Integrating genetic and oral histories of Southwest Indian populations

Arjun Biddanda[1†‡], Esha Bandyopadhyay[1†], Constanza de la Fuente Castro[1†], David Witonsky[1], Nagarjuna Pasupuleti[2], Renée Fonseca[1], Suzanne Freilich[1,3], Hannah M. Moots[1], Jovan Stanisavic[1], Tabitha Willis[1], Anoushka Menon[4§], Mohammed S. Mustak[2], Chinnappa Dilip Kodira[5], Anjaparavanda P. Naren[6‖], Mithun Sikdar[7], Niraj Rai[8*], Maanasa Raghavan[1*]

[1] Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA.
[2] Department of Applied Zoology, Mangalore University, Mangalagangothri, Karnataka, 574199, India.
[3] Department of Anthropology, University of Vienna, Vienna, 1090, Austria.
[4] Department of Archaeology, University of Cambridge, Cambridge CB2 3DZ, UK.
[5] PureTech Health, 6 Tide Street, Boston, MA 02210, USA.
[6] Division of Pulmonary Medicine, Cystic Fibrosis Research Center, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH 45229, USA.
[7] Anthropological Survey of India, Mysore, Karnataka, 570026, India.
[8] Birbal Sahni Institute of Palaeosciences, Uttar Pradesh, Lucknow, Uttar Pradesh, 226007, India.
‡ Present address: 54Gene, Washington, DC., USA.
§ Present address: Thriva Health, London, UK.
‖ Present address: Division of Pulmonary and Critical Care Medicine, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA 90048, USA.
† These authors contributed equally to this manuscript.
* Correspondence to: mraghavan@uchicago.edu and nirajrai@bsip.res.in.

Abstract

Several Indian populations have genetic contributions from ancient hunter-gatherer, Iranian, and Eurasian Steppe lineages. However, sparse sampling across finer-scale regions impedes our understanding of the emergence of present-day population structure. Moreover, oral histories maintained by many Indian populations are only partially documented. To address these gaps and investigate the interplay between oral and genetic histories, we analyzed whole-genome sequences from 158 Southwest Indian individuals identifying as Bunt, Kodava, Nair, and Kapla. From community interviews and historical literature, we found that all four populations have references to non-Indian origins. We observed that while the Kapla are genetically differentiated from the Bunt, Kodava, and Nair, all four populations are genetically most similar to other Indian populations. Our results demonstrate that oral and genetic histories may not always provide consistent accounts of population origins and suggest community-engaged, multi-disciplinary investigations to understand their relationship.

Introduction

India has a complex history of human migrations and admixture, as well as a rich set of varied socio-cultural practices, all of which have contributed to the extensive cultural and genetic diversity in this region. Previous studies of genetic diversity in India identified population genetic structure that broadly tracks with geography and language (*1–7*). These studies showed that many Indian populations can be modeled along a genetic cline bounded by two statistical constructs: first, Ancestral South Indians (ASI), with genetic components related to Indigenous Andamanese and ancient Iranian farmers and second, Ancestral North Indians (ANI), modeled as a mix of ASI and a genetic component related to Middle-to-Late Bronze Age (MLBA) groups in the Central Steppe region (*8*, *9*); (*3*). In addition to these larger-scale ancestry gradients, at a finer-scale, genetic structure in India has been impacted by founder events and endogamous practices (*2*, *9–12*). Despite substantial diversity, characterization of human genetic variation in India has been disproportionately limited. While recent studies have made substantial inroads into characterizing the genetic diversity of populations across South Asia (*1*, e.g. *7*, *8*), many populations in India have not yet been characterized at the genetic level, leaving gaps in our understanding of sub-regional population structure.

While such gaps in the literature contribute to the motivation behind genomic research in underrepresented populations, including in India, it is imperative to be mindful of the ethical and socio-political ramifications of the research process for marginalized populations (e.g. *13*). While institutional ethics boards, such as institutional review boards (IRBs), exist, the push for diversification in genomics has, in several regions, not been matched with accountability measures to safeguard the interests of participating populations. A lack of equitable partnerships in the research process often leads to an extractive and imbalanced power dynamic between researchers and participants (*14–16*). Moreover, the conflation between self-identities, derived from a myriad of sources and life experiences, and genetics can lead to biases, misinterpretations, and confusion. At its most extreme, such conflation between genetic ancestry and self-identity can potentially impact a population's status and recognition by local governments and afforded rights and privileges (*17*). Instead of attempting to reconcile genetic and cultural histories, it may be more valuable to recognize the complexities of both forms of knowledge (*18–20*). Unfortunately, in India, as in other regions of the world, oral traditions (e.g. folk songs, stories) that underpin self-identities and local concepts of ethnogenesis in many populations are dwindling with limited written records to preserve this unique heritage. All of this calls for an effort to document oral traditions and seek out more interdisciplinary and sensitive ways for geneticists to engage with populations and their cultural histories.

This study aims to develop a fine-scale characterization of population structure in Southwest India by recognizing the unique positions held by inferences from genetic data and oral histories and by building a community-informed collaborative framework. We generated and analyzed whole-genome sequences from individuals identifying as Kodava, Bunt, Nair, and Kapla and, in conjunction with published genome-wide sequences from worldwide populations, investigated genetic histories and population structure in present-day Southwest India. We concurrently present community-engaged anthropological documentation of oral histories surveyed by the research team via conversations and observations. We found that the Bunt, Kodava, and Nair have strong self-identities as well as unique cultural traits and oral histories reflected in their origin narratives, some of which are shared between these populations. Compared to these three populations, there is relatively little known about the Kapla based on the anthropological literature and/or community accounts. Available information from communities and historical records reference non-local origins and contacts for all four populations. Here, we considered genetic and oral histories from our investigations within an integrated framework to motivate future research on the unique oral histories of these populations and, more broadly, on thoughtful ways of bringing together their unique self-identities and genetic histories.

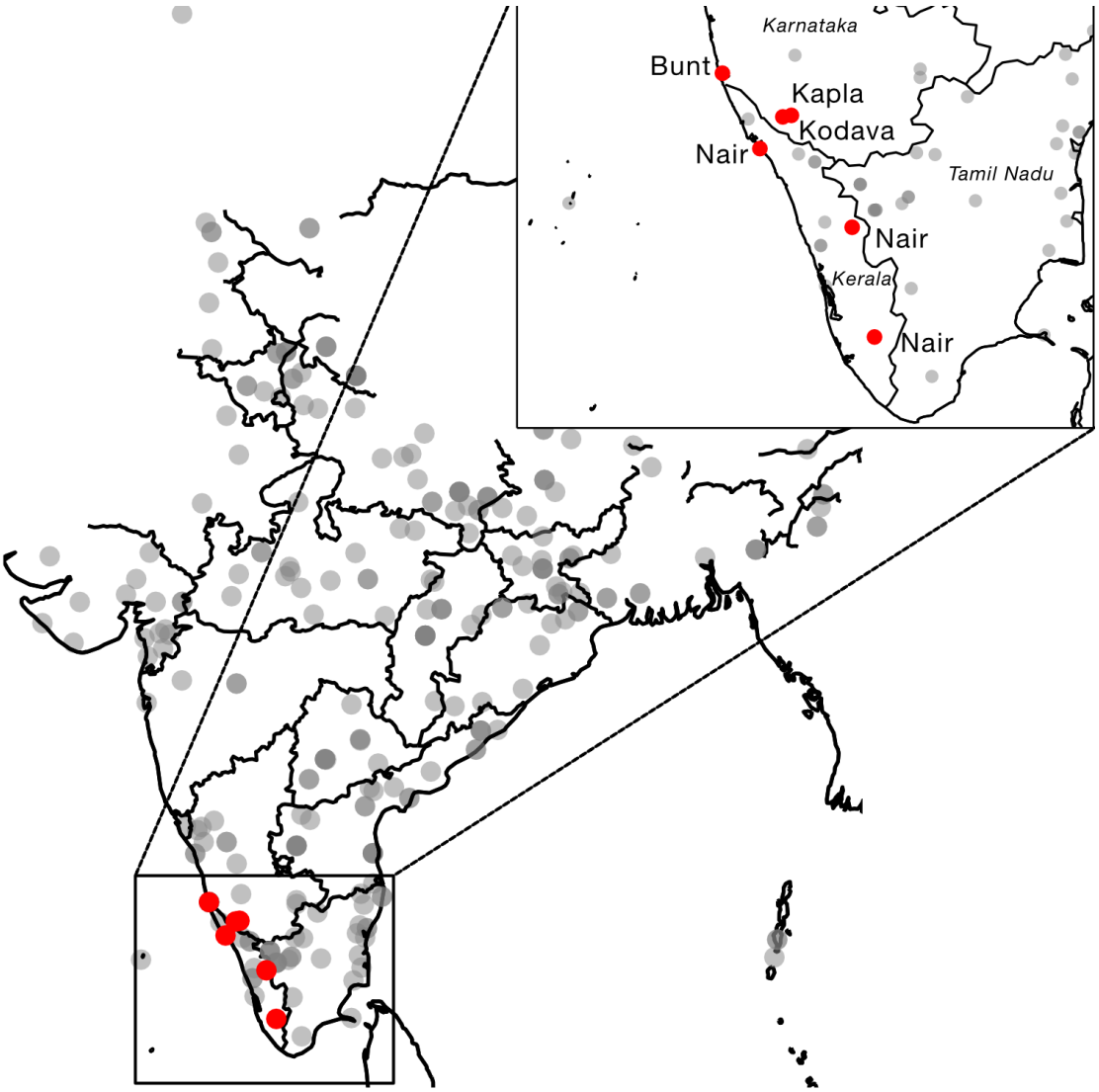

Figure 1. Map of newly sampled populations from this study. Gray points represent regional populations sampled from previous studies (*2*, *7*).

Results

Anthropological surveys of population oral histories

Community engagement in this study included information sessions in India and the US prior to and at the start of the project. At these sessions and during subsequent interactions, the research team received valuable input from members and leaders of the study populations on their social organization, cultural traits, oral histories as well as their interest in genetic studies and expectations, which enriched the research process. Results of this engagement process are reported in this section, augmented by published historical and limited anthropological works.

Study participants self-identified as Kodava and Kapla from the state of Karnataka, Nair from the state of Kerala, and Bunt from both states, and were sampled in India in 2018 (Figure 1 and Table S1). In addition, in 2019, we were approached by a larger set of individuals self-identifying as Kodava, denoted as Kodava_US in the genetic analyses to differentiate them from the Kodava sampled in India (denoted as Kodava), who recently immigrated to the United States and were interested in learning more about their genetic past. While the Kodava, Bunt, and Nair have strong self-identities and associated oral histories, relatively little is documented in the ethnographic literature (*21*–*25*). These populations are all speakers of Dravidian languages. Moreover, there are notable overlaps in these populations' traditions and cultural traits, which may stem from geographical proximity and historical contacts that are also referenced in these populations' oral histories (*25*). These include historical 'warrior' status designations/identities and matrilineal descent in the Bunt and Nair, with the latter additionally practicing matrilocality. Moreover, the Nair have a complex and longstanding social system that consists of several subgroups (*26*). Some of these socio-cultural characteristics are noted in the anthropological literature (*27*, *28*) and were additionally shared by donors and community representatives with the research team during fieldwork. Similarly, there are anecdotal accounts, for instance, noted in blogs maintained by community members and relayed to members of the research team during sampling and subsequent population interactions, of phenotypic characteristics that members of these populations draw on to support possible non-local origins. From our interactions with these populations, these narratives are used by population members to contemplate unique phenotypes and customs such as music, diet, religious affiliation, and traditional wear that set them apart from neighboring populations. To our knowledge, there is no anthropological research following up on these oral accounts and whether the nature of the non-local population interactions that they invoke was socio-economic, genetic, or both. Moreover, the exact mode and timing of past population contacts may not always be explicitly stated in oral histories.

Some of the proposed links to non-local populations, such as to the Scythians, ancient nomadic people inhabiting the Eurasian Steppe during the Iron Age, are shared across the oral histories of the Bunt, Kodava, and Nair. In addition, the Kodava have oral histories suggesting links to the Kurdish Barzani tribe through the latter's involvement in Alexander the Great's campaign into India. The following is noted in (*29*): "*In those days, when the army advanced, their families of fighting men too moved behind them, as camp followers. After Alexander turned back, some tribes in his army who had no energy to get back to their homeland, stayed back in India…...Our ancestors* [Kodava, are] *believed to have taken a southernly route along with Western Ghats in search of better prospects* [and] *eventually settled in Kodagu (Coorg) which was then an unnamed, inhospitable and extremely rugged hilly region*". Height and nose shape were cited during our interactions with Kodava_US community members as examples of phenotypic characteristics that contribute to the physical distinction between the Kodava and neighboring populations and as evidence of a non-local origin.

Field observations by members of the research team suggested that the Kapla are socio-culturally different from the other populations sampled in this study. To our knowledge, there is very little documentation on the Kapla in the anthropological literature. The Kapla population lives geographically close to the Kodava in the Kodagu region of southern Karnataka. They still practice a form of hunting and gathering, but their subsistence is transitioning slowly with increased contacts with the Kodava, on whose coffee plantations Kapla men work during the harvesting season. They speak a mixture of Tulu and Kodava languages, both members of the Dravidian language family. Historical records suggest (*30*): "*The Kaplas, who live near Nalkanad palace seem to be mixed descendants of the Siddis - the Coorg Rajahs' Ethiopian bodyguard - as their features resemble the Ethiopian type. They have landed property of their own near the palace, given by the Rajahs, and work also as day laborers with the Coorgs. Their number consists of only 15 families*". Siddis are historical migrants from Africa who live in India and Pakistan (*31*). Anthropological information gathered by the research team from long-term community contacts suggests that once the Kapla settled in their present location in Kodagu, they were isolated from neighboring

populations. Furthermore, the Kapla follow a patriarchal and patrilocal system, with a preference towards marriages involving MBD (mother's brother's daughter), and junior sororate marriage practices.

