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], and Peristera Paschou[*,3]
[1] Computational Genomics, IBM T.J Watson Research Center, Yorktown Heights, NY 10598
[2] Computer Science Department, Purdue University, West Lafayette, IN 47907
[3] Department of Biological Sciences, Purdue University, West Lafayette, IN 47907
***Corresponding author:** E-mail: ppaschou@purdue.edu

Associate Editor: XXXX

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 dataset 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 using a ridge leverage score statistic. Integrating data from India with a dataset of additional 1,323 individuals from 50 Eurasian populations we find that Indo-European and Dravidian speakers of India show shared genetic drift with Europeans, whereas the Tibeto-Burman speaking tribal groups have maximum shared genetic drift with East Asians.

Key words: India; Population structure; South Asia; Genomics; Algorithms; Data Mining

Article

INTRODUCTION

The genetic structure of human populations reflects gene flow around and through geographic, linguistic, cultural, and social barriers (Cavalli-Sforza et al. 1988; Sokal 1991). The intricate tapestry of population substructure and complexity in India undoubtedly showcases the interplay among them. The Indian subcontinent encompasses 3,200 km from North to South, complex topography with elements ranging from the Himalayas to the Thar desert, plateaux and rain forests, almost 800 spoken languages, a long history of migrations and invasions and a strict caste system imposing endogamy.

The strata within India can be summarized into the so-called backward castes and forward castes (Desai and Dubey 2012), 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, a matrimonial rule governing mate-exchange within local groups (Vidyarthi and Rai 1977). On the other hand, the caste system is a rigorous social hierarchy of endogamous groups in which individuals are born (Olcott 1944; Wooding et al. 2004). 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 (Moorjani et al. 2013). Historically, the so-called forward castes have been associated with socio-economic privileges while the backward castes and tribal groups faced social segregation (Desai and Dubey 2012). 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 (Thapar 1990).

Numerous studies have attempted to dissect the genetic components and origins of Indian populations (Bamshad et al. 2001; Majumder 2001; Roychoudhury et al. 2001; Basu et al. 2003; Brahmachari et al. 2005; Reich et al. 2009; Metspalu et al. 2011; ArunKumar et al. 2012; Moorjani et al. 2013; Basu et al. 2016; Silva et al. 2017; Pathak et al. 2018) along with ancient individuals from Central and South Asia (Narasimhan et al. 2019). 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, Jarawa and Onge, from Andaman (AND) archipelago (Basu et al. 2016). Great Andamanese is considered as the sixth language family of India, being a linguistic isolate, typologically and genealogically different from other AND languages (Abbi, 2009). 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 (Bradburd et al. 2013) or did not provide any statistical significance or feature selection to identify the most relevant structure-related factors (Schlebusch et al. 2012). 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.

NEW APPROACHES

In order to understand the genetic substructure of India, considering the strongly endogamous social structure as well as the presence of multiple language families and their geographical distribution, we developed COGG (Correlation Optimization of Genetics and Geodemographics). COGG is a deterministic algorithm that may be used to simultaneously correlate genome-wide genotypes, with multiple factors that may have acted to shape population genetic substructure. In the context of this study, we correlate genetic structure as depicted by the top two Principal Components (PCs) with geography (longitude and latitude) and sociolinguistic factors (social and language group information in this case) as shown in Equation 1. We encoded four language groups AA, DR, IE and TB as well as the social group information as indicator variables i,e. if a sample belongs to a social or language group we use 1 and 0 otherwise. We refrain from using terms that could be considered socially stigmatizing and instead refer to Social Group A (SGA) for forward castes and Social Group B (SGB) for backward castes, respectively. For the semi-nomadic tribes in India, we assign Social Group C (SGC) (see more details in Supplementary notes).

Given information on m samples, the objective of COGG is to maximize the correlation between $\mathbf{u}$, the genetic component as represented by either of the top two PCs of the genetic covariance matrix formed by the genotype data and a geodemographic matrix $\mathbf{G}\in\mathbb{R}^{m\times k}$ where k is the number of demographic features.

$$\begin{array}{cc} & \begin{array}{ccccccccc} G_1 & G_2 & G_3 & G_4 & G_5 & G_6 & G_7 & G_8 & G_9 \end{array} \\ \mathbf{G}= & \left[\begin{array}{c|c|c|c|c|c|c|c|c} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Latitude & Longitude & SGA & SGB & SGC & AA & DR & IE & TB \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{array}\right] \end{array} \tag{1}$$

Therefore, COGG solves the following optimization problem,

$$\max_{\mathbf{a}} \mathbf{Corr}\left(\mathbf{u},\sum_{i=1}^{k} a_i\mathbf{G}_i\right) \tag{2}$$

where $\mathbf{a}$ be the k-dimensional vector whose elements are $a_1,\cdots,a_k$ ($k=9$ in this case). Recall that $\mathbf{G}_i$ denotes the i-th column vector of $\mathbf{G}$. Let $d_i=\mathbf{u}^T\mathbf{G}_i/\sqrt{\mathbf{Var}[\mathbf{u}]}$ for $i=1\ldots k$ 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 $i,j=1\ldots k$ 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 conducting 1,000 permutation tests on the sociolinguistic variables in **G**. On top of COGG we used a greedy feature selection algorithm to select the most significant factors which influence genetic variation in India.

To further study the interplay between these factors, we propose a simple analytic procedure using the so-called Ridge Leverage Score (RLS) statistic that highlights the significant populations capturing genetic diversity in India. The RLS of the i-th row of any matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is defined as

$$\tau_i{}^{\lambda}(\mathbf{A}) = \left(\mathbf{A}\mathbf{A}^\top \left(\mathbf{A}\mathbf{A}^\top + \lambda \mathbf{I}_n \right)^{-1} \right)_{ii}, \qquad (3)$$

where $\lambda > 0$ is the regularization parameter.

