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

India is home to thousands of ethno-linguistically distinct groups, many maintaining strong self-identities that derive from oral traditions and histories. However, these traditions and histories are only partially documented and are in danger of being lost over time. More recently, genetic studies have established the existence of ancestry gradients derived from both western and eastern Eurasia as well as evidence of practices such as endogamy and consanguinity, revealing complexity in the regional population structure with consequences for the health landscape of local populations. Despite the increase in genome-wide data from India, there is still sparse sampling across finer-scale geographic regions leading to gaps in our understanding of how and when present-day genetic structure came into existence. To address the gaps in genetic and oral histories, we analyzed whole-genome sequences of 70 individuals from Southwest India identifying as Bunt, Kodava, and Nair—populations that share unique oral histories and origin narratives—and 78 recent immigrants to the United States with Kodava ancestry as part of a community-led initiative. We additionally generated genome-wide data from 10 individuals self-identifying as Kapla, a population from the same region that is socio-culturally different to the other three study populations. We supplemented existing but limited anthropological records on these populations with oral history accounts narrated by community members and non-member contacts during sampling and subsequent community engagement. Overall, we find that components of genetic ancestry are relatively homogeneous among the Bunt, Kodava, and Nair populations and comparable to neighboring populations in India, which motivates further investigation of non-local origin narratives referenced in their oral histories. A notable exception is the Kapla population, with a higher proportion of ancestry represented in the Onge from the Andaman Islands, similar to several South Indian tribal populations. Utilizing haplotype-based methods, we find latent genetic structure across South India, including the sampled populations

from Southwest India, suggesting more recent population structure between geographically proximal populations in the region. This study represents an attempt for community-engaged anthropological and genetic investigations in India and presents results from both sources, underscoring the need to recognize that oral and genetic histories should not be expected to overlap. Ultimately, oral traditions and unique self-identities, such as those held close by some of the study populations, warrant more community-driven anthropological investigations to better understand how they originate and their relationship to genetic histories.

Introduction

Indian populations are characterized by a complex history of human migrations and admixture, as well as a variety of traditional socio-cultural practices, all of which have contributed to extensive cultural and genetic diversity. Previous studies have attempted to characterize genetic diversity in India, identifying population genetic structure broadly concomitant with geography and language (Basu, Sarkar-Roy, and Majumder 2016; Nakatsuka et al. 2017; Narasimhan et al. 2019; Tätte et al. 2019; Pathak et al. 2018; Metspalu, Mondal, and Chaubey 2018; GenomeAsia100K Consortium 2019). These studies highlight that the genetic structure of many Indian populations can be modeled along an admixture cline bounded by two statistical constructs, namely Ancestral North Indians (ANI), related to present-day western Eurasians, and Ancestral South Indians (ASI), related to Indigenous Andamanese that are typically modeled using the Onge population (Reich et al. 2009; Moorjani et al. 2013). The admixture landscape of Indian populations has been further refined using ancient samples as potential sources of western Eurasian ancestry, suggesting that both ANI and ASI groups carry western Eurasian ancestry related to ancient Iranian farmers (Indus Periphery Cline described in (Narasimhan et al. 2019). The ANI group has also been inferred to carry an additional western Eurasian component via admixture with Middle-to-Late Bronze Age (MLBA) groups in the Central Steppe region and the ANI and ASI groups may have only formed in entirety sometime after the second millennium BCE (Narasimhan et al. 2019).

In addition to these larger-scale ancestry gradients, at a finer-scale, genetic structure has been impacted by unique social structures and endogamous practices over the past ~2000 years (Moorjani et al. 2013; Debortoli et al. 2020). The extent of endogamy and, consequently, recessive disease risks, varies greatly across populations and can be a valuable research focus to aid in community health endeavors (Nakatsuka et al. 2017; Arciero et al. 2020; Finer et al. 2020). In spite of the diversity at the genetic and socio-cultural levels within India, the sampling efforts of human genetic diversity have been disproportionately small. Recent efforts have made substantial inroads to characterizing the broader genetic diversity of populations across South Asia (e.g. GenomeAsia100K Consortium 2019; Basu, Sarkar-Roy, and Majumder 2016; Reich et al. 2009). However, many populations in India still have not been characterized at the genetic level, leaving gaps in our understanding of sub-regional differences in population structure.

While such gaps in the population and medical genetics literature contribute to the motivation behind genomic research in underrepresented populations, it is imperative to be mindful of the ethical and socio-political ramifications of the research process, and the genetic narratives that emerge from it, for marginalized populations (e.g. Silva et al. 2022). While institutional ethical frameworks 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 (Hwang 2008; Haelewaters, Hofmann, and Romero-Olivares 2021; Argüelles, Fuentes, and Yáñez 2022). 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 genetics and self-identity can potentially impact a populations' status and recognition by local governments and afforded rights and privileges (Prince and Berkman 2018). Instead of attempting to reconcile genetic and cultural histories, it may be more valuable to recognize the complexities of both forms of knowledge (Crellin and Harris 2020; Donovan and Nehm 2020; TallBear 2013). 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 genome-wide data and oral histories and by building a community-informed collaborative framework. In this study, we generate and analyze whole-genome sequences from individuals identifying as Kodava, Bunt, Nair, and Kapla and, in conjunction with published genome-wide sequences from worldwide populations, investigate genetic histories and population structure in present-day Southwest India. We concurrently present community-engaged anthropological documentation of dwindling oral histories surveyed via conversations and observations conducted by the research team. The Bunt, Kodava, and Nair have strong self-identities as well as unique cultural traits and oral histories reflecting origin narratives, some of which are shared between these populations. By virtue of being largely undocumented, these oral traditions are in constant danger of being lost over time. There is comparatively less known about the Kapla through anthropological and/or community accounts. We discuss genetic and oral histories derived from our investigations within an integrated framework and motivate future research on the unique oral histories of these populations and, more broadly, consider the distinct positioning of genetic and self identities.

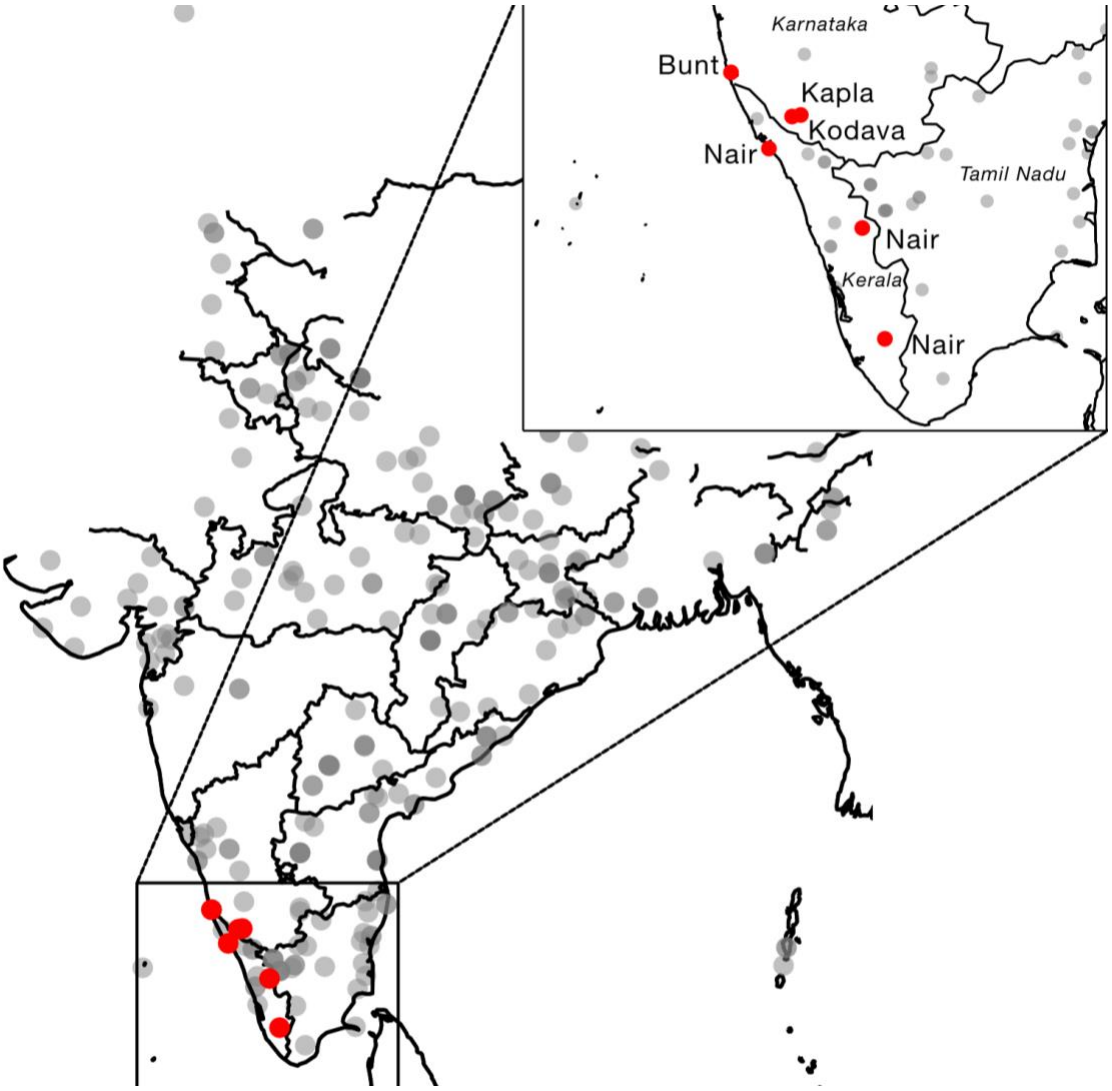

Figure 1. Map of newly sampled populations from this study. Gray points represent populations sampled from previous studies (Nakatsuka et al. 2017; GenomeAsia100K Consortium 2019).