Broad-scale population structure in Southwest India

We used principal component analysis (PCA) to explore the broader population structure within South Asia, with a particular focus on Southwest India. The main axes of genetic variation highlight the ANI-ASI genetic cline (Figure 2A, B). The Kodava, Kodava_US, Bunt, and Nair populations are placed in PCA-space close to other populations that are geographically proximal (e.g., Iyer, Iyangar, and individuals from Urban Bangalore from (*2*, *7*)). This pattern was also captured by genetic ancestry clusters from ADMIXTURE where, at K = 7, the Kodava, Kodava_US, Bunt, and Nair shared genetic components (colored pink and gray in Figure 2C, Figure S1) with many other populations from South Asia (Figure 2C, Figure S1). Moreover, they displayed components maximized in the Kalash (colored orange) and in Europe (colored red) that were also observed in populations from Central Asia and the Middle East as well as in most North Indian and some South Indian populations. Notably, the proportion of all these components in the Bunt, Kodava, Kodava_US, and Nair were in line with the proportions displayed by geographical neighbors such as Iyer and individuals from Urban Bangalore and Urban Chennai from (*7*). Furthermore, we did not detect any structure within the Nair despite sampling broadly across the state of Kerala or within the Kodava_US donors who, while being recent migrants to the US, originate from various locations within the Kodagu district in southern Karnataka. Based on these analyses, we also did not observe any notable differences in genetic ancestry between Kodava sampled in India and the Kodava sampled in the US (Kodava_US). These results are consistent with the diaspora being too recent for appreciable genetic differentiation and that recent migrants to the US are a representative random sample of the Indian population with respect to genetic ancestry. On the other hand, the Kapla showed genetic similarity to populations with higher ASI-related genetic ancestry from South India, such as the Ulladan and Paniya, displaying pink and gray components found in most South Asians but substantially lower orange component maximized in the Kalash and found in populations with higher ANI-related genetic ancestry (Figure 2A, B, C). Our results showed little qualitative support for substantial genetic similarity between the Kapla and the Siddi or tested African individuals (Figure 2C, Figure S3), as suggested in historical records.

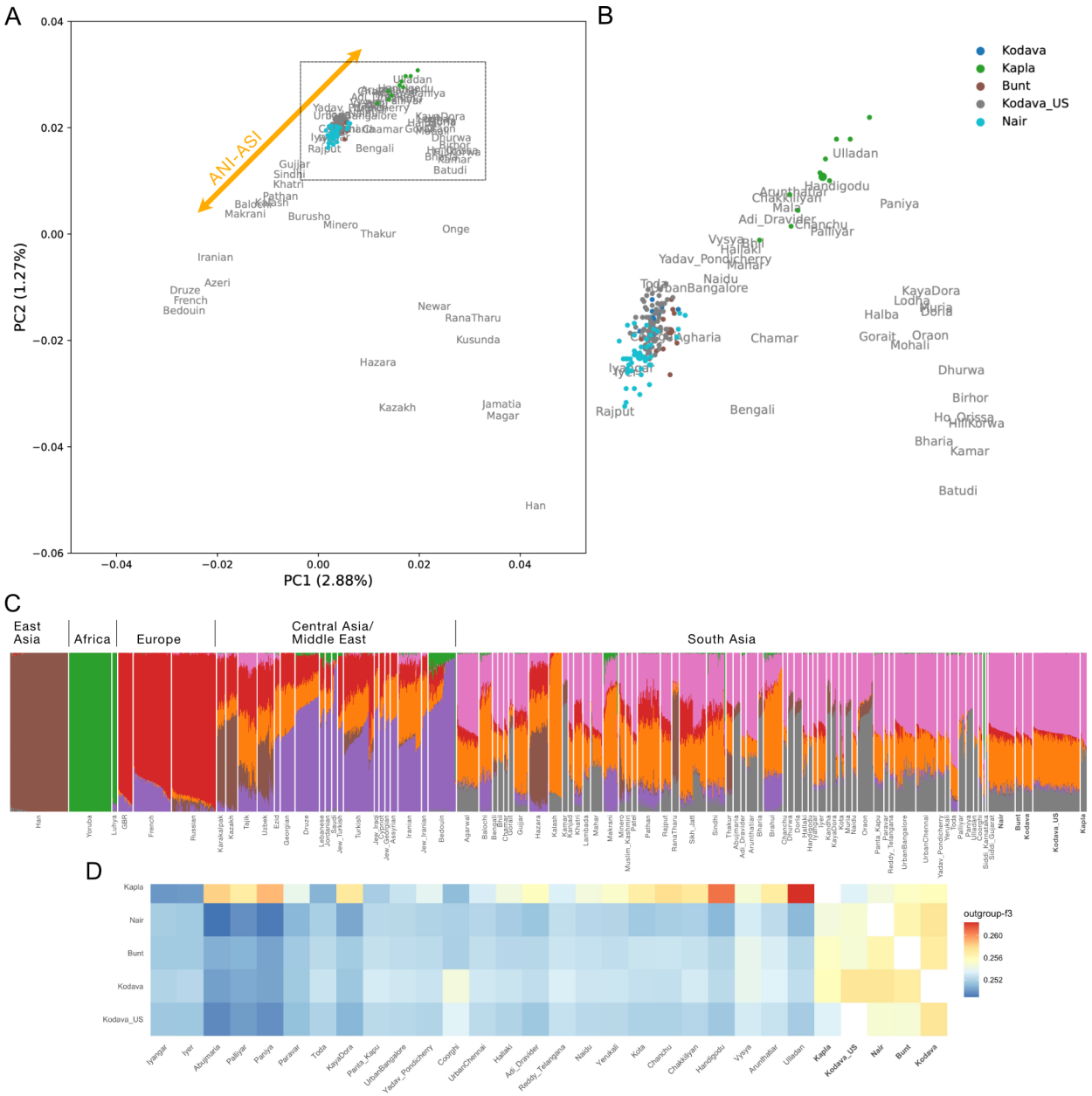

Figure 2. (A) Principal component analysis of merged dataset with Human Origins samples (see Methods), with (B) representing a zoomed inset focusing on the study populations and their position on the ANI-ASI genetic ancestry cline. Population labels in gray represent the mean location of all samples in PCA. For populations sampled specifically in this study, mean locations are represented as larger circles of the same color. (C) ADMIXTURE results (*K=7)* showcasing genetic similarity across Southwest Indian populations in relation to African, East Asian, Central Asian, Middle Eastern, and other South Asian populations. We have removed some population labels for visual clarity (see Figure S1 for a full set of labels). Populations sampled specifically in this study are represented using bolded population labels. (D) Outgroup f_3-statistics results for the study populations relative to other South Indian populations. See Figure S4 for an extended panel consisting of worldwide populations.

We evaluated shared genetic drift, reflecting genetic affinities, between the study populations from Southwest India and several Eurasian populations using an outgroup f_3-statistic of the form f_3*(Study_populations, X; Mbuti)*, where *Study_populations* correspond to one of the Southwest Indian populations sequenced in this study and *X* to a set of Eurasian populations from the dataset described

in Methods. We observed a higher genetic affinity between all our study populations and other Indian populations (Figure 2D, Figure S4A-E), particularly from South India, which is consistent with the results of PCA and ADMIXTURE. More specifically, patterns of shared drift between the study populations and other South Indian populations differed, as seen in previous analyses. The Kodava, Kodava_US, Bunt, and Nair shared higher genetic affinity with each other and, to an extent, with populations that derive genetic ancestry from both ANI- and ASI-related sources, while the Kapla were genetically similar to populations with more ASI-related genetic ancestry such as tribal populations like Ulladan from Kerala (*32*) and individuals from Handigodu village of Shimoga District of Karnataka (*2*).

We constructed a maximum-likelihood tree using Treemix and modeled up to ten admixture events using a subset of present-day and ancient populations from Eurasia and an African outgroup (Mbuti), excluding populations with known recent admixture from Africa and East Asia (Table S2 and Figures S5A-K). Without migration edges, we observed expected broad-level relationships between the populations, with Western and Eastern Eurasian populations forming distinct clusters. South Asian populations fell between these clusters, with populations that maximized ANI-related genetic ancestry falling closer to Western Eurasians and populations that maximized ASI-related genetic ancestry falling closer to the Onge and East Asians. The Kodava, Kodava_US, Bunt, and Nair were closely related to each other and to other South Indian populations such as Iyer and Iyangar. The Kapla fell closer to populations with a higher proportion of ASI-related ancestry, such as Paniya and Palliyar, consistent with previous analyses (Figure 2). Few early migration edges (Table S3) recapitulated previously reported events such as the Steppe-related gene flow, represented by ancient Central Steppe MLBA individuals from Russia, into the French. We postulated that subsequent tree configurations and associated migration edges represent attempts by the algorithm to optimize the placement of South Asian populations along the Eastern-Western Eurasian genetic ancestry gradient.

Estimation of admixture proportions and timing

To directly model the study populations as a mixture of genetic ancestral components, we used both the f_4 ratio test and qpAdm. A two-way admixture model tested using the f_4 ratio method has been shown to be overly simplified for most Indian populations (*3*, *9*, *33*, *34*), and we used it primarily to assess the relative proportions of Central Steppe MLBA related genetic ancestry, a proxy for ANI-related genetic ancestry, in the study populations (Figure S6). The proportion of Central Steppe MLBA in the Nair, Bunt, Kodava, and Kodava_US ranged between 45% (+/- 1.4%) and 48% (+/- 1.3%), similar to the proportion found in other populations with higher ANI-related genetic ancestry, including some South Indian and the majority of North Indian populations in India. In contrast, the proportion of Central Steppe MLBA in Kapla was genetically similar to neighboring South Indian tribal populations with higher ASI-related genetic ancestry (27.17% +/- 2.1%), which was lower than the other study populations (Figure S8).

Next, we employed qpAdm to model genetic ancestry proportions in the study populations using a more complex mix of genetic sources (*35*, *36*) (see Methods), particularly as a three-way mixture of genetic ancestries related to Central Steppe MLBA, Indus Periphery Cline (predominantly, ancient Iranian farmer-related genetic ancestry), and Onge as proposed by (*3*) (Table S5). We noted that several South Asian populations, including a majority of those sequenced in the study, had a p-value < 0.05 (*35*), which was also observed in the original study (*3*), suggesting that this demographic model may not be accurately capturing the complexities of genetic ancestries represented by many populations on the ANI-ASI cline. In particular, we observed that this model was plausible (p-value > 0.05) for those South Asian populations with very little Central Steppe MLBA genetic ancestry. This observation was validated by Kapla being the only study population with a p-value over 0.05 and could be modeled as having very little (4.6%) Central Steppe MLBA-related genetic ancestry compared to genetic ancestry related to Indus Periphery Cline (40.9%) and Onge (54.5%). These genetic ancestry proportions were similar to Palliyar,

Paniya, and Ulladan, which are South Indian tribal groups with higher ASI-related genetic ancestry and also yielded p-values greater than 0.05 in this analysis. On the other hand, we failed to recover well-supported qpAdm models for those populations with appreciable Central Steppe MLBA-related genetic ancestry, including the Bunt, Kodava (both groups), and Nair, as well as previously published data from populations like the Iyangar from South India and most tested North Indian populations. While we could not accept a working model for these populations based strictly on p-values, we noted that qpAdm does infer very similar proportions of the three ancestral genetic sources in the Bunt, Kodava (both groups), and Nair, with higher proportions of genetic ancestries related to Central Steppe MLBA and the Indus Periphery Cline in these populations compared to the Kapla (Table S5). Broadly, the genetic source proportions in the study populations were within the range displayed by other South Asian populations, though the genetic ancestries related to Central Steppe MLBA and Indus Periphery were qualitatively at the higher end of the range in the Nair and Kodava compared to those observed in neighboring Indian populations. The relative amounts of allele sharing of our study and other South Asian populations with the three genetic sources, determined using *D*-statistics (Figure S9A-C), were in agreement with these results and our other analyses. More genetic data, especially from ancient samples from the region, can potentially shed further light on the complexities in genetic ancestry that may be missing in the current model.

To estimate the timing of the introduction of genetic ancestral source(s) related to ANI in South Asia, we ran ALDER on select South Asian populations (*3*, *9*, *37*). Importantly, we did not establish *a priori* present-day sources, but instead used PC-derived weights reflective of the ANI-ASI ancestry cline to anchor the ancestry sources (see methods; (*9*). The date estimates for populations with higher Central Steppe MLBA-related genetic ancestry, typically from North India, ranged from 48.04 ± 22.87 to 166.66 ± 26.94 generations. South Indian populations with higher ASI-related genetic ancestry tended to display slightly older dates than the former group, ranging from 74.19 ± 36.47 to 234.96 ± 97.71 generations (Table S7), in agreement with previous studies (*3*, *9*). Excluding the Kapla that did not yield a significant result (z-score >=2, corresponding to a p-value < 0.05), the ANI-ASI admixture time estimates for the populations sequenced in this study ranged between 102.61 ± 30.32 and 138.8 ± 29.99 generations, which was within the range displayed by other populations included in the analysis (*37*) (Table S7). As pointed out previously , these dates represent a complex series of admixture events and, particularly for populations with both the Indus Periphery Cline and Central Steppe MLBA-related genetic ancestries, may reflect an average of these events (*3*, *9*).

Characterizing fine-scale genetic admixture using haplotype-based analyses

While allele-frequency based methods can be useful for highlighting population structure, they are limited in that they can only show structure emerging at the time-scale of allele frequency drift. Therefore, it is also useful to take advantage of patterns of linked variation in haplotypes, since haplotypes are broken up by recombination events over shorter time-scales than allele frequency drift, providing clearer signal for more recent and finer-scale patterns of population structure (*38*, *39*).

To evaluate population structure and genetic affinities at a finer scale, we implemented a haplotype-based analysis using ChromoPainter and fineSTRUCTURE (*38*). In this analysis, we first investigated general patterns of haplotype similarities followed by regional level patterns, including only South Indian populations. The approach taken here derives from the Li and Stephens “haplotype copying model” (*40*) where each individual (recipient) is modeled as a mosaic of haplotypes copied from a set of haplotypes that derive from other individuals in the dataset (donors). Therefore, the degree to which an individual copies their haplotypes from another individual reflects the genetic similarity between those individuals at a haplotypic level.

At a broad level, we observed that Kodava (both groups), Bunt, Nair, and Kapla copied most of their haplotypes from South Asian donors, particularly South Indian populations, consistent with previous analyses (Figure S10-11). At a regional scale, the populations sequenced in this study clustered into three groups: i) Nair and Bunt, ii) Kodava, Kodava_US and Coorghi, and iii) Kapla (Figure 3 and Figure S11). Within the first two clusters, we could not identify further sub-structure. In the case of Kapla, we observed haplotype similarity with Ulladan and Handigodu village donors. The additional sub-structure uncovered through haplotype-based methods across the Nair, Bunt, and Kodava groups suggests more recent population structure in these populations.

