, similar to that within Iranian Jews (0.046–0.308), Kalash (0.02–0.233), Brahui (0.04–0.213), Makrani (0.05–0.2773), and Tiwari (0.0397–0.1605) and hence justifying their merging into a single group to increase power in subsequent analyses (similar to the approach taken by Leslie et al.<sup>30</sup>).

The composition of the remaining genetically homogeneous clusters can be found in [Figure 1A](#) and [Table S2](#).

Among our sampled individuals from Armenia, India, Iran, and Pakistan, we measured genetic distance between pairs of groups using two different techniques: (1) the commonly used, allele frequency-based measure  $F_{ST}$ <sup>33</sup> and (2) the haplotype-based measure TVD<sup>30</sup> ([Table S3](#), [Figures S2](#) and [S3](#)). Genetic distance among groups is not large overall, with e.g., typically  $F_{ST} < 0.04$ . However, similar to Jewish groups from these regions, the Onge from the isolated Andaman Islands, and the Indian Kharia, an indigenous tribal ethnic group that has been isolated from other groups,<sup>57</sup> Zoroastrians were relatively strongly genetically differentiated from non-Zoroastrians under each of these measures, agreeing with previous work.<sup>10,11,58</sup> For example, the genetic distance between Iranian Zoroastrians and non-Jewish, non-Zoroastrians from Iran ranged from 0.015 to 0.029 for  $F_{ST}$  and from 0.544 to 0.551 for TVD, with each distance measure larger than the maximum such measure between any two non-Zoroastrian, non-Jewish Iranian groups (0.011 and 0.164, respectively) ([Table S3](#)). Similarly, excluding the Onge and Kharia, the genetic distances between Parsis and non-Zoroastrians from India ranged from 0.014 to 0.028 for  $F_{ST}$  and from 0.221 to 0.278 for TVD, with each measure larger than the maximum distance between any two other non-Zoroastrian Indian groups (0.002–0.008 and 0.058–0.122, respectively). Therefore, in both Iran and India, these

**Table 1. Measuring within Group Homogeneity: Segment Size (CHROMOPAINTER), FIBD, and PI\_HAT**

Cluster	Segment (95% CI)	FIBD (95% CI)	PI_HAT (95% CI)
Indian_A	0.108 (0.103–0.117)	0.063 (0.051–0.078)	0.302 (0.297–0.309)
Indian_B	0.106 (0.104–0.128)	0.063 (0.052–0.075)	0.301 (0.296–0.308)
Indian_C	0.115 (0.106–0.14)	0.06 (0.05–0.096)	0.304 (0.299–0.312)
Indian Zoroastrians	0.161 (0.12–0.184)	0.113 (0.074–0.145)	0.312 (0.299–0.324)
Iranian_A	0.106 (0.103–0.11)	0.06 (0.047–0.08)	0.301 (0.294–0.306)
Iranian Zoroastrians	0.212 (0.171–0.271)	0.148 (0.098–0.301)	0.326 (0.31–0.372)
Kharia	0.134 (0.091–0.223)	0.075 (0.052–0.318)	0.323 (0.311–0.412)
Mala_Vishwabrahmin	0.104 (0.101–0.112)	0.061 (0.049–0.089)	0.308 (0.301–0.319)

CHROMOPAINTER's inferred median haplotype segment sizes (in cM) copied intact from a single donor, when allowing 13 randomly sampled individuals from each cluster to copy from the other 12 individuals assigned to the same cluster, using 50 steps of expectation-maximization (E-M). IBD values inferred by fastIBD (FIBD) implemented in BEAGLE v.3.3.2 using the same 13 randomly sampled individuals. PI\_HAT values inferred by PLINK v.1.9 across the same 13 randomly sampled individuals after sub-sampling SNPs to remove those in high linkage disequilibrium are also reported. Median and empirical quantile values across the 13 individuals are given for each metric for each cluster.

results indicate a high degree of genetic distance between the Zoroastrians in these countries relative to most other sampled individuals from their respective countries.

Relatively recent isolation may enhance the observed genetic differences between Zoroastrians and other groups, e.g., reflected in how individuals in each Zoroastrian group are inferred to share more recent ancestors with individuals of the same population label than with individuals from other populations under our all donors painting (Figure S1). Therefore, to mitigate genetic differences between Zoroastrians and other groups that are attributable to recent isolation, we performed additional analyses that compared haplotype patterns in our Iranian and Indian groups only to those in groups outside of their respective countries (i.e., the non-Iranian/Parsi donors painting for Iranian populations and non-Indian donors painting for Indian populations; see **Subjects and Methods**). Relative to our all donors painting, these analyses likely reflect the sharing of ancestors that lived further back in time. Critically, two groups in Iran or India that descend from a similar ancestral source(s) should match to similar outside groups at similar proportions under these analyses, even if those two groups have high  $F_{ST}$  between them due to a lack of recent intermixing.<sup>29</sup> For these analyses, we formally compared inferred matching patterns between groups using the haplotype-based genetic distance measure,  $F_{XY}$ ,<sup>29</sup> see **Subjects and Methods**), which uses independent drift effects across chromosomes to further subtract out genetic differentiation due to recent isolation (see **Subjects and Methods**).