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.

Donors who contributed to this study self-identify 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, hereafter denoted as Kodava_US to differentiate them from the Kodava sampled in India (hereafter 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, very little is documented in the ethnographic literature (Panikkar 1918; Srinivas 1965; Thruston 1909; Schneider 1962). These populations are all speakers of Dravidian languages. Moreover, there are notable overlaps in these populations' traditions and cultural traits, many of which are unique to each population and may stem from geographical proximity and historical contacts that are also referenced in these populations' oral histories. These include historical 'warrior' status designations and 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 (Fuller 1975). Some of these socio-cultural characteristics are noted in the anthropological literature (Menon 2018; Fuller 1976) 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 population members and relayed to members of the research team during sampling and subsequent population interactions, of phenotypic characteristics that members of these populations use to support possible non-local origins. 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. While the timing and exact mode of past population contacts may not always be explicitly stated in oral histories, these narratives are passed down through the generations. 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.

Some of the proposed links to non-local populations, such as the Scythians, ancient nomadic people inhabiting the Eurasian Steppe during the Iron Age, are shared across the oral histories of these three populations. In addition, the Kodava have oral histories suggesting links to western Asian and southern European populations such as Greeks and Iranians. The following is noted in (Karumbaya 2018): "*the ancestors of Kodava were part of the war-like Brazana tribe originally hailing from the Kurdish area of present-day Turkey, Iran and Iraq, which is a hilly region like Kodagu. They entered India during 320 BC in the pre-Islamic era, as a part of the Iranian contingent which had joined Emperor Alexander's invading army. 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…..the Kodavas who are believed to have taken a southernly route along with Western Ghats in search of better prospects, eventually settled in Kodagu which was 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 suggest 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 live 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 and speak a mixture of Tulu and Kodava languages, both members of the Dravidian language family. Historical records suggest (Richter 1984): "*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 (Shah et al. 2011). Anthropological information gathered by the research team from long-term community contacts suggests that once they were relocated to their present location in Kodagu (*Coorg*), 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 use principal components 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 cline commonly seen in populations with South Asian ancestry (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 and individuals from Urban Bangalore from (Nakatsuka et al. 2017; GenomeAsia100K Consortium 2019)). This pattern is also captured within ancestry clusters from ADMIXTURE where, at K = 9, these populations share the South Asian-specific components (colored blue and gray in Figure 2C, Figure S1) with most other populations from the region (Figure 2C, Figure S1). They also share a component (colored pink) with most South Asian and some populations in western and central Asia. Moreover, they display a component maximized in the Kalash (colored orange) and a small proportion of components (colored red and brown) present in broader western Eurasia 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 are in line with the proportions displayed by geographical neighbors such as Iyer and individuals from Urban Bangalore and Urban Chennai from (GenomeAsia100K Consortium 2019). Furthermore, we do 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. We also do not observe any notable differences in 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 show genetic similarity to populations with higher ASI ancestry from South India, such as the Ulladan and Paniya, displaying blue, 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 ancestry (Figure 2A, B, C). Our results show little support for the Kapla having substantial Siddi or African genetic ancestry (Figure S3), as suggested in historical records.

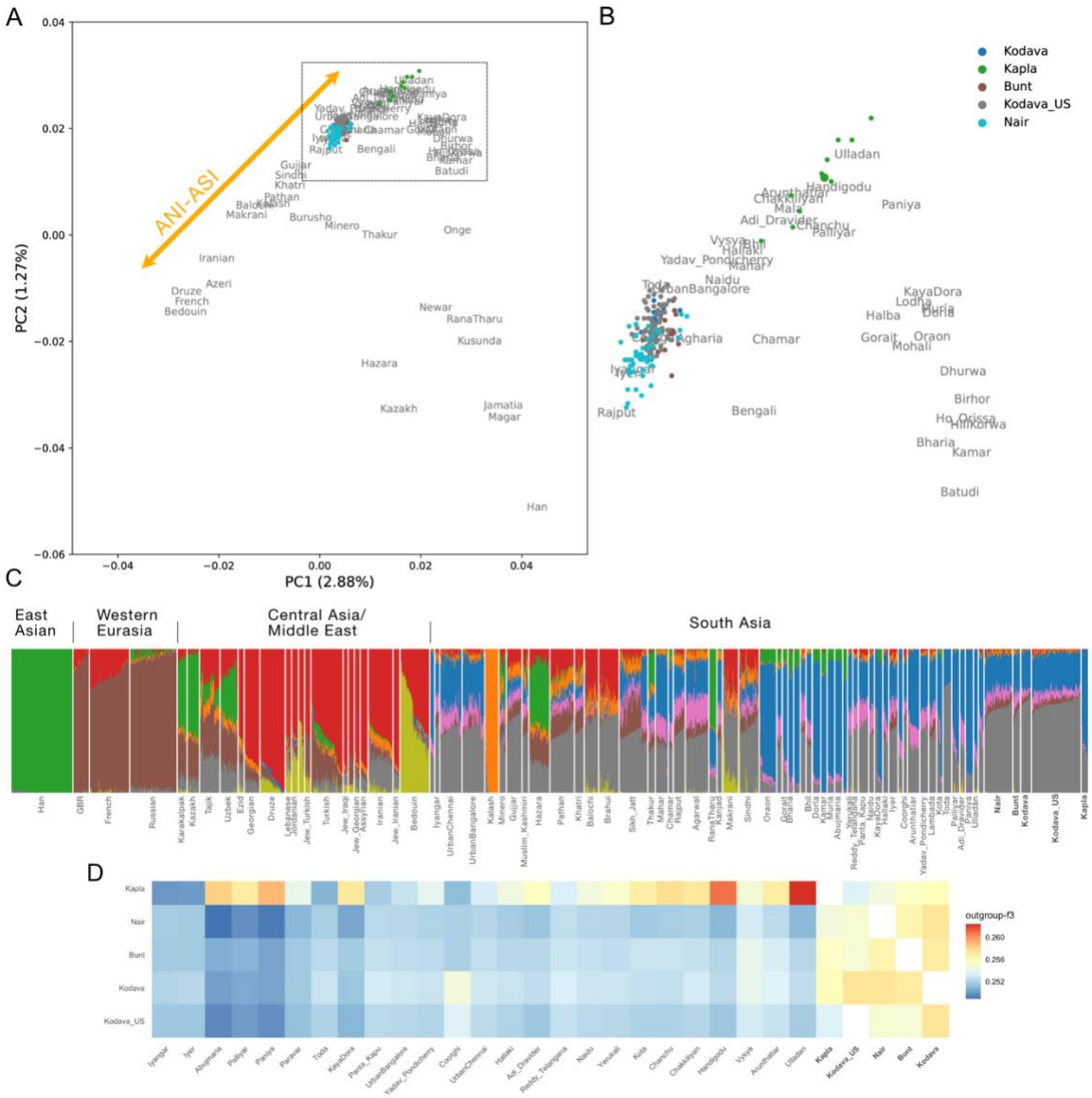

Figure 2. (A) Principal component analysis of merged dataset with Human Origins samples (see Methods), with (B) representing a zoomed inset focusing on newly sampled populations from this study and their position on the ANI-ASI ancestry cline. Population labels in gray represent the mean location of 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=9)* showcasing similarity across Southwest Indian populations in relation to East Asian, Western Eurasian, Middle Eastern, and other South Asian populations. We have removed some population labels for visual clarity (see Figure S1 for full set of labels). Populations sampled specifically in this study are represented using bolded population labels. (D) Outgroup f_3-statistics results for the newly sampled populations relative to other South Indian populations.