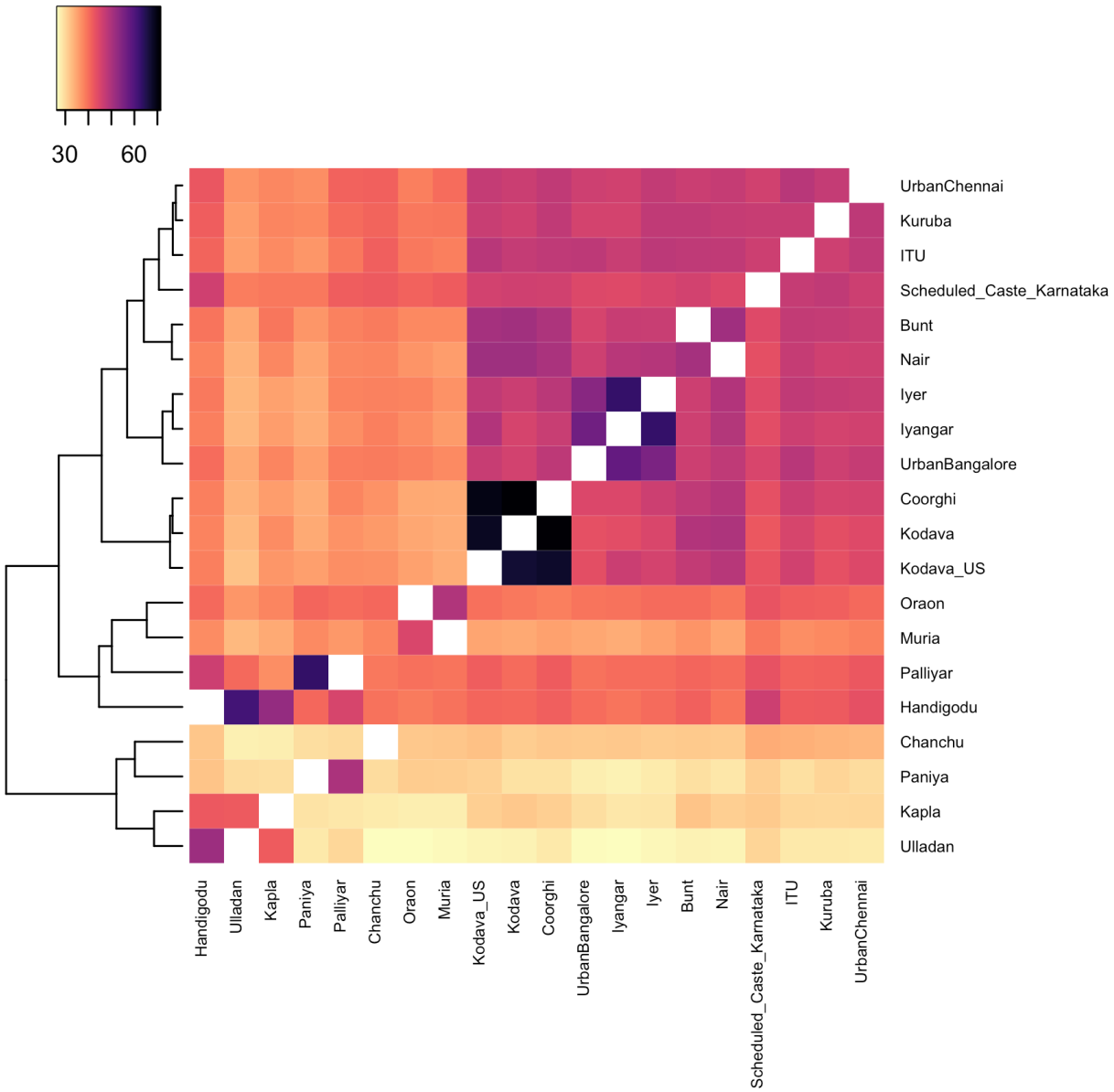

Figure 3: Haplotype-copying co-ancestry matrix across the regional level representing South Indian populations. Color bar represents the average length of haplotypes shared between populations.

Genetic variation within the Kapla population

Our analyses suggested that in addition to being genetically differentiated from the other populations sampled in this study, the Kapla also exhibited substantial spread along the first PC (Figure 2B). In particular, the ten Kapla individuals fell along the ANI-ASI cline, with some individuals closer to the ASI end of the cline while others were slightly closer to other South Asian populations with higher ANI ancestry. While this may be due to sampling bias, we noted that other South Indian populations maximizing ASI ancestry like Ulladan, Paniya, and Vysya did not display as high variance on the PCA as the Kapla (Figure S12). As the ADMIXTURE results did not suggest major genetic differences within Kapla, and all individuals showed little genetic contribution from the Indus Periphery Cline and Central Steppe MLBA-related genetic ancestries (Figure 2C, Figure S1), we formally evaluated the PCA observation by separating Kapla individuals into two groups based on the PCA results where: Kapla_A included 6 individuals closer to the ASI end of the cline and Kapla_B included the other 4 individuals closer to the ANI end. We acknowledge that these groups may not represent true structure within the population and resulted from an arbitrary division to evaluate potential differences observed in the PCA. We estimated *D*-statistics of the form *D(Kapla_B, Kapla_A; H3, Mbuti)*, where H3 denotes populations from Europe and the Middle East, and those populations from South Asia that fall on the ANI-ASI cline (see Methods). In agreement with the PCA, we found that most tested European and Middle Eastern populations were significantly closer to the four Kapla individuals closer to the ANI end (Kapla_B). In contrast, a few tested populations that maximized ASI ancestry, such as Ulladan, Paniya, and Palliyar, were significantly closer to Kapla_A (Table S8). To infer the admixture proportions of the three sources (Central Steppe MLBA, Indus Periphery Cline and Onge) in Kapla_A and Kapla_B, we implemented qpAdm using sources and outgroups from (*3*) (Table S5). We observed significant p-values in these tests, with Kapla_A exhibiting a slightly higher proportion of Onge-related ancestry (57.6%) compared to Kapla_B (49.7) and Kapla_B exhibiting a slightly higher proportion of the other two components: 7.6% Central Steppe MLBA versus 2.7% in Kapla_A and 42.7% Indus Periphery Cline versus 39.6% in Kapla_A. This latter signal may explain the PCA observation. We could not further narrow down the mode or timing of the within-population differentiation we observed above (e.g., via recent admixture or more ancient structure) due to the small sample size and substantial ancestry sharing between Kapla and neighboring South Indian populations included in our tests.

Analysis of uniparental markers

The populations sequenced in this study exhibited a mix of maternal lineages primarily found today in South Asia (haplogroup M) and Western Eurasia (haplogroups R, J, U, HV) (Figure 4A, Table S9). Macro-haplogroup M makes up more than half the maternal lineages in India today, ranging between 58% and 72%, with tribal groups and South Indian populations at the higher end of this range (*41–43*). In the study populations, we observed those subclades of haplogroup M previously reported primarily in India (*42*, *44*). Macro-haplogroup R is predominantly observed in Eurasia (*45*) and is reported to make up more than 40% of the maternal lineages in India (*43*). In the study populations, we primarily observed lineages of haplogroup R that are reported to have originated in South Asia (*41*, *45–47*). Macro-haplogroup U is one of the oldest clades of haplogroup R (*44*) and there are multiple subclades associated with specific regions across Eurasia (*48*). We found lineages of this haplogroup in all the study populations except Kapla, especially those previously reported in ancient and present-day individuals from the Near East, South Asia, Central Asia, and Southeast Asia (*3*, *48*, *49*). One of the individuals identifying as a Kodava_US displayed a subclade of macro-haplogroup J (haplogroup J1c1b1a). This subclade has been reported in ancient individuals in Central and Eastern Europe while, more broadly, other lineages of sub-haplogroup J1c have been observed in Central and Eastern Europe, the Balkans, and, at a lower frequency, in the Near East (*3*, *34*, *50*, *51*). To our knowledge, the subclade of haplogroup J that we observed has not been reported in South Asia thus far. In addition, some sequenced individuals

identifying as Kodava, Kodava_US, and Nair belonged to subclades of macro-haplogroup HV, a major subclade of haplogroup R0 (*52*). We observed subclades of this haplogroup that have been reported in ancient and present-day individuals from South Asia, Central Asia, the Middle East, West Asia, Central and Eastern Europe, and the Caucasus (*3*, *52–56*). One individual identifying as Kodava_US was assigned to haplogroup A1a, a subclade of haplogroup A primarily observed in East Asia (*57*) that has, to our knowledge, not been reported in Indian populations.

We observed a total of eight Eurasian Y-chromosome haplogroups (R, H, L, G, N, J, E, Q) (Figure 4B, Table S10). Most Y-chromosome haplogroups in India and Pakistan are assigned to haplogroup R (*58*), followed by haplogroup H, which is suggested to have originated in South-Central Asia or the Middle East and later spread into India (*58*, *59*). Haplogroup L has a frequency of 7-15% in India and 28% in Balochistan and western parts of Pakistan (*58*). In our data, we observed those lineages of R, H, and L in all populations that have been reported in present-day and ancient individuals from South Asia, Southeast Asia, Central Asia, Arabian Peninsula, Europe, and present-day Turkey (*3*, *60–64*). Haplogroup G is found at high frequencies in the Middle East and southern Europe and, at a low frequency, in South Asia (*65*). In our data, we found a lineage of this haplogroup in one Nair individual that is distributed in South Asia, the Middle East, and Eastern Europe (*66*). Haplogroup N is prominent throughout Northern Eurasia (*67*). We observed lineage N1 (N-L735), a less common lineage of this haplogroup, in one of the individuals identifying as Kodava_US. The lineages of haplogroup J that we observed in Kodava, Kodava_US, and Nair populations are reported in ancient and present-day individuals from South and Central Asia (*3*, *68*). One of the individuals identifying as Nair displayed macro-haplogroup E-M96, which has been reported in ancient and present-day individuals in Eastern Europe and South Asia (*3*, *56*). The lineages of haplogroup Q that we observed in Kodava, Kodava_US, and Nair populations have been reported in ancient and present-day individuals from across Eurasia (*3*, *51*, *69*, *70*).

While we found a few mitochondrial and Y-chromosome haplogroups in the study populations that have not been reported previously in South Asian populations, autosomal analyses suggest that these individuals are not genetic outliers relative to other individuals from the same populations. This may suggest female-mediated genetic admixture in the past, ancient mitochondrial DNA structure, or dilution of putative non-local admixing sources by male-mediated local admixture.

We used the proportion of uniquely observed Y-chromosome haplogroups and uniquely observed mitochondrial haplogroups for each of the study population as a proxy for the relative diversity displayed at these two loci in our data. We omitted Bunt from this analysis due to the small sample size, especially for the Y-chromosome haplogroup assignments. The ratio of unique Y-chromosome haplogroups to the mitochondrial haplogroups for the study populations were 0.8 for Kodava, 0.6 for Kodava_US, 1.0 for Nair, and 2.3 for Kapla. This implies a slightly elevated haplogroup diversity associated with the Y-chromosome compared to the mitochondria in the Nair and the Kapla, while the opposite is true for the two Kodava groups.

To explore the relationship between uniparental markers and matrilocality for the Nair, we compared the haplotype and pairwise diversity on the mitochondrial DNA (mtDNA) similar to analyses from (*71*) (Table S11). For groups with similar sample sizes, we found the strongest difference in haplotype diversity to be between the Bunt and Kodava samples ($p = 0.19$; $t = -0.88$), which does not reject the null hypothesis of similar haplotype diversity matrilocal and patrilocal groups in our dataset.

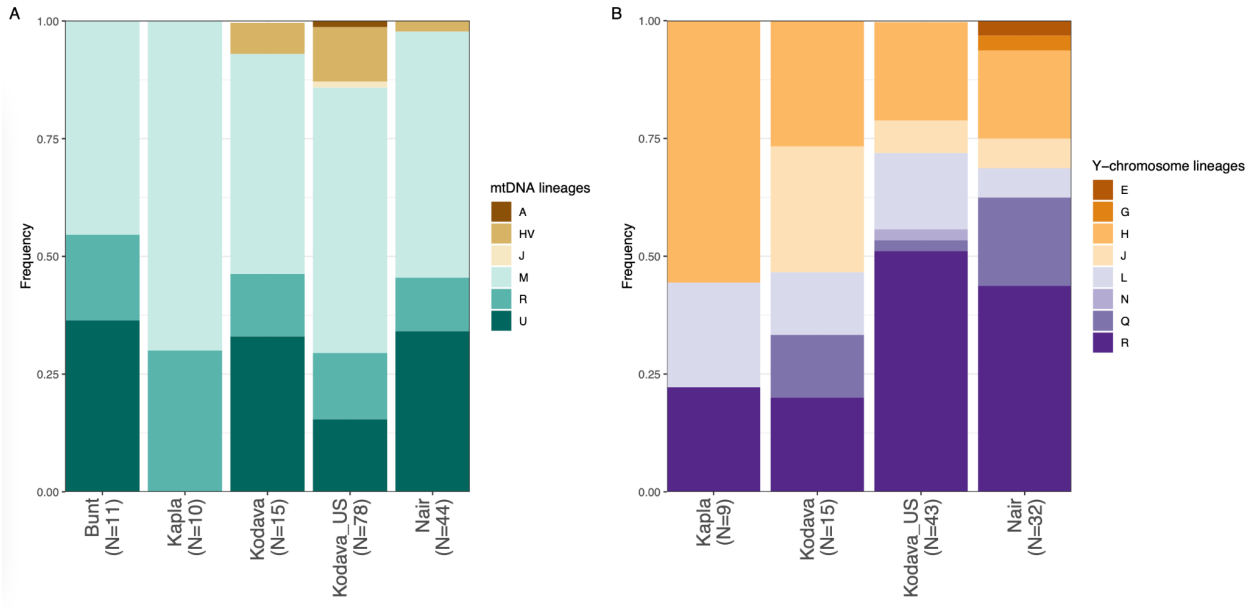

Figure 4: Distribution of the major mtDNA **(A)** and Y chromosome **(B)** macro-haplogroups from the analysis of 158 complete mitochondrial genomes and 99 complete Y chromosomes from this study. Bunt were not included in the Y chromosome analysis due to a single male in the dataset.

Endogamy in Southwest Indian populations

Founder effects and endogamy are prominent demographic forces shaping genetic diversity across populations of South Asian genetic ancestries (*2*, *8*, *72*). We sought to characterize rates of endogamy in the study populations relative to other populations sampled in India as endogamy can increase the frequency of recessive disease alleles. We calculated the Identity-by-Descent (IBD) score (see Methods), which measures the extent to which unrelated individuals in a population share segments of the genome identical-by-descent (Figure 5).