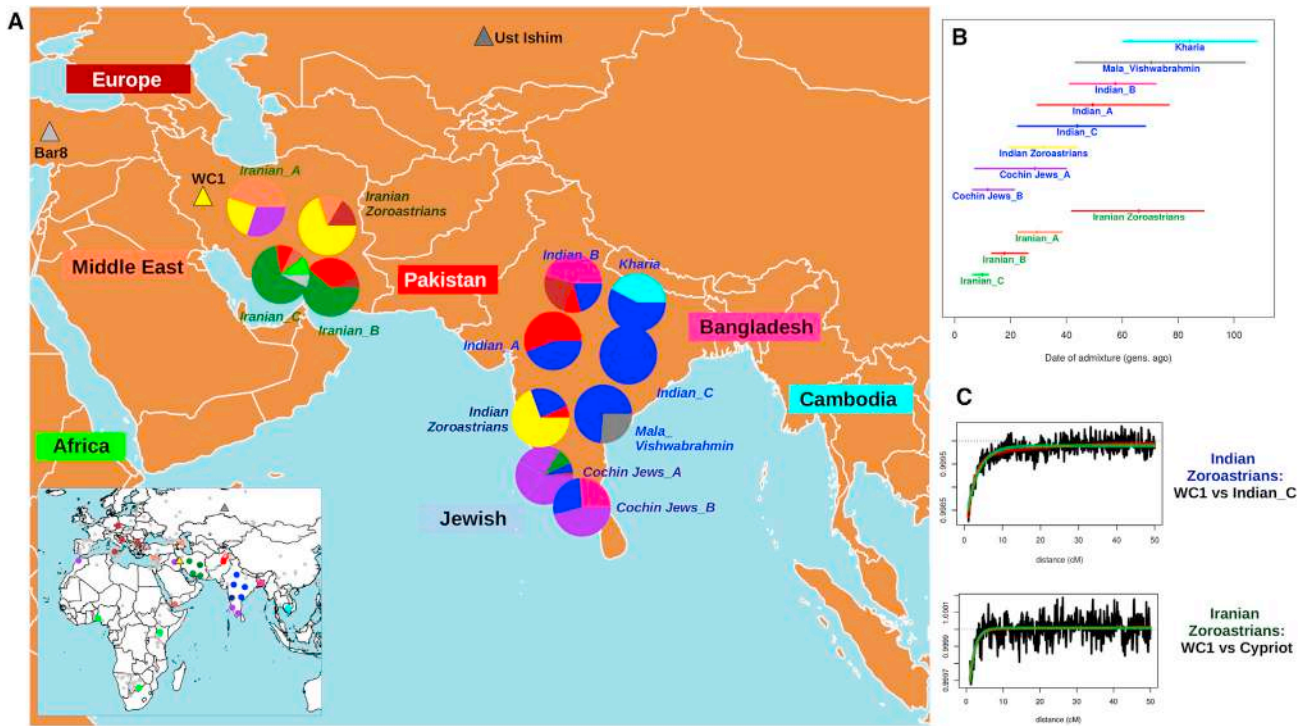
Under this  $F_{XY}$  measure, Iranian Zoroastrians showed a much-reduced genetic distance to other Iranian groups (Figure 1D), e.g., with Zoroastrians and the Iranian\_A cluster having the lowest  $F_{XY}$  value out of all comparisons of Iranian groups (Table S3). In contrast to results using our  $F_{ST}$  and TVD measures, genetic dissimilarities measured by  $F_{XY}$  among the non-Zoroastrian Iranian groups are higher. However, the  $F_{XY}$  scores are not noticeably lower

between the Parsis and non-Zoroastrian groups from India, with in general the Parsis showing a similar relatively high amount of genetic differentiation as the Kharia, Onge, and Indian Cochin Jewish groups to all other Indian groups (Table S3), mimicking our results when comparing these groups using  $F_{ST}$  and TVD (Figure 1C).

#### Genetic Homogeneity Is Higher in Zoroastrian Groups, Consistent with Increased Endogamy Relative to Non-Zoroastrians in Iran and India

Compared to non-Zoroastrian groups, we found that each of Iranian Zoroastrians and Parsis shared relatively longer haplotype segments with members of their own group (Figures 1B and S6, Table 1), reflecting a higher degree of genetic similarity within each Zoroastrian group relative to non-Zoroastrian groups. This is consistent with both Iranian and Indian Zoroastrians being genetically isolated from non-Zoroastrian groups. This is true under two distinct homogeneity estimators that use haplotype information. The first approach, FastIBD,<sup>34</sup> compares the DNA of pairwise combinations of a group's individuals, and here gave median shared haplotype lengths of 0.148 cM and 0.113 cM across pairwise combinations of Iranian Zoroastrians and Parsis, respectively, relative to 0.075 for the third largest value in the Kharia. The second approach, CHROMOPAINTER<sup>28</sup> (under our within-group-diversity painting; see **Subjects and Methods**), compares the DNA of all of a group's individuals jointly, and here gave median shared haplotype lengths across individuals of 0.212 and 0.161 cM for Iranian Zoroastrians and Parsis, respectively, relative to 0.134 for the third largest value in the Kharia. Conflicting slightly with this, we note that the PI\_HAT value from PLINK v.1.9,<sup>20</sup> which is based on an alternative technique that ignores haplotype information when measuring homogeneity, infer the Kharia to have more homogeneity than Parsis, giving median values of 0.323 and 0.312 across pairwise combinations of Kharia and Parsis, respectively. This perhaps results from a





**Figure 2. Recent Admixture in India and Iran**

(A) Inferred recent admixture in India and Iran under the all donors painting, using admixture surrogates from Europe (brown), Middle East (orange; Yemen in dark orange), Africa (light green), Pakistan (red), Bangladesh (pink), Cambodia (cyan), Iran (dark green), and India (blue) and of Jewish heritage (purple), plus the ancient samples WC1 (yellow), Ust'Ishim (dark gray), and Bar8 (light gray). Inferred proportions of haplotype sharing with each surrogate group are represented in the pie graphs, with all contributing groups highlighted in non-gray in the map in the left bottom corner.

(B) Dates of admixture (dots) and 95% confidence intervals (bars) inferred by GLOBETROTTER, colored according to the surrogate that best reflects the minor contributing admixture source.

(C) GLOBETROTTER coancestry curves, illustrating the weighted probability (black lines) that DNA segments separated by distance  $x$  (in cM) match to the two admixture surrogates labeled here, for Parsis (surrogates WC1 versus Indian\_C) and Iranian Zoroastrians (surrogates WC1 versus Cypriot), along with the best fitting exponential distributions (green lines) using the inferred date from (B) for each.