We evaluated shared genetic drift reflecting genetic affinities between the sampled populations from Southwest India and several Eurasian populations using an outgroup f_3-statistic of the form $f_3(Target, X; Mbuti)$, where *Target* corresponds 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 observe a higher genetic affinity between all our target populations and other Indian

populations (Figure 2D), particularly from South India, which is consistent with the results of PCA and ADMIXTURE. Within South India, patterns of shared drift differ, as seen in previous analyses, with the Kodava, Kodava_US, Bunt, and Nair sharing higher genetic affinity with each other and, to an extent, with populations that derive ancestry from both ANI and ASI, while the Kapla show higher genetic affinity to populations with more ASI ancestry such as tribal populations like Ulladan from Kerala (Scheduled Tribes Development Department 2022) and individuals from Handigodu village of Shimoga District of Karnataka (Nakatsuka et al. 2017).

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 S4-S25). Without migration edges, we observed expected broad-level relationships between the populations, with western and eastern Eurasian populations forming distinct clusters. South Asian populations fall between these clusters, with ANI-rich populations falling closer to western Eurasians and ASI-rich populations closer to populations such as the Onge. The Kodava, Kodava_US, Bunt, and Nair are closely related to each other and to other South Indian populations such as Iyer and Iyangar. The Kapla fall close to ASI-rich South Asians, consistent with showing higher levels of ASI ancestry (Figure 2C). Few early migration edges (Table S3) recapitulate previously-reported events such as the Steppe-related gene flow, represented by ancient Central Steppe MLBA individuals from Russia, into the French. We postulate 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 ancestry gradient.

Estimation of admixture proportions and timing

To directly model target populations as a mixture of 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 (Narasimhan et al. 2019; Lazaridis et al. 2014; Allentoft et al. 2015; Moorjani et al. 2013), and we used it primarily to assess the relative proportions of Central Steppe MLBA related ancestry, a proxy for ANI ancestry in the target populations (Figure S26). The proportion of Central Steppe MLBA in the target populations Nair, Bunt, Kodava, and Kodava_US ranges between 45% (+/- 1.4%) and 48% (+/- 1.3%), similar to the proportion found in other populations with higher ANI ancestry, including some South Indian and the majority of North Indian populations in India. In contrast, the proportion of Central Steppe MLBA in Kapla is similar to neighboring South Indian tribal populations with higher ASI ancestry (27.17% +/- 2.1%), which is comparatively lower than the other target populations (Figure S28).

Next, we employed qpAdm to model ancestry proportions in the target populations using a more complex mix of sources (Harney et al. 2021; Haak et al. 2015) (see Methods), particularly as a three-way mixture of ancestries related to Central Steppe MLBA, Indus Periphery Cline, and Onge as proposed by (Narasimhan et al. 2019) (Table S5). We note that several South Asian populations, including a majority of those sequenced in the study, did not pass the p-value threshold, also observed in the original study (Narasimhan et al. 2019), suggesting that this model may not be accurately capturing the complexities of ancestries represented by many populations on the ANI-ASI cline. In particular, we observe that this model is plausible (p-value > 0.05) for those South Asian populations with very little Central Steppe MLBA ancestry. This observation is validated by Kapla being the only study population with a p-value over 0.05 and can be modeled as having very little (4.6%) Central Steppe MLBA-related ancestry compared to ancestry related to Indus Periphery Cline (40.9%) and Onge (54.5%). These ancestry proportions are similar to

Palliyar, Paniya, and Ulladan, which are South Indian tribal groups with higher ASI ancestry and also yield p-values greater than 0.05 in this analysis. On the other hand, we fail to recover well-supported qpAdm models for those populations with appreciable Central Steppe MLBA-related ancestry, including the Bunt, Kodava (both groups), and Nair that were sequenced in this study and previously published data from populations like the Iyangar from South India and most tested North Indian populations. While we cannot accept a working model for these populations based strictly on p-values, we note that qpAdm does infer very similar proportions of the three ancestral sources in the Bunt, Kodava (both groups), and Nair, with higher proportions of the two western Eurasian-related sources in these populations compared to the Kapla (Table S5). Broadly, the source proportions in the study populations are within the range displayed by other South Asian populations, though the ancestries related to Central Steppe MLBA and Indus Periphery are qualitatively at the higher end of the range in the Nair and Kodava compared to those observed in neighboring populations. The relative affinities of our study and other South Asian populations to the three sources, determined using *D* statistics (Figure S29-31), are 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 ancestry that may be missing in the current model .

To estimate the timing of the introduction of the two western Eurasian ancestral sources in South Asia, we ran ALDER on select South Asian populations (Moorjani et al. 2013; Narasimhan et al. 2019; Loh et al. 2013). We focus our discussion on populations that yielded significant results. The date estimates for populations with higher Steppe-related ancestry, typically from North India, range from 48.04 ± 22.87 to 166.66 ± 26.94 generations. South Indian populations with higher ASI ancestry tend to display slightly older dates than the former group, ranging from 74.19 ± 36.47 to 234.96 ± 97.71 generations (Table S7), which is in agreement with previous studies (Moorjani et al. 2013; Narasimhan et al. 2019). Excluding the Kapla that did not yield a significant result, the admixture time estimates for the populations sequenced in this study range between 102.61 ± 30.32 and 138.8 ± 29.99 generations, which is within the range displayed by other populations included in the analysis (Table S7). As pointed out previously (Moorjani et al. 2013; Narasimhan et al. 2019), these dates represent a complex series of admixture events and, particularly for populations with both Iranian- and Steppe-related ancestry, may reflect an average of these events.

Characterizing fine-scale 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 (Lawson et al. 2012; Han et al. 2017).

To evaluate population structure and affinities at a finer scale, we implemented a haplotype-based analysis using ChromoPainter and fineSTRUCTURE (Lawson et al. 2012). This analysis first targeted 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" (Li and Stephens 2003) 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, Bunt, Nair, and Kapla copied most of their haplotypes from South Asian donors, particularly South Indian populations, consistent with previous analyses (Figure S32-S33). 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). 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 subtle and recent population structure in these populations.

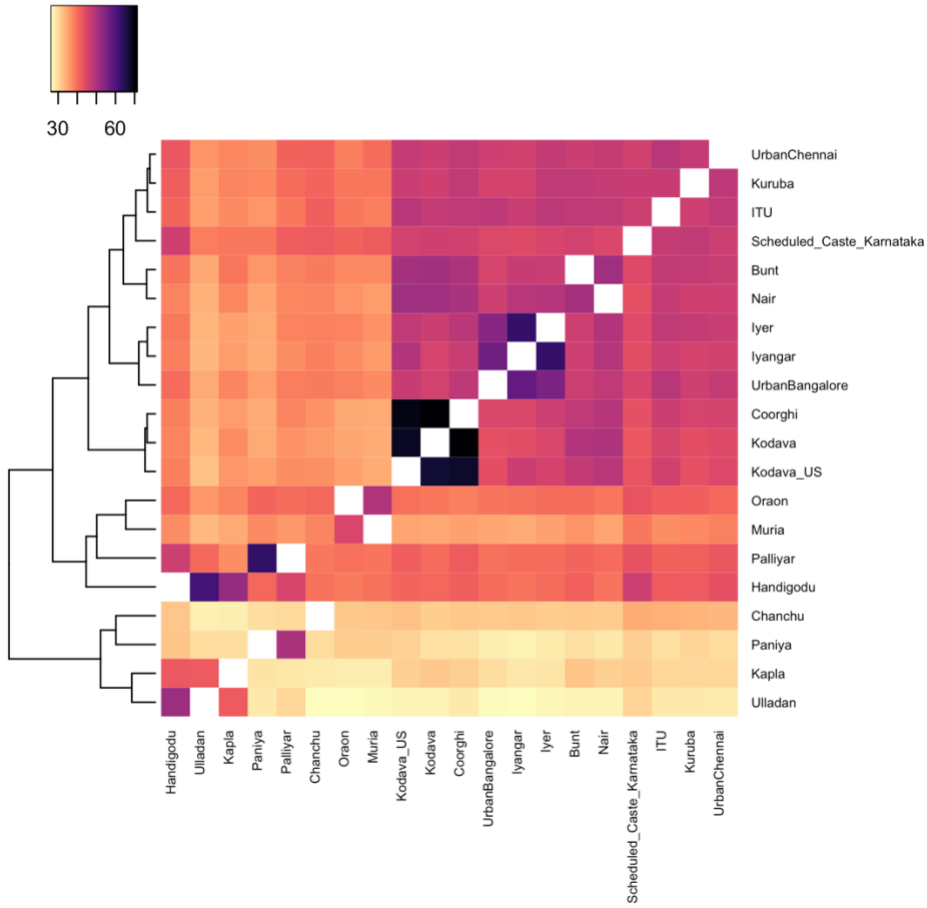

Figure 3: Haplotype-copying co-ancestry matrix across the regional level representing South Indian populations. Color bar represents the average length of haplotypes shared between populations.

