Integrating linguistics, social structure, and geography to model genetic diversity within India

Aritra Bose[1], Daniel E. Platt[1], Laxmi Parida[1], Petros Drineas[2], Peristera Paschou[3,*],

1 Computational Genomics, IBM T.J Watson Research Center, Yorktown Heights, NY
2 Computer Science Department, Purdue University, West Lafayette, IN
3 Department of Biological Sciences, Purdue University, West Lafayette, IN

*** ppaschou@purdue.edu**

Abstract

India represents an intricate tapestry of population substructure shaped by geography, language, culture and social stratification. While geography closely correlates with genetic structure in other parts of the world, the strict endogamy imposed by the Indian caste system and the large number of spoken languages add further levels of complexity to understand Indian population structure. To date, no study has attempted to model and evaluate how these factors have interacted to shape the patterns of genetic diversity within India. We merged all publicly available data from the Indian subcontinent into a data set of 891 individuals from 90 well-defined groups. Bringing together geography, genetics and demographic factors, we developed COGG (Correlation Optimization of Genetics and Geodemographics) to build a model that explains the observed population genetic substructure. We show that shared language along with social structure have been the most powerful forces in creating paths of gene flow in the subcontinent. Furthermore, we discover the ethnic groups that best capture the diverse genetic substructure highlighted by COGG. Integrating data from India with a data set of additional 1,323 individuals from 50 populations we find that Europeans show shared genetic drift with the Indo-European and Dravidian speakers of India, whereas the East Asians have the maximum shared genetic drift with Tibeto-Burman speaking tribal groups.

Introduction

The genetic structure of human populations reflects gene flow around and through geographic, linguistic, cultural, and social barriers [16], [59]. The intricate tapestry of population substructure and complexity in India undoubtedly showcases the interplay among them; 3,200 km from North to South, complex topography with elements ranging from the Himalayas to the Thar desert, plateaus and rain forests, almost 800 spoken languages, a long history of migrations and invasions and a strict caste system imposing endogamy are factors that have shaped extant human genetic diversity within India.

The strata within India can be summarized into the so-called backward castes and forward castes ([24]), while 8.2% of the total population belongs to tribes (1991 census) representing minorities that are unassimilated into the caste system. The tribes in India continue to live in forest hills and naturally isolated regions with a largely

hunting-gathering subsistence mode. They practice endogamy, governing mate-exchange within local groups [64]. On the other hand, the caste system is a rigorous social hierarchy of endogamous groups in which individuals are born into [43], [68]. Prior to the establishment of the caste system there was wide admixture among them, which came to an abrupt end 1,900 to 4,200 years before present [39]. Historically, the so-called forward castes have been associated with socio-economic privileges while the backward castes and tribal groups faced social segregation [24]. Although discrimination on the basis of caste was abolished by the Indian constitution in 1950, this strict social structure has existed for thousands of years [63].

Numerous studies have attempted to dissect the genetic components and origins of Indian populations [7], [34], [55], [8], [14], [53], [37], [4], [39], [9], [58], [46] along with ancient individuals from Central and South Asia [40]. Studies of Indian populations based on groupings of tribal versus non-tribal, geographic regions, or linguistic affiliation have shown that the observed genetic structure resulted from admixture of five ancestral populations. These are Ancestral North Indians, which loosely captures Indo-European (IE) speakers in Northern India; Ancestral South Indians, who are mostly Dravidian (DR) speakers of Southern India; Ancestral Austroasiatic with Austroasiatic (AA) speakers of Central and Eastern India; Ancestral Tibeto-Burman speakers constituted of Tibeto-Burman (TB) speakers in Northeast and the tribal populations from Andaman archipelago (AND) [9]. Although, the Great Andamanese are language isolates, typologically and geneaologically different from other Andamanese populations, Jarawa and Onge and their spoken language is considered as the sixth language family of India [1]. However, to date, no study has attempted to model how different spatio-cultural features acted in concert in order to create the observed genetic structure across the Indian subcontinent and to evaluate the relative contribution of each factor.

Earlier attempts to investigate the covariance of allele frequencies and non-genetic factors on genetic structure, either depended heavily on assumptions and a computationally expensive Bayesian framework [13] or did not provide any statistical significance or feature selection to identify the most relevant structure-related factors [57]. To dissect the population substructure in Indian populations, we designed a quantitative framework for the evaluation of the relative contribution of geodemographic features such as geography, spoken language and social structure to the architecture of the genetic pool of human populations. Our work provides a general model that may be used to study the significance of each underlying factor on the genetic substructure of a given population.

Materials and Methods

Study design and datasets

We used PLINK1.9 ([17]) to assemble genome-wide data for 891 samples from 90 well-defined sociolinguistic groups (**Figure 1**; **Supplementary Table S1**) genotyped on 47,283 autosomal SNPs. These samples were collected from various sources ([53]; [37]; [19]; [39]; [9]) with the consent of the corresponding authors. We created subsets of this dataset in order to construct an equal representation of social groups, language families and geographical locations for this study and tested for correlation between genetics and geography along with sociolinguistic features. The normalized subset (See **Supplementary Notes** for details) for which we have reported results on COGG, contains 368 samples from 33 populations genotyped on 47,283 SNPs (**Supplementary Table S1B**). We converted all data to the same build (hg19) using LiftOver from the UCSC Genome Browser ([30]) before merging the data. Further quality control such as filtering out variants with missing call rates exceeding 5% and

minor allele frequency (MAF) of at least 5% was performed in PLINK ([50]; [17]).

We merged 1,323 individuals across 50 populations from Eurasia and Southeast Asia, collected from various publicly available sources such as HGDP ([15]), the Estonian Biocenter ([10]; [69]; [25]; [27]; [32]; [51]; [70]) and the Allele Frequency Database (ALFRED) ([52]) (**Supplementary Table S1C**) with our normalized Indian dataset to create a merged data set of 1,691 samples from 83 population groups genotyped on 42,975 SNPs overlapping between all data sets (**Supplementary Table 1**).

Mantel Tests

We ran Mantel tests on the normalized Indian data set by computing pairwise F_{ST} distances between 33 groups using PLINK 1.9 ([17]). We performed 10,000 permutations and estimated Spearman's correlation, acknowledging the caveat of overestimation of p-values obtained from the tests ([28]).

PCA and LDA

We used TeraPCA ([12]) to perform PCA on our datasets after pruning for LD structure by setting `--indep-pairwise 50 10 0.4` in PLINK 1.9 ([17]). We checked for outliers (using EIGENSTRAT's ([49]) outlier detection method) in the PCA plot (**Supplementary Figure 2A**) and removed three outliers, each one from TB speakers Jamatia, Tripuri and Sherpa.

We implemented Rao's Discriminant Analysis which is directly based on Fisher's Linear Discriminant Analysis (**Supplementary Note**).

COGG and feature selection using Orthogonal Matching Pursuit

COGG stands for Correlation Optimization of Genetics and Geodemographics and maximizes the correlation between one of the top two principal components (for more PCs see CCA section in **Supplementary Note**) and the Geodemographic matrix, aimed to model genetic structure within India, consists of nine features (columns) corresponding to geographical coordinates, social groups and language information encoded as indicator variables. We restrict our encoding into three social groups (A, B and C) as described before. $\mathbf{u}$ is the vector containing either one of the top two principal components, computed by TeraPCA ([12]); the Geodemographic matrix is mean-centered and denoted as $\mathbf{G} \in \mathbb{R}^{m \times k}$, including k features across m individuals. The social group (A, B, C) and language (AA, DR, IE, TB) encoding was performed as 1 if the sample belongs to that social group (or language) and 0, otherwise.

Let $\mathbf{a}$ be the k-dimensional vector whose elements are $a_1 \dots a_k$ (in our case, $k = 9$). COGG solves the following optimization problem (see **Supplementary Note** for details):

$$\max_{\mathbf{a}} \mathbf{Corr}\left(\mathbf{u}, \sum_{i=1}^{k} a_i \mathbf{G}_i\right).$$

Recall that $\mathbf{G}_i$ denotes the i-th column of $\mathbf{G}$ as a column vector. Let $d_i = \mathbf{u}^T \mathbf{G}_i / \sqrt{\mathbf{Var}\left[\mathbf{u}\right]}$ for $i = 1 \dots k$ and let $\mathbf{d}$ be the vector of the d_i's. Also, let $M_{ij} = \mathbf{G}_i^T \mathbf{G}_j$ for all $i, j = 1 \dots k$ and let $\mathbf{M}$ be the matrix of M_{ij}. Then the optimizer for COGG is given by

$$\mathbf{a}_{\max} = \mathbf{M}^{-1}\mathbf{d}.$$

We also check for statistical significance of the maximum squared Pearson correlation coefficient r^2, returned by COGG, by randomly permuting the columns corresponding

to social factors and languages in $\mathbf{G}$ in 1,000 iterations and calculating $\mathbf{a}_{\max}$ for each iteration; we report the histogram of the resulting r^2 values (**Supplementary Figure 6**).