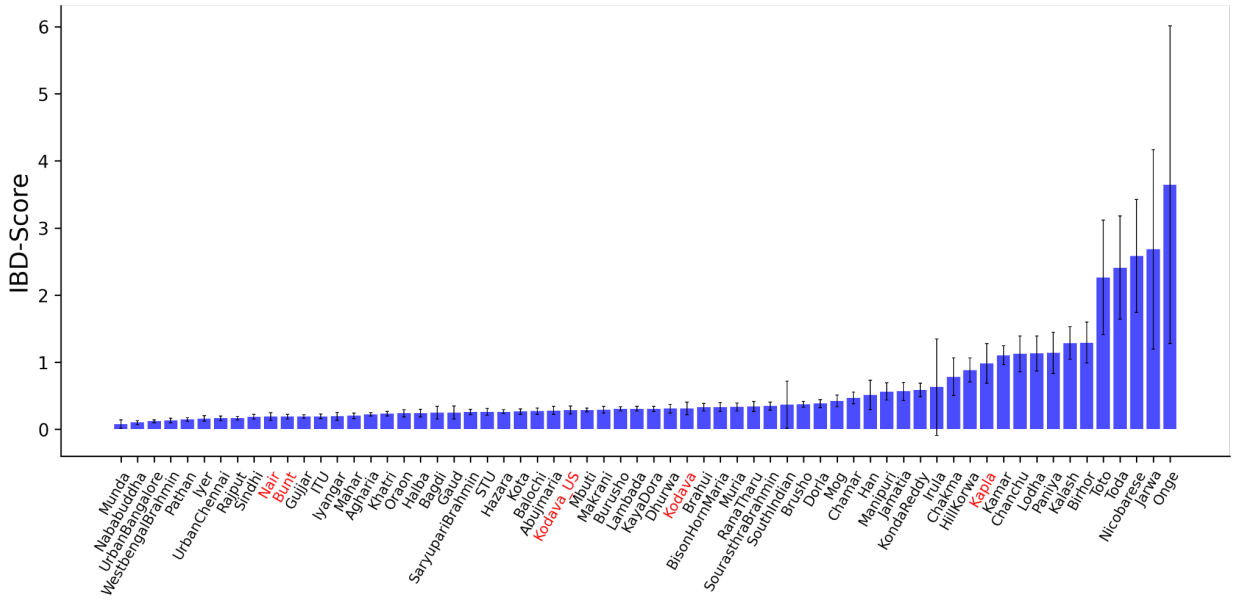

Figure 5: IBD score across multiple populations in South Asia. Populations that were sequenced in this study are highlighted in red.

We found that the Nair and Bunt populations had similar levels of endogamy as captured by the IBD score. Additionally, both Kodava populations had higher IBD scores than the Bunt and Nair but were consistent between themselves. We found the Kapla to have elevated levels of endogamy relative to geographically neighboring populations such as the Kodava.

We estimated runs of homozygosity (ROH) to further explore endogamy in the study populations (see Methods). In agreement with the IBD scores, we observed the Kapla population had an increased number of ROH (NROH) and generally longer ROH (>10Mb) (Figures S13 and S14), consistent with smaller effective population size and higher levels of consanguinity (*73*).

Novel variant discovery in Southwest Indian populations

We generated high-coverage genomes (>30x) for a total of nine individuals representing the Bunt, Kapla, Kodava_US, and Nair (see Methods), and evaluated the presence of variants not previously categorized in existing catalogs of genetic diversity. After alignment, quality control filters, and genotype calling across these whole-genome sequencing (WGS) samples, we ran Ensembl Variant Effect Predictor (VEP) v105.0 (see Methods) to assess whether each variant was found in pre-existing catalogs of variation and their putative functional consequences. We used 1000 Genomes, GnomAD, and GenomeAsia as catalogs of allelic variation against which to compare and establish novel occurrences of variants. We observed 11,132,650 total variants (8,562,627 SNPs, 2,570,023 indels) and, of these, 208,041 (151,137 SNPs, 56,904 indels) are not found in existing catalogs of variation, which we term as “novel variants”. Of these novel variants, we subsequently addressed the proportion of SNPs and indels that are in specific functional impact classes as defined by VEP (modifier, low, medium, high). For novel SNPs, there were 150,315 modifier, 349 low, 444 moderate, and 29 high impact variants. For novel indels, the proportions are quite similar with 56,714 modifier, 49 low, 36 moderate, and 105 high impact variants. We also noted that the majority of the putatively high impact variants were singletons, i.e. only found on a single haplotype out of the 18 that are tested, which is to be expected based on the action of negative selection (*74*).

We also explored the burden of moderate and high effect variants within this small set of high-coverage genomes, given previous evidence of endogamy in Indian populations influencing the burden of deleterious mutations (*2*, *7*). When looking at only observed mutations with moderate or high predicted effects based on VEP, we found that there is an enrichment of deleterious variants in the Kodava_US population ($p < 1e-12$; Mann-Whitney U: 24e6 between Kodava_US and Bunt), which does not support a direct relationship between predicted endogamy (Figure 5) and burden of deleterious variants (Figure S15A). However, this is confounded by the fact that the Kodava_US samples had ~77x coverage compared to 30x coverage in the high-coverage genomes from the other populations, so there was a higher absolute number of variants discovered in the Kodava (US) individuals.

We hypothesized that populations with higher fractions of ASI ancestry would have a higher fraction of novel variants due to undersampling of ASI-related genetic ancestry in known catalogs of human allelic variation. We found a strong enrichment of novel variants in the Kapla, supporting our initial hypothesis ($p < 1e-12$; Mann-Whitney U: 25e6 between Kapla and Bunt). In absolute terms, this is a difference of ~ 3,000 variants per haplotype (Figure S15B). Overall, these findings motivate conducting whole-genome surveys of genomic variation in underrepresented genetic ancestries both for discovery of functional genetic variation and improving catalogs of allelic diversity.

Discussion

This study engages with the genetic and oral histories of populations from Southwest India. By analyzing whole-genome sequences from participants in India and the US, we reconstruct the broader and fine-scale genetic relationship of the study populations to worldwide and other South Indian populations, respectively. We find the Bunt, Kodava, and Nair to share close genetic ancestry with each other and with other neighboring South Indian populations. These populations, along with several other Indian populations, carry genetic ancestries related to Onge, Indus Periphery Cline, and Central Steppe MLBA. This mix of genetic ancestries is reflected in the uniparental markers that also show a mix of South Asian and Western Eurasian mtDNA and Y chromosome haplogroups. Future studies are needed to better understand the distribution ranges of haplogroups not typical of this region, such as mitochondrial haplogroup A1a.

Our result of not detecting additional sources of Western Eurasian ancestry in the Bunt, Kodava, and Nair motivates anthropological follow-up to better understand the connections in their oral histories to non-local populations such as ancient Scythians and Alexander's armies. It is possible that our sampling scheme and/or analytical power limits the detection of such a signal in the genetic data, particularly if subsequent admixture with local Indian populations caused a dilution of non-local genetic signatures in these populations. Moreover, the timing of ANI-ASI admixture, reported to be between 2,000-4,000 years before present (*9*), may confound detection of admixture events noted in oral histories, such as that between local Indian populations and members of Alexander's army sometime after 327 BCE. In fact, the Nair and Kodava individuals sequenced in this study have a slightly higher proportion of Western Eurasian ancestry compared to neighboring populations in South India, though comparable to and/or lower than most North Indian populations. The available data in this study precludes determination of the reason for this slightly elevated signal in these populations, including potential sampling bias and long-standing endogamy. Still, the observed discordance between oral and genetic histories leads to a crucial aspect involving the danger of conflating self-identities with genetic ancestries. For example, could these origin stories be reflective of past cultural and/or economic contacts with the latter populations rather than involving gene flow? Moreover, are there aspects of these populations' oral histories that are suggestive of the temporal context of the non-local contacts, and should these necessarily be expected to coincide with admixture times estimated from genetic data? In studies of demographic histories, the expectation

of origin stories arising from oral traditions to converge with genetic histories, by using the former to motivate hypotheses that are then either accepted or rejected by the genetic data, tends to suggest that genetic and oral histories should be alignable. On the contrary, they represent distinct facets of individual and group identities that should not be forced to reconcile by pitting one against another (*18*, *20*) or be assumed to occur across similar timeframes. Taking a similar stance, we stress that our findings based on genetic data should not be used to validate or reject community origin stories based on oral traditions. Instead, we encourage further anthropological research on the nature of the relationships with non-local populations prevalent in the oral histories of the populations included in our study. This would serve to enhance our understanding of the cultural and oral legacy of these populations and potentially illuminate new dimensions of the regional history. Such conversations are especially pertinent today as the genomics revolution coupled with direct-to-consumer initiatives over the past decade has brought conversations around genetics and ancestry squarely into the public domain and calls for increased engagement from researchers to explain the interpretive scope of the data (*75*).

At a more fine-scale level, haplotype-based analysis suggests more recent genetic contacts between Bunt and Nair populations. Moreover, the close genetic relationship observed between the Kodava from India and Kodava_US supports these individuals deriving from the same Kodava population in India. Although we lack socio-cultural context for the Coorghi individuals genotyped in (*2*), we conjecture that the genetic similarity of the Coorghi and the two Kodava datasets from this study may, again, be capturing their common origin since Coorghi is an anglicized version of Kodava that was adopted during the colonial times. Overall, close genetic affinities between the Bunt, Kodava, and Nair do, in fact, complement their oral histories that speak to historical cultural contacts as well as overlapping self-identities, likely facilitated by their geographical proximity to one another. Our genetic analyses do not detect any discernible sub-structure within the Bunt, Kodava, and Nair despite the Nair and Kodava donors originating in different locations within Kerala and Kodagu, respectively, and the complex socio-cultural subgroups recognized within the Nair population. This may suggest that geography and social stratification have not produced long-term reductions to gene flow within these populations.

We also find little statistical support from genetic diversity statistics (Table S11) to distinguish between groups anthropologically associated to be matrilocal. From haplotype and nucleotide diversity on mtDNA, we did not find statistically significant lower levels of these metrics expected for matrilocal groups, especially the Nair, compared to patrilocal groups. From haplogroup diversity on the Y chromosome relative to the mtDNA, the Nair do display a slightly higher diversity than the Kodava, which may be expected under matrilocality. However, this result is qualitative and it is likely that larger sample sizes per group would be required in order to address this in a stronger quantitative sense. There may be other possible explanations for a lack of clear signal in uniparental markers between matrilocal and patrilocal populations, such as recent shifts from strict matrilocality that could potentially dilute the signal in the mtDNA genetic diversity, or smaller long-term population size that could result in shifts in diversity which may not be well-accounted for in our analyses. We leave these as future areas of work bridging matrilocal and patrilocal population dynamics to observed genetic data.

In contrast to the other study populations, the Kapla are genetically more similar to tribal South Indian populations, which have higher proportions of Onge-related genetic ancestry and negligible amounts of the Central Steppe MLBA genetic component. In this regard, the Kapla may represent one of several close genetic descendants of the ASI group (*3*). A close link of the Kapla to the ASI group is also evident in the uniparental markers, which are enriched for haplogroups with proposed South Asian origins. The lack of substantial support for a Siddi or African origin for the Kapla in the genetic data raises an interesting follow up to these links proposed in historical records (Figure S3). Furthermore, despite their geographical proximity to the Kodava and socio-economic relationships between these two populations observed by the research team, stark differences in lifestyle between the two populations may have resulted in more long-term genetic isolation of the Kapla from the Kodava. Their genetic isolation and

elevated IBD score may be reflected in historical narratives, both published accounts that state the Kapla "*consist of only 15 families*" (*30*) and were relocated to their present location by a local ruler, invoking a founder population, and anecdotal accounts that suggest they were then isolated from neighboring populations. However, we do detect slightly higher genetic ancestry related to Central Steppe MLBA and Indus Periphery Cline in four Kapla individuals relative to the others, which may be due to either recent contacts with Kodava and possibly other neighboring populations, necessitated by socio-economic contacts between these populations, or more ancient structure. Future studies with increased sample sizes are needed to further characterize the genetic variability within the Kapla more accurately. Once again, this opens up an anthropological avenue to document the histories and lifestyles of populations such as the Kapla, especially in the face of socio-cultural transitions and changing traditions, and to observe whether such ongoing processes are accompanied by genetic admixture.

Although our analyses were not designed to make claims regarding complex traits or disease, we expect to find an increased prevalence of recessive genetic disorders in populations with higher rates of endogamy (*2*, *8*, *76*). While follow-up studies with detailed phenotype information across populations in Southwest India will bring more evidence to bear on the relationship between endogamy and disease risk, particularly for recessive diseases, we find all four study populations to be within the range of IBD scores and ROH displayed by other Indian populations. Notably, the Kapla registered a higher IBD score and increased number and length of ROH compared to the Bunt, Kodava, and Nair, which suggests elevated endogamy in the Kapla, presumably due to their isolation and smaller population size. Our analysis of deleterious variants in a smaller, but higher-coverage, whole-genome dataset found that the Kapla were not the population with the highest burden of deleterious mutations. However, the Kapla did contain significantly higher rates of novel variants not found in existing public catalogs of allelic variation (Figure S15B).

The results from this study provide an impetus for follow-up studies to further characterize the levels of genetic variation in these and other populations in India through increased sample sizes and phenotype collection. Additionally, our results suggest ways for future studies of population histories to engage more closely with the populations and with relevant fields such as anthropology. Importantly, such work should be conducted with substantial community engagement to ensure accurate recording of information on those aspects of their history that the population would like to further follow up on, given the sensitivities around identities stemming from oral histories.