Starting from the mean-centered (subtracting each column by its respective mean) genotype matrix $\mathbf{Z} \in \mathbb{R}^{m \times n}$ where n is the number of markers for each of m samples and **G** as described above, we compute population level RLS (median RLS of the samples in the population) for each matrix (details in **Materials and Methods** and **Supplementary Note**). Thereafter, we compute an additive RLS statistic for each population highlighting the ethnic groups which represent and capture the greatest portion of observed genetic diversity across India. Our analysis aims to better understand the intricate details of admixture, substructure, and genetic variation across social and language groups in the Indian subcontinent. The need for methods such as COGG has been previously underlined by many studies (Bamshad et al. 2001; Roychoudhury et al. 2001; Basu et al. 2003; Majumder 2010; Basu et al. 2016). The ability to correlate genomic background with geographic, sociolinguistic and cultural differences opens new avenues to study genomic structure of extant human populations.

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 S1**). We initially compiled a pan-Indian dataset of 891 individuals across 90 populations (**Supplementary Table S1A** and **Supplementary Figure S1A**) and 47,283 SNPs from various sources (Reich et al. 2009; Chaubey et al. 2011; Metspalu et al. 2011; Moorjani et al. 2013; Basu et al. 2016). This dataset 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 S1B** and **Figure 1A**) in which four language families AA, DR, IE, and TB are represented (**Supplementary Note**) and used it for 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 dataset. Furthermore, in order to interrogate the shared ancestry between Indian sociolinguistic groups and Eurasia, we merged the normalized subset with 1,323 individuals from 50 populations and 42,975 SNPs across Eurasia (**Supplementary Table S1C**). For the outgroup f_3 analysis we present later in this section, we used 124 samples of Yorubans in Nigeria (YRI) from the 1000 Genomes phase 3 dataset (Auton et al. 2015) and merged it with the Eurasian dataset.

Geography versus population structure within India

Studies of populations in different parts of the world have shown that when top two 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 (Lao et al. 2006; Rosenberg et al. 2006; Chen et al. 2009; Paschou et al. 2010). For instance, Novembre and Stephens (2008) showed that within Europe, the Pearson correlation coefficient (r^2) (hereafter r^2) between PC1 vs. latitude (North-South) is equal to 0.77 and 0.78 for PC2 vs. longitude (East-West). In order to explore whether Indian genetic information mirrors geography, we computed Principal Component Analysis (PCA) on the normalized dataset of 33 Indian populations and plotted the top two PCs (**Figure 1B and C**, **Supplementary Figure S1B** for language, sociolinguistic and geographical groupings, respectively). The first three PCs explained 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 on PC2 (**Figure 1B**). Next, we computed r^2 between the top two PCs 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 (Reich et al. 2009). We note that IE and DR speakers also share significant ancestry among SGA and SGB groups as indicated by the result of ADMIXTURE analysis (**Supplementary Figure S3**). Interestingly, we observe clusters of sociolinguistic groups which become more prominent in the second and third PCs (**Supplementary Figure S4**) with the