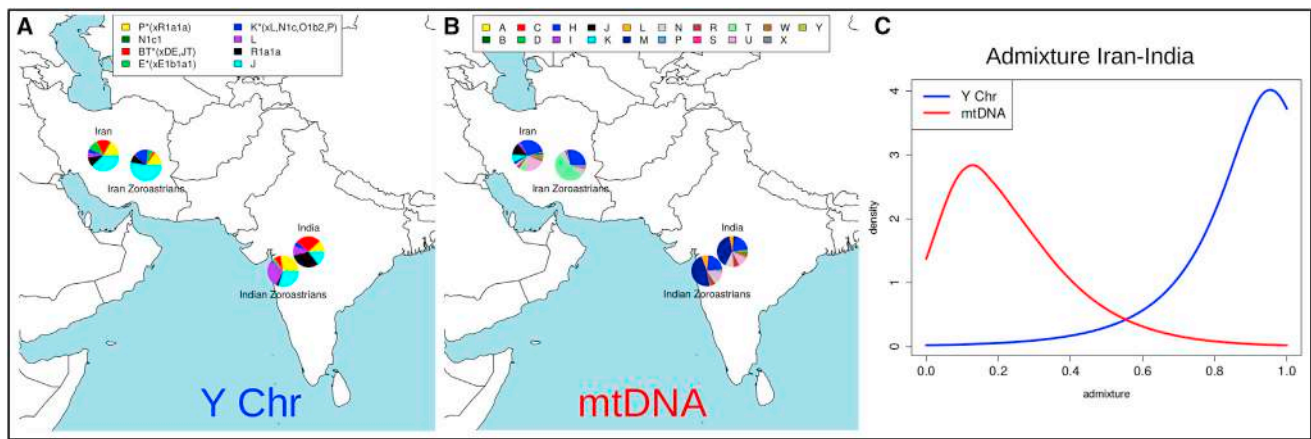
decreased resolution when not exploiting linkage disequilibrium information, at least when using ascertained SNPs.<sup>34</sup>

#### Evidence for Admixture in Zoroastrian Groups with Different Sources and Times using Nuclear Data

We calculated  $f_3$  statistics using autosomal DNA from the Iranian Zoroastrians and Parsis as targets and all pairwise combinations of the other modern and ancient groups as sources, reporting all pairwise combinations that gave a negative  $f_3$  value with a Z score  $> |2|$  for the Parsis in Table S4. In all cases one source of admixture is best represented by a modern-day Indian population. The second source is generally represented by an ancient Neolithic sample from Europe or Anatolia or a modern group close to Iran such as Armenia, Lebanon, or Iraqi\_Jews, suggesting an Iranian-like source. In the case of the Iranian Zoroastrians, no admixture events were inferred with any group present in the dataset, consistent with previous reports of  $f_3$  statistics sometimes having decreased power to detect admixture in isolated groups with, for example, bottleneck or founder effects.<sup>32</sup>

Additionally, we dated admixture events in both Parsis and Iranian Zoroastrians using the haplotype-based software GLOBETROTTER as described in Subjects and Methods. Analyses using (1) only modern groups and (2) both ancient and modern groups as possible surrogates gave largely corroborating results, e.g., with confidence intervals for dates overlapping when admixture is inferred for the same target group (Figures 2 and S7, Tables S5–S7). However, test (2) was sometimes more sensitive as noted below.

In (1) and (2), we detected admixture in the Parsis dated to 27 (range: 17–38) and 32 (19–44) generations ago, respectively, in each case between one predominantly Indian-like source and one predominantly Iranian-like source. This large contribution from an Iranian-like source (~64%–76%) is not seen in any of our other 7 Indian clusters, though we detect admixture in each of these 7 groups from wide-ranging sources related to modern day individuals from Bangladesh, Cambodia, Europe, Pakistan, or of Jewish heritage (Figures 2 and S7, Tables S5–S7). For Iranian Zoroastrians, we detect admixture only under analysis (2), occurring 66 (42–89) generations ago between a source



**Figure 3. mtDNA and Y Chromosome Variability in Iran and India**

(A and B) NRY (A) and mtDNA (B) macrohaplogroup frequencies in India and Iran.

(C) Posterior distribution of Iranian-like admixture proportions in lay Parsis assuming non-Zoroastrian Indian and Iranian lay Zoroastrian surrogate groups, using observed iMhg and Yhg values.

best genetically explained as a mixture of modern-day Croatian and Cypriot samples, and a second source matching to the Neolithic Iranian farmer WC1. We infer admixture in all three other non-Jewish Iranian groups, though consistently more recent (<38 generations ago) with contributions from sources related to modern-day groups from Turkey, Pakistan, and East Africa (Figures 2 and S7, Tables S5–S7). The two Iranian Zoroastrians that had been excluded as outliers exhibited admixture patterns more similar to the Lebanese, Turkish Jews, or Iranian Bandari individuals than to Zoroastrians (Table S8).

We also ran TreeMix in order to infer a bifurcating tree relating five sampled groups: our two Zoroastrian groups, one other Indian group (Indian\_C), one other Iranian group (Iranian\_A), and Yoruba as an outgroup (Figure S8). This TREEMIX analysis inferred the highest drift value in the Iranian Zoroastrians, in agreement with our analyses described above (Figures 1B and S6, Table S1). Inclusion of a single migration event improved residuals of the fitted tree (Figure S9) but suggested admixture from Parsis into other Indian groups rather than the other way around. This likely reflects the challenge in accurately identifying and describing admixture events in some cases when not directly measuring the decay of linkage disequilibrium that is expected in genuine admixture signals.<sup>31,59</sup>

#### Evidence for Sex-Biased Admixture in Parsis using Y Chromosome and mtDNA Data

Y chromosome binary polymorphism defined haplogroups (Yhg) and inferred mtDNA haplogroups (iMhg) showed that the Parsi priests sample has the lowest gene diversity values of all population samples studied for both Y and mtDNA (Tables S1, S9, and S10), although we did not have enough data from Iranian Zoroastrian priests to make any analogous observation. The iMhg and Yhg frequency-based pairwise  $F_{ST}$  values indicate that through the male line the lay Parsis have a closer relationship to

the lay Iranian Zoroastrians, but through the female line they have a closer relationship to the non-Zoroastrians from India (Figure S10). However, no shared Y chromosome STR+biallelic marker or mtDNA control region sequence haplotypes were shared between the Parsi priest and Iranian Zoroastrian priest samples, and all  $F_{ST}$  p values and exact tests, whether based on Yhg, Y-haplotype, iMhg, or mtDNA haplotype frequencies, indicated significant differentiation between these two groups.

Using the likelihood-based estimation of admixture (LEA) method of Chikhi et al.<sup>50</sup> as implemented in the LEA software<sup>60</sup> on Yhg and iMhg data, with the non-Zoroastrian Indians and Iranian lay Zoroastrians as surrogates for the two admixing source populations, we infer the most probable Iranian lay Zoroastrian contribution to the lay Parsis Y chromosomes to be 96% (Figure 3C; median = 86%, mean = 82%, 95% CI = 41% to 99%), whereas the most probable Iranian lay Zoroastrian contribution to Parsis mtDNA is 13% (Figure 3C; median = 20%, mean = 24%, 95% CI = 0.05% to 78%). More than 90% of posterior estimates for Y chromosome Iranian lay Zoroastrian contribution to the lay Parsis were higher than the posterior estimates for mtDNA Iranian lay Zoroastrian contribution to the lay Parsis in random samples drawn from each distribution.