A note on genetic variation within the Kapla population

Our analyses suggest that the Kapla, in addition to being genetically different from the other populations sampled in this study, are substantially spread along the first PC (Figure 2B). In particular, the ten Kapla individuals fall along the ANI-ASI cline, with some individuals closer to the ASI end while others are slightly closer to other South Asian populations with higher ANI ancestry. While this may be due to sampling bias, we note that other South Indian populations maximizing ASI ancestry like Ulladan, Paniya, and Vysya do not display as high variance on the PCA as the Kapla (Figure S34). As the ADMIXTURE results do not suggest major differences

within Kapla, and all individuals show little contribution from western Eurasian ancestries (Figure 2C, Figure S1), we formally evaluate this difference by separating Kapla individuals into two groups based on the PCA results where: Kapla_A includes 6 individuals closer to the ASI end of the cline and Kapla_B includes the other 4 individuals closer to the ANI end. We acknowledge that these groups may not represent true structure within the population and result from an arbitrary division to evaluate potential differences observed in the PCA. We estimate a *D* statistic 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 find that most tested western Eurasian populations are significantly closer to the four Kapla individuals closer to the ANI end (Kapla_B) (Peter 2022). In contrast, a few tested populations that maximize ASI ancestry, such as Ulladan, Paniya, and Palliyar are 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 (Narasimhan et al. 2019) (Table S5). We observe 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 western Eurasian 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 exact 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

We find that the populations sequenced in this study exhibit 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 (Metspalu et al. 2004; Chandrasekar et al. 2009; Herrera and Garcia-Bertrand 2018). In our data, we observe those subclades of haplogroup M that have previously been reported primarily in India (Chandrasekar et al. 2009; Larruga et al. 2017), in all populations. Macro-haplogroup R is predominantly reported in Eurasia (Sylvester et al. 2019) and makes up more than 40% of the maternal lineages in India (Herrera and Garcia-Bertrand 2018). In our data, we primarily observe lineages of haplogroup R in all populations that are reported to have originated in South Asia (Quintana-Murci et al. 2004; Metspalu et al. 2004; Chaubey et al. 2008; Sylvester et al. 2019). Macro-haplogroup U is one of the oldest clades of haplogroup R (Larruga et al. 2017) and there are multiple subclades associated with specific regions across Eurasia (Sahakyan et al. 2017). In our data, we find lineages of this haplogroup in all populations, except Kapla, previously reported in ancient and present-day individuals from the Near East, South Asia, Central Asia, and Southeast Asia (Narasimhan et al. 2019; Sahakyan et al. 2017; Kutanan et al. 2018). One of the individuals identifying as a Kodava_US displays 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 (Pala et al. 2012; Damgaard et al. 2018; Allentoft et al. 2015; Narasimhan et al. 2019). To our knowledge, the subclade of haplogroup J that we observe has not been reported in South Asia thus far. In addition, some sequenced individuals identifying as Kodava, Kodava_US and Nair belong to subclades of macro-haplogroup HV, a major subclade of haplogroup R0 (Shamoon-Pour, Li, and Merriwether 2019). We observe 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 (Narasimhan et al. 2019; Derenko et al. 2013; Shamoon-Pour, Li, and Merriwether 2019; Al-Zahery et al. 2003; Yaka et al. 2018; Mathieson et al. 2018). One individual identifying as Kodava_US is assigned to haplogroup A1a, a subclade of haplogroup A primarily observed in East Asia (Tanaka et al. 2004) that has, to our knowledge, not been reported in Indian populations.

We observe 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 (Mahal and Matsoukas 2018), followed by haplogroup H, which is suggested to have originated in South-Central Asia or the Middle East and later spread into India (Wells 2007; Mahal and Matsoukas 2018). Haplogroup L has a frequency of 7-15% in India and 28% in Balochistan and western parts of Pakistan (Mahal and Matsoukas 2018). In our data, we observe 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 (Kivisild et al. 2003; Kerchner 2013; Brunelli et al. 2017; Haber et al. 2020; Narasimhan et al. 2019; Scorrano et al. 2017). Haplogroup G is found at high frequencies in the Middle East and southern Europe and, at a low frequency, in South Asia (Rootsi et al. 2012). In our data, we find the lineage in Nair that is distributed in India, Pakistan, Iraq, Ukraine, Saudi Arabia (Isogg n.d.). Haplogroup N is prominent throughout northern Eurasia ("ISOGG 2007 Y-DNA Haplogroup Tree"). The lineage N1 (N-L735), a less common lineage of this haplogroup, is observed in one of the individuals identifying as Kodava_US. The lineages of haplogroup J that we see in Kodava, Kodava_US and Nair populations are reported in ancient and present-day individuals from South and Central Asia (Narasimhan et al. 2019; Sengupta et al. 2006). One of the individuals identifying as Nair displays macro-haplogroup E-M96, which has been reported in ancient and present-day individuals in eastern Europe and South Asia (Mathieson et al. 2018; Narasimhan et al. 2019). The lineages of haplogroup Q that we see in in Kodava, Kodava_US and Nair populations have been reported in ancient and present-day individuals from across Eurasia (Narasimhan et al. 2019; Damgaard et al. 2018; Huang et al. 2018; AlSafar et al. 2019). We note that our autosomal-based analyses do not suggest that those individuals with haplogroups not reported thus far in South Asia are genetic outliers relative to the other individuals from the populations.

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 Y-chromosome to mitochondrial unique haplogroup ratios for the sampled populations are 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 mitochondrion in the Nair and the Kapla, while the opposite is true for the two Kodava groups.

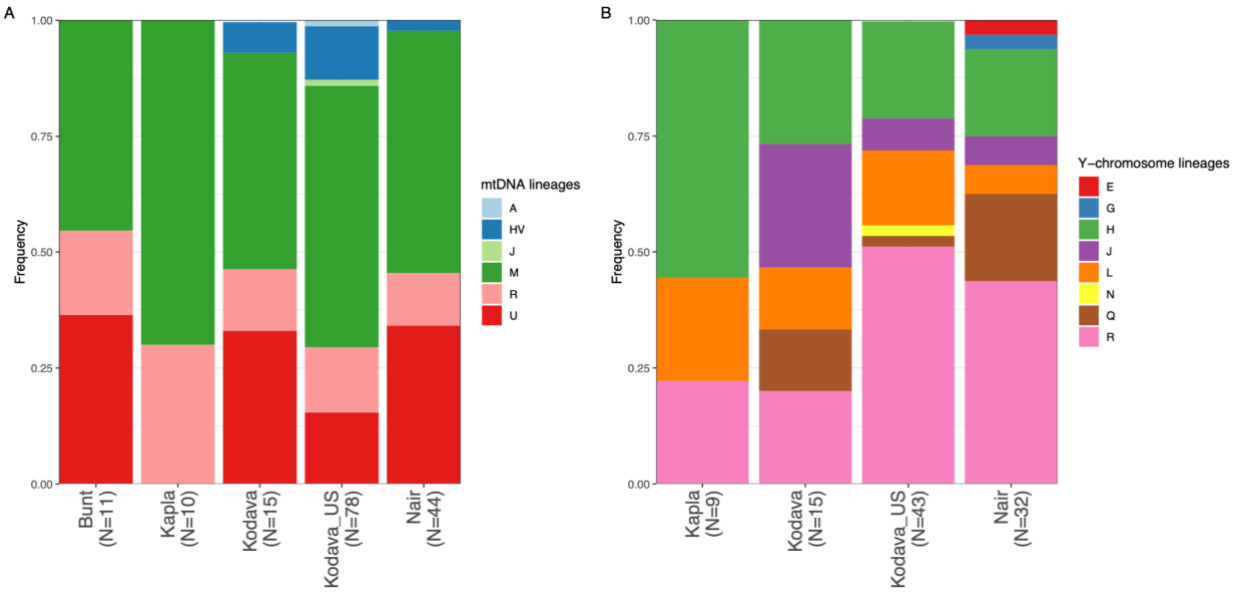

Figure 4: Distribution of the major mtDNA **(A)** and Y chromosome **(B)** macro-haplogroups from the analysis of 158 complete mitochondrial genomes and 99 complete Y chromosomes from this study. Bunt were included in this figure due to a single male in the dataset.