Methods

Sampling and community engagement

This project was approved by the Mangalore University Institutional Review Board (IRB) (MU-IHEC-2020-4), the Institutional Ethical Committee of Birbal Sahni Institute of Palaeosciences, Lucknow (BSIP/Ethical/2021), and the University of Chicago Biological Sciences Division IRB (IRB18-1572). Sampling was conducted in two phases. In phase one, community contact and sampling of whole blood were performed in January 2018 in South India in the states of Karnataka, Kerala, and Tamil Nadu. Prior to sampling, project aims, methods, and other relevant details were explained to interested community members, and informed consent principles were observed during enrollment in the study. In several locations, local translators, community members, and long-term community contacts joined the research team to facilitate communication with the broader communities. Members of the research team traveled to Kodagu, southern Karnataka, to conduct sampling of Kodava community members (N = 15). Here, the team also met with members of the Kapla community, primarily men, who had come down from a mountain settlement close by to work in the Kodagu coffee plantations. Their isolated lifestyle, compared to the Kodava, motivated engagement with Kapla community members and their inclusion in this study

(N = 10). The team also traveled to Mangalore, southern Karnataka, to conduct sampling of Bunt community members originating in the districts of Udupi and Dakshina Kannada in southern Karnataka and Kasaragod district in northern Kerala (N = 11). Moreover, the team visited four locations in Kerala and Tamil Nadu to conduct sampling of Nair community members (N = 44) with origins in the districts of Pathanamthitta, Palakkad, Kozhikode, and Kannur in Kerala. The sampling in India included engagement with individuals and smaller groups of community members who chose whether or not to participate in the study after the team's presentation of the project and responses to their questions. The research team also collected snippets of oral histories and information on social structures and unique cultural traits from several participants from each community. Subsequent anthropological interviews were conducted by co-author M.S. in India involving members of the study populations, in particular to follow up on specific origin narratives that were noted by the research team during initial sampling. In phase two, the research team in the US was approached by members of the Kodava diaspora in the US to partner up on a project investigating their population history. Community engagement was conducted through two community representatives in the US (coauthors C.D.K. and A.P.N.), who initiated contacts with other community members. Saliva kits, consent forms, and a brief questionnaire were mailed out to interested participants, and 105 participants returned all three to the research team and were enrolled in the study. This dataset was subsequently merged with phase one data, given overlapping aims, with the consent of members of the Kodava_US community. During the course of the project, the research team remained in contact with members of the Kodava_US community, providing updates on the study progress as well as collecting more information on oral traditions. Results of this study were disseminated to community members of all the study populations, both in India and in the US. Results were returned through a combination of written reports, presentations by the researchers, and a newspaper article, with special efforts taken to make the information more accessible to a non-scientific audience. Dissemination was performed in English and local languages.

Sample processing and genotype calling

For the Kodava, Nair, Bunt, and Kapla populations from India, DNA extracts from whole blood were performed in India and were sent to MedGenome Inc, Bangalore, India, for sequencing on Illumina HiSeq X Ten. They were sequenced to an average autosomal depth of 2.5x. For members of the Kodava_US population, saliva samples were collected using the Oragene DNA self-collection kit and extraction was performed using the QIAcube standard protocol in Chicago. The extracts were sequenced on the Illumina NovaSeq 6000 at Novogene, Sacramento, USA in four batches as noted in Table S1, to an average autosomal depth of 5.5X. Four samples were re-extracted due to low DNA concentration in the first extraction round and are denoted as ID*_b in Table S1. We re-sequenced three Kodava_US individuals to 77x average depth and two individuals each from Bunt, Kapla, and Nair populations to 30x average depth for variant discovery and for validation of concordance in our subsequent genotype calling pipeline. Table S1 summarizes the sequencing results. All reads were subsequently filtered to have a mapping quality score greater than 30 and were aligned to the human reference genome build 37 (hg19) and the revised Cambridge Reference Sequence (rCRS) build 17, for autosomal + Y-chromosome and mtDNA variation, respectively.

Genotypes were called on all high and low coverage samples jointly using the GATK (v4.0) germline short variant discovery workflow on the 8,183,696 sites from the Genome Asia panel (*7*). After this initial phase of genotype calling, the GATK phred-scaled genotype likelihoods were converted to genotype probabilities. IMPUTE2 (v2.3.2) along with the 1000 Genome Phase 3 reference panel were used to perform genotype refinement (*77*, *78*). After genotype refinement, hard-called genotypes were set to the genotype with the maximum probability and genotypes with a maximum probability less than 0.9 were considered missing. Following genotype calling and refinement, chromosomes were phased using SHAPEIT4 (v4.2.2) along with the 1000 Genome Phase3 reference panel (*79*).

Merging with external datasets and quality control

We merged our sequence data with publicly available data from the GenomeAsia 100K consortium (*7*) and the Human Genome Diversity Project (*80*). To provide additional context in South Asia, especially to increase representation of Indian populations, we further merged this dataset with genotype data from (*2*). Following these merges, we retain 425,620 SNPs for analysis of population structure.

In order to compare genetic affinities with ancient samples, we merged samples from the Allen Ancient DNA Resource (*81*), which contains genotypes across ~1.23 million sites represented on the '1240k capture panel'. We also merged the publicly available Human Origins dataset from (*33*), which contains present-day worldwide samples. As most of the variants covered in the Human Origins dataset are also represented in the 1240k panel, we retain the same number of variants for downstream analysis. We used this final merged dataset, consisting of our study samples and five external SNP datasets, for all downstream analyses in this manuscript.

Following these merging steps, we filtered samples according to relatedness (up to 2nd degree) using KING (*82*) and removed samples with > 5% SNP missingness for all downstream population genetic analyses. We did not apply this filter to data from ancient samples to retain sparse genotyping data for these individuals. Applying these criteria to our data led to the retention of all individuals as described in the *Sampling and community engagement* section, except for the Kodava_US, from which we excluded 27 samples for a final sample number of 78 from the 105 sampled initially.

Principal Component Analysis and ADMIXTURE

We performed a principal component analysis (PCA) using smartpca v7.3.0 (*83*, *84*). We filtered the merged datasets according to linkage disequilibrium using PLINK 1.9 [`--indep-pairwise 200 25 0.4`] (*85*). We chose to plot only population median locations in PCA-space (as gray text labels) of populations from external datasets for ease of visualization. We represent population medians for populations exclusive to this study by larger dots of the same color (Figure 2B).

This same set of LD-pruned variants was used to run ADMIXTURE v1.30 (*86*), with K = 6 to 11 (Figure S1). The value of K=7 on the largest merged dataset minimized the cross-validation error (Figure S2). We restricted visualization of clustering results in the main text figure to representative populations from each geographic region with at least 8 samples.

Admixture history using Treemix

We modeled the relationship between the study populations in Southwest India and a subset of present-day and ancient groups from Eurasia using Treemix (*87*). Positions were filtered by missingness and LD-pruned with PLINK 1.9 as in PCA. The analysis was run, estimating up to 10 migration edges with Mbuti as the outgroup.

Ancestry estimation using *f*-statistics

For calculating *f*-statistics defined in (*88*), we used the ADMIXTOOLS v7.0.2 software. The first statistic we used was the outgroup f_3-statistic, which measures the shared drift or genetic similarity between populations (A, B) relative to an outgroup population (O) and was calculated using the `qp3pop` program. We always use the Mbuti population for our outgroup (*80*), unless otherwise noted.

To model the study populations (*Study*) as mixtures of ANI and ASI ancestries, we used the f_4-ratio test (*8*). We measured the proportion of Central Steppe MLBA (α) genetic ancestry (*3*) via the ratio (see Figure S8):

$$\alpha = \frac{f4(Mbuti, Karelians\,;\, Study, Onge)}{f4(Mbuti, Karelians\,;\, MLBA, Onge)}$$

We selected populations to include in the above topologies based on the scheme in (*9*). Briefly, we ran a *D*-statistic using the `qpDstat` program of the form *Dstat(Central_Steppe_MLBA, Iran_GanjDareh; H3; Ethiopian4500)* to evaluate the degree of allele sharing between Central Steppe MLBA and H3. As H3, we selected a large array of present-day populations from Eurasia. We identified the population with the highest value in this test (Karelia) as the closest population to Central Steppe MLBA (Figure S28).

We computed allele sharing of Kapla_A and Kapla_B with different present-day populations (H3) from Europe, the Middle East, and South Asia on the ANI-ASI cline using *D*-statistics *(88)* using the `qpDstat` program of the form *D(Kapla_B, Kapla_A; H3, Mbuti)*. For all computed *D*-statistics, we used the (`f4mode: NO, printsd: YES`) options in `qpDstat`.

Estimation of ancestry proportions based on qpAdm

To model the ancestry proportions of the study and other select Indian populations for comparison, we used qpAdm (*35*, *36*) implemented in ADMIXTOOLS v7.0.2 software with (`allsnps: YES, oldallsnpsmode: NO, inbreed: NO, details: YES`). As many present-day South Asians on the ANI-ASI cline can be modeled as a mix of Indus Periphery Cline, Onge, and/or Central Steppe MLBA related genetic ancestries (*3*), we considered these as our genetic sources for the test populations included in the analysis (Table S4). For the outgroups, we considered the list outlined in the proximal model of present-day South Asians by (*3*) (Table S4). However, most of the populations, including the study populations, had p-values less than 0.05 (*35*) with this set of outgroups (Table S5), suggesting that the proposed model may require more complex admixture patterns in the South Asian populations. This phenomena is also seen in (*3*).

Admixture dates based on LD decay in ALDER

To estimate the timing of ANI-related genetic admixture in the study populations, we used the weighted admixture linkage disequilibrium decay method in ALDER (*37*). The true ancestral populations for South Asian populations are unknown, so we followed the approach implemented in (*3*, *9*). Briefly, this modified approach approximates the spectrum of ANI-ASI admixture using PCA-derived SNP-weights to capture ancestry-associated differences in allele frequency without requiring explicitly labeled source populations. Following (*9*), we included primarily Dravidian- and Indo-European-speaking populations in India that fall on the ANI-ASI cline on the PCA and have a sample size greater than five, and Basque individuals from Europe to function as an anchor for Western Eurasian ancestry when constructing the SNP-weights (Table S6-S7). We estimated the admixture timings inferred by the program in generations and in years, assuming 1 generation is 28 years (*3*) (Table S6-S7). We did not include individuals from the study population when running PCA. For this reason, we also excluded the Coorghi from the reference panel and the test group to avoid skewing the analysis due to historical connections between the names 'Coorgh' and 'Kodagu', referring to the same region in Southwest India. Based on historical accounts, Coorghi and Kodava may be the same or closely related communities. Unless otherwise stated, we use

Z-score >= 2 (p-value < 0.05) across ALDER tests as indicative of a significant signal of admixture (*9*, *37*).

Haplotype-based Estimation of population structure using fineStructure

We implemented a haplotype-based approach using the software ChromoPainter and fineSTRUCTURE (*38*). Briefly, this approach estimates a co-ancestry matrix based on haplotype sharing (ChromoPainter), which can be used later to identify population structure (fineSTRUCTURE). We first estimated the recombination scaling constant Ne and the mutation parameter θ in a subset of chromosomes (1, 5, 10, and 20) with 10 E-M iterations. The average value for both parameters was estimated across chromosomes and used for the subsequent runs (Ne = 401.352, θ = 0.0002). We standardized the number of individuals per population to 5, filtering out groups with a sample size below that threshold and randomly sampling five individuals from larger groups. The co-ancestry matrix obtained from ChromoPainter was then used for inferences in fineSTRUCTURE. The analysis was run at different levels of population composition; the first included all Eurasian populations in the database, followed by South Asia-specific runs.

mtDNA and Y-chromosome analyses

The alignment files for each sample were used to retrieve reads mapping specifically to the mitochondrial genome using samtools (*89*). Variants from mtDNA reads for each sample were called against the revised Cambridge Reference Sequence (rCRS) using `bcftools mpileup` and `bcftools call`, requiring a mapping quality of 30 and base quality of 20. The 158 mtDNA Variant Call Format (VCF) files were used as input for assigning mitochondrial DNA (mtDNA) haplogroups using the program Haplogrep2 (*90*) with PhyloTree mtDNA tree Build 17 (*91*). The final haplogroups reported in the study were picked based on the highest quality score and rank assigned by the program (Table S9).

For assigning Y-chromosome haplogroups, we first extracted reads mapping to the Y-chromosome using samtools. Variants from Y-chromosome reads were then called using `bcftools mpileup` and `bcftools call`, requiring a mapping quality of 30 and base quality of 20 (`-d 2000 -m 3 -EQ 20 -q 30`). The 100 Y-chromosome VCF files were then used to call haplogroups using the program Yhaplo (*92*) (Table S10).

To calculate the relative genetic diversity across the two uniparental loci in the study populations, we computed the proportion of uniquely observed Y-chromosome and mtDNA haplogroups in each population (e.g. 21 unique Y-chromosome haplogroups out of 32 sampled male Nair individuals results in a proportion of 0.7 and 29 unique mtDNA haplogroups out of 44 sampled Kodava individuals results in a proportion of 0.7). We, then, calculated the ratio of these two estimates to check if any of the study populations had elevated levels of genetic diversity, as measured by haplogroups here, on one of the two loci.

To support our investigation of matrilocality and patrilocality influencing uni-parentally inherited markers, we calculated haplotype diversity (*71*) and nucleotide diversity on the mitochondrial DNA (*93*). We estimated standard deviations of haplotype diversity within populations using 100 bootstrap samples and used one-sided t-tests to evaluate differences between haplotype diversity.

Estimation of IBD Scores

For IBD-based analyses, we wanted to measure the extent of endogamy in each population and computed the IBD score from (*2*). The IBD score is defined as the sum of IBD between 3 cM and 20 cM

detected between individuals of the same population, divided by $\frac{2n(2n-1)}{2} - n$, where "n" is the number of individuals. We used GERMLINE2 (*94*) to call IBD segments with the flag `-m 3` to only consider IBD segments at least 3 cM in length, using the deCODE genetic map coordinates for genome build hg19.

Following (*2*), we also removed individuals related up to second degree using KING (*82*) and removed IBD segments > 20 cM. In (*2*), the authors normalized to individuals of Finnish and Ashkenazi Jewish ancestry as they had access to these samples, and they are both examples where there exists a higher recessive disease burden. However, we have left the IBD scores as raw values, so we only interpret the results relatively across the other Indian populations we tested against.

Runs of Homozygosity

We estimate runs of homozygosity (ROH) using PLINK 1.9 following the recommendations from (*95*) and (*73*), using the parameters: `--homozyg-window-snp 50 --homozyg-snp 50 --homozyg-kb 1500 --homozyg-gap 1000 --homozyg-density 50 --homozyg-window-missing 5 --homozyg-window-het 1`. To increase SNP density, we limited our analysis to a panel including only whole-genome sequencing data from GenomeAsia 100K (*7*) and the Human Genome Diversity Project (*80*). After filtering positions with more than 0.5% missing data and MAF of 5%, the final database has 3,288,336 variants.

Novel Variant Discovery

To perform variant discovery, we relied on nine high-coverage WGS samples (coverage > 30x) from four population groups. In total, we had three Kodava_US, two Bunt, two Kapla, and two Nair individuals in this smaller set of individuals with WGS data. Alignment to GrCh37 (hg19) was performed using bwa-mem (*96*) (Table S1). We called genotypes within these individuals using GATK v.3.0 best practices and filtered to SNPs and indels with a phred-scaled quality score of > 30 to carry forward for variant annotation.