#### Origins of the Parsi Priests

Y/mtDNA data defined 8 Y chromosome haplogroups and 182 total Y chromosome haplotypes when using biallelic and STR loci (Tables S9 and S10) and 240 mtDNA haplotypes that clustered into 14 haplogroups using key HVS-1 mutations. The Parsi priests sample has the lowest gene diversity values of all samples studied for both Y and mtDNA, with the majority of the Parsi priest's Y chromosomes (86%) falling into either Yhg-P\*(xR1a1a) or Yhg-L (as defined in Figure S11). The distribution of STR-defined haplotypes within these haplogroups is characterized by

the presence of a high-frequency modal haplotype (>50%), with the remaining haplotypes being only a small number of mutation steps different from the modal haplotype (Figure S11). The exception to this is one “outlier” Yhg-L chromosome that was found to be nine mutation steps different from the nine-microsatellite defined Yhg-L modal haplotype. These data are consistent with the majority of Parsi priests being patrilineal descendants of two male founders in the relatively recent past. Assuming that with the exception of the one Yhg-L outlier, the modals are the ancestral haplotypes<sup>53</sup> to all other chromosomes within each Yhg, we estimate the coalescence dates for these Parsi Yhg-P\*(xR1a1a) and Yhg-L chromosomes to be 37 generations (95% CI 19 to 61 generations) and 31 generations (95% CI 18 to 46 generations), respectively. Assuming a generation time of 28 years, this translates to 1,036 years (95% CI 532 to 1,708 years) and 868 years (95% CI 504 to 1,288 years), respectively. Noting that these two coalescence date estimates are not significantly different (only 63% of simulated dates for Yhg-P\*(xR1a1a) are older than those for Yhg-L), we re-estimated the coalescence date assuming that both lineages originated at the same time by finding the mean ASD from the respective modal haplotypes for both clusters. This gave a combined within haplogroup coalescence date estimate of 923 years (95% CI 597 to 1,277 years). When uncertainty in the mutation rate estimate is taken into account, the 95% CI widens to 501 to 1,782 years.

#### Identifying Genetic Regions Showing Evidence of Selection in Zoroastrians Relative to Non-Zoroastrians

We calculated XP-EHH values for Iranian Zoroastrians and Parsis using other Iranians and Indians as reference populations (Figure S12). Tables S11 and S12 provide details for all the SNPs below and above quantiles 0.0001 and 0.9999 of the permutation-based empirical distribution, respectively (see [Subjects and Methods](#)), including the genes within those regions, or the flanking genes in the case of intergenic SNPs. However, we note that our permutation procedure shows a very similar distribution of XP-EHH values as those observed in the real data (Figure S13), suggesting that these reported signals should be interpreted with caution.

In the case of the Iranian Zoroastrians, most of the regions with the strongest signals of selection (positive XP-EHH values) are located in intergenic or intronic regions. Among these, some of the most significant SNPs ( $p < 0.0001$  based on a permutation procedure; see [Subjects and Methods](#)) are located upstream of gene *SLC39A10* (Solute Carrier Family 39 Member 10 [MIM: 608733]) with an important role in humoral immunity<sup>61</sup> or in *CALB2* (Calbindin 2 [MIM: 114051]), which plays a major role in the cerebellar physiology.<sup>62</sup>

With regard to the positive selection tests on Parsis versus India Hindu/Gujarati groups, the most significant SNPs were embedded in *WWOX* (WW domain-containing oxidoreductase [MIM: 605131]), associated with neurolog-

ical disorders like early epilepsy (MIM: 616211),<sup>63</sup> and in a region in chromosome 20 some core domains from the *WFDC* (acidic protein WAP four-disulfide core domain containing) family and other genes like *SPINT4* (serine peptidase inhibitor, Kunitz type 4), *SNX21* (sorting nexin family member 21), or *TNNC2* (troponin C2, fast skeletal type [MIM: 191039]) (see [Table S11](#) for a complete list). On the other hand, among the SNPs showing signatures of positive selection in the reference Indian population, two highly significant selection signals were identified: *LOC102467224* and *LOC283177*, with unknown functions.

#### Discussion

Though recent studies have investigated the origins of different Jewish populations from India, like the Cochin Jews or the Bene Israel,<sup>64–66</sup> little is known about the genetic structure of the relatively isolated populations found mainly in India and Iran that practice Zoroastrianism, one of the oldest extant religions in the world. We present genome-scale genetic analyses of Zoroastrians from Iran and India and provide genetic evidence for their historical exodus.<sup>6</sup>

Zoroastrians in both Iran and India are genetically differentiated from other groups in these countries, in Y chromosome, mtDNA, and autosomal patterns of variation (Figures 1, 3, S1–S5, and S10, Tables S2, S3, S9, and S10). However, we found two Iranian Zoroastrian individuals that were genetically distinct from the other Zoroastrians (Figures S4 and S5), suggesting that these individuals were possibly mislabelled or recently converted to Zoroastrianism. The latter would indicate that present-day Zoroastrians in Iran are not as closed a group today as previously reported.<sup>9</sup>

Excluding these two Iranian Zoroastrians, the remaining Zoroastrians in both Iran and India display a high level of genetic homogeneity, greater than any other Iranian and Indian group examined in this study (Figure 1B), despite being sampled from multiple locations. This is likely attributable to founder effects, bottlenecks, and/or some endogamy throughout the last millennium and up to the present day. These factors likely played a major role in the observed differences in autosomal DNA patterns between Iranian Zoroastrians and non-Zoroastrians from Iran, as analyses that attempt to mitigate these genetic isolation effects notably decrease the observed genetic differences between Iranian Zoroastrians and non-Zoroastrian Iranians (Figures 1D and S2, Table S3). In contrast, our analyses to mitigate isolation effects do not drastically affect observed genetic differences between the Indian Zoroastrians (Parsis) and non-Zoroastrian groups from India, suggesting that the different admixture histories of different Indian groups play a major role in shaping observed genetic differences among these Indian groups today (Figures 1C and S3, Table S3).