Endogamy in Southwest Indian populations

Founder effects and endogamy are prominent demographic forces shaping genetic diversity across populations of South Asian ancestry (Nakatsuka et al. 2017; Reich et al. 2009; Angural et al. 2020). We sought to characterize rates of endogamy in our newly sampled populations relative to other populations sampled in India as it can increase the frequency of recessive disease alleles. We calculated the 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).

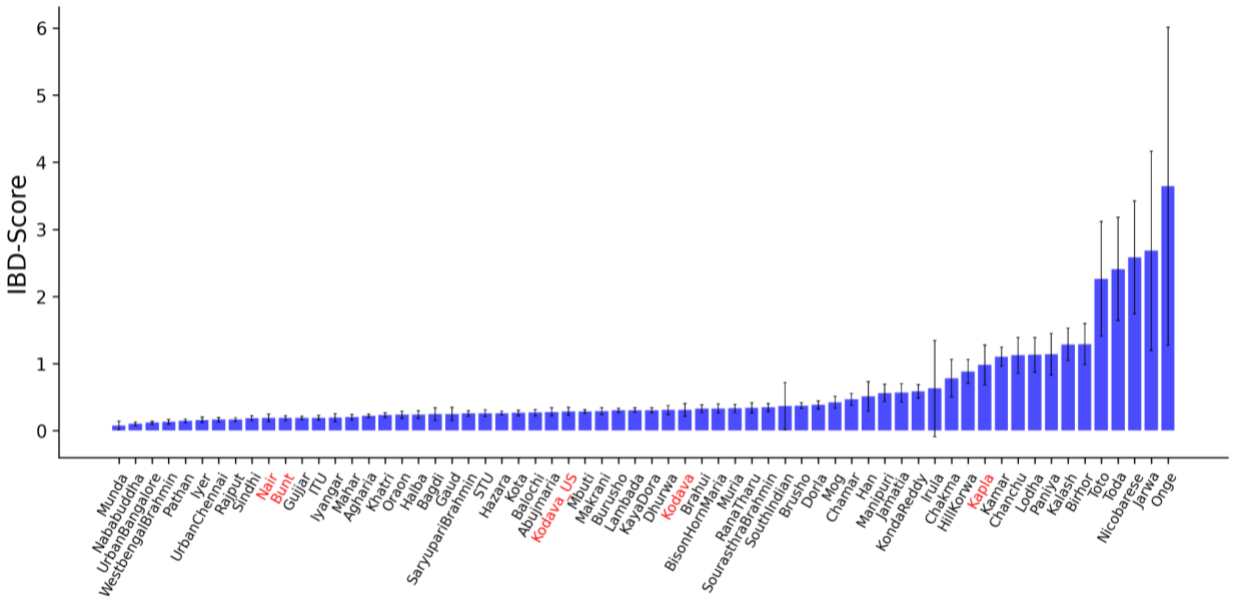

Figure 5: IBD score across multiple populations in South Asia. Populations that were sequenced as a part of this study are highlighted in red.

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

We estimate runs of homozygosity (ROH) to further explore the endogamy in the study populations (see Methods). In agreement with the IBD scores, we observe a pattern where the Kapla population has an increased number of ROH (NROH) and generally longer ROH (>10Mb) (Figures S35 and S36), consistent with smaller population size and higher levels of consanguinity (Ceballos et al. 2018).

Discussion

In this study, we engage with both genetic and oral histories of populations from Southwest India. By analyzing whole-genome sequences from donors in India and the US identifying as members of one of the four focal populations, we reconstruct the relationship of the four studied populations to the broader and fine-scale genetic structure in Southwest India. 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 ancestries related to Onge, Indus Periphery Cline, and Central Steppe MLBA. This mix of 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 these populations compared to neighboring populations raises an interesting anthropological follow-up to better understand the connections in their oral histories to non-local populations such as Iranian, Greek, and ancient Scythians. It is possible that our sampling scheme and/or analytical power limits the detection of such a signal in the genetic data, particularly if admixture levels were low to begin with, diluted over time due to local admixture, or overlapped with other closely related admixture events such as between the formation and admixture between ANI and ASI sometime after the second millennium BCE and the invasion of Alexander's army in 327 BCE, the latter mentioned in origin stories of the Kodava. In fact, the Nair and Kodava individuals sequenced in this study may have a slightly higher proportion of Western Eurasian ancestry compared to neighboring populations in South India, though comparable 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 identities. 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, often 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 (Crellin and Harris 2020; TallBear 2013) 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.

At a more fine-scale level, haplotype-based analysis suggests more recent genetic contacts between Bunt and Nair populations. 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 sequenced in (Nakatsuka et al. 2017), 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 these populations 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. Interestingly, the Nair and Kapla display a slightly higher diversity in Y chromosome haplogroups relative to mitochondrial haplogroups than the Kodava. While this may be due to imbalanced sample sizes across the study populations, such a pattern is expected of matrilocal groups of which the Nair are a prime example. While there is limited anthropological information on the Kapla, our own anthropological investigation suggests that they follow a patriarchal and patrilocal system. The diversity skew towards Y chromosome haplogroups in the Kapla could again be driven by low sample sizes or other demographic processes such as reduced maternal genetic diversity in the founder population, more recent male-mediated admixture, or extreme genetic drift due to small population size to which uniparental markers are more susceptible.

In contrast to the other sampled populations, the Kapla are genetically closer to tribal South Indian populations, which have higher proportions of Onge-related ancestry and negligible amounts of the Central Steppe MLBA component. In this regard, the Kapla may represent one of several close genetic descendants of the ASI group (Narasimhan et al. 2019). A close link 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 former. 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" (Richter 1984) 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 western Eurasian ancestry 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 (Nakatsuka et al. 2017; Reich et al. 2009; GUaRDIAN Consortium, Sivasubbu, and Scaria 2019). 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.

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 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, the project's 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 and community members 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, they 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. 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, 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. Anthropological interviews were conducted in India involving both members of the study populations and non-member contacts with a similar aim. Efforts to disseminate the results of this study are underway, both in India and in the US. Results are being returned to members of the study populations through a combination of written reports and presentations by the researchers making the information more accessible to a non-scientific audience.

Sample processing and genotype calling

For the Kodava, Nair, Bunt, and Kapla populations from India, DNA extracts from whole blood were sent to MedGenome Inc, Bangalore, India, for sequencing on Illumina HiSeq X Ten. They were sequenced to an average 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. The extracts were sequenced on the Illumina NovaSeq 6000 at Novogene, Sacramento, USA, to an average autosomal depth of 5.5X. We also re-sequenced three Kodava_US individuals to 77x average depth at Novogene and two individuals each from Bunt, Kapla, and Nair populations to 30x average depth at the University of Chicago DNA Sequencing Facility, Chicago, USA, 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 (GenomeAsia100K Consortium 2019). 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 (Howie et al. 2012; 1000 Genomes Project Consortium et al. 2015). 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 (Delaneau et al. 2019).

Merging with external datasets and quality control

We merged our sequence data with publicly available data from the GenomeAsia 100K consortium (GenomeAsia100K Consortium 2019) and the Human Genome Diversity Project (Bergström et al. 2020). To provide additional context in Southwest India, we further merged this dataset with genotype data from (Nakatsuka et al. 2017). Following these merges, we retain 425,620 SNPs for analysis of population structure.

In order to compare genetic affinities with ancient samples, we also merged samples from the Allen Ancient DNA Resource ("Allen Ancient Genome Diversity Project / John Templeton Ancient DNA Atlas"), which contains genotypes across ~1.23 million sites represented on the '1240k capture panel'. We also merged the publicly available Human Origins dataset from (Lazaridis et al. 2014), which contains present-day worldwide samples from central Asia. 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 (Manichaikul et al. 2010) and removed samples with > 5% SNP missingness for all downstream population genetic inferences. We note that 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.

Population Structure Analyses

Principal Component Analysis and ADMIXTURE

We performed a principal component analysis (PCA) using smartpca v7.3.0 (Price et al. 2006; Patterson, Price, and Reich 2006). We filtered the merged datasets according to linkage disequilibrium using PLINK 1.9 [`--indep-pairwise 200 25 0.4`] (Chang et al. 2015). 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 (Alexander, Novembre, and Lange 2009), with K = 5 to 12 (Figure S1). The value of K = 9 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 samples from each geographic region or language group category with at least 8 samples.

Admixture history using Treemix

We modeled the relationship between our target populations in Southwest India and a subset of present-day and ancient groups from Eurasia using Treemix (Pickrell and Pritchard 2012).