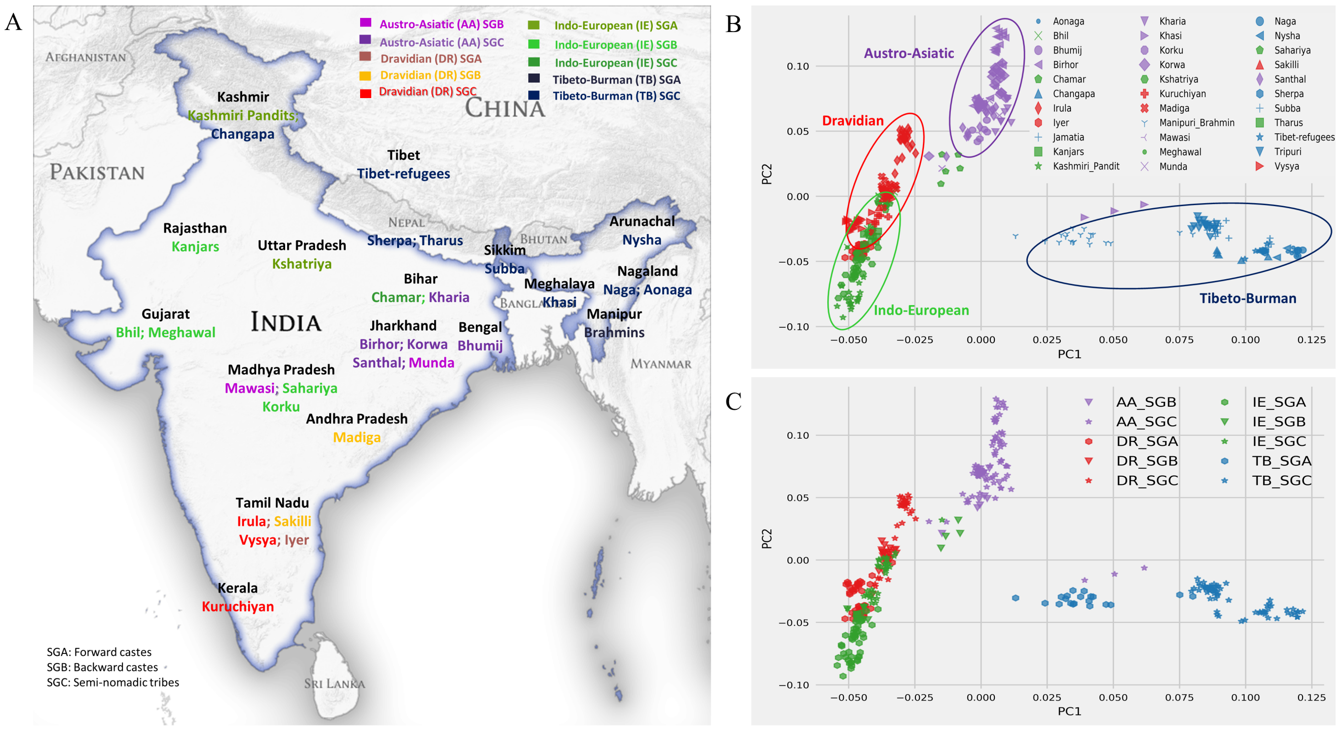

FIG. 1. A map of locations of the 33 populations in the normalized set and the results of principal component analysis. **A**. Map of India showing the locations of the 368 individuals in the normalized subset across 33 well-defined populations, 47,283 SNPs (see Supplementary Figure S1A for the pan-Indian dataset of 90 ethnic groups and Supplementary Figure S2 for the corresponding PCA plot). The populations are colored by their sociolinguistic group. **B**. Top two PCs of the normalized dataset show clustering by language groups. **C**. PCA plot colored and marked by sociolinguistic groups shows the genetic structure stratified by sociolinguistic groups.

SGCs distinguished from SGA and SGB within their language group.

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 dataset with 33 groups. These findings are in sharp contrast with findings within the European continent (Novembre and Stephens 2008; Drineas et al. 2010) and highlight the need for social and linguistic factors to be accounted for, as noted in prior work (Bamshad et al. 2001; Roychoudhury et al. 2001; Brahmachari et al. 2005; Majumder 2010; Basu et al. 2016). We performed Linear Discriminant Analysis (LDA) (**Supplementary Figures S5**) 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 dataset with the language groups set as classes (**Supplementary Figures S5A**) followed by the geographic regions (**Supplementary Figure S5B**). 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 almost perfect correlation with PC1 and PC2 representing the genetic structure of the Indian subcontinent using the geodemographic matrix $\mathbf{G}$ instead of just longitude and latitude: r^2 increases from 0.6 to 0.93 ($p<10^{-22}$) for PC1 vs. $\mathbf{G}$ and from 0.06 to 0.85 ($p<10^{-15}$) for PC2 vs. $\mathbf{G}$.

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 (Cann et al. 2002; Novembre and Stephens 2008; Auton et al. 2015). Our results are statistically significant (**Supplementary Figure S6**) 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. We note that these eight PCs capture, collectively, 89% of the variance of the genetic matrix.

Running COGG-CCA on these inputs returns a statistically significant (**Supplementary Figure S7**) r^2 equal to 0.94 ($p<10^{-16}$) which is well above the $r^2=0.74$ obtained when COGG-CCA is run without including the sociolinguistic factors (See **Supplementary Note** for details).

Identifying the features that drive population structure within India

In order to formally investigate which of the nine features in the geodemographic matrix $\mathbf{G}$ contribute more in the optimization problem posed by COGG (Equation 2), we used the sparse approximation framework and the Orthogonal

Matching Pursuit (OMP) algorithm from applied mathematics (Natarajan 1995) (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:

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

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.

Ethnic groups capturing genetic diversity across India

We developed a simple approach based on the Ridge Leverage Score (RLS) statistic (Alaoui and Mahoney 2015) (**Materials and Methods**) 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 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 AND represent distinct ethnic groups and outliers with respect to mainland Indian populations (**Supplementary Figure S2B**). Great Andamanese are also linguistically divergent from Jarawa (Abbi 2009); (2) Vysyas, who underwent a founder event going back 100 generations, due to the strong imposition of endogamy (Reich et al. 2009); (3) Language isolates Vedda from Sri Lanka (Chaubey 2014); (4) Minicoy from Lakshadweep archipelago with strong founder effects and diverse mixture due to the archipelago being a popular destination

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 (Moorjani et al. 2013)).

Population	State/Territory	Language family	Social group
Great Andamenese	Andaman and Nicobar islands	Great Andamanese	SGC
Minicoy	Lakshadweep 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

for maritime sailors (Samuel et al. 2009); (5) AA speaking Mundas who have Ancestral North and South Indian ancestry and an Ancestral Southeast Asian component (Tätte et al. 2019); (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 the language families and (7) TB speaking Changpas, who are semi-nomadic pastoralists 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 AND isolates (Thangaraj et al. 2003; Mondal et al. 2016) in order to highlight the cryptic relatedness among ethnic groups that COGG posits.

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 (Paschou et al. 2014) based on weights that reflect shared ancestry (**Supplementary Table S2**) as computed by meta-analysis of ADMIXTURE results (Alexander et al. 2009) (see **Materials and Methods** and **Supplementary Note** for details). The shared ancestry network, revealed four major clusters (ie 1. IE & DR, 2. AA, 3. TB and 4. AND) and a few exceptions as outlined in detail below.

IE and DR populations across social groups

A cluster of IE and DR speakers across social groups resembling a nearly complete graph with over 60% of all possible edges was observed (**Figure 2**). This was further supported by a similar pattern of strong shared ancestry in outgroup f_3 statistics (Patterson et al. 2012) using YRI from the 1000 Genomes dataset as the outgroup (Auton et al. 2015) as well as in f_3 tests for signs of admixture. We find that most IE and DR populations share more alleles with each other (**Supplementary Figure S8**) and are admixed with each other (**Supplementary Table 3**). IE speakers share above 70% average ancestry with DR_SGA and DR_SGB (**Supplementary Figure S3B**) in the meta-analysis of ADMIXTURE. This supports the notion that there was mixture between IE and DR speakers across SGA and SGB around 1,900 to 4,200 years ago (Moorjani et al. 2013) and that the caste system originated in a "classless" semi-nomadic society, which became hierarchical with the knowledge of agriculture (Kosambi 1964; Majumder 2001). Furthermore, it provides a possible explanation for DR loanwords appearing in early Hindu texts which are not found in IE languages outside the Indian subcontinent (Mallory and Adams 1997; Witzel 2001; Moorjani et al. 2013). The high relatedness between SGA and SGC across IE and DR speakers barring a few exceptions (**Supplementary Figure S9**), also provides genetic evidence to the claim that although the caste system was formally defined and observed to be stringent, it was broken in some cases, allowing mixture between SGC and SGA (Thapar 2014).