We used a greedy feature selection algorithm described in ([41]) to select features of the Geodemographic matrix $\mathbf{G}$. We obtain two sets, S_1 and S_2 of the three most significant quantitative features from the nine features in $\mathbf{G}$, one for PC1 and the other for PC2. The algorithm is described in detail in the **Supplementary Note**. In short, it selects the column which results in the maximum r^2 value from $\mathbf{G}$ and then projects $\mathbf{G}$ (and $\mathbf{u}$) on the subspace perpendicular to the selected column in order to form $\mathbf{G}'$ (and $\mathbf{u}'$) . We iterate the process until we remove the required number of features from $\mathbf{G}$.

All the values returned by this method are statistically significant, as random permutations of the elements of the two feature sets return negligible r^2. We considered all $\binom{9}{3}$ possible sets of three features exhaustively and concluded that, out of all possible sets, using only S_1 and S_2 would return the maximum correlation to PC1 and PC2 respectively.

Ridge Leverage Scores

We devised a simple method based on the Ridge Leverage Score (RLS) statistic in order to identify Indian populations that maximally contribute to the genetic diversity within the Indian sub-continent. We considered the genotype data, denoted by mean-centered (by SNPs) matrix $\mathbf{Z} \in \mathbb{R}^{m \times n}$ where m is the number of individuals and n is the number of markers in the pan-Indian data set of 90 Indian populations (891 individuals) and 47,283 SNPs. Since we are interested in the median RLS statistic as the representative of a population, including groups of larger sample size would not introduce any biases, so there was no need for normalization. We also considered the mean-centered Geodemographic matrix $\mathbf{G}$. Our analysis procedure based on the RLS statistic has four steps:

- We apply the RLS algorithm separately to the matrices $\mathbf{Z}$ and $\mathbf{G}$ to find their corresponding row ridge leverage scores, denoted by $\tau_i^\lambda(\mathbf{Z})$ and $\tau_i^\lambda(\mathbf{G})$, respectively, for $i = 1 \ldots m$.
- We grouped the RLSs by population groups to obtain a single score (median RLS) per group. If there are $T = \{t_1, t_2, \ldots, t_T\}$ populations in the entire set of the Indian populations ($|T| = 90$ in this case), then we obtain $|T|$ RLSs in this manner, one per population t_i, defined as the $|T| \times 1$ vectors $\bar{\tau}^\lambda(\mathbf{Z})$ and $\bar{\tau}^\lambda(\mathbf{G})$.
- Next, we compute an additive RLS for each population after normalizing the vectors obtained in the last step. This additive RLS highlights the significant rows (in our case, Indian populations), across both the genotype and the Geodemographic matrices. We define this consolidated additive RLS as,

$$\widetilde{\tau} = \bar{\tau}^\lambda(\mathbf{Z}) + \bar{\tau}^\lambda(\mathbf{G}).$$

- Finally, we sort the entries of $\widetilde{\tau}$ in descending order to obtain a set of representative populations.

Estimating population admixture and meta-analysis

We used the ADMIXTURE v1.22 software ([3]) for all admixture analyses. Prior to running ADMIXTURE, we pruned for LD using PLINK ([17]) by setting `--indep-pairwise 50 10 0.8`. To determine the optimal number of ancestral

populations (K), we varied K between two and eight performing iterations until convergence for each value of K (performing this procedure eight times).

We also performed a quantitative analysis (**Supplementary Note**) of ADMIXTURE's output using a method described and implemented in ([60]). To compute the shared ancestry between populations $\mathbf{X}$ and $\mathbf{Y}$, we create two matrices $\mathbf{P_X} \in \mathbb{R}^{x \times K}$ and $\mathbf{P_Y} \in \mathbb{R}^{y \times K}$ containing the estimates from ADMIXTURE, where x and y are the numbers of samples in $\mathbf{X}$ and $\mathbf{Y}$ respectively. Thereafter, we project $\mathbf{P_X}$ onto the subspace spanned by $\mathbf{P_Y}$. In other words, we take the top p eigenvectors of $\mathbf{P_X}$, $\mathbf{V_X}$ and perform the following to find the shared ancestry between $\mathbf{X}$ and $\mathbf{Y}$,

$$\frac{\|\mathbf{P_Y}\mathbf{V_X}\|_F^2}{\|\mathbf{P_X}\|_F^2}$$

We designed a color-coding scheme to better visualize the results, where the highest shared ancestry between two populations is black and the lowest shared ancestry is white. All intermediate values of shared ancestry follow a gradient from white to black.

Three population statistics and network analysis

We used ADMIXTOOLS ([47]) to compute f_3 statistics and used YRI (Yorubans) as an outgroup population. We set the significance thresholds for z-score as $|Z| > 3$.

To better visualize and understand the connection between the populations included in our study, we performed a network analysis on the meta-analysis results of ADMIXTURE, using a method presented by a previous study ([44]). We varied the nearest neighbors (NN) of the graph (**Figure 2**) from four to eight for NN = 5 . As we can have $\binom{m}{2}$ number of edges for an undirected graph with m nodes, we allow edges to the graph (**Figure 2**) until all the n populations (nodes) appear in the graph with their corresponding NNs sorted by decreasing edge weight (shared ancestry). For Figure 2 we obtained the top 40% of all edges using this criterion.

Results and Discussion

Description of Compiled Data Sets

We begin by briefly introducing the different data sets that are presented throughout our analysis (**Supplementary Table 1**). We initially compiled a pan-Indian data set of 891 individuals across 90 population groups (**Supplementary Table 1A**) and 47,283 SNPs from various sources [53], [37], [19], [39], [9]. This data set presented unequal representations of the five language families IE, DR, AA, TB and AND as well as uneven distribution across social groups and geographical regions. To create a normalized subset across these spatio-cultural features we selected a subset of 33 populations spanning 368 individuals (**Supplementary Table 1B**) from the pan-Indian data set and used it for the COGG and subsequent feature selection analyses. For other analyses such as the RLS statistic identifying representative ethnic groups contributing to the genetic diversity in India and relationship between sociolinguistic groups, we used the pan-Indian data set. Furthermore, in order to interrogate the shared ancestry between Eurasia and the Indian sociolinguistic groups, we merged the normalized subset with 1,323 individuals from 50 populations and 42,975 SNPs across Eurasia (**Supplementary Table 1C**). For the outgroup f3 analysis we present later in this section, we also used 124 samples of Yorubans in Nigeria (YRI) in 1000 Genomes data set [5] and merged it with the Eurasian data set.

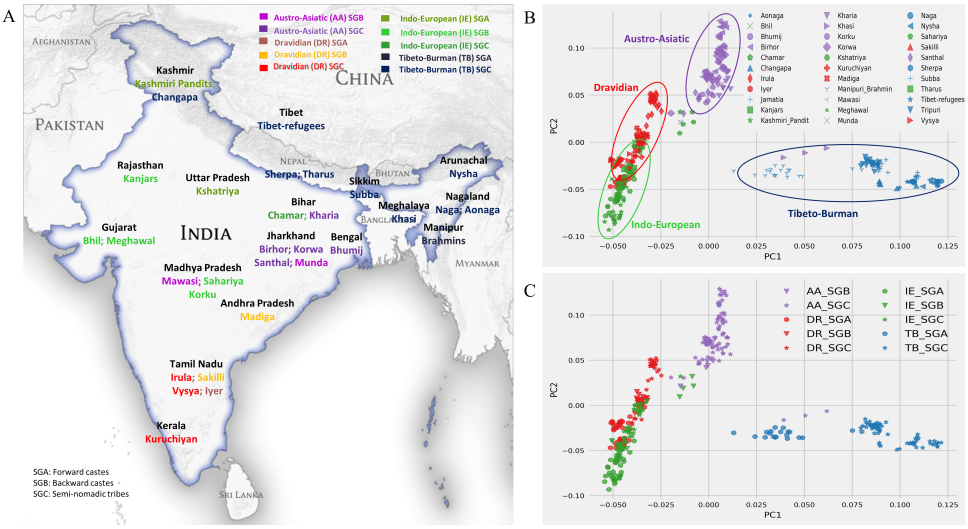

Figure 1. **A**. Map of India showing the pan-Indian locations of the 368 individuals in the normalized subset across 33 well-defined population groups, 47,283 SNPs, (see Supplementary Figure 1A for the entire data set of 90 ethnic groups across India). The populations are colored by their sociolinguistic group. **B**. Top two PCs of the normalized data set show clustering by language groups. **C**. PCA plot colored and marked by sociolinguistic groups affirming the stratified gene flow in Indian subcontinent.

