# Dissecting Population Substructure in India via Correlation Optimization of Genetics and Geodemographics

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

[1]Computer Science Department, Purdue University, West Lafayette, IN 47907.
[2]Computational Biology Center, IBM TJ Watson Research Center, Yorktown Heights, NY 10598.
[3]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907.
[*]Corresponding authors: ppaschou@purdue.edu; pdrineas@purdue.edu.

India represents an intricate tapestry of population sub-structure shaped by geography, language, culture and social stratification operating in concert [1–3]. To date, no study has attempted to model and evaluate how these evolutionary forces have interacted to shape the patterns of genetic diversity within India. Geography has been shown to closely correlate with genetic structure in other parts of the world [4,5]. However, the strict endogamy imposed by the Indian caste system, and the large number of spoken languages add further levels of complexity. We merged all publicly available data from the Indian subcontinent into a dataset of 835 individuals across 48,373 SNPs from 84 well-defined groups [2, 6–9]. Bringing together geography, sociolinguistics and genetics, we developed COGG (Correlation Optimization of Genetics and Geodemographics) in order to build a model that optimally explains the observed population genetic sub-structure. We find that shared language rather than geography or social structure has been the most powerful force in creating paths of gene flow within India. Further investigating the origins of Indian substructure, we create population genetic networks across Eurasia. We observe two major corridors towards mainland India; one through the Northwestern and another through the Northeastern frontier with the Uygur population acting as a bridge across the two routes. Importantly, network, ADMIXTURE analysis and $f_3$ statistics support a far northern path connecting Europe to Siberia and gene flow from Siberia and Mongolia towards Central Asia and India.

The genetic structure of human populations reflects gene flow around and through geographic, linguistic, cultural, and social barriers. We set out to explore how the complex interplay of these factors may shape the patterns of genetic variation focusing on India, a country of intriguing levels of population structure complexity. The caste system in India has been documented since 1500-1000 BC and imposes strict rules of endogamy over the past several thousands of years. Social stratification within India may be summarised into the so-called Forward Castes and the Backward Castes [10], while 8.2% of the total population belongs to Scheduled Tribes and represents minorities that lie outside the caste system, still largely based on hunting, gathering and unorganized agriculture, with no written form of language [11]. Furthermore, there are 22 official languages within India, also following a distinctive geographic spread. The Dravidian (DR) speaking groups inhabit southern India, Indo-European (IE) speakers inhabit primarily northern India (but also parts of west and east India as well) and Tibeto-Burman (TB) speakers are mostly confined to northeastern India. The numerically small group of Austro-Asiatic (AA) speakers, who are exclusively tribal and are thought to be the original inhabitants of mainland India, inhabit fragmented geographical areas of eastern and central India. Previous studies have

uncovered four ancestral components within India [2,8,9], representing Northern India, Southern India, Austroasiatic speakers and Tibeto-Burman speakers. Furthermore, it has been shown that prior to the establishment of the caste system, there was wide admixture across tribes and castes in India which came to an abrupt end 1,900 to 4,200 years before present [8].

Starting from all publicly available data from the Indian subcontinent (835 individuals, see **Figure 1A** and **Supplementary Table 1**) and unlike prior studies [9,12], we created a normalized data set over castes, tribes, geographical locations, and language families that guarantees an approximately equal representation of endogamous populations, geographical locations, and language groups (a total of 368 individuals from 33 populations genotyped across 48,373 SNPs). In other regions of the world, it has often been observed that individuals from the same geographical region cluster together and it is often the case that the top two principal components are well-correlated with geography, namely longitude and latitude [13,14]. For instance, within Europe, the squared Pearson-correlation coefficient $r^2$ between the top singular vector of the genetic co-variance matrix vs. latitude (north-south) was 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 principal components using EIGENSTRAT [15] and plotted the top two left singular vectors of the resulting genetic covariance matrix (**Figure 1B**). It is straight-forward to observe that the IE and DR speaking populations form a long cline, while the AA and TB speakers form separate clusters. 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 and we observed $r^2 = 0.604$ for PC1 vs. longitude and $r^2 = 0.065$ for PC2 vs. latitude. Thus, PC1 recovers a relatively significant fraction of the longitude, but PC2 essentially entirely fails to recover the latitude. These findings are in sharp contrast with findings within the European continent [4,9,16]. ADMIXTURE analysis is consistent with previous studies, showing high degrees of shared ancestry across castes, but also across castes and tribes, thus supporting the notion that a demographic shift from wide admixture to endogamy occurred recently in Indian history (**Figure 2**, **Supplementary Figure 1**). Our meta-analysis of the ADMIXTURE output [17] shows that the IE and DR populations across castes shared very high ancestry, indicating the autochthonous origin of the caste system in India (**Figure 2**). $f_3$ statistics show that most of the castes and tribes in India are admixed, with contributions from other castes and/or tribes, across languages affiliations (**Supplementary Table 4** and **Supplementary Note**). The geographically isolated Tibeto-Burman tribes and the Dravidian speaking tribes appear to be the most isolated in India. Linear Discriminant Analysis on the normalized data set clearly supports genetic stratification by castes and languages in the Indian sub-continent (**Supplementary Figures 3A and 3B**).