In particular, we detect an admixture event in the Parsis dated to around 1030 CE (690–1390), between a source genetically similar to modern Indian groups and a second source best represented genetically by a ~9,500-year-old Neolithic farmer from Iran (Figure 2, Table S6). This Iranian source of introgression differs from the sources of admixture inferred in all other sampled Indian groups (Figure 2, Table S6). Our admixture date matches the historical records of a large-scale migration of Zoroastrians to India beginning around 785 CE<sup>5</sup> or 936 CE,<sup>6</sup> providing genetic evidence for this period of migration and suggesting that the migrants mixed with locals soon upon arrival. Our results suggest that these migrations may have resulted in a single “pulse” of admixture occurring around 1030 CE, though our dates are also consistent with multiple episodes of migration from around 690 CE to 1390 CE, which is difficult to disentangle given these sample sizes.<sup>31</sup> However, we see evidence of Iranian-like origins only in Parsis and in no additional sampled non-Zoroastrian groups from India, which strongly indicates our admixture signal is due to the migration of Zoroastrians from Iran rather than being related to historically documented trade between present-day Iran and India<sup>8</sup> that would likely have included mixture among non-Zoroastrian groups.

That our approach inferred the Neolithic Iranian sample WC1 to be a better surrogate for the Iranian admixing source in the Parsis than any modern Iranian groups (including Iranian Zoroastrians) (Figure 2, Table S6) likely results from strong bottleneck effects and/or recent admixture events that have made modern Iranian groups look more genetically differentiated from the Zoroastrian source group that migrated to India ~17–44 generations ago. For example, when performing an alternative analysis that attempts to mitigate genetic isolation effects within each modern Iranian and Indian group by disallowing genetic matching to members from the same assigned cluster (i.e., the non-Indian/Iranian donors painting; see [Subjects and Methods](#)), this high aDNA contribution to Parsis is replaced by the modern Iranian Zoroastrians (Table S7, Figure S7). If we instead use the original approach that does not mitigate these isolation effects (i.e., the all donors painting in Figure 2, Table S6) but exclude WC1 as a surrogate, the highest contributing Iranian group to the Parsis is Iranian\_A and not the Iranian Zoroastrians (Table S5). The fact that Iranian Zoroastrians are favored as the source of admixture in Parsis only after mitigating isolation effects suggests that at least some of these drift effects in the Iranian Zoroastrians have occurred more recently than the migrations of Parsis to India ~600–1,300 years ago. In contrast, for the Parsis it is difficult to discern the extent to which their relative genetic homogeneity (e.g., Figure 1B) reflects recent isolation since admixture or isolation occurring in their ancestry source from Persia prior to this admixture event.

Patterns of mtDNA and NRY variation also provide clear evidence of contrasting maternal and paternal ancestry in Parsis, consistent with previous studies that suggest that

migration of the ancestors of the present-day Parsi population from Iran to India was largely sexually asymmetrical.<sup>67</sup> In particular, it supports Zoroastrianism being brought from Iran to India by a group of males, and/or that gene flow into the Parsi community from the neighboring Indian population was mainly female mediated. Consistent with this, with the genetically estimated and historically attested arrival date of Parsis in India and with the claim of patrilineal descent among Parsi priests, we infer that the majority of Parsi priests are descended from two male founders 923 years (95% CI 597 to 1,277 years) ago. This parallels the Jewish *kohanim* patrilineal priesthood, who claim descent from Moses’ brother Aaron and display low Y chromosome diversity, with most Y chromosome STR haplotypes either belonging to or being only a small number of mutation steps away from a modal haplotype.<sup>53</sup>

In Iranian Zoroastrians, we inferred a relatively old admixture event between sources best represented genetically by the Neolithic Iranian WC1 and modern-day Cypriots occurring around 70 CE (range: 570 BCE–750 CE). While we infer admixture in each of the three other non-Jewish Iranian groups (Figures 2 and S7, Tables S5–S7), this admixture date in the Zoroastrians is significantly older, consistent with their long-standing isolation. The date uncertainty and ancient nature of this event prevents interpretation in a clear historical context, but one intriguing possibility is that it reflects mixture among groups joined via the allegiance of the Cypriots with Alexander the Great to help conquer the Persian Empire in 332 BCE. At any rate, interestingly our date range corresponds closely to that spanning the three major Persian empires (Achaemenid, Parthian, Sasanian) for which Zoroastrianism acted as official state religion (559 BCE–651 CE). Ancient DNA from these regions related to these ancient groups and others will greatly enhance our understanding of this older signal.

Notably, when using only modern groups as surrogates and excluding WC1, GLOBETROTTER was not able to detect this older admixture event (Table S5). In this latter analysis, our model considered the Iranian Zoroastrians to be sufficiently genetically matched to a single modern group (Iranian\_A) without requiring any other ancestry sources. Presumably this is because Iranian\_A has similar genetic patterns to the Iranian Zoroastrians, with GLOBETROTTER inferring similar (but more recent) admixture 20–38 generations ago in Iranian\_A between sources best represented by WC1 and modern-day Turkish groups. Our results here suggest that this similarity masks the older DNA contributions to the Zoroastrians. However, the combination of WC1 and other modern groups provides a better match to an ancestral source of the Iranian Zoroastrians than using only Iranian\_A, enabling a clear signal of admixture (Figures 2 and S7, Tables S6 and S7). This reveals how adding even small numbers of ancient samples, particularly those less affected by recent admixture, can increase power and insights in population genetic



history inference, even if those ancient samples are substantially older than the time period under study, as is the case here with WC1 living more than 7,000 years earlier.

Our analyses suggest that present-day Iranian Zoroastrians ceased intermixing with other groups shortly after the Arab conquest of Persia during 633–654 CE.<sup>68</sup> In contrast, we infer more recent events in each of our three other non-Jewish Iranian groups (Figures 2 and S7, Tables S5–S7). For example, in Iranian\_A we infer an admixture event with an ancestry contribution related to a Turkish-like source group, dated to 1222 CE (range of 95% CI across all 3 GLOBETROTTER analyses: 1026–1362 CE), and hence overlapping with the period of the Seljuq Empire (1037–1194 CE) that spanned parts of present-day Turkey,<sup>69</sup> while in Iranian\_B we infer admixture dated to 1418 CE (1194–1558 CE) from a Pakistan-like source that falls into the period of the Timurid Empire (1370–1507 CE) that extended into present-day Pakistan.<sup>70</sup> Finally, we infer our most recent Iranian admixture event in our two Iranian\_C individuals, with contributions from an East Africa-like source and an inferred date of 1642 CE (1558–1754 CE) that overlaps with the Safavid Empire (1501–1736 CE<sup>71</sup>) and could be related to the Arab slave trade, as previously observed.<sup>31</sup> More work is required to infer the precise historical nature of these admixture events.