Geography versus population structure within India

Studies of populations in different parts of the world, have previously shown that, when top PCs are extracted from genome-wide genotypes, individuals from the same geographic region cluster together with the PCs being well correlated with geographic coordinates, namely longitude and latitude [33], [54], [22], [45]. For instance, Novembre et al. (2008) [42] showed that within Europe, the squared Pearson-correlation coefficient r^2 between the top singular vector of the genetic covariance matrix vs. latitude (North-South) is equal to 0.77 and 0.78 for the second singular vector of the same matrix vs. longitude (East-West). In order to explore whether Indian genetic information mirrors geography, we computed the top two PCs using TeraPCA [12] on 33 Indian populations normalized over social groups, language families and their geographical distribution (details in **Materials and Methods** and **Supplementary Note**) and plotted the top two left singular vectors of the resulting genetic covariance matrix (**Figure 1** and **Supplementary Figures 1 and 2** for the entire dataset), with the first three PCs explaining 32%, 15% and 10% of the total variance, respectively. Along PC1, we observed a separation of TB speakers from the rest of the Indian populations. On the other hand, the IE and DR speaking populations formed a cline separated from AA speakers in PC2. Next, we computed the Pearson correlation coefficient (r^2) between the top two left singular vectors (we will denote them by PC1 and PC2) of the covariance matrix and the geographic coordinates (longitude and latitude) of the samples under study. We observed $r^2 = 0.604$ ($p < 10^{-9}$) for PC1 vs. longitude and $r^2 = 0.065$ ($p < 10^{-9}$) for PC2 vs. latitude. Thus, PC1 correlates well with longitude due to the East-West cline of language families with IE and TB speakers in Northwestern and Northeastern frontiers, respectively and AA speakers dwelling in the forests of Central India between them. However, PC2 only minimally correlates with latitude, just barely picking up a previously reported North-South cline of IE and

DR speakers [53] (**Figure 1B**). We note that IE and DR speakers also share significant ancestry among SGA and SGB groups (**Supplementary Figure 3**). Interestingly, in the second and third PCs, we observed further sociolinguistic stratification with SGCs distinguished from SGA and SGB within their language group (**Supplementary Figure 4**).

This weak correlation between geography and genetics in Indian context is confirmed by Mantel tests between genetic (F_{ST}) and geographic distances which returned a low $r^2 = 0.17$ ($p = 0.0001$, $Z = 5.71$) when run on the normalized data set with 33 groups. These findings are in sharp contrast with findings within the European continent [42], [26] and highlight the need for social and linguistic factors to be accounted for, as noted in prior work [7], [55], [14], [35], [9]. We performed Linear Discriminant Analysis (LDA) (**Supplementary Figures 5**) in order to gain further understanding of the relationship between genetics, geography, language and social groups in shaping the structure of the data. We run LDA on the normalized subset with the classes set as language groups (**Supplementary Figures 5A**) and then as geographic regions (**Supplementary Figures 5B**). In the LDA performed by language group, three separate clusters capturing IE social groups (SGA, SGB and SGC) appear in one axis of variation. The second axis captures the rest of the language groups again stratified by social group. In the LDA performed by geography, we see an east-west cline with TB speakers in the left and IE speakers in the right along the first discriminant. However, the second discriminant does not pick up the north-south cline as was expected, further indicating confounding by sociolinguistic groups.

Correlation Optimization of Genetics and Geodemographics

Having shown that geography alone cannot explain the genetic structure within India, we applied COGG to explore whether integrating information on spoken language and social structure as shaped by endogamy can lead to an improved model. Indeed, solving the optimization problem that underlies COGG (see **Materials and Methods** and **Supplementary Note** for the exact formulation) and plugging in the solution, we observe a significant increase in a Pearson correlation coefficient (r^2), from 0.6 to 0.93 ($p < 10^{-22}$) for PC1 vs. **G** and from 0.06 to 0.85 ($p < 10^{-15}$) for PC2 vs. **G**. Our results clearly show that endogamy and language families are pivotal in studying the genetic stratification of Indian populations. This is in sharp contrast to what has been seen in other parts of the world where geography is a major contributor in shaping genetic structure of populations [15], [42], [5]. Our results are statistically significant (**Supplementary Figure 6**) over 1,000 iterations with permutation of the variables related to social factors and languages (see **Supplementary Note**).

We further explored an extension of COGG in order to jointly analyze multiple PCs simultaneously and not just each component individually. To do this, we employed Canonical Correlation Analysis (CCA), a well-studied statistical technique, which maximizes the correlation between the genetic and the Geodemographic matrices by jointly finding linear combinations of the variables in each matrix. We used the top eight PCs of the genetic matrix as the results did not improve significantly, beyond that (**Supplementary Figure 7**). We note that these eight principal components capture, collectively, 89% of the variance of the genetic matrix.

Running COGG-CCA on these inputs returns a statistically significant (**Supplementary Figure 7**) r^2 equal to 0.94 (which is well above the $r^2 = 0.74$ obtained when COGG-CCA is run without including the sociolinguistic factors and significant with $p < 10^{-16}$. See **Supplementary Note** for details.)

Identifying the features that drive population structure within India

In order to formally investigate which of these nine features in the Geodemographic matrix **G** contribute more in the optimization problem posed by COGG, we used the sparse approximation framework and the Orthogonal Matching Pursuit (OMP) algorithm from applied mathematics ([41]) (see **Supplementary Note**). Running OMP on our dataset we obtain two sets of three features each, S_1 and S_2, for PC1 and PC2 respectively:

$$S_1 = \{\text{AA, TB, SGA}\}, \quad \textit{and}$$
$$S_2 = \{\text{AA, Latitude, SGA}\}.$$

Plugging in S_1 as the reduced feature space in COGG resulted in $r^2 = 0.92$ ($p < 10^{-15}$) for PC1 vs. S_1 and 0.85 ($p < 10^{-12}$) for PC2 vs. S_2. These values capture over 99% of the correlation returned by COGG when all the features in G are included. Membership to the AA and TB language groups which are identified among the top significant features correspond mostly to tribal nomadic hunter gatherers dwelling in the hills and forests of Central East and North East India, respectively. Thus, the AA and TB language groups automatically capture SGC. On the other hand, membership to SGA, which is the other top significant feature that we identified, spans most of the IE and DR speakers found across Northern and Southern India. Thus, these three features appear to encompass most of the geographic, social and linguistic diversity found in the Indian subcontinent and highlight their interplay.

Table 1. Top ten significant ethnic groups in India capturing the genetic structure of the subcontinent as reflected by the RLS statistic (* Vysyas are classified as in between SGA and SGB [39].

Population group	State/Territory	Language family	Social group
Great Andamenese	Andaman and Nicobar islands	Great Andamanese	SGC
Minicoy	Lakshwadeep islands	IE	SGB
Vedda	Sri Lanka	IE	SGC
Vysya	Andhra Pradesh	DR	SGA *
Palliyar	Tamil Nadu	DR	SGC
Munda	Madhya Pradesh	AA	SGC
Changpas	Jammu and Kashmir	TB	SGC
Manipuri Brahmins	Manipur	TB	SGA
Meghawal	Rajasthan	IE	SGB
Jarawa	Andaman and Nicobar islands	Ongan	SGC

Ethnic groups capturing genetic diversity across India

We developed a simple approach based on the Ridge Leverage Score (RLS) statistic ([2]) (see Equation **??** and **Supplementary Note** for details) to identify influential (from a genetic perspective) Indian populations which represent and capture the greatest portion of observed genetic diversity across India. Here, we analyzed the full pan-Indian data set of 90 populations (details in **Materials and Methods**).