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).

Clique of TB speakers

TB speakers from North East India form a strongly connected cluster with the Khasis (AA speakers residing in North East India) who also clustered together with TB speakers in the scatter plot of the top two PCs (**Figure 1B**). The cluster also contain Manipuri Brahmins (TB_SGA), who are known to have significant admixture from IE_SGA (see **Supplementary Table S3**) and Tharus (IE_SGC) (Chaubey et al. 2014) from Tarai region in Nepal and eastern India.

Isolated AND groups

The AND groups Jarawa and Onge diverge from the rest of the Indian populations. This has also been shown in (Thangaraj et al. 2003; Reich et al. 2009; Basu et al. 2016; Mondal et al. 2016). They belong to the Ongan language family which has a debatable connection with Austronesian languages (Blevins 2007), showing divergence from all language families in mainland India.

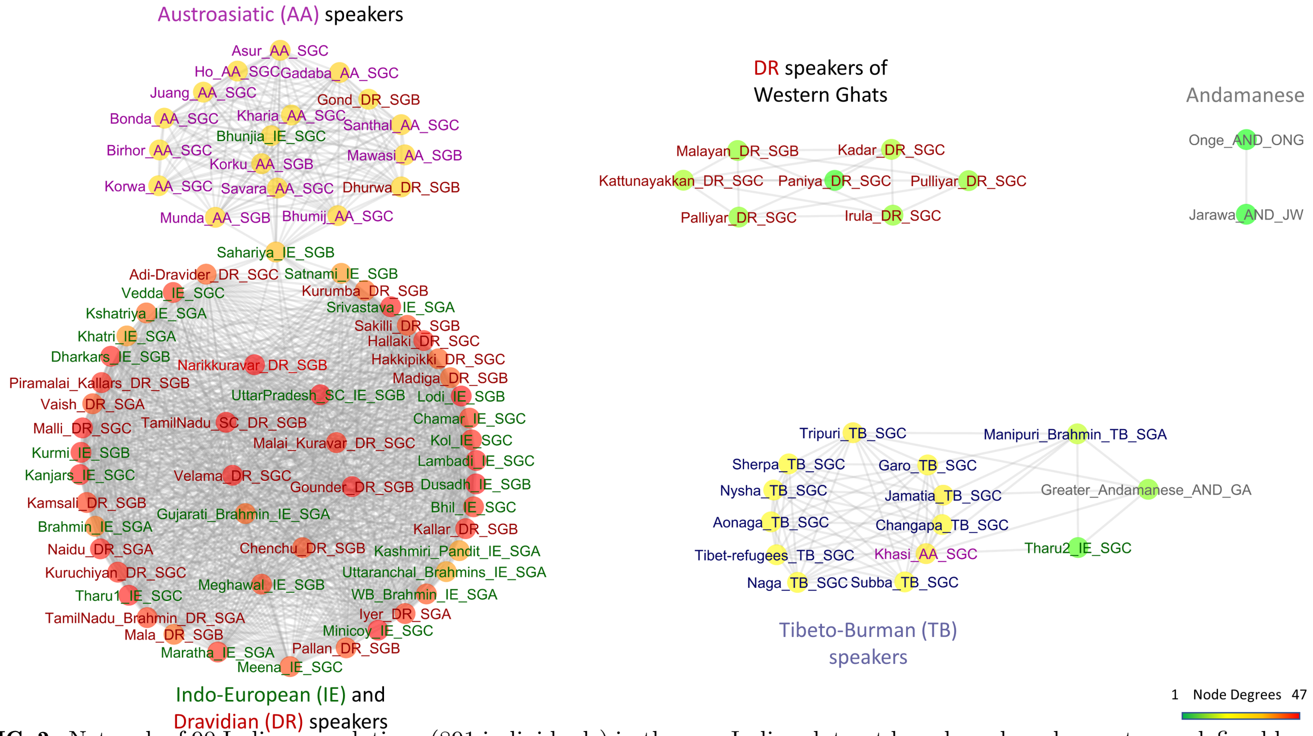

FIG. 2. Network of 90 Indian populations (891 individuals) in the pan-Indian dataset based on shared ancestry as defined by meta-analysis of ADMIXTURE results. Only the top 40% of edges (most related) populations are shown here (see Materials and Methods for details). The node labels are colored by their corresponding language groups as shown in Figure 1.

Populations outside major clusters

Above, we describe four major clusters each capturing the majority of individuals from different language groups: 1. The IE & DR cluster with 81% of IE and 69% of DR, 2. The AA cluster, capturing 93% of AA, 3. TB cluster with 73% of TB, and 4. a main AND cluster with 66% of AND populations. However, in each case, we also observed some exceptions revealing cryptic relatedness among ethnic groups which we outline here.

Few DR_SGC groups such as Kadar, Irula, Palliyar, and Paniya (which contain the lowest levels of Ancestral North Indian ancestry among Indian populations (Moorjani et al. 2013)) 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 S9**).