In order to understand the genetic substructure of India, considering the strongly endogamous social structure as well as the presence of multiple language families, we developed COGG (Correlation Optimization of Genetics and Geodemographics). COGG is a novel method that correlates genomewide genotypes, as represented by the top two principal components, with geography (longitude and latitude) and sociolinguistic factors (caste and language information in this case). The need for such methods has been pointed out by many studies [3,9,18–26]. Given information on $m$ samples, the objective of COGG is to maximize the correlation between the genetic component as represented by the top singular vectors of the genetic covariance matrix formed by the genotypic data and a matrix containing information on geography, castes, tribes, and languages for each sample. More precisely, let $\mathbf{u}$ be the $m$-dimensional vector that represents either PC1 or PC2. Let $\mathbf{G}$ be the Geodemographic Matrix (an $m \times k$ matrix, where $k$ is the

number of geodemographic attributes that will be studied). Then, COGG seeks to maximize

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

In the above, $\mathbf{a}$ is an (unknown) $k$-dimensional vector whose elements are the $a_i$'s; we use $\mathbf{G}_i$ to denote the $i$-th column of the matrix $\mathbf{G}$ as a column vector. In our experiment, $\mathbf{G}$ has nine columns (i.e., $k = 9$): longitude and latitude are represented as numeric values, but caste/tribe/language information are encoded as zero-one indicator variables. We analytically solved the optimization problem of eqn. (1) to obtain a closed form solution for $\mathbf{a}_{\max}$ (see **Supplementary Note**). Plugging in the solution for $\mathbf{a}_{\max}$ in our data, we obtain a Pearson correlation coefficient $r^2 = 0.93$ for PC1 vs. $\mathbf{G}$ and $r^2 = 0.85$ for PC2 vs. $\mathbf{G}$. Thus, we are recovering almost all of the genetic structure of the Indian subcontinent using the Geodemographic matrix $\mathbf{G}$ instead of just longitude and latitude: the values of $r^2$ increase from 0.6 to 0.93 for PC1 and from 0.06 to 0.85 for PC2. This massive improvement came from considering endogamy and language families, two attributes that are pivotal in study the genetic stratification of Indian populations and is statistically significant (**Figure 3**).

In order to formally investigate which of the nine features (columns) in the geodemographic matrix $\mathbf{G}$ contribute more in the optimization problem of eqn. (1) we used the sparse approximation framework and the Orthogonal Matching Pursuit (OMP) algorithm from applied mathematics [27] (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:

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

Plugging in $S_1$ as the reduced feature space in COGG resulted in $r^2 = 0.92$ for PC1 vs. $S_1$ and $r^2 = 0.85$ for PC2 vs. $S_2$; these values are capturing approximately over 99% of the values returned by COGG when all the features in $G$ are included. Our feature selection approach for COGG explains the influence of sociolinguistics in shaping the genetic structure of the region, identifying membership to the AA or TB language group (which mostly consists of Backward Caste and Tribal groups), Forward Caste (who are usually found in IE and DR language groups), and latitude as the most significant geodemographic features that correlate to genetic structure within India, highlighting the language-caste interplay.