The RLS statistic highlights ethnic groups in the Indian subcontinent who either are quite distinct (e.g. underwent a founder event, or practiced endogamy and maintained isolation from other groups) or populations that show signs of admixture from distinctly different language families (**Table 1**). Such populations create a mesh of complex layers

of admixture across language and social barriers. We observe mostly SGB and SGC populations across all the language families in India encapsulate much of its genetic structure. Some of the highlighted populations are: (1) Great Andamanese and Jarawas from Andaman and Nicobar archipelago, which represent distinct ethnic groups and outliers with respect to mainland Indian populations (**Supplementary Figure 2B**). Great Andamanese are also linguistically divergent from Jarawa ([1]); (2) Vysyas, who underwent a founder event going back 100 generations, due to the strong imposition of endogamy ([53]); (3) Language isolates Vedda from Sri Lanka ([18]); (4) Minicoy from Lakhswadeep archipelago with strong founder effects and diverse mixture due to the archipelago being a popular destination for maritime sailors ([56]); (5) AA speaking Mundas who have Ancestral North and South Indian ancestry and an Ancestral Southeast Asian component ([61]); (6) Manipuri Brahmins (TB_SGA) who show high shared ancestry with IE_SGA as well as TB_SGC (**Supplementary Table S2**), since they are at the junction of both language families and (7) TB speaking Changpas, who are semi-nomadic pastoralits dwelling in the high altitudes of Tibet and Ladakh in India.

Relationship between sociolinguistic groups

Our analyses using COGG clearly support the fact that language families and endogamy within social groups have played a significant role in shaping the genetic structure of the Indian subcontinent. Here, we further dissect the relationship between the endogamous social groups including the Andaman isolates [62], [38] in order to highlight the cryptic relatedness among ethnic groups that COGG posits.

The SGA populations across language groups share approximately 85% average ancestry with SGC belonging to the same language group (**Supplementary Figure 3B**). The IE and DR language groups show more homogeneity in shared ancestry than the TB and AA groups. This supports the notion that there was mixture between IE and DR speakers across SGA and SGB (**Supplementary Figure 8**) around 1,900 to 4,200 years ago ([39]) and that the caste system originated from a "classless" society, which became hierarchical with the knowledge of agriculture [31], [34].

To better illustrate the intricacies in the relationships between the social groups in India, we constructed a network of all the 90 populations across India (**Figure 2**). The network was built as we have previously described [44] based on weights that reflect shared ancestry (**Supplementary Table S2**) as computed by meta-analysis of ADMIXTURE results [3] (see **Materials and Methods** and **Supplementary Note** for details). The shared ancestry network, revealed four clusters, mostly owing to the sociolinguistic stratification.

IE and DR populations across social groups

A cluster of IE and DR speakers across social groups resembles a nearly complete graph with over 60% of all possible edges presented (**Figure 2**). We observe a similar pattern of strong shared ancestry in outgroup f_3 statistics [47], [48] using Yorubans in Nigeria (YRI) from the 1000 Genomes phase3 dataset as the outgroup [5]. We find that IE and DR populations share more alleles with each other (**Supplementary Figure 9**) with some exceptions such as Paniyas, Kadar, and Irula (also observed in **Figure 2**).

f_3 tests for signs of admixture (**Supplementary Table S3**), and meta-analyses of ADMIXTURE (**Supplementary Figure 8**) revealed similar patterns, with IE speakers showing more homogeneity across social groups than DR speakers (see also **Supplementary Note** for more details).

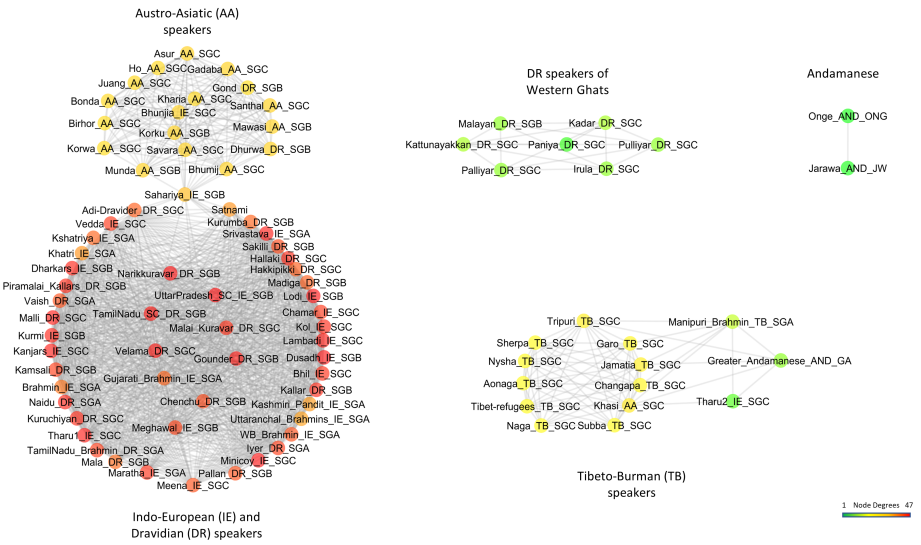

Figure 2. Network of 90 Indian populations (891 individuals) based on shared ancestry as defined by meta-analysis of ADMIXTURE results. Only the top 40% of edges (most related) populations are shown here (see Methods for details).

Isolation of DR speakers of Western Ghats

Few DR_SGC groups such as Kadar, Irula, Palliyar, and Paniya (which contain the lowest levels of Ancestral North Indian ancestry among Indian populations ([39])) formed a connected component, isolated from the main IE-DR cluster. They are hunter gatherer populations dwelling in the forests of Western Ghats in Southern India, isolated from the rest of the DR_SGCs and very low shared ancestry with IE_SGC (**Supplementary Figure 8**).

AA speakers forming a clique

Almost all AA populations from Central and East India tightly cluster together with fellow Central Indian groups such as Bhunjia (IE_SGC), Gonds (DR_SGB) and Sahariya (IE_SGB).

The Gonds and Sahariyas are a candidate mosaic Indian population, which is also reflected by their location as bridge nodes in the graph. They contain high AA, DR and IE ancestry (**Supplementary Figures 8 and 9** and **Supplementary Table S2**), which can be attributed to their central location in India [21] and their long history of exogamy.

Clique of TB speakers connected to Great Andamanese

TB speakers from North East India forms a cluster connected to the Khasis (AA speakers residing in North East India). The cluster also contain Manipuri Brahmins (TB_SGA), who are known to have significant admixture from IE_SGA (see **Supplementary Table S3**) and Tharus [20] from Tarai region in Nepal and eastern India.

The cluster reveals interaction of TB speakers of North East India with the Great Andamanese with $\sim 50\%$ shared ancestry (**Supplementary Table S2**) as well as showing strong shared genetic drift with respect to outgroup f_3 statistics (**Supplementary Figure 9**). The Great Andamanese is known to be

genetically divergent from other Andamanese groups Jarawa and Onge [62], [1]. To the best of our knowledge, this is the first observed interaction of the group with rest of mainland Indian populations.

Isolated Andamanese groups

The other Andamanese groups Jarawa and Onge diverge from the rest of the Indian populations. This has also been shown in [62], [53], [9], [38]. They belong to the Ongan language family which has a debatable connection with Austronesian languages [11], showing divergence from all language families in mainland India.

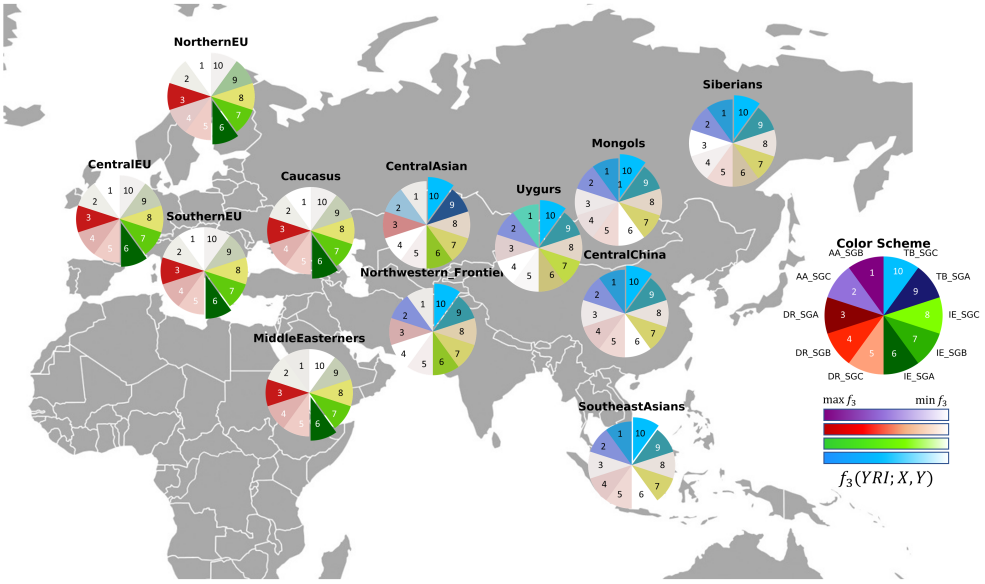

Figure 3. Shared genetic drift between 33 Indian populations (denoted by X) and 50 Eurasian/East Asian populations (denoted by Y) as estimated by f_3 statistics with Yoruba as an outgroup f_3(YRI;X,Y). Europeans show shared genetic drift with the IE and DR speakers of India, whereas the East Asians have the maximum shared genetic drift with TB speakers from SGC (tribal groups). The darkest colors correspond to greatest portions of shared genetic drift with Indian populations. Full results can be found in Supplementary Table S4.