The Gonds and Sahariyas are candidate mosaic Indian populations, which is also reflected by their location as bridge nodes between the AA and IE-DR cliques. They contain high AA, DR and IE ancestry (**Supplementary Figures S8 and S9** and **Supplementary Table S2**), which can be attributed to their central location in India (Chaubey et al. 2017) and their long history of exogamy.

We also found the Great Andamenese to be connected to TB speakers of North East India, rather than other AND populations. They share approximately 50% shared ancestry

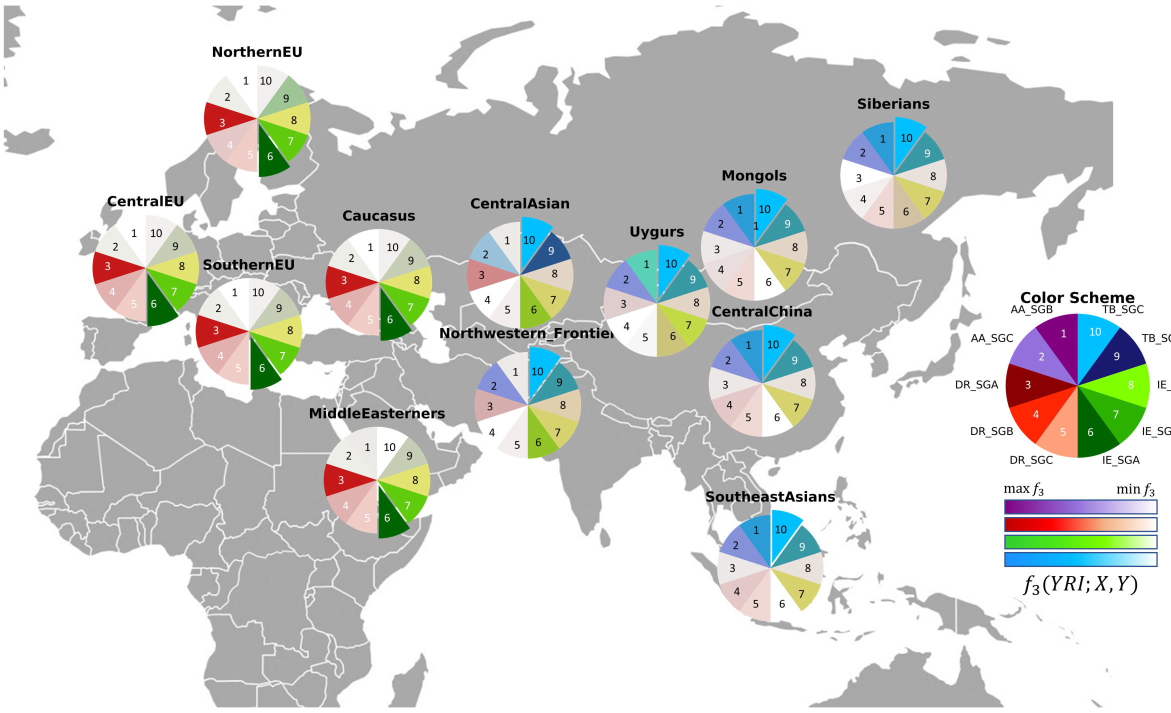

FIG. 3. Shared genetic drift between 33 Indian populations (denoted by X) and 50 Eurasian/East Asian populations (denoted by Y) as estimated by f_3 statistics with Yoruba as an outgroup f_3(YRI;X,Y). The darkest colors correspond to greatest portions of shared genetic drift with Indian populations. Full results can be found in Supplementary Table S4.

(**Supplementary Table S2**) as well as showing strong shared genetic drift with respect to outgroup f_3 statistics (**Supplementary Figure S9**). The Great Andamanese are known to be genetically divergent from other AND groups Jarawa and Onge (Thangaraj et al. 2003; Abbi 2009). To the best of our knowledge, this is the first observed interaction of the group to the rest of mainland Indian speakers based on autosomal markers and should be interpreted with caution due to small samples sizes of all groups involved. However, a study focused on the mitochondrial haplogroup M31 showed that with the exception of M31a1 (specific to AND), lineages M31a2, M31b and M31c are prevalent in North East India and surrounding regions (Wang et al. 2011). 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) (Voris 2000; Clark et al. 2009).

The mosaic of Indian sociolinguistics in the context of Eurasia

Indian populations from diverse sociolinguistic groups have different genetic affinities towards Eurasian populations. Outgroup f_3 statistics between the sociolinguistic groups and European populations with YRI as outgroup, reveal greater shared genetic drift between IE speakers (across social groups) and DR_SGA with European and Middle Eastern populations (**Supplementary Table S2**).

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 (Tätte et al. 2019). Our results clearly show two paths with a gradient of decreasing shared genetic drift from India and Eurasia: one from North East India towards China, Mongolia and Siberia and the other from North West India towards Central Asia, Uygurs, Middle Easterners and Europeans (**Figure 3**). This is concordant with our findings from network analysis with respect to connections with possible gateways to and from the Indian subcontinent (**Supplementary Figure S10**).

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 correlation 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 highlighting the need for population-specific studies.

We find evidence of wide mixture across all the social groups (tribal and non-tribal) for IE speakers and across SGA and SGB for DR speakers. We also provide further support for broad admixture and a long contact between IE and DR speakers in India. Our analysis also identifies finer substructure and population relationships within Indian sociolinguistic groups as well as their relatedness with various Eurasian populations. Interestingly, we find stronger shared ancestry between the Great Andamenese with TB speakers of North East India than other mainland speakers, a relationship which is observed for the first time using autosomal markers.