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 (Patterson et al. 2012), 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 (Bergström et al. 2020), unless otherwise noted.

To model our target populations as mixtures of ANI and ASI ancestries, we used the *f4*-ratio test (Reich et al. 2009). We measured the proportion of Central Steppe MLBA (α) (Narasimhan et al. 2019) via the ratio (see Figure S25):

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

We selected populations to include in the above topologies based on the scheme in (Moorjani et al. 2013). 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 S26).

We computed allele sharing of Kapla_A and Kapla_B sub-populations with different modern populations (H3) from Europe, the Middle East, and South Asia on the ANI-ASI cline using *D-statistics (Patterson et al. 2012)* 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`(Table S8) .

Estimation of ancestry proportions based on qpAdm

To model the ancestry proportions of the target and other select Indian populations for comparison, we used qpAdm (Harney et al. 2021; Haak et al. 2015) implemented in ADMIXTOOLS v7.0.2 software with (`allsnps: YES, oldallsnpsmode: NO, inbreed: NO, details: YES`). As many of the present-day South Asians on the ANI-ASI gradient can be modeled as a mix of Indus Periphery Cline, Onge and/or Central Steppe MLBA related ancestries (Narasimhan et al. 2019), we considered these as our sources for the test populations included in the analysis (Table S4). For the outgroups, we first considered the list outlined in the proximal model of Modern South Asians by (Narasimhan et al. 2019) (Table S4). However, most of the populations, including our target populations, did not pass the significant threshold of 5% with this set of outgroups (Table S5), suggesting more complex admixture patterns in the South Asian populations.

Admixture dates based on LD decay in ALDER

To estimate the timing of admixture of ancestral components in the study populations, we used the weighted admixture linkage disequilibrium decay method in ALDER (Loh et al. 2013). The true ancestral populations for South Asia are unknown, so we followed the approach implemented in (Moorjani et al. 2013; Narasimhan et al. 2019). 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 (Moorjani et al. 2013), we included primarily Dravidian- and Indo-European-speaking populations on 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 estimate the admixture timings inferred by the program in generations and in years, assuming 1 generation is 28 years (Narasimhan et al. 2019) (Table S6-S7). We did not include individuals from the test 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.

Haplotype-based Estimation of Population Structure using fineStructure

We implemented a haplotype-based approach using the software ChromoPainter and fineSTRUCTURE (Lawson et al. 2012). 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 EM 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 (Danecek et al. 2021). 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 159 mtDNA VCF files were used as input for assigning mitochondrial DNA (mtDNA) haplogroups using the program Haplogrep2 (Weissensteiner et al. 2016) with PhyloTree mtDNA tree Build 17 (Van Oven 2015). 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 (Poznik 2016) (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 here by haplogroups, on one of the two loci.

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 (Nakatsuka et al. 2017). 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 (Nait Saada et al. 2020) 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 (Nakatsuka et al. 2017), we also removed individuals related up to second degree using KING (Manichaikul et al. 2010) and removed IBD segments > 20 cM. In (Nakatsuka et al. 2017), 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 (Joshi et al. 2015) and (Ceballos et al. 2018), 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 (GenomeAsia100K Consortium 2019) and the Human Genome Diversity Project (Bergström et al. 2020). After filtering positions with more than 0.5% of missing data and MAF of 5%, the final database has 3,288,336 variants.

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. 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.

Data Sharing

In compliance with the informed consent associated with the donors who participated in the study, raw data (fastq files), alignments (bam files) and variant calls (VCFs) are available under data access agreement with the corresponding authors M.R. and N.R.

References

1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, et al. 2015. “A Global Reference for Human Genetic Variation.” *Nature* 526 (7571): 68–74.

Alexander, David H., John Novembre, and Kenneth Lange. 2009. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research* 19 (9): 1655–64.

“Allen Ancient Genome Diversity Project / John Templeton Ancient DNA Atlas.” Accessed February 21, 2022. https://reich.hms.harvard.edu/ancient-genome-diversity-project.

Allentoft, Morten E., Martin Sikora, Karl-Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, et al. 2015. “Population Genomics of Bronze Age Eurasia.” *Nature* 522 (7555): 167–72.

AlSafar, Habiba S., Mariam Al-Ali, Gihan Daw Elbait, Mustafa H. Al-Maini, Dymitr Ruta, Braulio Peramo, Andreas Henschel, and Guan K. Tay. 2019. “Introducing the First Whole Genomes of Nationals from the United Arab Emirates.” *Scientific Reports* 9 (1): 14725.

Al-Zahery, N., O. Semino, G. Benuzzi, C. Magri, G. Passarino, A. Torroni, and A. S. Santachiara-Benerecetti. 2003. “Y-Chromosome and mtDNA Polymorphisms in Iraq, a Crossroad of the Early Human Dispersal and of Post-Neolithic Migrations.” *Molecular Phylogenetics and Evolution* 28 (3): 458–72.

Angural, Arshia, Akshi Spolia, Ankit Mahajan, Vijeshwar Verma, Ankush Sharma, Parvinder Kumar, Manoj Kumar Dhar, Kamal Kishore Pandita, Ekta Rai, and Swarkar Sharma. 2020. “Review: Understanding Rare Genetic Diseases in Low Resource Regions like Jammu and Kashmir - India.” *Frontiers in Genetics* 11 (April): 415.

Arciero, Elena, Sufyan A. Dogra, Massimo Mezzavilla, Theofanis Tsismentzoglou, Qin Qin Huang, Karen A. Hunt, Dan Mason, et al. 2020. “Fine-Scale Population Structure and Demographic History of British Pakistanis.” *bioRxiv*. https://doi.org/10.1101/2020.09.02.279190.

Argüelles, Juan Manuel, Agustín Fuentes, and Bernardo Yáñez. 2022. “Analyzing Asymmetries and Praxis in aDNA Research: A Bioanthropological Critique.” *American Anthropologist* 124 (1): 130–40.

Basu, Analabha, Neeta Sarkar-Roy, and Partha P. Majumder. 2016. “Genomic Reconstruction of the History of Extant Populations of India Reveals Five Distinct Ancestral Components and a Complex Structure.” *Proceedings of the National Academy of Sciences of the United States of America* 113 (6): 1594–99.

Bergström, Anders, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, et al. 2020. “Insights into Human Genetic Variation and Population History from 929 Diverse Genomes.” *Science* 367 (6484). https://doi.org/10.1126/science.aay5012.

Brunelli, Andrea, Jatupol Kampuansai, Mark Seielstad, Khemika Lomthaisong, Daoroong Kangwanpong, Silvia Ghirotto, and Wibhu Kutanan. 2017. “Y Chromosomal Evidence on the Origin of Northern Thai People.” *PloS One* 12 (7): e0181935.

Ceballos, Francisco C., Peter K. Joshi, David W. Clark, Michèle Ramsay, and James F. Wilson. 2018. “Runs of Homozygosity: Windows into Population History and Trait Architecture.” *Nature Reviews. Genetics* 19 (4): 220–34.

Chandrasekar, Adimoolam, Satish Kumar, Jwalapuram Sreenath, Bishwa Nath Sarkar, Bhaskar Pralhad Urade, Sujit Mallick, Syam Sundar Bandopadhyay, et al. 2009. “Updating Phylogeny of Mitochondrial DNA Macrohaplogroup M in India: Dispersal of Modern Human in South Asian Corridor.” *PloS One* 4 (10): e7447.

Chang, Christopher C., Carson C. Chow, Laurent Cam Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (February): 7.

Chaubey, Gyaneshwer, Monika Karmin, Ene Metspalu, Mait Metspalu, Deepa Selvi-Rani, Vijay Kumar Singh, Jüri Parik, et al. 2008. “Phylogeography of mtDNA Haplogroup R7 in the Indian Peninsula.” *BMC Evolutionary Biology* 8 (August): 227.

Crellin, Rachel J., and Oliver J. T. Harris. 2020. “Beyond Binaries. Interrogating Ancient DNA.” *Archaeological Dialogues* 27 (1): 37–56.

Damgaard, Peter de Barros, Nina Marchi, Simon Rasmussen, Michaël Peyrot, Gabriel Renaud, Thorfinn Korneliussen, J. Víctor Moreno-Mayar, et al. 2018. “137 Ancient Human Genomes from across the Eurasian Steppes.” *Nature* 557 (7705): 369–74.

Danecek, Petr, James K. Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O. Pollard, Andrew Whitwham, et al. 2021. “Twelve Years of SAMtools and BCFtools.” *GigaScience* 10 (2). https://doi.org/10.1093/gigascience/giab008.