For variant annotation, we used Ensembl VEP v105.0 with custom annotations derived from GnomAD genomes site-level VCF files and GenomeAsia site-level VCF files. In addition, we used the `--af_gnomad --af_1kg` flags in order to detect whether the alleles were found in the 1000 Genomes project or in GnomAD exomes. Therefore, our definition of novel variants only includes variants that were not found in any of the aforementioned datasets. When assessing variant effects, we used the broad categories provided by VEP (MODIFIER, LOW, MODERATE, HIGH).

References

1. A. Basu, N. Sarkar-Roy, P. P. Majumder, Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 1594–1599 (2016).

2. N. Nakatsuka, P. Moorjani, N. Rai, B. Sarkar, A. Tandon, N. Patterson, G. S. Bhavani, K. M. Girisha, M. S. Mustak, S. Srinivasan, A. Kaushik, S. A. Vahab, S. M. Jagadeesh, K. Satyamoorthy, L. Singh, D. Reich, K. Thangaraj, The promise of discovering population-specific disease-associated genes in South Asia. *Nat. Genet.* **49**, 1403–1407 (2017).

3. V. M. Narasimhan, N. Patterson, P. Moorjani, N. Rohland, R. Bernardos, S. Mallick, I. Lazaridis, N. Nakatsuka, I. Olalde, M. Lipson, A. M. Kim, L. M. Olivieri, A. Coppa, M. Vidale, J. Mallory, V. Moiseyev, E. Kitov, J. Monge, N. Adamski, N. Alex, N. Broomandkhoshbacht, F. Candilio, K. Callan, O. Cheronet, B. J. Culleton, M. Ferry, D. Fernandes, S. Freilich, B. Gamarra, D. Gaudio, M. Hajdinjak, É. Harney, T. K. Harper, D. Keating, A. M. Lawson, M. Mah, K. Mandl, M. Michel, M. Novak, J. Oppenheimer, N. Rai, K. Sirak, V. Slon, K. Stewardson, F. Zalzala, Z. Zhang, G. Akhatov, A. N. Bagashev, A. Bagnera, B. Baitanayev, J. Bendezu-Sarmiento, A. A. Bissembaev, G. L. Bonora, T. T. Chargynov, T. Chikisheva, P. K. Dashkovskiy, A. Derevianko, M. Dobeš, K. Douka, N. Dubova, M. N. Duisengali, D. Enshin, A. Epimakhov, A. V. Fribus, D. Fuller, A. Goryachev, A. Gromov, S. P. Grushin, B. Hanks, M. Judd, E. Kazizov, A. Khokhlov, A. P. Krygin, E. Kupriyanova, P. Kuznetsov, D. Luiselli, F. Maksudov, A. M. Mamedov, T. B. Mamirov, C. Meiklejohn, D. C. Merrett, R. Micheli, O. Mochalov, S. Mustafokulov, A. Nayak, D. Pettener, R. Potts, D. Razhev, M. Rykun, S. Sarno, T. M. Savenkova, K. Sikhymbaeva, S. M. Slepchenko, O. A. Soltobaev, N. Stepanova, S. Svyatko, K. Tabaldiev, M. Teschler-Nicola, A. A. Tishkin, V. V. Tkachev, S. Vasilyev, P. Velemínský, D. Voyakin, A. Yermolayeva, M. Zahir, V. S. Zubkov, A. Zubova, V. S. Shinde, C. Lalueza-Fox, M. Meyer, D. Anthony, N. Boivin, K. Thangaraj, D. J. Kennett, M. Frachetti, R. Pinhasi, D. Reich, The formation of human populations in South and Central Asia. *Science*. **365** (2019), doi:10.1126/science.aat7487.

4. K. Tätte, L. Pagani, A. K. Pathak, S. Kõks, B. Ho Duy, X. D. Ho, G. N. N. Sultana, M. I. Sharif, M. Asaduzzaman, D. M. Behar, Y. Hadid, R. Villems, G. Chaubey, T. Kivisild, M. Metspalu, The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. *Sci. Rep.* **9**, 3818 (2019).

5. A. K. Pathak, A. Kadian, A. Kushniarevich, F. Montinaro, M. Mondal, L. Ongaro, M. Singh, P. Kumar, N. Rai, J. Parik, E. Metspalu, S. Rootsi, L. Pagani, T. Kivisild, M. Metspalu, G. Chaubey, R. Villems, The Genetic Ancestry of Modern Indus Valley Populations from Northwest India. *Am. J. Hum. Genet.* **103**, 918–929 (2018).

6. M. Metspalu, M. Mondal, G. Chaubey, The genetic makings of South Asia. *Curr. Opin. Genet. Dev.* **53**, 128–133 (2018).

7. GenomeAsia100K Consortium, The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature*. **576**, 106–111 (2019).

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

9. P. Moorjani, K. Thangaraj, N. Patterson, M. Lipson, P.-R. Loh, P. Govindaraj, B. Berger, D. Reich, L. Singh, Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93**, 422–438 (2013).

10. G. Debortoli, C. Abbatangelo, F. Ceballos, C. Fortes-Lima, H. L. Norton, S. Ozarkar, E. J. Parra, M. Jonnalagadda, Novel insights on demographic history of tribal and caste groups from West Maharashtra (India) using genome-wide data. *Sci. Rep.* **10**, 10075 (2020).

11. E. Arciero, S. A. Dogra, M. Mezzavilla, T. Tsismentzoglou, Q. Q. Huang, K. A. Hunt, D. Mason, D. A. van Heel, E. Sheridan, J. Wright, N. Small, S. Carmi, M. M. Iles, H. C. Martin, Fine-scale population structure and demographic history of British Pakistanis. *bioRxiv* (2020), p. 2020.09.02.279190.

12. S. Finer, H. C. Martin, A. Khan, K. A. Hunt, B. MacLaughlin, Z. Ahmed, R. Ashcroft, C. Durham, D.

G. MacArthur, M. I. McCarthy, J. Robson, B. Trivedi, C. Griffiths, J. Wright, R. C. Trembath, D. A. van Heel, Cohort Profile: East London Genes & Health (ELGH), a community-based population genomics and health study in British Bangladeshi and British Pakistani people. *Int. J. Epidemiol.* **49**, 20–21i (2020).

13. C. P. Silva, C. de la Fuente Castro, T. González Zarzar, M. Raghavan, A. Tonko-Huenucoy, F. I. Martínez, N. Montalva, The Articulation of Genomics, Mestizaje, and Indigenous Identities in Chile: A Case Study of the Social Implications of Genomic Research in Light of Current Research Practices. *Front. Genet.* **13** (2022), doi:10.3389/fgene.2022.817318.

14. K. Hwang, International Collaboration in Multilayered Center-Periphery in the Globalization of Science and Technology. *Sci. Technol. Human Values*. **33**, 101–133 (2008).

15. D. Haelewaters, T. A. Hofmann, A. L. Romero-Olivares, Ten simple rules for Global North researchers to stop perpetuating helicopter research in the Global South. *PLoS Comput. Biol.* **17**, e1009277 (2021).

16. J. M. Argüelles, A. Fuentes, B. Yáñez, Analyzing asymmetries and praxis in aDNA research: A bioanthropological critique. *Am. Anthropol.* **124**, 130–140 (2022).

17. A. E. R. Prince, B. E. Berkman, Reconceptualizing harms and benefits in the genomic age. *Per. Med.* **15**, 419–428 (2018).

18. R. J. Crellin, O. J. T. Harris, Beyond binaries. Interrogating ancient DNA. *Archaeological Dialogues*. **27**, 37–56 (2020).

19. B. Donovan, R. H. Nehm, Genetics and Identity. *Science & Education*. **29**, 1451–1458 (2020).

20. K. TallBear, Genomic articulations of indigeneity. *Soc. Stud. Sci.* **43**, 509–533 (2013).

21. K. M. Panikkar, Some Aspects of Nayar Life. *The Journal of the Royal Anthropological Institute of Great Britain and Ireland*. **48**, 254–293 (1918).

22. M. N. Srinivas, *Religion and society among the Coorgs of South India* (Asia Publishing House., Bombay, 1965).

23. E. Thruston, *The Caste and Tribes of Southern India* (Cosmo Publications, New Delhi. , 1909), *Reprint*.

24. D. M. Schneider, *Matrilineal Kinship, Edited by David M. Schneider and Kathleen Gough* (University of California Press, 1962 C1961, 1962).

25. M. Kushalappa, *The Early Coorgs: A History of Early Kodagu and Its People* (Createspace Independent Pub, 2013).

26. C. J. Fuller, The Internal Structure of the Nayar Caste. *Journal of Anthropological Research*. **31** (1975), pp. 283–312.

27. A. Menon, thesis, University of Cambridge (2018).

28. C. J. Fuller, *The Nayars Today* (Cambridge University Press, 1976).

29. C. Karumbaya, "Kodavas through the ages" in *Are Kodavas (Coorgs) Hindus?*, P. T. Bopanna, Ed.

(Rolling stone Publications, Bangalore, 2018).

30. G. Richter, *Manual of Coorg: A Gazetteer of the natural features of the Country and the social and political condition of its inhabitants* (Basel Mission Book Depository [Published by C Stolz], Mangalore, India, 1870).

31. A. M. Shah, R. Tamang, P. Moorjani, D. S. Rani, P. Govindaraj, G. Kulkarni, T. Bhattacharya, M. S. Mustak, L. V. K. S. Bhaskar, A. G. Reddy, D. Gadhvi, P. B. Gai, G. Chaubey, N. Patterson, D. Reich, C. Tyler-Smith, L. Singh, K. Thangaraj, Indian Siddis: African descendants with Indian admixture. *Am. J. Hum. Genet.* **89**, 154–161 (2011).

32. Scheduled Tribes Development Department, *Scheduled Tribes Development Department Website* (2022), (available at https://www.stdd.kerala.gov.in).

33. I. Lazaridis, N. Patterson, A. Mittnik, G. Renaud, S. Mallick, K. Kirsanow, P. H. Sudmant, J. G. Schraiber, S. Castellano, M. Lipson, B. Berger, C. Economou, R. Bollongino, Q. Fu, K. I. Bos, S. Nordenfelt, H. Li, C. de Filippo, K. Prüfer, S. Sawyer, C. Posth, W. Haak, F. Hallgren, E. Fornander, N. Rohland, D. Delsate, M. Francken, J.-M. Guinet, J. Wahl, G. Ayodo, H. A. Babiker, G. Bailliet, E. Balanovska, O. Balanovsky, R. Barrantes, G. Bedoya, H. Ben-Ami, J. Bene, F. Berrada, C. M. Bravi, F. Brisighelli, G. B. J. Busby, F. Cali, M. Churnosov, D. E. C. Cole, D. Corach, L. Damba, G. van Driem, S. Dryomov, J.-M. Dugoujon, S. A. Fedorova, I. Gallego Romero, M. Gubina, M. Hammer, B. M. Henn, T. Hervig, U. Hodoglugil, A. R. Jha, S. Karachanak-Yankova, R. Khusainova, E. Khusnutdinova, R. Kittles, T. Kivisild, W. Klitz, V. Kučinskas, A. Kushniarevich, L. Laredj, S. Litvinov, T. Loukidis, R. W. Mahley, B. Melegh, E. Metspalu, J. Molina, J. Mountain, K. Näkkäläjärvi, D. Nesheva, T. Nyambo, L. Osipova, J. Parik, F. Platonov, O. Posukh, V. Romano, F. Rothhammer, I. Rudan, R. Ruizbakiev, H. Sahakyan, A. Sajantila, A. Salas, E. B. Starikovskaya, A. Tarekegn, D. Toncheva, S. Turdikulova, I. Uktveryte, O. Utevska, R. Vasquez, M. Villena, M. Voevoda, C. A. Winkler, L. Yepiskoposyan, P. Zalloua, T. Zemunik, A. Cooper, C. Capelli, M. G. Thomas, A. Ruiz-Linares, S. A. Tishkoff, L. Singh, K. Thangaraj, R. Villems, D. Comas, R. Sukernik, M. Metspalu, M. Meyer, E. E. Eichler, J. Burger, M. Slatkin, S. Pääbo, J. Kelso, D. Reich, J. Krause, Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*. **513**, 409–413 (2014).

34. M. E. Allentoft, M. Sikora, K.-G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, T. Ahlström, L. Vinner, A.-S. Malaspinas, A. Margaryan, T. Higham, D. Chivall, N. Lynnerup, L. Harvig, J. Baron, P. Della Casa, P. Dąbrowski, P. R. Duffy, A. V. Ebel, A. Epimakhov, K. Frei, M. Furmanek, T. Gralak, A. Gromov, S. Gronkiewicz, G. Grupe, T. Hajdu, R. Jarysz, V. Khartanovich, A. Khokhlov, V. Kiss, J. Kolář, A. Kriiska, I. Lasak, C. Longhi, G. McGlynn, A. Merkevicius, I. Merkyte, M. Metspalu, R. Mkrtchyan, V. Moiseyev, L. Paja, G. Pálfi, D. Pokutta, Ł. Pospieszny, T. D. Price, L. Saag, M. Sablin, N. Shishlina, V. Smrčka, V. I. Soenov, V. Szeverényi, G. Tóth, S. V. Trifanova, L. Varul, M. Vicze, L. Yepiskoposyan, V. Zhitenev, L. Orlando, T. Sicheritz-Pontén, S. Brunak, R. Nielsen, K. Kristiansen, E. Willerslev, Population genomics of Bronze Age Eurasia. *Nature*. **522**, 167–172 (2015).

35. É. Harney, N. Patterson, D. Reich, J. Wakeley, Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*. **217** (2021), doi:10.1093/genetics/iyaa045.

36. W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, K. Stewardson, Q. Fu, A. Mittnik, E. Bánffy, C. Economou, M. Francken, S. Friederich, R. G. Pena, F. Hallgren, V. Khartanovich, A. Khokhlov, M. Kunst, P. Kuznetsov, H. Meller, O. Mochalov, V. Moiseyev, N. Nicklisch, S. L. Pichler, R. Risch, M. A. Rojo Guerra, C. Roth, A. Szécsényi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, D.

Reich, Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*. **522**, 207–211 (2015).

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

38. D. J. Lawson, G. Hellenthal, S. Myers, D. Falush, Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

39. E. Han, P. Carbonetto, R. E. Curtis, Y. Wang, J. M. Granka, J. Byrnes, K. Noto, A. R. Kermany, N. M. Myres, M. J. Barber, K. A. Rand, S. Song, T. Roman, E. Battat, E. Elyashiv, H. Guturu, E. L. Hong, K. G. Chahine, C. A. Ball, Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nat. Commun.* **8**, 14238 (2017).