Genetic isolation and endogamous practices can be associated with higher frequencies of disease prevalence. For example, there are reports claiming a high recurrence of diseases such as diabetes<sup>72</sup> among the Iranian Zoroastrians, and Parkinson,<sup>73</sup> colon cancer,<sup>74</sup> or the deficiency of G6PD,<sup>75</sup> an enzyme that triggers the sudden reduction of red blood cells, among the Parsis. Researchers have argued that in addition to these demographic effects, selection can also play a role in the incidence of rare disorders or other phenotypes, as has been previously reported for Ashkenazi Jews.<sup>76,77</sup> Therefore, identifying regions under positive selection in the Zoroastrian populations may be helpful to understand the prevalence of diseases or distinct phenotypic traits in the community. Using XP-EHH<sup>36</sup> comparing Zoroastrians to non-Zoroastrians, we have identified some regions that might have been under selection specifically in the Zoroastrians ( $p < 0.0001$  based on a permutation procedure; see **Subjects and Methods**), as well as in the non-Zoroastrian reference groups (Tables S11 and S12). Some of these regions contain genes that have been associated with different diseases (see **Results**). However, a permutation study that re-assigned Zoroastrians and non-Zoroastrians randomly to two groups and then tested for selection between these groups gave very similar magnitudes of XP-EHH scores to that seen in our non-permuted data (Figure S13), warranting caution in interpreting these findings and illustrating the challenges in identifying selection signals using genome-wide scans. A larger cohort would be needed to corroborate their significance, coupled with exhaustive epidemiological studies. Nonetheless, they represent a first insight into understanding genetic

predisposition and/or resistance to disease in these groups and could form the basis for targeted medical approaches in these isolated groups.

In summary, we explore the genetic landscape and structure of Indian and Iranian Zoroastrians and provide genome-wide genetic evidence that the Parsis descend from an admixture event between ancestral groups consisting predominantly of males with Iranian-related ancestry and females with Indian-related ancestry. We date this event in ancestral Parsis to around 1030 CE, in agreement with historical records. We also provide evidence of a much older admixture event in Iranian Zoroastrians dated to around 74 CE with an unknown historical explanation but overlapping the period where Zoroastrianism acted as state religion in the region. We also demonstrate that Zoroastrians in both countries are genetically homogeneous populations differentiated from other population living locally, likely in part due to religious rules that discourage intermixing with non-Zoroastrians. Further work is required to help understand whether the genetic differences attributable to this isolation correlate with observed differences in disease phenotypes between these communities and other local groups.

#### Accession Numbers

Genotypes for the Indian Zoroastrian and Indian Hindu individuals are available in plink format at <https://doi.org/10.6084/m9.figshare.5223583.v1>.

#### Supplemental Data

Supplemental Data include 13 figures and 12 tables and can be found with this article online at <http://dx.doi.org/10.1016/j.ajhg.2017.07.013>.

#### Acknowledgments

S.L. is supported by BBSRC (grant number BB/L009382/1). G.H. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (grant number 098386/Z/12/Z) and supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. M.G.T. was supported by Wellcome Trust Senior Investigator Award (grant number 100719/Z/12/Z). L.v.D. is supported by CoMPLEX via EPSRC (grant number EP/F500351/1). We are grateful to Zoroastrian Studies (Bombay) and in particular to Mr. Khojeste P. Mistree for the facilitation of the project in India and the introduction to priests in all the locations where samples were obtained. We also thank Mrs. Firoza Punthakey Mistree, Ms. Tashan Mistree, and the late Mrs. Shehnaz Neville Munshi for their involvement in the samples collection in India and Iran. We also thank the Children's Hospital of Philadelphia for genotyping the samples on the Affymetrix Human Origins array. G.H. is a founder and director of GENSCI and consultant to LivingDNA. M.E.W. is an employee of Genomics plc, a company providing genomic analysis services to the pharmaceutical and healthcare sectors. His involvement in the conduct of this research was solely in his

capacity as a Reader in Statistical Genetics at King's College London.

Received: April 17, 2017

Accepted: July 24, 2017

Published: August 24, 2017

## Web Resources

Genotype data for this paper in plink format, <https://doi.org/10.6084/m9.figshare.5223583.v1>