Debortoli, Guilherme, Cristina Abbatangelo, Francisco Ceballos, Cesar Fortes-Lima, Heather L. Norton, Shantanu Ozarkar, Esteban J. Parra, and Manjari Jonnalagadda. 2020. “Novel Insights on Demographic History of Tribal and Caste Groups from West Maharashtra (India) Using Genome-Wide Data.” *Scientific Reports* 10 (1): 10075.

Delaneau, Olivier, Jean-François Zagury, Matthew R. Robinson, Jonathan L. Marchini, and Emmanouil T. Dermitzakis. 2019. “Accurate, Scalable and Integrative Haplotype Estimation.” *Nature Communications* 10 (1): 5436.

Derenko, Miroslava, Boris Malyarchuk, Ardeshir Bahmanimehr, Galina Denisova, Maria Perkova, Shirin Farjadian, and Levon Yepiskoposyan. 2013. “Complete Mitochondrial DNA Diversity in Iranians.” *PloS One* 8 (11): e80673.

Donovan, Brian, and Ross H. Nehm. 2020. “Genetics and Identity.” *Science & Education* 29 (6): 1451–58.

Finer, Sarah, Hilary C. Martin, Ahsan Khan, Karen A. Hunt, Beverley MacLaughlin, Zaheer Ahmed, Richard Ashcroft, et al. 2020. “Cohort Profile: East London Genes & Health (ELGH), a Community-Based Population Genomics and Health Study in British Bangladeshi and British Pakistani People.” *International Journal of Epidemiology* 49 (1): 20–21i.

Fuller, C. J. 1975. “The Internal Structure of the Nayar Caste.” *Journal of Anthropological Research*. https://doi.org/10.1086/jar.31.4.3629883.

———. 1976. *The Nayars Today*. Cambridge University Press.

GenomeAsia100K Consortium. 2019. “The GenomeAsia 100K Project Enables Genetic Discoveries across Asia.” *Nature* 576 (7785): 106–11.

GUaRDIAN Consortium, Sridhar Sivasubbu, and Vinod Scaria. 2019. “Genomics of Rare Genetic Diseases-Experiences from India.” *Human Genomics* 14 (1): 52.

Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. “Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe.” *Nature* 522 (7555): 207–11.

Haber, Marc, Joyce Nassar, Mohamed A. Almarri, Tina Saupe, Lehti Saag, Samuel J. Griffith, Claude Doumet-Serhal, et al. 2020. “A Genetic History of the Near East from an aDNA Time Course Sampling Eight Points in the Past 4,000 Years.” *American Journal of Human*

Genetics 107 (1): 149–57.
Haelewaters, Danny, Tina A. Hofmann, and Adriana L. Romero-Olivares. 2021. "Ten Simple Rules for Global North Researchers to Stop Perpetuating Helicopter Research in the Global South." *PLoS Computational Biology* 17 (8): e1009277.
Han, Eunjung, Peter Carbonetto, Ross E. Curtis, Yong Wang, Julie M. Granka, Jake Byrnes, Keith Noto, et al. 2017. "Clustering of 770,000 Genomes Reveals Post-Colonial Population Structure of North America." *Nature Communications* 8 (February): 14238.
Harney, Éadaoin, Nick Patterson, David Reich, and John Wakeley. 2021. "Assessing the Performance of qpAdm: A Statistical Tool for Studying Population Admixture." *Genetics* 217 (4). https://doi.org/10.1093/genetics/iyaa045.
Herrera, Rene J., and Ralph Garcia-Bertrand. 2018. "Dispersals Into India." *Ancestral DNA, Human Origins, and Migrations*. https://doi.org/10.1016/b978-0-12-804124-6.00007-0.
Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. 2012. "Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing." *Nature Genetics* 44 (8): 955–59.
Huang, Yun-Zhi, Horolma Pamjav, Pavel Flegontov, Vlastimil Stenzl, Shao-Qing Wen, Xin-Zhu Tong, Chuan-Chao Wang, et al. 2018. "Dispersals of the Siberian Y-Chromosome Haplogroup Q in Eurasia." *Molecular Genetics and Genomics: MGG* 293 (1): 107–17.
Hwang, Kumju. 2008. "International Collaboration in Multilayered Center-Periphery in the Globalization of Science and Technology." *Science, Technology & Human Values* 33 (1): 101–33.
"ISOGG 2007 Y-DNA Haplogroup Tree." Accessed March 16, 2022. https://isogg.org/tree/2007/index07.html.
Isogg, Copyright 2019-2020. "ISOGG 2019 Y-DNA Haplogroup Tree." Accessed March 16, 2022. https://isogg.org/tree/index.html.
Joshi, Peter K., Tonu Esko, Hannele Mattsson, Niina Eklund, Ilaria Gandin, Teresa Nutile, Anne U. Jackson, et al. 2015. "Directional Dominance on Stature and Cognition in Diverse Human Populations." *Nature* 523 (7561): 459–62.
Karumbaya, Codanda. 2018. "Kodavas through the Ages." In *Are Kodavas (Coorgs) Hindus?*, edited by P. T. Bopanna. Rolling stone Publications, Bangalore.
Kerchner, C. F. 2013. "YDNA Haplogroup Descriptions & Information Links."
Kivisild, T., S. Rootsi, M. Metspalu, S. Mastana, K. Kaldma, J. Parik, E. Metspalu, et al. 2003. "The Genetic Heritage of the Earliest Settlers Persists Both in Indian Tribal and Caste Populations." *American Journal of Human Genetics* 72 (2): 313–32.
Kutanan, Wibhu, Jatupol Kampuansai, Andrea Brunelli, Silvia Ghirotto, Pittayawat Pittayaporn, Sukhum Ruangchai, Roland Schröder, et al. 2018. "New Insights from Thailand into the Maternal Genetic History of Mainland Southeast Asia." *European Journal of Human Genetics: EJHG* 26 (6): 898–911.
Larruga, Jose M., Patricia Marrero, Khaled K. Abu-Amero, Maria V. Golubenko, and Vicente M. Cabrera. 2017. "Carriers of Mitochondrial DNA Macrohaplogroup R Colonized Eurasia and Australasia from a Southeast Asia Core Area." *BMC Evolutionary Biology* 17 (1): 115.
Lawson, Daniel John, Garrett Hellenthal, Simon Myers, and Daniel Falush. 2012. "Inference of Population Structure Using Dense Haplotype Data." *PLoS Genetics* 8 (1): e1002453.
Lazaridis, Iosif, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, et al. 2014. "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans." *Nature* 513 (7518): 409–13.
Li, Na, and Matthew Stephens. 2003. "Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data." *Genetics* 165 (4): 2213–33.
Loh, Po-Ru, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K. Pickrell, David Reich, and Bonnie Berger. 2013. "Inferring Admixture Histories of Human Populations Using Linkage