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

MATERIALS AND METHODS

Study design and datasets

We used PLINK 1.9 (Chang et al. 2015) to assemble genome-wide data for 891 samples from 90 well-defined sociolinguistic groups

(**Figure 1A**; **Supplementary Table S1**) genotyped on 47,283 autosomal SNPs. These samples were collected from various sources (Reich et al. 2009; Chaubey et al. 2011; Metspalu et al. 2011; Moorjani et al. 2013; Basu et al. 2016) 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 (Hinrichs 2006) before merging the data. Further quality control such as filtering out variants with missing call rates $>5\%$ and minor allele frequency (MAF) <0.05 was performed in PLINK (Purcell et al. 2007; Chang et al. 2015).

We merged 1,323 individuals across 50 populations from Eurasia and Southeast Asia, collected from various publicly available sources such as HGDP (Cann et al. 2002), the Estonian Biocenter (Behar et al. 2010; Yunusbayev et al. 2012; Di Cristofaro et al. 2013; Fedorova et al. 2013; Kovacevic et al. 2014; Raghavan et al. 2014; Yunusbayev et al. 2015) and the Allele Frequency Database (ALFRED) (Rajeevan et al. 2003) (**Supplementary Table S1C**) with our normalized Indian dataset to create a merged data set of 1,691 samples from 83 populations genotyped on 42,975 SNPs overlapping between all data sets.

PCA and LDA

We used TeraPCA (Bose et al. 2019) to perform PCA on our datasets after pruning for LD structure by setting `--indep-pairwise 50 10 0.4` in PLINK 1.9. We checked for outliers (using EIGENSTRAT's (Price et al. 2006) outlier detection method) in the PCA plot (**Supplementary Figure S2A**) 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**).

Mantel Tests

We computed pairwise F_{ST} distances between 33 Indian populations in the normalized dataset using PLINK 1.9. Thereafter, we computed the correlation between the F_{ST} and the distance matrix based on the geodemographic variables using the Mantel test function in Python's scikit-bio package. We performed 10,000 permutations and estimated Spearman's correlation, acknowledging the caveat of overestimation of p-values obtained from the tests (Guillot and Rousset 2013).

COGG and feature selection using Orthogonal Matching Pursuit

Aimed to model genetic structure within India, COGG maximizes the correlation between the top two PCs (for more PCs see CCA section in **Supplementary Note**) and the geodemographic matrix which consists of nine variables (columns) corresponding to geographical coordinates (latitude and longitude), social groups and language information encoded as indicator variables. COGG is explained in detail in **New Approaches** and **Supplementary Note**.

On top of COGG, we used a greedy feature selection algorithm described in (Natarajan 1995) to select features of the geodemographic matrix $\mathbf{G}$. We obtain two sets, S_1 and S_2 of the three most significant features from $\mathbf{G}$, for PC1 and PC2, respectively. 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 have removed the required number of features from $\mathbf{G}$ (details in **Supplementary Note**).

All the values returned by this method are statistically significant. When COGG was run with random permutations of the elements of S_1 and S_2, it returned negligible r^2. We also considered all $\binom{9}{3}$ combinations of three feature sets and concluded that, out of all possible sets, only S_1 and S_2 return maximum correlation with 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 bias, 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 (**Supplementary Note**) 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 populations 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 geodemographic matrices $\mathbf{Z}$ and $\mathbf{G}$. 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 (Alexander et al. 2009) for all admixture analyses. Prior to running ADMIXTURE, we pruned for LD using PLINK 1.9 by setting `--indep-pairwise 50 10 0.8`. We used eight fold Cross-Validation (CV) 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 and selected the one with the lowest CV error.

We also performed a quantitative analysis (**Supplementary Note**) of ADMIXTURE's output as shown in (Stamatoyannopoulos et al. 2017). 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 compute the shared ancestry values for each K, by varying it from four to eight and report the mean shared ancestry across these ancestral components. Furthermore, we designed a color-coding scheme for better visualization. The highest and lowest shared ancestry correspond to black and white respectively, and all intermediate values follow a gradient from black to white.

Three population statistics

f_3 tests are conducted for checking whether a target population (Z) is admixed between two source populations (X and Y) or to measure the shared drift between two test populations (X and Y) from an outgroup (Z).

$$f_3(X,Y;Z) = \mathbf{E}[(p_Z - p_X)(p_Z - p_Y)]$$

where p_i is the allele frequency for a given site in population i (see (Patterson et al. 2012; Peter 2016) for a detailed exposition on f_3 tests). We employ both these tests using ADMIXTOOLS (Patterson et al. 2012) to find signs of admixture and shared genetic drift within Indian populations as well as to find shared drift between Indian sociolinguistic groups and Eurasian populations using YRI as an outgroup.

We set the significance thresholds for z-score as $|Z| > 3$.

Network analysis

To better visualize and understand the connection between the populations included in our study, we performed a network analysis where the nodes represent each of 90 Indian populations and the edge weights correspond to the mean shared ancestry computed by meta-analysis results of ADMIXTURE (varying K from four to eight), as shown in a previous study (Paschou et al. 2014). 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 nearest neighbors (NN) sorted by decreasing edge weight (shared ancestry). Using this method with 3 NN, we obtained the top 40% of all edges for Figure 2.

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 at Molecular Biology and Evolution online..

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) and 378 samples from Moorjani et al. (2013). We also thank P. P. Majumder who allowed us to use the genotypic data from 367 samples from (Basu et al. 2016.

References

Abbi A. 2009. Is great Andamanese genealogically and typologically distinct from onge and jarawa? Lang Sci. 31:791–812.

Alaoui AE, Mahoney MW. 2015. Fast randomized kernel ridge regression with statistical guarantees. In: Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1. Cambridge, MA, USA: MIT Press, NIPS'15, pp. 775–783.