ISOGG, <https://isogg.org/>

OMIM, <http://www.omim.org/>

## References

1. Lamberg-Karlovsky, C.C. (2005). Archaeology and language: the case of the Bronze Age Indo-Iranians. In *The Indo-Aryan Controversy. Evidence and Inference in Indian History*, E.F. Bryant and L.L. Patton, eds. (London, New York: Taylor & Francis Group), pp. 142–178.
2. Mallory, J.P. (2002). Archaeological models and Asian Indo-Europeans. In *Indo-Iranian Languages and Peoples*, N. Sims-Williams, ed. (British Academy), pp. 19–42.
3. Skjærvø, P.O. (1996). Zarathustra in the Avesta and in Manichaeism. *Irano-Manichaica III*. In *La Persia e l'Asia centrale da Alessandro al X secolo* (Accademia Nazionale dei Lincei), pp. 597–628.
4. Stewart, S. (2013). *The Everlasting Flame: Zoroastrianism in History and Imagination* (I.B. Tauris).
5. Modi, J.J. (1905). *A Few Events in the Early History of the Parsis and Their Dates* (Bombay: K.R. Cama Oriental Institute).
6. Hodivala, S.H. (1920). *Studies in Parsi History* (Bombay).
7. Williams, A. (2009). *The Zoroastrian Myth of Migration from Iran and Settlement in the India Diaspora* (Brill Academic Publishers).
8. Wink, A. (1990). *Al-Hind: The Making of the Islamic World, Volume 1* (Brill Academic Publishers).
9. Boyce, M. (2001). *Zoroastrians: Their Religious Beliefs and Practices* (Routledge).
10. Farjadian, S., Sazzini, M., Tofanelli, S., Castrì, L., Taglioli, L., Pettener, D., Ghaderi, A., Romeo, G., and Luiselli, D. (2011). Discordant patterns of mtDNA and ethno-linguistic variation in 14 Iranian ethnic groups. *Hum. Hered.* 72, 73–84.
11. Lashgary, Z., Khodadadi, A., Singh, Y., Houshmand, S.M., Mahjoubi, F., Sharma, P., Singh, S., Seyedin, M., Srivastava, A., Ataee, M., et al. (2011). Y chromosome diversity among the Iranian religious groups: a reservoir of genetic variation. *Ann. Hum. Biol.* 38, 364–371.
12. Broushaki, E., Thomas, M.G., Link, V., López, S., van Dorp, L., Kirsanow, K., Hofmanová, Z., Diekmann, Y., Cassidy, L.M., Díez-Del-Molino, D., et al. (2016). Early Neolithic genomes from the eastern fertile crescent. *Science* 353, 499–503.
13. Stepaniants, M. (2002). *The Encounter of Zoroastrianism with Islam* (University of Hawai'i Press).
14. al-Magthteh, M., Ray, V., Mastana, S.S., Garralda, M.D., Bhattacharya, S.S., and Papiha, S.S. (1993). Variation in DNA polymorphisms of the short arm of the human X chromosome: genetic affinity of Parsi from western India. *Hum. Hered.* 43, 239–243.
15. Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C., and Mehdi, S.Q. (2002). Y-chromosomal DNA variation in Pakistan. *Am. J. Hum. Genet.* 70, 1107–1124.
16. Mohyuddin, A., and Mehdi, S.Q. (2005). HLA analysis of the Parsi (Zoroastrian) population in Pakistan. *Tissue Antigens* 66, 691–695.
17. Hinnells, J. (1996). *Zoroastrians in Britain* (Clarendon Press Oxford).
18. Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet.* 38, 1251–1260.
19. Hellenthal, G., Auton, A., and Falush, D. (2008). Inferring human colonization history using a copying model. *PLoS Genet.* 4, e1000078.
20. Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga-science* 4, 7.
21. Lazaridis, I., Nadel, D., Rollefson, G., Merrett, D.C., Rohland, N., Mallick, S., Fernandes, D., Novak, M., Gamarra, B., Sirak, K., et al. (2016). Genomic insights into the origin of farming in the ancient Near East. *Nature* 536, 419–424.
22. Hofmanová, Z., Kreutzer, S., Hellenthal, G., Sell, C., Diekmann, Y., Díez-Del-Molino, D., van Dorp, L., López, S., Kousathanas, A., Link, V., et al. (2016). Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. USA* 113, 6886–6891.
23. Jones, E.R., Gonzalez-Forbes, G., Connell, S., Siska, V., Eriksson, A., Martiniano, R., McLaughlin, R.L., Gallego Llorente, M., Cassidy, L.M., Gamba, C., et al. (2015). Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nat. Commun.* 6, 8912.
24. Gamba, C., Jones, E.R., Teasdale, M.D., McLaughlin, R.L., Gonzalez-Forbes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., et al. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* 5, 5257.
25. Gallego Llorente, M., Jones, E.R., Eriksson, A., Siska, V., Arthur, K.W., Arthur, J.W., Curtis, M.C., Stock, J.T., Coltorti, M., Pieruccini, P., et al. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science* 350, 820–822.
26. Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S.M., Bondarev, A.A., Johnson, P.L., Aximu-Petri, A., Prüfer, K., de Filippo, C., et al. (2014). Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* 514, 445–449.
27. Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. *Nat. Methods* 9, 179–181.
28. Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. *PLoS Genet.* 8, e1002453.
29. van Dorp, L., Balding, D., Myers, S., Pagani, L., Tyler-Smith, C., Bekele, E., Tarekegn, A., Thomas, M.G., Bradman, N., and Hellenthal, G. (2015). Evidence for a common origin of blacksmiths and cultivators in the Ethiopian Ari within the last 4500 years: lessons for clustering-based inference. *PLoS Genet.* 11, e1005397.
30. Leslie, S., Winney, B., Hellenthal, G., Davison, D., Boumertit, A., Day, T., Hutnik, K., Royrvik, E.C., Cunliffe, B., Lawson, D.J., et al.; Wellcome Trust Case Control Consortium 2; and International Multiple Sclerosis Genetics Consortium (2015). The