Disequilibrium." *Genetics* 193 (4): 1233–54.
Mahal, David G., and Ianis G. Matsoukas. 2018. "The Geographic Origins of Ethnic Groups in the Indian Subcontinent: Exploring Ancient Footprints with Y-DNA Haplogroups." *Frontiers in Genetics* 9 (January): 4.
Manichaikul, Ani, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." *Bioinformatics* 26 (22): 2867–73.
Mathieson, Iain, Songül Alpaslan-Roodenberg, Cosimo Posth, Anna Szécsényi-Nagy, Nadin Rohland, Swapan Mallick, Iñigo Olalde, et al. 2018. "The Genomic History of Southeastern Europe." *Nature* 555 (7695): 197–203.
Menon, Anoushka. 2018. "Searching for Preliminary Mitochondrial and Phenotypic Signatures That May Underpin the 'Self-Identities' of Three Distinct South Indian Warrior Populations: The Nairs, the Bunts & the Kodava." MSc thesis, University of Cambridge.
Metspalu, Mait, Toomas Kivisild, Ene Metspalu, Jüri Parik, Georgi Hudjashov, Katrin Kaldma, Piia Serk, et al. 2004. "Most of the Extant mtDNA Boundaries in South and Southwest Asia Were Likely Shaped during the Initial Settlement of Eurasia by Anatomically Modern Humans." *BMC Genetics* 5 (August): 26.
Metspalu, Mait, Mayukh Mondal, and Gyaneshwer Chaubey. 2018. "The Genetic Makings of South Asia." *Current Opinion in Genetics & Development* 53 (December): 128–33.
Moorjani, Priya, Kumarasamy Thangaraj, Nick Patterson, Mark Lipson, Po-Ru Loh, Periyasamy Govindaraj, Bonnie Berger, David Reich, and Lalji Singh. 2013. "Genetic Evidence for Recent Population Mixture in India." *American Journal of Human Genetics* 93 (3): 422–38.
Nait Saada, Juba, Georgios Kalantzis, Derek Shyr, Fergus Cooper, Martin Robinson, Alexander Gusev, and Pier Francesco Palamara. 2020. "Identity-by-Descent Detection across 487,409 British Samples Reveals Fine Scale Population Structure and Ultra-Rare Variant Associations." *Nature Communications* 11 (1): 6130.
Nakatsuka, Nathan, Priya Moorjani, Niraj Rai, Biswanath Sarkar, Arti Tandon, Nick Patterson, Gandham Srilakshmi Bhavani, et al. 2017. "The Promise of Discovering Population-Specific Disease-Associated Genes in South Asia." *Nature Genetics* 49 (9): 1403–7.
Narasimhan, Vagheesh M., Nick Patterson, Priya Moorjani, Nadin Rohland, Rebecca Bernardos, Swapan Mallick, Iosif Lazaridis, et al. 2019. "The Formation of Human Populations in South and Central Asia." *Science* 365 (6457). https://doi.org/10.1126/science.aat7487.
Pala, Maria, Anna Olivieri, Alessandro Achilli, Matteo Accetturo, Ene Metspalu, Maere Reidla, Erika Tamm, et al. 2012. "Mitochondrial DNA Signals of Late Glacial Recolonization of Europe from near Eastern Refugia." *American Journal of Human Genetics* 90 (5): 915–24.
Panikkar, K. M. 1918. "Some Aspects of Nayar Life." *The Journal of the Royal Anthropological Institute of Great Britain and Ireland* 48: 254–93.
Pathak, Ajai K., Anurag Kadian, Alena Kushniarevich, Francesco Montinaro, Mayukh Mondal, Linda Ongaro, Manvendra Singh, et al. 2018. "The Genetic Ancestry of Modern Indus Valley Populations from Northwest India." *American Journal of Human Genetics* 103 (6): 918–29.
Patterson, Nick, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. 2012. "Ancient Admixture in Human History." *Genetics* 192 (3): 1065–93.
Patterson, Nick, Alkes L. Price, and David Reich. 2006. "Population Structure and Eigenanalysis." *PLoS Genetics* 2 (12): e190.
Peter, Benjamin M. 2022. "A Geometric Relationship of F2, F3 and F4-Statistics with Principal Component Analysis." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 377 (1852): 20200413.
Pickrell, Joseph K., and Jonathan K. Pritchard. 2012. "Inference of Population Splits and

Mixtures from Genome-Wide Allele Frequency Data." *PLoS Genetics* 8 (11): e1002967.
Poznik, G. David. 2016. "White Paper 23-13 yHaplo| Identifying Y-Chromosome Haplogroups in Arbitrarily Large Samples of Sequenced or Genotyped Men." https://permalinks.23andme.com/pdf/23-13_paternal_haplogroups_yHaplo.pdf.
Price, Alkes L., Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9.
Prince, Anya E. R., and Benjamin E. Berkman. 2018. "Reconceptualizing Harms and Benefits in the Genomic Age." *Personalized Medicine* 15 (5): 419–28.
Quintana-Murci, Lluis, Raphaelle Chaix, R. Spencer Wells, Doron M. Behar, Hamid Sayar, Rosaria Scozzari, Chiara Rengo, et al. 2004. "Where West Meets East: The Complex mtDNA Landscape of the Southwest and Central Asian Corridor." *American Journal of Human Genetics* 74 (5): 827–45.
Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh. 2009. "Reconstructing Indian Population History." *Nature* 461 (7263): 489–94.
Richter, Georg. 1984. *Gazetteer of Coorg: Natural Features of the Country and the Social and Political Condition of Its Inhabitants*. Vol. Reprint. BR Publishing Corporation.
Rootsi, Siiri, Natalie M. Myres, Alice A. Lin, Mari Järve, Roy J. King, Ildus Kutuev, Vicente M. Cabrera, et al. 2012. "Distinguishing the Co-Ancestries of Haplogroup G Y-Chromosomes in the Populations of Europe and the Caucasus." *European Journal of Human Genetics: EJHG* 20 (12): 1275–82.
Sahakyan, Hovhannes, Baharak Hooshiar Kashani, Rakesh Tamang, Alena Kushniarevich, Amirtharaj Francis, Marta D. Costa, Ajai Kumar Pathak, et al. 2017. "Origin and Spread of Human Mitochondrial DNA Haplogroup U7." *Scientific Reports* 7 (April): 46044.
Scheduled Tribes Development Department. 2022. Scheduled Tribes Development Department Website. 2022. https://www.stdd.kerala.gov.in.
Schneider, David Murray. 1962. *Matrilineal Kinship, Edited by David M. Schneider and Kathleen Gough*. University of California Press, 1962 C1961.
Scorrano, Gabriele, Andrea Finocchio, Flavio De Angelis, Cristina Martínez-Labarga, Jelena Šarac, Irene Contini, Giuseppina Scano, Natalija Novokmet, Domenico Frezza, and Olga Rickards. 2017. "The Genetic Landscape of Serbian Populations through Mitochondrial DNA Sequencing and Non-Recombining Region of the Y Chromosome Microsatellites." *Collegium Antropologicum* 41 (3): 275–96.
Sengupta, Sanghamitra, Lev A. Zhivotovsky, Roy King, S. Q. Mehdi, Christopher A. Edmonds, Cheryl-Emiliane T. Chow, Alice A. Lin, et al. 2006. "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*. https://doi.org/10.1086/499411.
Shah, Anish M., Rakesh Tamang, Priya Moorjani, Deepa Selvi Rani, Periyasamy Govindaraj, Gururaj Kulkarni, Tanmoy Bhattacharya, et al. 2011. "Indian Siddis: African Descendants with Indian Admixture." *American Journal of Human Genetics* 89 (1): 154–61.
Shamoon-Pour, Michel, Mian Li, and D. Andrew Merriwether. 2019. "Rare Human Mitochondrial HV Lineages Spread from the Near East and Caucasus during Post-LGM and Neolithic Expansions." *Scientific Reports* 9 (1): 14751.
Silva, Constanza P., Constanza de la Fuente Castro, Tomás González Zarzar, Maanasa Raghavan, Ayelén Tonko-Huenucoy, Felipe I. Martínez, and Nicolás Montalva. 2022. "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." *Frontiers in Genetics* 13. https://doi.org/10.3389/fgene.2022.817318.
Srinivas, M. N. 1965. *Religion and Society among the Coorgs of South India*. Bombay: Asia Publishing House.

Sylvester, Charles, J. S. Rao, Adimoolam Chandrasekar, and M. S. Krishna. 2019. “An Updated Phylogeny of mtDNA Haplogroup R8 Based on Complete Mitogenomes.” *Journal of the Anthropological Survey of India* 68 (1): 114–22.

TallBear, Kim. 2013. “Genomic Articulations of Indigeneity.” *Social Studies of Science* 43 (4): 509–33.

Tanaka, Masashi, Vicente M. Cabrera, Ana M. González, José M. Larruga, Takeshi Takeyasu, Noriyuki Fuku, Li-Jun Guo, et al. 2004. “Mitochondrial Genome Variation in Eastern Asia and the Peopling of Japan.” *Genome Research* 14 (10A): 1832–50.

Tätte, Kai, Luca Pagani, Ajai K. Pathak, Sulev Kõks, Binh Ho Duy, Xuan Dung Ho, Gazi Nurun Nahar Sultana, et al. 2019. “The Genetic Legacy of Continental Scale Admixture in Indian Austroasiatic Speakers.” *Scientific Reports* 9 (1): 3818.

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

Van Oven, Mannis. 2015. “PhyloTree Build 17: Growing the Human Mitochondrial DNA Tree.” *Forensic Science International: Genetics Supplement Series* 5: e392–94.

Weissensteiner, Hansi, Dominic Pacher, Anita Kloss-Brandstätter, Lukas Forer, Günther Specht, Hans-Jürgen Bandelt, Florian Kronenberg, Antonio Salas, and Sebastian Schönherr. 2016. “HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing.” *Nucleic Acids Research* 44 (W1): W58–63.

Wells, Spencer. 2007. *Deep Ancestry: Inside the Genographic Project*. National Geographic Books.

Yaka, Reyhan, Ayşegül Birand, Yasemin Yılmaz, Ceren Caner, Sinan Can Açan, Sidar Gündüzalp, Poorya Parvizi, Aslı Erim Özdoğan, İnci Togan, and Mehmet Somel. 2018. “Archaeogenetics of Late Iron Age Çemialo Sırtı, Batman: Investigating Maternal Genetic Continuity in North Mesopotamia since the Neolithic.” *American Journal of Physical Anthropology* 166 (1): 196–207.