The mosaic of Indian sociolinguistics in the context of Eurasia

Indian populations from diverse sociolinguistic groups have different genetic affinites towards Eurasian populations [53], [39], [29], [9], [46], [40]. To understand their relationship to each group, we analyzed the outgroup f_3 statistics, using YRI as the outgroup (**Supplementary Table S1C**) as done in previous studies on South Asia [6, 46]. We test the extent of allele sharing between the defined social groups of AA, DR, IE and TB speakers in India and Eurasian populations. We conduct f_3 tests such as $f_3(YRI; X, Y)$, where X is an Indian group and Y is a Eurasian group. Outgroup f_3 statistics reveal European populations showing greater shared genetic drift, with the IE social groups, along with DR_SGA.

The East Asian populations have more shared drift with the TB speakers along with some affinity with AA speakers, which is in agreement with a previous study [61]. Our

results clearly show a gradient of genetic drift from Siberia, then Mongolia, splitting, on one hand, towards China and North East India and, on the other hand, towards the Uygurs, Central Asia, Middle Easterners, and Europeans. This is concordant with our findings from network analysis with respect to connections with possible gateways to and from the Indian subcontinent (**Supplementary Figure 10**).

Conclusion

India represents a country of great social and linguistic complexity. We established a quantitative deterministic and non-parametric framework called COGG, aiming to evaluate the relative contribution of language, social structure and geography in shaping the Indian gene pool. COGG resulted in a dramatic increase in the Pearson correlation coefficient (r^2) between top PCs depicting genomic structure and the geodemographic factors that we investigated. We applied a feature selection algorithm to identify the most important factors shaping genomic structure in India, as well as a RLS statistic to highlight ethnic groups in India that best capture its diverse gene pool. Intriguingly, our study shows that spoken language seems to have been the major force bringing people together in India, across geographic and social barriers. This is in sharp contrast with what has been shown for other parts of the world where geography seems to have a major impact [15], [45], [5] and highlights the need for population-specific studies.

We find evidence for wide mixture across all the social groups (tribal and non-tribal) for IE speakers and across SGA and SGB for DR speakers in India. This is in concordance with a previous study [34] that indicates the origin of the caste system from a semi-nomadic society which became hierarchical with the knowledge of agriculture [31], [34]. As shown previously, this wide admixture came to an abrupt end around 1,900 to 4,200 years ago [39].

IE languages (primarily spoken in Northern India) are part of a larger language family that includes a great majority of European languages. In contrast, DR languages (primarily spoken in Southern India) are not closely related to languages outside of South Asia. Nevertheless, the earliest Hindu text (the Rig Veda, written in archaic Sanskrit) contains DR loanwords that are not found in IE languages outside the Indian subcontinent [36], [39], [67]. Further supporting the long contact between IE and DR speakers in India, our network analysis and f_3 tests identify a large cluster consisting of IE and DR populations which resembles an almost complete graph with almost all pairs of populations connected to each other.

We observe a stronger shared ancestry between the Great Andamenese with TB speakers of North East India than other mainland speakers. This is the first time this observation is made based on autosomal markers and should be interpreted with caution due to small sample sizes of all the groups involved. However, a study focused on the mitochondrial haplogroup M31 showed that with the exception of M31a1 (specific to Andaman archipelago), lineages M31a2, M31b and M31c are prevalent in North East India and surrounding regions [66]. The authors concluded with time estimation that the Andaman archipelago was likely settled by modern humans from North East India *via* the land-bridge connecting Andaman archipelago and Myanmar around Last Glacial Maximum (LGM) [65], [23].

The framework we have developed and applied in order to understand genetic structure within the Indian subcontinent can be applied more broadly in order to model the interaction between different factors that may have shaped genetic diversity and population structure. The possibility to correlate genomic background to geographic, social and cultural differences opens new avenues for understanding how human history and mating patterns translate into the genomic structure of extant human populations

Data availability

Data used in this manuscript is available from the respective corresponding authors. Code for COGG and COGG-CCA is available here: `https://github.com/aritra90/COGG`.

Supplementary Material

Supplementary note, tables S1 - S4 and figures S1 - S10 are available as attachments.

Acknowledgments

This study was supported by NSF IIS-1319280, NSF IIS-1661760, and IBM. We thank D. Reich and P. Moorjani for sharing genotypic data of 248 samples from Reich et al. (2009) [53] and 378 samples from Moorjani et al. (2013) [39]. We also thank P. P. Majumder who allowed us to use the genotypic data from 367 samples from Basu et al. (2015) [9].

References

1. A. Abbi. Is great andamanese genealogically and typologically distinct from onge and jarawa? *Lang Sci.*, 31(6):791–812, 2009.

2. A. E. Alaoui and M. W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 775–783, Cambridge, MA, USA, 2015. MIT Press.

3. D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664, Sep 2009. 19648217[pmid].

4. G. ArunKumar, D. F. Soria-Hernanz, V. J. Kavitha, V. S. Arun, A. Syama, K. S. Ashokan, K. T. Gandhirajan, K. Vijayakumar, M. Narayanan, M. Jayalakshmi, J. S. Ziegle, A. K. Royyuru, L. Parida, R. S. Wells, C. Renfrew, T. G. Schurr, C. T. Smith, D. E. Platt, and R. Pitchappan. Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural Expansions Predating the Caste System. *PLoS One*, 7(11), 2012.

5. A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, J. Wang, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter,

A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L. J. M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Bhatia, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, A. G. Clark, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel, T. Rausch, M. H. Fritz, A. M. Stütz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. S. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, B. Barnes, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, E. R. Mardis, L. Ding, D. C. Koboldt, D. Larson, K. Ye, S. Gravel, A. Swaroop, E. Chew, T. Lappalainen, Y. Erlich, M. Gymrek, T. Frederick Willems, J. T. Simpson, M. D. Shriver, J. A. Rosenfeld, C. D. Bustamante, S. B. Montgomery, F. M. De La Vega, J. K. Byrnes, A. W. Carroll, M. K. DeGorter, P. Lacroute, B. K. Maples, A. R. Martin, A. Moreno-Estrada, S. S. Shringarpure, F. Zakharia, E. Halperin, Y. Baran, C. Lee, E. Cerveira, J. Hwang, A. Malhotra, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, F. C. L. Hyland, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, A. A. Kurdoglu, S. A. Sinari, K. Squire, S. T. Sherry, C. Xiao, J. Sebat, D. Antaki, M. Gujral, A. Noor, K. Ye, E. G. Burchard, R. D. Hernandez, C. R. Gignoux, D. Haussler, S. J. Katzman, W. James Kent, B. Howie, A. Ruiz-Linares, E. T. Dermitzakis, S. E. Devine, G. R. Abecasis, H. Min Kang, J. M. Kidd, T. Blackwell, S. Caron, W. Chen, S. Emery, L. Fritsche, C. Fuchsberger, G. Jun, B. Li, R. Lyons, C. Scheller, C. Sidore, S. Song, E. Sliwerska, D. Taliun, A. Tan, R. Welch, M. Kate Wing, X. Zhan, P. Awadalla, A. Hodgkinson, Y. Li, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, J. L. Marchini, S. Myers, C. Churchhouse, O. Delaneau, A. Gupta-Hinch, W. Kretzschmar, Z. Iqbal, I. Mathieson, A. Menelaou, A. Rimmer, D. K. Xifara, T. K. Oleksyk, Y. Fu, X. Liu, M. Xiong, L. Jorde, D. Witherspoon, J. Xing, E. E. Eichler, B. L. Browning, S. R. Browning, F. Hormozdiari, P. H. Sudmant, E. Khurana, R. M. Durbin, M. E. Hurles, C. Tyler-Smith, C. A. Albers, Q. Ayub, S. Balasubramaniam, Y. Chen, V. Colonna, P. Danecek, L. Jostins, T. M. Keane, S. McCarthy, K. Walter, Y. Xue, M. B. Gerstein, A. Abyzov, S. Balasubramanian, J. Chen, D. Clarke, Y. Fu, A. O. Harmanci, M. Jin, D. Lee, J. Liu, X. Jasmine Mu, J. Zhang, Y. Zhang, Y. Li, R. Luo, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P.