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

ArunKumar G, Soria-Hernanz DF, Kavitha VJ, Arun VS, Syama A, Ashokan KS, Gandhirajan KT, Vijayakumar K, Narayanan M, Jayalakshmi M, et al. 2012. Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural Expansions Predating the Caste System. PLoS One. 7:e50269.

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. Nature. 526:68–74.

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BVR, Reddy PG, Rasanayagam A, et al. 2001. Genetic evidence on the origins of Indian caste populations. Genome Res. 11:994–1004.

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, et al. 2003. Ethnic India: A genomic view, with special reference to peopling and structure. Genome Res. 13:2277–2290.

Basu A, Sarkar-Roy N, Majumder PP. 2016. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proc Natl Acad Sci USA. 113:1594–1599.

Behar DM, Yunusbayev B, Metspalu M, Metspalu E, Rosset S, Parik J, Rootsi S, Chaubey G, Kutuev I, Yudkovsky G, et al. 2010. The genome-wide structure of the Jewish people. Nature. 466:238–242.

Blevins J. 2007. A long lost sister of proto-austronesian?: Proto-ongan, mother of jarawa and onge of the andaman islands. Ocean Linguist. 46:154–198.

Bose A, Kalantzis V, Kontopoulou EM, Elkady M, Paschou P, Drineas P. 2019. Terapca: a fast and scalable software package to study genetic variation in tera-scale genotypes. Bioinformatics. 35:3679–3683.

Bradburd GS, Ralph PL, Coop GM. 2013. Disentangling the effects of geographic and ecological isolation on genetic differentiation. Evolution (N Y). 67:3258–3273.

Brahmachari SK, Singh L, Sharma A, Mukerji M, Ray K, Roychoudhury S, Chandak GR, Thangaraj K, Habib S, Parmar D, et al. 2005. The Indian Genome Variation database (IGVdb): A project overview. Hum Genet. 118:1–11.

Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, et al. 2002. A human genome diversity cell line panel. Science. 296:261–262.

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

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation plink: rising to the challenge of larger and richer datasets. GigaScience. 4:7.

Chaubey G. 2014. 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. 59:61–63.

Chaubey G, Metspalu M, Choi Y, Mägi R, Romero IG, Soares P, Van Oven M, Behar DM, Rootsi S, Hudjashov G, et al. 2011. Population genetic structure in Indian austroasiatic speakers: The role of landscape barriers and sex-specific admixture. Mol Biol Evol. 28:1013–1024.

Chaubey G, Singh M, Crivellaro F, Tamang R, Nandan A, Singh K, Sharma VK, Pathak AK, Shah AM, Sharma V, et al. 2014. Unravelling the distinct strains of Tharu ancestry. Eur J Hum Genet. 22:1404–1412.

Chaubey G, Tamang R, Pennarun E, Dubey P, Rai N, Upadhyay RK, Meena RP, Patel JR, van Driem G, Thangaraj K, et al. 2017. Reconstructing the population history of the largest tribe of India: the Dravidian speaking Gond. Eur J Hum Genet. 25:493–498.

Chen J, Zheng H, Bei JX, Sun L, Jia Wh, Li T, Zhang F, Seielstad M, Zeng YX, Zhang X, et al. 2009. Genetic structure of the han chinese population revealed by genome-wide snp variation. Am J Hum Genet. 85:775–785.

Clark PU, Dyke AS, Shakun JD, Carlson AE, Clark J, Wohlfarth B, Mitrovica JX, Hostetler SW, McCabe AM. 2009. The last glacial maximum. Science. 325:710–714.

Desai S, Dubey A. 2012. Caste in 21st century India: Competing narratives. Econ Polit Wkly. 46:40–49.

Di Cristofaro J, Pennarun E, Mazières S, Myres NM, Lin AA, Temori SA, Metspalu M, Metspalu E, Witzel M, King RJ, et al. 2013. Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. PLoS One. 8:e76748.

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

Fedorova SA, Reidla M, Metspalu E, Metspalu M, Rootsi S, Tambets K, Trofimova N, Zhadanov SI, Kashani BH, Olivieri A, et al. 2013. Autosomal and uniparental portraits of the native populations of Sakha (Yakutia): implications for the peopling of Northeast Eurasia. BMC Evo Biol. 13:1–18.

Guillot G, Rousset F. 2013. Dismantling the mantel tests. Methods Ecol Evol. 4:336–344.

Hinrichs AS. 2006. The UCSC Genome Browser Database: update 2006. Nucleic Acids Res. 34:D590–D598.

Kosambi D. 1964. The culture and civilisation of Ancient India in Historical Outline. New Delhi: Vikas Publishing House Pvt. Ltd.

Kovacevic L, Tambets K, Ilumäe AM, Kushniarevich A, Yunusbayev B, Solnik A, Bego T, Primorac D, Skaro V, Leskovac A, et al. 2014. Standing at the Gateway to Europe - The Genetic Structure of Western Balkan Populations Based on Autosomal and Haploid Markers. PLoS One. 9:e105090.

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

Majumder PP. 2001. Indian caste origins: Genomic insights and future outlook. Genome Res. 11:931–932.

Majumder PP. 2010. The Human Genetic History of South Asia. Curr Biol. 20:R184–R187.

Mallory J, Adams D. 1997. Encyclopedia of Indo-European Culture. London and Chicago: Fitzroy Dearborn.

Metspalu M, Romero IG, Yunusbayev B, Chaubey G, Mallick CB, Hudjashov G, Nelis M, Mägi R, Metspalu E, Remm M, et al. 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. Am J Hum Genet. 89:731–744.

Mondal M, Casals F, Xu T, Dall'Olio GM, Pybus M, Netea MG, Comas D, Laayouni H, Li Q, Majumder PP, et al. 2016. Genomic analysis of Andamanese provides insights into ancient human migration into asia and adaptation. Nat Genet. 48:1066.