40. N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*. **165**, 2213–2233 (2003).

41. M. Metspalu, T. Kivisild, E. Metspalu, J. Parik, G. Hudjashov, K. Kaldma, P. Serk, M. Karmin, D. M. Behar, M. T. P. Gilbert, P. Endicott, S. Mastana, S. S. Papiha, K. Skorecki, A. Torroni, R. Villems, Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet.* **5**, 26 (2004).

42. A. Chandrasekar, S. Kumar, J. Sreenath, B. N. Sarkar, B. P. Urade, S. Mallick, S. S. Bandopadhyay, P. Barua, S. S. Barik, D. Basu, U. Kiran, P. Gangopadhyay, R. Sahani, B. V. R. Prasad, S. Gangopadhyay, G. R. Lakshmi, R. R. Ravuri, K. Padmaja, P. N. Venugopal, M. B. Sharma, V. R. Rao, Updating phylogeny of mitochondrial DNA macrohaplogroup m in India: dispersal of modern human in South Asian corridor. *PLoS One*. **4**, e7447 (2009).

43. R. J. Herrera, R. Garcia-Bertrand, Dispersals Into India. *Ancestral DNA, Human Origins, and Migrations* (2018), pp. 207–255.

44. J. M. Larruga, P. Marrero, K. K. Abu-Amero, M. V. Golubenko, V. M. Cabrera, Carriers of mitochondrial DNA macrohaplogroup R colonized Eurasia and Australasia from a southeast Asia core area. *BMC Evol. Biol.* **17**, 115 (2017).

45. C. Sylvester, J. S. Rao, A. Chandrasekar, M. S. Krishna, An Updated Phylogeny of mtDNA Haplogroup R8 Based on Complete Mitogenomes. *Journal of the Anthropological Survey of India*. **68**, 114–122 (2019).

46. L. Quintana-Murci, R. Chaix, R. S. Wells, D. M. Behar, H. Sayar, R. Scozzari, C. Rengo, N. Al-Zahery, O. Semino, A. S. Santachiara-Benerecetti, A. Coppa, Q. Ayub, A. Mohyuddin, C. Tyler-Smith, S. Qasim Mehdi, A. Torroni, K. McElreavey, Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor. *Am. J. Hum. Genet.* **74**, 827–845 (2004).

47. G. Chaubey, M. Karmin, E. Metspalu, M. Metspalu, D. Selvi-Rani, V. K. Singh, J. Parik, A. Solnik, B. P. Naidu, A. Kumar, N. Adarsh, C. B. Mallick, B. Trivedi, S. Prakash, R. Reddy, P. Shukla, S. Bhagat, S. Verma, S. Vasnik, I. Khan, A. Barwa, D. Sahoo, A. Sharma, M. Rashid, V. Chandra, A. G. Reddy, A. Torroni, R. A. Foley, K. Thangaraj, L. Singh, T. Kivisild, R. Villems, Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. *BMC Evol. Biol.* **8**, 227 (2008).

48. H. Sahakyan, B. Hooshiar Kashani, R. Tamang, A. Kushniarevich, A. Francis, M. D. Costa, A. K.

Pathak, Z. Khachatryan, I. Sharma, M. van Oven, J. Parik, H. Hovhannisyan, E. Metspalu, E. Pennarun, M. Karmin, E. Tamm, K. Tambets, A. Bahmanimehr, T. Reisberg, M. Reidla, A. Achilli, A. Olivieri, F. Gandini, U. A. Perego, N. Al-Zahery, M. Houshmand, M. H. Sanati, P. Soares, E. Rai, J. Šarac, T. Šarić, V. Sharma, L. Pereira, V. Fernandes, V. Černý, S. Farjadian, D. P. Singh, H. Azakli, D. Üstek, N. Ekomasova Trofimova, I. Kutuev, S. Litvinov, M. Bermisheva, E. K. Khusnutdinova, N. Rai, M. Singh, V. K. Singh, A. G. Reddy, H.-V. Tolk, S. Cvjetan, L. B. Lauc, P. Rudan, E. N. Michalodimitrakis, N. P. Anagnou, K. I. Pappa, M. V. Golubenko, V. Orekhov, S. A. Borinskaya, K. Kaldma, M. A. Schauer, M. Simionescu, V. Gusar, E. Grechanina, P. Govindaraj, M. Voevoda, L. Damba, S. Sharma, L. Singh, O. Semino, D. M. Behar, L. Yepiskoposyan, M. B. Richards, M. Metspalu, T. Kivisild, K. Thangaraj, P. Endicott, G. Chaubey, A. Torroni, R. Villems, Origin and spread of human mitochondrial DNA haplogroup U7. *Sci. Rep.* **7**, 46044 (2017).

49. W. Kutanan, J. Kampuansai, A. Brunelli, S. Ghirotto, P. Pittayaporn, S. Ruangchai, R. Schröder, E. Macholdt, M. Srikummool, D. Kangwanpong, A. Hübner, L. Arias, M. Stoneking, New insights from Thailand into the maternal genetic history of Mainland Southeast Asia. *Eur. J. Hum. Genet.* **26**, 898–911 (2018).

50. M. Pala, A. Olivieri, A. Achilli, M. Accetturo, E. Metspalu, M. Reidla, E. Tamm, M. Karmin, T. Reisberg, B. Hooshiar Kashani, U. A. Perego, V. Carossa, F. Gandini, J. B. Pereira, P. Soares, N. Angerhofer, S. Rychkov, N. Al-Zahery, V. Carelli, M. H. Sanati, M. Houshmand, J. Hatina, V. Macaulay, L. Pereira, S. R. Woodward, W. Davies, C. Gamble, D. Baird, O. Semino, R. Villems, A. Torroni, M. B. Richards, Mitochondrial DNA signals of late glacial recolonization of Europe from near eastern refugia. *Am. J. Hum. Genet.* **90**, 915–924 (2012).

51. P. de B. Damgaard, N. Marchi, S. Rasmussen, M. Peyrot, G. Renaud, T. Korneliussen, J. V. Moreno-Mayar, M. W. Pedersen, A. Goldberg, E. Usmanova, N. Baimukhanov, V. Loman, L. Hedeager, A. G. Pedersen, K. Nielsen, G. Afanasiev, K. Akmatov, A. Aldashev, A. Alpaslan, G. Baimbetov, V. I. Bazaliiskii, A. Beisenov, B. Boldbaatar, B. Boldgiv, C. Dorzhu, S. Ellingvag, D. Erdenebaatar, R. Dajani, E. Dmitriev, V. Evdokimov, K. M. Frei, A. Gromov, A. Goryachev, H. Hakonarson, T. Hegay, Z. Khachatryan, R. Khaskhanov, E. Kitov, A. Kolbina, T. Kubatbek, A. Kukushkin, I. Kukushkin, N. Lau, A. Margaryan, I. Merkyte, I. V. Mertz, V. K. Mertz, E. Mijiddorj, V. Moiyesev, G. Mukhtarova, B. Nurmukhanbetov, Z. Orozbekova, I. Panyushkina, K. Pieta, V. Smrčka, I. Shevnina, A. Logvin, K.-G. Sjögren, T. Štolcová, A. M. Taravella, K. Tashbaeva, A. Tkachev, T. Tulegenov, D. Voyakin, L. Yepiskoposyan, S. Undrakhbold, V. Varfolomeev, A. Weber, M. A. Wilson Sayres, N. Kradin, M. E. Allentoft, L. Orlando, R. Nielsen, M. Sikora, E. Heyer, K. Kristiansen, E. Willerslev, 137 ancient human genomes from across the Eurasian steppes. *Nature*. **557**, 369–374 (2018).

52. M. Shamoon-Pour, M. Li, D. A. Merriwether, Rare human mitochondrial HV lineages spread from the Near East and Caucasus during post-LGM and Neolithic expansions. *Sci. Rep.* **9**, 14751 (2019).

53. M. Derenko, B. Malyarchuk, A. Bahmanimehr, G. Denisova, M. Perkova, S. Farjadian, L. Yepiskoposyan, Complete mitochondrial DNA diversity in Iranians. *PLoS One*. **8**, e80673 (2013).

54. N. Al-Zahery, O. Semino, G. Benuzzi, C. Magri, G. Passarino, A. Torroni, A. S. Santachiara-Benerecetti, Y-chromosome and mtDNA polymorphisms in Iraq, a crossroad of the early human dispersal and of post-Neolithic migrations. *Mol. Phylogenet. Evol.* **28**, 458–472 (2003).

55. R. Yaka, A. Birand, Y. Yılmaz, C. Caner, S. C. Açan, S. Gündüzalp, P. Parvizi, A. Erim Özdoğan, İ. Togan, M. Somel, Archaeogenetics of Late Iron Age Çemialo Sırtı, Batman: Investigating maternal genetic continuity in north Mesopotamia since the Neolithic. *Am. J. Phys. Anthropol.* **166**, 196–207

(2018).

56. I. Mathieson, S. Alpaslan-Roodenberg, C. Posth, A. Szécsényi-Nagy, N. Rohland, S. Mallick, I. Olalde, N. Broomandkhoshbacht, F. Candilio, O. Cheronet, D. Fernandes, M. Ferry, B. Gamarra, G. G. Fortes, W. Haak, E. Harney, E. Jones, D. Keating, B. Krause-Kyora, I. Kucukkalipci, M. Michel, A. Mittnik, K. Nägele, M. Novak, J. Oppenheimer, N. Patterson, S. Pfrengle, K. Sirak, K. Stewardson, S. Vai, S. Alexandrov, K. W. Alt, R. Andreescu, D. Antonović, A. Ash, N. Atanassova, K. Bacvarov, M. B. Gusztáv, H. Bocherens, M. Bolus, A. Boroneanţ, Y. Boyadzhiev, A. Budnik, J. Burmaz, S. Chohadzhiev, N. J. Conard, R. Cottiaux, M. Čuka, C. Cupillard, D. G. Drucker, N. Elenski, M. Francken, B. Galabova, G. Ganetsovski, B. Gély, T. Hajdu, V. Handzhyiska, K. Harvati, T. Higham, S. Iliev, I. Janković, I. Karavanić, D. J. Kennett, D. Komšo, A. Kozak, D. Labuda, M. Lari, C. Lazar, M. Leppek, K. Leshtakov, D. L. Vetro, D. Los, I. Lozanov, M. Malina, F. Martini, K. McSweeney, H. Meller, M. Menđušić, P. Mirea, V. Moiseyev, V. Petrova, T. D. Price, A. Simalcsik, L. Sineo, M. Šlaus, V. Slavchev, P. Stanev, A. Starović, T. Szeniczey, S. Talamo, M. Teschler-Nicola, C. Thevenet, I. Valchev, F. Valentin, S. Vasilyev, F. Veljanovska, S. Venelinova, E. Veselovskaya, B. Viola, C. Virag, J. Zaninović, S. Zäuner, P. W. Stockhammer, G. Catalano, R. Krauß, D. Caramelli, G. Zariņa, B. Gaydarska, M. Lillie, A. G. Nikitin, I. Potekhina, A. Papathanasiou, D. Borić, C. Bonsall, J. Krause, R. Pinhasi, D. Reich, The genomic history of southeastern Europe. *Nature*. **555**, 197–203 (2018).

57. M. Tanaka, V. M. Cabrera, A. M. González, J. M. Larruga, T. Takeyasu, N. Fuku, L.-J. Guo, R. Hirose, Y. Fujita, M. Kurata, K.-I. Shinoda, K. Umetsu, Y. Yamada, Y. Oshida, Y. Sato, N. Hattori, Y. Mizuno, Y. Arai, N. Hirose, S. Ohta, O. Ogawa, Y. Tanaka, R. Kawamori, M. Shamoto-Nagai, W. Maruyama, H. Shimokata, R. Suzuki, H. Shimodaira, Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res.* **14**, 1832–1850 (2004).

58. D. G. Mahal, I. G. Matsoukas, The Geographic Origins of Ethnic Groups in the Indian Subcontinent: Exploring Ancient Footprints with Y-DNA Haplogroups. *Front. Genet.* **9**, 4 (2018).

59. S. Wells, *Deep Ancestry: Inside the Genographic Project* (National Geographic Books, 2007).

60. T. Kivisild, S. Rootsi, M. Metspalu, S. Mastana, K. Kaldma, J. Parik, E. Metspalu, M. Adojaan, H.-V. Tolk, V. Stepanov, M. Gölge, E. Usanga, S. S. Papiha, C. Cinnioğlu, R. King, L. Cavalli-Sforza, P. A. Underhill, R. Villems, The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am. J. Hum. Genet.* **72**, 313–332 (2003).

61. C. F. Kerchner, YDNA Haplogroup Descriptions & Information Links (2013).

62. A. Brunelli, J. Kampuansai, M. Seielstad, K. Lomthaisong, D. Kangwanpong, S. Ghirotto, W. Kutanan, Y chromosomal evidence on the origin of northern Thai people. *PLoS One*. **12**, e0181935 (2017).

63. M. Haber, J. Nassar, M. A. Almarri, T. Saupe, L. Saag, S. J. Griffith, C. Doumet-Serhal, J. Chanteau, M. Saghieh-Beydoun, Y. Xue, C. L. Scheib, C. Tyler-Smith, A Genetic History of the Near East from an aDNA Time Course Sampling Eight Points in the Past 4,000 Years. *Am. J. Hum. Genet.* **107**, 149–157 (2020).

64. G. Scorrano, A. Finocchio, F. De Angelis, C. Martínez-Labarga, J. Šarac, I. Contini, G. Scano, N. Novokmet, D. Frezza, O. Rickards, The genetic landscape of Serbian populations through mitochondrial DNA sequencing and non-recombining region of the Y chromosome microsatellites. *Coll. Antropol.* **41**, 275–296 (2017).

65. S. Rootsi, N. M. Myres, A. A. Lin, M. Järve, R. J. King, I. Kutuev, V. M. Cabrera, E. K. Khusnutdinova, K. Varendi, H. Sahakyan, D. M. Behar, R. Khusainova, O. Balanovsky, E. Balanovska, P. Rudan, L. Yepiskoposyan, A. Bahmanimehr, S. Farjadian, A. Kushniarevich, R. J. Herrera, V. Grugni, V. Battaglia, C. Nici, F. Crobu, S. Karachanak, B. Hooshiar Kashani, M. Houshmand, M. H. Sanati, D. Toncheva, A. Lisa, O. Semino, J. Chiaroni, J. Di Cristofaro, R. Villems, T. Kivisild, P. A. Underhill, Distinguishing the co-ancestries of haplogroup G Y-chromosomes in the populations of Europe and the Caucasus. *Eur. J. Hum. Genet.* **20**, 1275–1282 (2012).