Garrison, D. Kural, W.-P. Lee, A. N. Ward, J. Wu, M. Zhang, S. A. McCarroll, R. E. Handsaker, D. M. Altshuler, E. Banks, G. del Angel, G. Genovese, C. Hartl, H. Li, S. Kashin, J. C. Nemesh, K. Shakir, S. C. Yoon, J. Lihm, V. Makarov, J. Degenhardt, J. O. Korbel, M. H. Fritz, S. Meiers, B. Raeder, T. Rausch, A. M. Stütz, P. Flicek, F. Paolo Casale, L. Clarke, R. E. Smith, O. Stegle, X. Zheng-Bradley, D. R. Bentley, B. Barnes, R. Keira Cheetham, M. Eberle, S. Humphray, S. Kahn, L. Murray, R. Shaw, E.-W. Lameijer, M. A. Batzer, M. K. Konkel, J. A. Walker, L. Ding, I. Hall, K. Ye, P. Lacroute, C. Lee, E. Cerveira, A. Malhotra, J. Hwang, D. Plewczynski, K. Radew, M. Romanovitch, C. Zhang, D. W. Craig, N. Homer, D. Church, C. Xiao, J. Sebat, D. Antaki, V. Bafna, J. Michaelson, K. Ye, S. E. Devine, E. J. Gardner, G. R. Abecasis, J. M. Kidd, R. E. Mills, G. Dayama, S. Emery, G. Jun, X. Shi, A. Quitadamo, G. Lunter, G. A. McVean, K. Chen, X. Fan, Z. Chong, T. Chen, D. Witherspoon, J. Xing, E. E. Eichler, M. J. Chaisson, F. Hormozdiari, J. Huddleston, M. Malig, B. J. Nelson, P. H. Sudmant, N. F. Parrish, E. Khurana, M. E. Hurles, B. Blackburne, S. J. Lindsay, Z. Ning, K. Walter, Y. Zhang, M. B. Gerstein, A. Abyzov, J. Chen, D. Clarke, H. Lam, X. Jasmine Mu, C. Sisu, J. Zhang, Y. Zhang, R. A. Gibbs, F. Yu, M. Bainbridge, D. Challis, U. S. Evani, C. Kovar, J. Lu, D. Muzny, U. Nagaswamy, J. G. Reid, A. Sabo, J. Yu, X. Guo, W. Li, Y. Li, R. Wu, G. T. Marth, E. P. Garrison, W. Fung Leong, A. N. Ward, G. del Angel, M. A. DePristo, S. B. Gabriel, N. Gupta, C. Hartl, R. E. Poplin, A. G. Clark, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, R. E. Smith, X. Zheng-Bradley, D. G. MacArthur, E. R. Mardis, R. Fulton, D. C. Koboldt, S. Gravel, C. D. Bustamante, D. W. Craig, A. Christoforides, N. Homer, T. Izatt, S. T. Sherry, C. Xiao, E. T. Dermitzakis, G. R. Abecasis, H. Min Kang, G. A. McVean, M. B. Gerstein, S. Balasubramanian, L. Habegger, H. Yu, P. Flicek, L. Clarke, F. Cunningham, I. Dunham, D. Zerbino, X. Zheng-Bradley, K. Lage, J. Berg Jespersen, H. Horn, S. B. Montgomery, M. K. DeGorter, E. Khurana, C. Tyler-Smith, Y. Chen, V. Colonna, Y. Xue, M. B. Gerstein, S. Balasubramanian, Y. Fu, D. Kim, A. Auton, A. Marcketta, R. Desalle, A. Narechania, M. A. Wilson Sayres, E. P. Garrison, R. E. Handsaker, S. Kashin, S. A. McCarroll, J. L. Rodriguez-Flores, P. Flicek, L. Clarke, X. Zheng-Bradley, Y. Erlich, M. Gymrek, T. Frederick Willems, C. D. Bustamante, F. L. Mendez, G. David Poznik, P. A. Underhill, C. Lee, E. Cerveira, A. Malhotra, M. Romanovitch, C. Zhang, G. R. Abecasis, L. Coin, H. Shao, D. Mittelman, C. Tyler-Smith, Q. Ayub, R. Banerjee, M. Cerezo, Y. Chen, T. W. Fitzgerald, S. Louzada, A. Massaia, S. McCarthy, G. R. Ritchie, Y. Xue, F. Yang, R. A. Gibbs, C. Kovar, D. Kalra, W. Hale, D. Muzny, J. G. Reid, J. Wang, X. Dan, X. Guo, G. Li, Y. Li, C. Ye, X. Zheng, D. M. Altshuler, P. Flicek, L. Clarke, X. Zheng-Bradley, D. R. Bentley, A. Cox, S. Humphray, S. Kahn, R. Sudbrak, M. W. Albrecht, M. Lienhard, D. Larson, D. W. Craig, T. Izatt, A. A. Kurdoglu, S. T. Sherry, C. Xiao, D. Haussler, G. R. Abecasis, G. A. McVean, R. M. Durbin, S. Balasubramaniam, T. M. Keane, S. McCarthy, J. Stalker, A. Chakravarti, B. M. Knoppers, G. R. Abecasis, K. C. Barnes, C. Beiswanger, E. G. Burchard, C. D. Bustamante, H. Cai, H. Cao, R. M. Durbin, N. P. Gerry, N. Gharani, R. A. Gibbs, C. R. Gignoux, S. Gravel, B. Henn, D. Jones, L. Jorde, J. S. Kaye, A. Keinan, A. Kent, A. Kerasidou, Y. Li, R. Mathias, G. A. McVean, A. Moreno-Estrada, P. N. Ossorio, M. Parker, A. M. Resch, C. N. Rotimi, C. D. Royal, K. Sandoval, Y. Su, R. Sudbrak, Z. Tian, S. Tishkoff, L. H. Toji, C. Tyler-Smith, M. Via, Y. Wang, H. Yang, L. Yang, J. Zhu, W. Bodmer, G. Bedoya, A. Ruiz-Linares, Z. Cai, Y. Gao, J. Chu, L. Peltonen, A. Garcia-Montero, A. Orfao, J. Dutil, J. C. Martinez-Cruzado, T. K. Oleksyk, K. C. Barnes, R. A. Mathias, A. Hennis, H. Watson, C. McKenzie,

F. Qadri, R. LaRocque, P. C. Sabeti, J. Zhu, X. Deng, P. C. Sabeti, D. Asogun, O. Folarin, C. Happi, O. Omoniwa, M. Stremlau, R. Tariyal, M. Jallow, F. Sisay Joof, T. Corrah, K. Rockett, D. Kwiatkowski, J. Kooner, T. Tñh Hiê'n, S. J. Dunstan, N. Thuy Hang, R. Fonnie, R. Garry, L. Kanneh, L. Moses, P. C. Sabeti, J. Schieffelin, D. S. Grant, C. Gallo, G. Poletti, D. Saleheen, A. Rasheed, L. D. Brooks, A. L. Felsenfeld, J. E. McEwen, Y. Vaydylevich, E. D. Green, A. Duncanson, M. Dunn, J. A. Schloss, J. Wang, H. Yang, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. Min Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, and G. R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

6. Q. Ayub, M. Mezzavilla, L. Pagani, M. Haber, A. Mohyuddin, S. Khaliq, S. Q. Mehdi, and C. Tyler-Smith. The kalash genetic isolate: ancient divergence, drift, and selection. *Am J Hum Genet.*, 96(5):775–783, 2015.

7. M. Bamshad, T. Kivisild, W. S. Watkins, M. E. Dixon, C. E. Ricker, B. B. Rao, J. M. Naidu, B. V. R. Prasad, P. G. Reddy, A. Rasanayagam, S. S. Papiha, R. Villems, A. J. Redd, M. F. Hammer, S. V. Nguyen, M. L. Carroll, M. A. Batzer, and L. B. Jorde. Genetic evidence on the origins of Indian caste populations. *Genome Res.*, 11(6):994–1004, 2001.

8. A. Basu, N. Mukherjee, S. Roy, S. Sengupta, S. Banerjee, M. Chakraborty, B. Dey, M. Roy, B. Roy, N. P. Bhattacharyya, S. Roychoudhury, and P. P. Majumder. Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res.*, 13(10):2277–2290, 2003.