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

Narasimhan VM, Patterson N, Moorjani P, Rohland N, Bernardos R, Mallick S, Lazaridis I, Nakatsuka N, Olalde I, Lipson M, et al. 2019. The formation of human populations in south and central asia. Science. 365:eaat7487.

Natarajan BK. 1995. Sparse approximate solutions to linear systems. SIAM J Comput. 24:227–234.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 40:646–649.

Olcott M. 1944. The caste system of India. Am Socio Rev. 9:648–657.

Paschou P, Drineas P, Yannaki E, Razou A, Kanaki K, Tsetsos F, Padmanabhuni SS, Michalodimitrakis M, Renda MC, Pavlovic S, et al. 2014. Maritime route of colonization of Europe. Proc Natl Acad Sci USA. 111:9211–9216.

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

Pathak AK, Kadian A, Kushniarevich A, Montinaro F, Mondal M, Ongaro L, Singh M, Kumar P, Rai N, Parik J, et al. 2018. The genetic ancestry of modern indus valley populations from northwest india. Am J Hum Genet. 103:918 – 929.

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

Peter BM. 2016. Admixture, population structure, and f-statistics. Genetics. 202:1485–1501.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38:904–909.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. Am J Hum Genet. 81:559–575.

Raghavan M, Skoglund P, Graf KE, Metspalu M, Albrechtsen A, Moltke I, Rasmussen S, Stafford Jr TW, Orlando L, Metspalu E, et al. 2014. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. Nature. 505:87–91.

Rajeevan H, Osier MV, Cheung KH, Deng H, Druskin L, Heinzen R, Kidd JR, Stein S, Pakstis AJ, Tosches NP, et al. 2003. ALFRED: The ALelle FREquency Database. Update. Nucleic Acids Research. 31:270–271.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature. 461:489–494.

Rosenberg NA, Mahajan S, Gonzalez-Quevedo C, Blum MGB, Nino-Rosales L, Ninis V, Das P, Hegde M, Molinari L, Zapata G, et al. 2006. Low levels of genetic divergence across geographically and linguistically diverse populations from India. PLoS Genet. 2:2052–2061.

Roychoudhury S, Roy S, Basu A, Banerjee R, Vishwanathan H, Usha Rani MV, Sil SK, Mitra M, Majumder PP. 2001. Genomic structures and population histories of linguistically distinct tribal groups of India. Hum Genet. 109:339–350.

Samuel PP, Krishnamoorthi R, Hamzakoya K, Aggarwal C, et al. 2009. Entomo-epidemiological investigations on chikungunya outbreak in the Lakshadweep islands, Indian ocean. Indian J Med Res. 129:442.

Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, Li S, De Jongh M, Singleton A, Blum MGB, et al. 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 338:374–379.

Silva M, Oliveira M, Vieira D, Brandão A, Rito T, Pereira JB, Fraser RM, Hudson B, Gandini F, Edwards C, et al. 2017. A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. BMC Evol Biol. 17:88.

Sokal RR. 1991. Ancient movement patterns determine modern genetic variances in Europe. Hum Biol. 84:553–554.

Stamatoyannopoulos G, Bose A, Teodosiadis A, Tsetsos F, Plantinga A, Psatha N, Zogas N, Yannaki E, Zalloua P, Kidd KK, et al. 2017. Genetics of the Peloponnesean populations and the theory of extinction of the medieval Peloponnesean Greeks. Eur J Hum Genet. 25:637–645.

Tätte K, Pagani L, Pathak AK, Kõks S, Duy BH, Ho XD, Sultana GNN, Sharif MI, Asaduzzaman M, Behar DM, et al. 2019. The genetic legacy of continental scale admixture in Indian Austroasiatic speakers. Sci Rep. 9:3818.

Thangaraj K, Singh L, Reddy AG, Rao VR, Sehgal SC, Underhill PA, Pierson M, Frame IG, Hagelberg E. 2003. Genetic affinities of the andaman islanders, a vanishing human population. Curr Biol. 13:86–93.

Thapar R. 1990. A history of India. Penguin Books, UK.

Thapar R. 2014. Can genetics help us understand Indian social history? Cold Spring Harb Perspect Biol. 6:a008599.

Vidyarthi LP, Rai BK. 1977. The tribal culture of India. New Delhi: Concept Publishing Company.

Voris HK. 2000. Maps of pleistocene sea levels in southeast asia: shorelines, river systems and time durations. J Biogeogr. 27:1153–1167.

Wang HW, Mitra B, Chaudhuri TK, gounder Palanichamy M, Kong QP, Zhang YP. 2011. Mitochondrial dna evidence supports northeast indian origin of the aboriginal Andamanese in the late Paleolithic. J Genet Genomics. 38:117–122.

Witzel M. 2001. Substrate languages in Old-Indo Aryan. IJDL. International Journal of Dravidian Linguistics. 30:1–94.

Wooding S, Ostler C, Prasad BVR, Watkins WS, Sung S, Bamshad M, Jorde LB. 2004. Directional migration in the hindu castes: inferences from mitochondrial, autosomal and y-chromosomal data. Hum Genet. 115:221–229.

Yunusbayev B, Metspalu M, Järve M, Kutuev I, Rootsi S, Metspalu E, Behar DM, Varendi K, Sahakyan H, Khusainova R, et al. 2012. The Caucasus as an asymmetric semipermeable barrier to ancient human migrations. Mol Biol Evol. 29:359–365.

Yunusbayev B, Metspalu M, Metspalu E, Valeev A, Litvinov S, Valiev R, Akhmetova V, Balanovska E, Balanovsky O, Turdikulova S, et al. 2015. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. PLoS Genet. 11:1–24.