- fine-scale genetic structure of the British population. *Nature* 519, 309–314.
31. Hellenthal, G., Busby, G.B.J., Band, G., Wilson, J.F., Capelli, C., Falush, D., and Myers, S. (2014). A genetic atlas of human admixture history. *Science* 343, 747–751.
  32. Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. (2012). Ancient admixture in human history. *Genetics* 192, 1065–1093.
  33. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution* 38, 1358–1370.
  34. Browning, B.L., and Browning, S.R. (2011). A fast, powerful method for detecting identity by descent. *Am. J. Hum. Genet.* 88, 173–182.
  35. Pickrell, J.K., and Pritchard, J.K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* 8, e1002967.
  36. Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al.; International HapMap Consortium (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature* 449, 913–918.
  37. Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., and Pritchard, J.K. (2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.* 19, 826–837.
  38. Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Mol. Biol. Evol.* 31, 2824–2827.
  39. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* 38, e164.
  40. Thomas, M.G., Bradman, N., and Flinn, H.M. (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum. Genet.* 105, 577–581.
  41. Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., et al. (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am. J. Hum. Genet.* 67, 1526–1543.
  42. Kayser, M., Caglià, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., et al. (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *Int. J. Legal Med.* 110, 125–133, 141–149.
  43. Thomas, M.G., Weale, M.E., Jones, A.L., Richards, M., Smith, A., Redhead, N., Torroni, A., Scozzari, R., Gratrix, F., Tarekgn, A., et al. (2002). Founding mothers of Jewish communities: geographically separated Jewish groups were independently founded by very few female ancestors. *Am. J. Hum. Genet.* 70, 1411–1420.
  44. Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., Sanger, F., et al. (1981). Sequence and organization of the human mitochondrial genome. *Nature* 290, 457–465.
  45. Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., and Kronenberg, F. (2011). HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* 32, 25–32.
  46. Nei, M. (1987). *Molecular Evolutionary Genetics* (Columbia University Press).
  47. Reynolds, J., Weir, B.S., and Cockerham, C.C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105, 767–779.
  48. Michalakis, Y., and Excoffier, L. (1996). A generic estimation of population subdivision using distances between alleles with special reference for microsatellite loci. *Genetics* 142, 1061–1064.
  49. Raymond, M., and Rousset, F. (1995). An exact test of population differentiation. *Evolution* 49, 1280–1283.
  50. Chikhi, L., Bruford, M.W., and Beaumont, M.A. (2001). Estimation of admixture proportions: a likelihood-based approach using Markov chain Monte Carlo. *Genetics* 158, 1347–1362.
  51. Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139, 457–462.
  52. Goldstein, D.B., Ruiz Linares, A., Cavalli-Sforza, L.L., and Feldman, M.W. (1995). An evaluation of genetic distances for use with microsatellite loci. *Genetics* 139, 463–471.
  53. Thomas, M.G., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N., and Goldstein, D.B. (1998). Origins of Old Testament priests. *Nature* 394, 138–140.
  54. Bianchi, N.O., Catanesi, C.I., Bailliet, G., Martinez-Marignac, V.L., Bravi, C.M., Vidal-Rioja, L.B., Herrera, R.J., and López-Camelo, J.S. (1998). Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations. *Am. J. Hum. Genet.* 63, 1862–1871.
  55. Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., and de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum. Mol. Genet.* 6, 799–803.
  56. Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Krüger, C., Krawczak, M., Nagy, M., Dobosz, T., et al. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am. J. Hum. Genet.* 66, 1580–1588.
  57. Vidyarthi, L.P., and Upadhyay, V.S. (1980). The Kharia: Then and Now: A Comparative Study of Hill, Dhelki, and Dudh Kharia of the Central-eastern region of India (Concept).
  58. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., and Singh, L. (2009). Reconstructing Indian population history. *Nature* 461, 489–494.
  59. Loh, P.R., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J.K., Reich, D., and Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Genetics* 193, 1233–1254.
  60. Langella, O., Chikhi, L., and Beaumont, M.A. (2001). LEA (Likelihood-based estimation of admixture): a program to simultaneously estimate admixture and the time since admixture. *Mol. Ecol. Resour.* 1, 357–358.
  61. Hojyo, S., Miyai, T., Fujishiro, H., Kawamura, M., Yasuda, T., Hijikata, A., Bin, B.H., Irié, T., Tanaka, J., Atsumi, T., et al. (2014). Zinc transporter SLC39A10/ZIP10 controls humoral immunity by modulating B-cell receptor signal strength. *Proc. Natl. Acad. Sci. USA* 111, 11786–11791.
  62. Schiffmann, S.N., Cheron, G., Lohof, A., d’Alcantara, P., Meyer, M., Parmentier, M., and Schurmann, S. (1999). Impaired motor coordination and Purkinje cell excitability in mice lacking calretinin. *Proc. Natl. Acad. Sci. USA* 96, 5257–5262.

63. Abdel-Salam, G., Thoenes, M., Affi, H.H., Körber, F., Swan, D., and Bolz, H.J. (2014). The supposed tumor suppressor gene *WWOX* is mutated in an early lethal microcephaly syndrome with epilepsy, growth retardation and retinal degeneration. *Orphanet J. Rare Dis.* 9, 12.
64. Waldman, Y.Y., Biddanda, A., Dubrovsky, M., Campbell, C.L., Oddoux, C., Friedman, E., Atzmon, G., Halperin, E., Ostrer, H., and Keinan, A. (2016). The genetic history of Cochin Jews from India. *Hum. Genet.* 135, 1127–1143.
65. Waldman, Y.Y., Biddanda, A., Davidson, N.R., Billing-Ross, P., Dubrovsky, M., Campbell, C.L., Oddoux, C., Friedman, E., Atzmon, G., Halperin, E., et al. (2016). The genetics of Bene Israel from India reveals both substantial Jewish and Indian ancestry. *PLoS ONE* 11, e0152056.
66. Chaubey, G., Singh, M., Rai, N., Kariappa, M., Singh, K., Singh, A., Pratap Singh, D., Tamang, R., Selvi Rani, D., Reddy, A.G., et al. (2016). Genetic affinities of the Jewish populations of India. *Sci. Rep.* 6, 19166.
67. McElreavey, K., and Quintana-Murci, L. (2005). A population genetics perspective of the Indus Valley through uniparentally-inherited markers. *Ann. Hum. Biol.* 32, 154–162.
68. Akram, A.I. (2009). *The Muslim Conquest of Persia* (Maktabah Publications).
69. Jackson, P. (2002). Review: the history of the Seljuq Turkmens. *J. Islamic Stud.* 13, 75–76.
70. Subtelny, M. (2007). *Timurids in Transition: Turko-Persian Politics and Acculturation in Medieval Iran* (BRILL).
71. Newman, A.J. (2006). *Safavid Iran: Rebirth of a Persian Empire* (I. B. Tauris).
72. Khalilzadeh, S., Afkhami-Ardekani, M., and Afrand, M. (2015). High prevalence of type 2 diabetes and pre-diabetes in adult Zoroastrians in Yazd, Iran: a cross-sectional study. *Electron. Physician* 7, 998–1004.
73. Bharucha, N.E., Bharucha, E.P., Bharucha, A.E., Bhise, A.V., and Schoenberg, B.S. (1988). Prevalence of Parkinson's disease in the Parsi community of Bombay, India. *Arch. Neurol.* 45, 1321–1323.
74. Jussawalla, D.J., and Gangadharan, P. (1977). Cancer of the colon: 32 years of experience in Bombay, India. *J. Surg. Oncol.* 9, 607–622.
75. Baxi, A.J., Balakrishnan, V., Undevia, J.V., and Sanghvi, L.D. (1963). Glucose-6-phosphate dehydrogenase deficiency in the Parsee community, Bombay. *Indian J. Med. Sci.* 17, 493–500.
76. Goodman, R.M., and Motulsky, A.G. (1979). *Genetic Diseases Among Ashkenazi Jews* (New York: Raven Press).
77. Bray, S.M., Mulle, J.G., Dodd, A.F., Pulver, A.E., Wooding, S., and Warren, S.T. (2010). Signatures of founder effects, admixture, and selection in the Ashkenazi Jewish population. *Proc. Natl. Acad. Sci. USA* 107, 16222–16227.