9. A. Basu, N. Sarkar-Roy, and 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.*, 113(6):1594–1599, 2016.

10. D. M. Behar, B. Yunusbayev, M. Metspalu, E. Metspalu, S. Rosset, J. Parik, S. Rootsi, G. Chaubey, I. Kutuev, G. Yudkovsky, E. K. Khusnutdinova, O. Balanovsky, O. Semino, L. Pereira, D. Comas, D. Gurwitz, B. Bonne-Tamir, T. Parfitt, M. F. Hammer, K. Skorecki, and R. Villems. The genome-wide structure of the jewish people. *Nature*, 466(7303):238–242, Jul 2010.

11. J. Blevins. A long lost sister of proto-austronesian?: Proto-ongan, mother of jarawa and onge of the andaman islands. *Oceanic Linguistics*, 46(1):154–198, 2007.

12. A. Bose, V. Kalantzis, E.-M. Kontopoulou, M. Elkady, P. Paschou, and P. Drineas. Terapca: a fast and scalable software package to study genetic variation in tera-scale genotypes. *Bioinformatics.*, 2019.

13. G. S. Bradburd, P. L. Ralph, and G. M. Coop. Disentangling the effects of geographic and ecological isolation on genetic differentiation. *Evolution (N Y).*, 67(11):3258–3273, 2013.

14. S. K. Brahmachari, L. Singh, A. Sharma, M. Mukerji, K. Ray, S. Roychoudhury, G. R. Chandak, K. Thangaraj, S. Habib, D. Parmar, P. P. Majumder, S. Sengupta, D. Bharadwaj, D. Dash, S. K. Rath, R. Shankar, J. Singh, K. Virdi, S. Bahl, V. R. Rao, S. Sinha, A. Singh, A. Mitra, S. K. Mishra, B. R. K. Shukla, Q. Pasha, S. Maiti, A. Sharma, J. Kumar, A. Ahsan, T. Stobdan, C. Chauhan, S. Malhotra, A. Vidhani, S. Siva, A. Baral, R. Pandey, R. Roy, M. Singh, S. P. Singh, N. Maurya, A. Bandyopadhyay, G. N. Jha, S. Dutta, G. Ghosh, T. Naiya, M. Jain, J. P. Srivatava, J. R. Gupta, V. Khanna, A. Dhawan, M. Anand, R. S.

Bharti, M. Singh, A. P. Singh, A. J. Khan, K. K. Bisht, A. Kumar, B. Ghosh, S. K. Das, T. Madan, C. Chauhan, R. Verma, U. Mittal, A. Mahajan, S. Chavali, R. Tabassum, V. Banerjee, J. Batra, R. Nagarkatti, S. Sharma, M. Sharma, R. Chatterjee, J. A. Paul, P. Srivastava, R. Chopra, A. Saxena, C. Rajput, P. K. Singh, M. Vaid, S. Das, K. Chaudhuri, R. Chowdhury, A. Mukhopadhyay, M. Acharya, A. Bhattacharyya, A. Saha, A. Biswas, M. Chaki, A. Gupta, S. Mukherjee, S. Mookherjee, I. Chattopadhyay, T. Banerjee, M. Chakravorty, C. Misra, G. Monadal, S. Sengupta, I. Deb, and A. Banerjee. The Indian Genome Variation database (IGVdb): A project overview. *Hum Genet.*, 118(1):1–11, 2005.

15. H. M. Cann, C. de Toma, L. Cazes, M.-F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. A human genome diversity cell line panel. *Science*, 296(5566):261–262, 2002.

16. L. L. Cavalli-Sforza, A. Piazza, P. Menozzi, and J. Mountain. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci.*, 85(16):6002–6006, 1988.

17. C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation plink: rising to the challenge of larger and richer datasets. *GigaScience*, 4(1):7, 2015.

18. G. Chaubey. Language isolates and their genetic identity: a commentary on mitochondrial dna history of sri lankan ethnic people: their relations within the island and with the indian subcontinental populations. *J Hum Genet.*, 59(2):61–63, 2014.

19. G. Chaubey, M. Metspalu, Y. Choi, R. Mägi, I. G. Romero, P. Soares, M. Van Oven, D. M. Behar, S. Rootsi, G. Hudjashov, C. B. Mallick, M. Karmin, M. Nelis, J. Parik, A. G. Reddy, E. Metspalu, G. Van Driem, Y. Xue, C. Tyler-Smith, K. Thangaraj, L. Singh, M. Remm, M. B. Richards, M. M. Lahr, M. Kayser, R. Villems, and T. Kivisild. Population genetic structure in indian austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Mol Biol Evol.*, 28(2):1013–1024, 2011.

20. G. Chaubey, M. Singh, F. Crivellaro, R. Tamang, A. Nandan, K. Singh, V. K. Sharma, A. K. Pathak, A. M. Shah, V. Sharma, V. K. Singh, D. Selvi Rani, N. Rai, A. Kushniarevich, A.-M. Ilumae, M. Karmin, A. Phillip, A. Verma, E. Prank, V. K. Singh, B. Li, P. Govindaraj, A. K. Chaubey, P. K. Dubey, A. G. Reddy, K. Premkumar, S. Vishnupriya, V. Pande, J. Parik, S. Rootsi, P. Endicott, M. Metspalu, M. M. Lahr, G. van Driem, R. Villems, T. Kivisild, L. Singh, and K. Thangaraj. Unravelling the distinct strains of tharu ancestry. *Eur J Hum Genet.*, 22(12):1404–1412, Dec 2014. Article.

21. G. Chaubey, R. Tamang, E. Pennarun, P. Dubey, N. Rai, R. K. Upadhyay, R. P. Meena, J. R. Patel, G. van Driem, K. Thangaraj, M. Metspalu, and R. Villems. Reconstructing the population history of the largest tribe of india: the dravidian speaking gond. *Eur J Hum Genet.*, 25(4):493–498, Apr 2017.

22. J. Chen, H. Zheng, J.-X. Bei, L. Sun, W.-h. Jia, T. Li, F. Zhang, M. Seielstad, Y.-X. Zeng, X. Zhang, et al. Genetic structure of the han chinese population revealed by genome-wide snp variation. *Am J Hum Genet.*, 85(6):775–785, 2009.

23. P. U. Clark, A. S. Dyke, J. D. Shakun, A. E. Carlson, J. Clark, B. Wohlfarth, J. X. Mitrovica, S. W. Hostetler, and A. M. McCabe. The last glacial maximum. *Science*, 325(5941):710–714, 2009.

24. S. Desai and A. Dubey. Caste in 21st century india: Competing narratives. *Econ Polit Wkly.*, 46(11):40–49, Mar 2012. 22736803[pmid].

25. J. Di Cristofaro, E. Pennarun, S. Mazières, N. M. Myres, A. a. Lin, S. A. Temori, M. Metspalu, E. Metspalu, M. Witzel, R. J. King, P. a. Underhill, R. Villems, and J. Chiaroni. Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. *PLoS One*, 8(10), 2013.

26. P. Drineas, J. Lewis, and P. Paschou. Inferring geographic coordinates of origin for europeans using small panels of ancestry informative markers. *PLOS ONE*, 5(8):1–6, 08 2010.

27. S. Fedorova et al. Autosomal and uniparental portraits of the native populations of sakha (yakutia): implications for the peopling of northeast eurasia. *BMC Evol Biol.*, 13(1):127, 2013.

28. G. Guillot and F. Rousset. Dismantling the mantel tests. *Methods Ecol Evol.*, 4(4):336–344, 2013.

29. 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. Banffy, 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. Szecsenyi-Nagy, J. Wahl, M. Meyer, J. Krause, D. Brown, D. Anthony, A. Cooper, K. W. Alt, and D. Reich. Massive migration from the steppe was a source for indo-european languages in europe. *Nature*, 522(7555):207–211, Jun 2015. Letter.

30. A. S. Hinrichs. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.*, 34(90001):D590–D598, 2006.

31. D. Kosambi. *The culture and civilisation of Ancient India in Historical Outline.* Vikas Publishing House Pvt. Ltd., 1964.

32. L. Kovacevic, K. Tambets, A.-M. Ilumäe, A. Kushniarevich, B. Yunusbayev, A. Solnik, T. Bego, D. Primorac, V. Skaro, A. Leskovac, Z. Jakovski, K. Drobnic, H.-V. Tolk, S. Kovacevic, P. Rudan, E. Metspalu, and D. Marjanovic. Standing at the Gateway to Europe - The Genetic Structure of Western Balkan Populations Based on Autosomal and Haploid Markers. *PLoS One*, 9(8):e105090, 2014.