66. C. 2019-2020 by Isogg, ISOGG 2019 Y-DNA Haplogroup Tree, (available at https://isogg.org/tree/index.html).

67. ISOGG 2007 Y-DNA Haplogroup Tree, (available at https://isogg.org/tree/2007/index07.html).

68. S. Sengupta, L. A. Zhivotovsky, R. King, S. Q. Mehdi, C. A. Edmonds, C.-E. T. Chow, A. A. Lin, M. Mitra, S. K. Sil, A. Ramesh, M. V. Usha Rani, C. M. Thakur, L. Luca Cavalli-Sforza, P. P. Majumder, P. A. Underhill, Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists. *The American Journal of Human Genetics*. **78** (2006), pp. 202–221.

69. Y.-Z. Huang, H. Pamjav, P. Flegontov, V. Stenzl, S.-Q. Wen, X.-Z. Tong, C.-C. Wang, L.-X. Wang, L.-H. Wei, J.-Y. Gao, L. Jin, H. Li, Dispersals of the Siberian Y-chromosome haplogroup Q in Eurasia. *Mol. Genet. Genomics*. **293**, 107–117 (2018).

70. H. S. AlSafar, M. Al-Ali, G. D. Elbait, M. H. Al-Maini, D. Ruta, B. Peramo, A. Henschel, G. K. Tay, Introducing the first whole genomes of nationals from the United Arab Emirates. *Sci. Rep.* **9**, 14725 (2019).

71. E. D. Gunnarsdóttir, M. R. Nandineni, M. Li, S. Myles, D. Gil, B. Pakendorf, M. Stoneking, Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nat. Commun.* **2**, 228 (2011).

72. A. Angural, A. Spolia, A. Mahajan, V. Verma, A. Sharma, P. Kumar, M. K. Dhar, K. K. Pandita, E. Rai, S. Sharma, Review: Understanding rare genetic diseases in low resource regions like Jammu and Kashmir - India. *Front. Genet.* **11**, 415 (2020).

73. F. C. Ceballos, P. K. Joshi, D. W. Clark, M. Ramsay, J. F. Wilson, Runs of homozygosity: windows into population history and trait architecture. *Nat. Rev. Genet.* **19**, 220–234 (2018).

74. M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S.-A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zöllner, J. C. Whittaker, S. L. Chissoe, J. Novembre, V. Mooser, An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science*. **337**, 100–104 (2012).

75. K. G. Claw, M. Z. Anderson, R. L. Begay, K. S. Tsosie, K. Fox, N. A. Garrison, Summer internship for INdigenous peoples in Genomics (SING) Consortium, A framework for enhancing ethical genomic research with Indigenous communities. *Nat. Commun.* **9**, 2957 (2018).

76. GUaRDIAN Consortium, S. Sivasubbu, V. Scaria, Genomics of rare genetic diseases-experiences from India. *Hum. Genomics*. **14**, 52 (2019).

77. B. Howie, C. Fuchsberger, M. Stephens, J. Marchini, G. R. Abecasis, Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

78. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature*. **526**, 68–74 (2015).

79. O. Delaneau, J.-F. Zagury, M. R. Robinson, J. L. Marchini, E. T. Dermitzakis, Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).

80. A. Bergström, S. A. McCarthy, R. Hui, M. A. Almarri, Q. Ayub, P. Danecek, Y. Chen, S. Felkel, P. Hallast, J. Kamm, H. Blanché, J.-F. Deleuze, H. Cann, S. Mallick, D. Reich, M. S. Sandhu, P. Skoglund, A. Scally, Y. Xue, R. Durbin, C. Tyler-Smith, Insights into human genetic variation and population history from 929 diverse genomes. *Science*. **367** (2020), doi:10.1126/science.aay5012.

81. Allen ancient genome diversity project / John Templeton ancient DNA atlas, (available at https://reich.hms.harvard.edu/ancient-genome-diversity-project).

82. A. Manichaikul, J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, W.-M. Chen, Robust relationship inference in genome-wide association studies. *Bioinformatics*. **26**, 2867–2873 (2010).

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

84. N. Patterson, A. L. Price, D. Reich, Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

85. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, J. J. Lee, Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. **4**, 7 (2015).

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

87. J. K. Pickrell, J. K. Pritchard, Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).

88. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, D. Reich, Ancient admixture in human history. *Genetics*. **192**, 1065–1093 (2012).

89. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience*. **10** (2021), doi:10.1093/gigascience/giab008.

90. H. Weissensteiner, D. Pacher, A. Kloss-Brandstätter, L. Forer, G. Specht, H.-J. Bandelt, F. Kronenberg, A. Salas, S. Schönherr, HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–63 (2016).

91. M. Van Oven, PhyloTree Build 17: Growing the human mitochondrial DNA tree. *Forensic Science International: Genetics Supplement Series*. **5**, e392–e394 (2015).

92. G. D. Poznik, White Paper 23-13 yHaplo| Identifying Y-chromosome haplogroups in arbitrarily large

samples of sequenced or genotyped men (2016) (available at https://permalinks.23andme.com/pdf/23-13_paternal_haplogroups_yHaplo.pdf).

93. K. L. Korunes, K. Samuk, pixy: Unbiased estimation of nucleotide diversity and divergence in the presence of missing data. *Mol. Ecol. Resour.* **21**, 1359–1368 (2021).

94. J. Nait Saada, G. Kalantzis, D. Shyr, F. Cooper, M. Robinson, A. Gusev, P. F. Palamara, Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat. Commun.* **11**, 6130 (2020).

95. P. K. Joshi, T. Esko, H. Mattsson, N. Eklund, I. Gandin, T. Nutile, A. U. Jackson, C. Schurmann, A. V. Smith, W. Zhang, Y. Okada, A. Stančáková, J. D. Faul, W. Zhao, T. M. Bartz, M. P. Concas, N. Franceschini, S. Enroth, V. Vitart, S. Trompet, X. Guo, D. I. Chasman, J. R. O'Connel, T. Corre, S. S. Nongmaithem, Y. Chen, M. Mangino, D. Ruggiero, M. Traglia, A.-E. Farmaki, T. Kacprowski, A. Bjonnes, A. van der Spek, Y. Wu, A. K. Giri, L. R. Yanek, L. Wang, E. Hofer, C. A. Rietveld, O. McLeod, M. C. Cornelis, C. Pattaro, N. Verweij, C. Baumbach, A. Abdellaoui, H. R. Warren, D. Vuckovic, H. Mei, C. Bouchard, J. R. B. Perry, S. Cappellani, S. S. Mirza, M. C. Benton, U. Broeckel, S. E. Medland, P. A. Lind, G. Malerba, A. Drong, L. Yengo, L. F. Bielak, D. Zhi, P. J. van der Most, D. Shriner, R. Mägi, G. Hemani, T. Karaderi, Z. Wang, T. Liu, I. Demuth, J. H. Zhao, W. Meng, L. Lataniotis, S. W. van der Laan, J. P. Bradfield, A. R. Wood, A. Bonnefond, T. S. Ahluwalia, L. M. Hall, E. Salvi, S. Yazar, L. Carstensen, H. G. de Haan, M. Abney, U. Afzal, M. A. Allison, N. Amin, F. W. Asselbergs, S. J. L. Bakker, R. G. Barr, S. E. Baumeister, D. J. Benjamin, S. Bergmann, E. Boerwinkle, E. P. Bottinger, A. Campbell, A. Chakravarti, Y. Chan, S. J. Chanock, C. Chen, Y.-D. I. Chen, F. S. Collins, J. Connell, A. Correa, L. A. Cupples, G. D. Smith, G. Davies, M. Dörr, G. Ehret, S. B. Ellis, B. Feenstra, M. F. Feitosa, I. Ford, C. S. Fox, T. M. Frayling, N. Friedrich, F. Geller, G. Scotland, I. Gillham-Nasenya, O. Gottesman, M. Graff, F. Grodstein, C. Gu, C. Haley, C. J. Hammond, S. E. Harris, T. B. Harris, N. D. Hastie, N. L. Heard-Costa, K. Heikkilä, L. J. Hocking, G. Homuth, J.-J. Hottenga, J. Huang, J. E. Huffman, P. G. Hysi, M. A. Ikram, E. Ingelsson, A. Joensuu, Å. Johansson, P. Jousilahti, J. W. Jukema, M. Kähönen, Y. Kamatani, S. Kanoni, S. M. Kerr, N. M. Khan, P. Koellinger, H. A. Koistinen, M. K. Kooner, M. Kubo, J. Kuusisto, J. Lahti, L. J. Launer, R. A. Lea, B. Lehne, T. Lehtimäki, D. C. M. Liewald, L. Lind, M. Loh, M.-L. Lokki, S. J. London, S. J. Loomis, A. Loukola, Y. Lu, T. Lumley, A. Lundqvist, S. Männistö, P. Marques-Vidal, C. Masciullo, A. Matchan, R. A. Mathias, K. Matsuda, J. B. Meigs, C. Meisinger, T. Meitinger, C. Menni, F. D. Mentch, E. Mihailov, L. Milani, M. E. Montasser, G. W. Montgomery, A. Morrison, R. H. Myers, R. Nadukuru, P. Navarro, M. Nelis, M. S. Nieminen, I. M. Nolte, G. T. O'Connor, A. Ogunniyi, S. Padmanabhan, W. R. Palmas, J. S. Pankow, I. Patarcic, F. Pavani, P. A. Peyser, K. Pietilainen, N. Poulter, I. Prokopenko, S. Ralhan, P. Redmond, S. S. Rich, H. Rissanen, A. Robino, L. M. Rose, R. Rose, C. Sala, B. Salako, V. Salomaa, A.-P. Sarin, R. Saxena, H. Schmidt, L. J. Scott, W. R. Scott, B. Sennblad, S. Seshadri, P. Sever, S. Shrestha, B. H. Smith, J. A. Smith, N. Soranzo, N. Sotoodehnia, L. Southam, A. V. Stanton, M. G. Stathopoulou, K. Strauch, R. J. Strawbridge, M. J. Suderman, N. Tandon, S.-T. Tang, K. D. Taylor, B. O. Tayo, A. M. Töglhofer, M. Tomaszewski, N. Tšernikova, J. Tuomilehto, A. G. Uitterlinden, D. Vaidya, A. van Hylckama Vlieg, J. van Setten, T. Vasankari, S. Vedantam, E. Vlachopoulou, D. Vozzi, E. Vuoksimaa, M. Waldenberger, E. B. Ware, W. Wentworth-Shields, J. B. Whitfield, S. Wild, G. Willemsen, C. S. Yajnik, J. Yao, G. Zaza, X. Zhu, T. B. J. Project, R. M. Salem, M. Melbye, H. Bisgaard, N. J. Samani, D. Cusi, D. A. Mackey, R. S. Cooper, P. Froguel, G. Pasterkamp, S. F. A. Grant, H. Hakonarson, L. Ferrucci, R. A. Scott, A. D. Morris, C. N. A. Palmer, G. Dedoussis, P. Deloukas, L. Bertram, U. Lindenberger, S. I. Berndt, C. M. Lindgren, N. J. Timpson, A. Tönjes, P. B. Munroe, T. I. A. Sørensen, C. N. Rotimi, D. K. Arnett, A. J. Oldehinkel, S. L. R. Kardia, B. Balkau, G. Gambaro, A. P. Morris, J. G. Eriksson, M. J. Wright, N. G. Martin, S. C. Hunt, J. M. Starr, I. J. Deary, L. R. Griffiths, H. Tiemeier, N. Pirastu, J. Kaprio, N. J. Wareham, L. Pérusse, J.

G. Wilson, G. Girotto, M. J. Caulfield, O. Raitakari, D. I. Boomsma, C. Gieger, P. van der Harst, A. A. Hicks, P. Kraft, J. Sinisalo, P. Knekt, M. Johannesson, P. K. E. Magnusson, A. Hamsten, R. Schmidt, I. B. Borecki, E. Vartiainen, D. M. Becker, D. Bharadwaj, K. L. Mohlke, M. Boehnke, C. M. van Duijn, D. K. Sanghera, A. Teumer, E. Zeggini, A. Metspalu, P. Gasparini, S. Ulivi, C. Ober, D. Toniolo, I. Rudan, D. J. Porteous, M. Ciullo, T. D. Spector, C. Hayward, J. Dupuis, R. J. F. Loos, A. F. Wright, G. R. Chandak, P. Vollenweider, A. Shuldiner, P. M. Ridker, J. I. Rotter, N. Sattar, U. Gyllensten, K. E. North, M. Pirastu, B. M. Psaty, D. R. Weir, M. Laakso, V. Gudnason, A. Takahashi, J. C. Chambers, J. S. Kooner, D. P. Strachan, H. Campbell, J. N. Hirschhorn, M. Perola, O. Polašek, J. F. Wilson, Directional dominance on stature and cognition in diverse human populations. *Nature*. **523**, 459–462 (2015).

96. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013), (available at http://arxiv.org/abs/1303.3997).

Acknowledgments

First and foremost, we would like to thank members of the Kodava (both in India and in the US), Bunt, Nair, and Kapla populations for their participation in this study and for graciously hosting research team members in their communities and sharing their oral histories. We also thank members of the sequencing facilities at MedGenome Inc, Bangalore (India), Novogene, Sacramento (USA), and the University of Chicago DNA Sequencing Facility, Chicago (USA). We are also extremely grateful to Anna Di Rienzo, John Novembre, Matthias Steinrücken, and Bridget Chak for their valuable feedback on the manuscript and to Kendra Kodira for creating awareness and promoting participation among the Kodava_US community.

Funding: This project was funded through the NIH Grant R35GM143094, University of Chicago start-up funds, Cincinnati Children's Endowment, and the Gibbs Travelling Research Fellowship from the Newnham College at the University of Cambridge.

Author contributions: M.R. and N.R. conceived the overall project. N.P., A.M., M.S.M., C.D.K., A.P.N., M.S., N.R., and M.R. participated in fieldwork, sampling, and community outreach in India and the US. N.P., C.F.C., and J.S. performed DNA extractions in India and the US. A.B., E.B., C.F.C., and D.W. performed most of the data analyses, with support from R.F., S.F., H.M.M., and T.W. A.B., E.B., C.F.C., D.W., and M.R. wrote the majority of the manuscript with critical input from H.M.M., M.S., N.R., R.F., and the other authors.

Competing interests: The authors declare that they have no competing interests.

Data availability: In compliance with the informed consent obtained from participants in the study, raw data (fastq files), alignments (bam files), and variant calls (VCF files) are available under data access agreement, requests for which should be submitted to: M.R. (mraghavan@uchicago.edu) and N.R. (nirajrai@bsip.res.in).