33. O. Lao, K. van Duijn, P. Kersbergen, P. de Knijff, and M. Kayser. Proportioning whole-genome single-nucleotide–polymorphism diversity for the identification of geographic population structure and genetic ancestry. *Am J Hum Genet.*, 78(4):680–690, 2006.

34. P. P. Majumder. Indian caste origins: Genomic insights and future outlook. *Genome Res.*, 11(6):931–932, 2001.

35. P. P. Majumder. The Human Genetic History of South Asia. *Curr Biol.*, 20(4):R184–R187, 2010.

36. J. Mallory and D. Adams. *Encyclopedia of Indo-European Culture*. Fitzroy Dearborn, 1997.

37. M. Metspalu, I. G. Romero, B. Yunusbayev, G. Chaubey, C. B. Mallick, G. Hudjashov, M. Nelis, R. Mägi, E. Metspalu, M. Remm, R. Pitchappan, L. Singh, K. Thangaraj, R. Villems, and T. Kivisild. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am J of Hum Genet.*, 89(6):731–744, 2011.

38. M. Mondal, F. Casals, T. Xu, G. M. Dall'Olio, M. Pybus, M. G. Netea, D. Comas, H. Laayouni, Q. Li, P. P. Majumder, et al. Genomic analysis of andamanese provides insights into ancient human migration into asia and adaptation. *Nat Genet.*, 48(9):1066, 2016.

39. P. Moorjani, K. Thangaraj, N. Patterson, M. Lipson, P. R. Loh, P. Govindaraj, B. Berger, D. Reich, and L. Singh. Genetic evidence for recent population mixture in India. *Am J of Hum Genet.*, 93(3):422–438, 2013.

40. 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. Dobes, 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, and D. Reich. The formation of human populations in south and central asia. *Science*, 365(6457), 2019.

41. B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J Comput.*, 24(2):227–234, 1995.

42. J. Novembre and M. Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet.*, 40(5):646–649, 2008.

43. M. Olcott. The caste system of india. *Am Socio Rev.*, 9(6):648–657, 1944.

44. P. Paschou, P. Drineas, E. Yannaki, A. Razou, K. Kanaki, F. Tsetsos, S. S. Padmanabhuni, M. Michalodimitrakis, M. C. Renda, S. Pavlovic, A. Anagnostopoulos, J. A. Stamatoyannopoulos, K. K. Kidd, and G. Stamatoyannopoulos. Maritime route of colonization of Europe. *Proc Natl Acad Sci.*, 111(25):9211–9216, 2014.

45. P. Paschou, J. Lewis, A. Javed, and P. Drineas. Ancestry informative markers for fine-scale individual assignment to worldwide populations. *J Med Genet.*, 47(12):835–847, 2010.

46. 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, and R. Villems. The genetic ancestry of modern indus valley populations from northwest india. *Am J Hum Genet.*, 103(6):918 – 929, 2018.

47. N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics.*, 192(3):1065–1093, 2012.

48. B. M. Peter. Admixture, population structure, and f-statistics. *Genetics.*, 202(4):1485–1501, 2016.

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

50. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet.*, 81(3):559–575, 2007.

51. M. Raghavan et al. Upper palaeolithic siberian genome reveals dual ancestry of native americans. *Nature*, 505(7481):87–91, Jan 2014. Letter.

52. H. Rajeevan, M. V. Osier, K. H. Cheung, H. Deng, L. Druskin, R. Heinzen, J. R. Kidd, S. Stein, A. J. Pakstis, N. P. Tosches, C. C. Yeh, P. L. Miller, and K. K. Kidd. ALFRED: The ALelle FREquency Database. Update. *Nucleic Acids Research*, 31(1):270–271, 2003.

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

54. N. A. Rosenberg, S. Mahajan, C. Gonzalez-Quevedo, M. G. B. Blum, L. Nino-Rosales, V. Ninis, P. Das, M. Hegde, L. Molinari, G. Zapata, J. L. Weber, J. W. Belmont, and P. I. Patel. Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet.*, 2(12):2052–2061, 2006.

55. S. Roychoudhury, S. Roy, A. Basu, R. Banerjee, H. Vishwanathan, M. V. Usha Rani, S. K. Sil, M. Mitra, and P. P. Majumder. Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum Genet.*, 109(3):339–350, 2001.

56. P. P. Samuel, R. Krishnamoorthi, K. Hamzakoya, C. Aggarwal, et al. Entomo-epidemiological investigations on chikungunya outbreak in the lakshadweep islands, indian ocean. *Indian J Med Res.*, 129(4):442, 2009.

57. C. M. Schlebusch, P. Skoglund, P. Sjödin, L. M. Gattepaille, D. Hernandez, F. Jay, S. Li, M. De Jongh, A. Singleton, M. G. B. Blum, H. Soodyall, and M. Jakobsson. Genomic variation in seven khoe-san groups reveals adaptation and complex african history. *Science*, 338(6105):374–379, 2012.

58. M. Silva, M. Oliveira, D. Vieira, A. Brandão, T. Rito, J. B. Pereira, R. M. Fraser, B. Hudson, F. Gandini, C. Edwards, M. Pala, J. Koch, J. F. Wilson, L. Pereira, M. B. Richards, and P. Soares. A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC Evol Biol.*, 17(1):88, 2017.

59. R. R. Sokal. Ancient movement patterns determine modern genetic variances in europe. *Hum Biol.*, pages 589–606, 1991.

60. G. Stamatoyannopoulos, A. Bose, A. Teodosiadis, F. Tsetsos, A. Plantinga, N. Psatha, N. Zogas, E. Yannaki, P. Zalloua, K. K. Kidd, B. L. Browning, J. Stamatoyannopoulos, P. Paschou, and P. Drineas. Genetics of the peloponnesean populations and the theory of extinction of the medieval peloponnesean Greeks. *Eur J Hum Genet.*, 25(5):637–645, 2017.

61. K. Tätte, L. Pagani, A. K. Pathak, S. Kõks, B. H. Duy, X. D. Ho, G. N. N. Sultana, M. I. Sharif, M. Asaduzzaman, D. M. Behar, et al. The genetic legacy of continental scale admixture in indian austroasiatic speakers. *Sci Rep.*, 9(1):3818, 2019.

62. K. Thangaraj, L. Singh, A. G. Reddy, V. R. Rao, S. C. Sehgal, P. A. Underhill, M. Pierson, I. G. Frame, and E. Hagelberg. Genetic affinities of the andaman islanders, a vanishing human population. *Curr Biol.*, 13(2):86–93, 2003.

63. R. Thapar. *A history of India*. Penguin UK, 1990.

64. L. P. Vidyarthi and B. K. Rai. *The tribal culture of India*. Concept Publishing Company, 1977.

65. H. K. Voris. Maps of pleistocene sea levels in southeast asia: shorelines, river systems and time durations. *J Biogeogr.*, 27(5):1153–1167, 2000.

66. H.-W. Wang, B. Mitra, T. K. Chaudhuri, M. gounder Palanichamy, Q.-P. Kong, and Y.-P. Zhang. Mitochondrial dna evidence supports northeast indian origin of the aboriginal andamanese in the late paleolithic. *J Genet Genomics.*, 38(3):117–122, 2011.

67. M. Witzel. Substrate languages in old-indo aryan. *IJDL. International journal of Dravidian linguistics*, 30(2):1–94, 2001.

68. S. Wooding, C. Ostler, B. V. R. Prasad, W. S. Watkins, S. Sung, M. Bamshad, and L. B. Jorde. Directional migration in the hindu castes: inferences from mitochondrial, autosomal and y-chromosomal data. *Hum Genet.*, 115(3):221–229, Aug 2004.

69. B. Yunusbayev, M. Metspalu, M. Järve, I. Kutuev, S. Rootsi, E. Metspalu, D. M. Behar, K. Varendi, H. Sahakyan, R. Khusainova, L. Yepiskoposyan, E. K. Khusnutdinova, P. A. Underhill, T. Kivisild, and R. Villems. The caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol Biol Evol.*, 29(1):359–365, 2012.

70. B. Yunusbayev, M. Metspalu, E. Metspalu, A. Valeev, S. Litvinov, R. Valiev, V. Akhmetova, E. Balanovska, O. Balanovsky, S. Turdikulova, D. Dalimova, P. Nymadawa, A. Bahmanimehr, H. Sahakyan, K. Tambets, S. Fedorova, N. Barashkov, I. Khidiyatova, E. Mihailov, R. Khusainova, L. Damba, M. Derenko, B. Malyarchuk, L. Osipova, M. Voevoda, L. Yepiskoposyan, T. Kivisild, E. Khusnutdinova, and R. Villems. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS Genet.*, 11(4):1–24, 2015.