We proceeded to explore the structure of the Indian sub-continent in relation to the rest of Eurasia analysing a dataset of 1,332 individuals over 42,975 SNPs (**Supplementary Table 1**), sampled from 73 populations. Meta-analysis of the ADMIXTURE output reveals that, overall, Indian populations share a great proportion of ancestry with the so-called Indian NorthWestern Frontier populations, namely the tribal populations spanning Afghanistan and Pakistan (**Figure 4**). In concordance with previous studies we find higher degrees of shared ancestry of Central Asian populations with IE and DR Forward Castes [12, 20, 28]. IE Forward Castes also share large amounts of ancestry with other IE speaking populations (ie Europeans). However, IE and TB speakers as well as DR speaking Castes also share considerable amounts of ancestry with the Uygurs. On the other hand, AA speakers, who have been suggested as the earliest settlers of India [20, 29], appear more isolated. TB speakers share very high amounts of ancestry with populations from China but also Mongolia and Siberia.

PCA uncovers a structure that resembles a triangle, with Europeans residing in one corner, the Chinese on another corner and the Dravidian and Austro-asiatic speaking tribal populations of India occupying the third corner (**Figure 5A**). Siberians, Mongols and Uygurs stretch towards

India's Northwestern Frontier, while Tibeto-Burman speaking Indians connect India to China. We employed a population network analysis approach [30] in order to trace the gene-flow paths towards the Indian subcontinent (**Figure 5B**). Within India, IE, TB and AA Tribes are major nodes connecting to multiple populations. Tibeto-Burman Tribes stand at the Northeastern gateway from China to India, while IE Forward Castes are at the entry-point from the Northwestern frontier. Considering the whole of Eurasia, we observe three major paths leading to the two entry points of India: from Europe to Central Asia and the Indian Northwerstern Frontier, from Northern Europe to Siberia, and then Mongolia, then splitting towards China and Northeast India on one hand or the Uygurs, Central Asia and Northwestern India on the other hand. $f_3$ tests [31] (**Figure 6**) and TreeMix [32] analyses also support the notion that IE and TB Forward Castes have arisen through admixture of populations originating from the Caucasus and Mongolia (**Supplementary Table 3**, **Supplementary Figure 4**, and **Supplementary Note**). Previous studies have also supported a north-western and north-eastern corridor of migration towards India. However, this is the first study to connect the two paths through the populations of Siberia and Mongolia.

In summary, we present a novel method building a model that correlates geography, social, cultural and linguistic factors to genetic structure. The method is of independent interest and can be used to analyze any dataset of genotypic data where side information (e.g., geographic locations and/or other demographic information) for the samples is known. We are thus able to uncover the major forces that have shaped population genetic structure within India. Furthermore, through population genetic networks, ADMIXTURE analysis and $f_3$ tests, we have drawn paths of migration and gene flow throughout Eurasia, bringing out the importance of an ancient northern route moving from Europe through Siberia, Mongolia and merging back towards Central Asia and India. 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.

**Code Availability.** All code (including source files) is available at `https://github.com/aritra90/COGG`.

**Data Availability.** We have used publicly available data sets along with data reported by other studies. Our data sets will be made available upon request to the corresponding authors.

**Acknowledgements.** This study was supported by NSF IIS-1319280, NSF IIS-1661760, and IBM. Part of this work was done at IBM TJ Watson Research Center where AB was an intern. We thank D. Reich and P. Moorjani for sharing genotypic data of 248 samples from [2] and 378 samples from [8]. We also thank P. P. Majumder who allowed us to use the genotypic data from 367 samples from [9].

**Author Contributions.** A.B., P.D. and P.P. conceived and designed the project. A.B. gathered samples from various sources and performed the data analyses after discussing with D.E.P., P.D. and P.P. D.E.P. performed and wrote the LDA analysis. L.P. participated in and discussed analyses. A.B., P.D. and P.P wrote the manuscript.

# Online Methods

## Samples

We used PLINK [33, 34] to assemble genome-wide data for 839 samples from 87 well-defined sociolinguistic groups (see **Supplementary Table 1**) genotyped on a 48,225 SNPs. These samples were collected from various sources [2, 6–9] with the consent of the corresponding authors. We created and tested subsets of this dataset in order to construct an equal representation of castes, tribes, language families and geographical locations for this study. The normalized subset for which we have reported results for the Indian populations contains 368 samples from 33 populations genotyped for 48,326 SNPs.

We merged reference populations from Eurasia and Southeast Asia, collected from various publicly available sources such as HGDP [35], the Estonian Biocenter [36–42] and the Allele Frequency Database (ALFRED) [43] with our normalized Indian dataset to create a merged data set of 1,332 samples from 73 population groups genotyped on 42,975 SNPs (**Supplementary Table 1**).

## PCA and LDA

We used the smartPCA program of the EIGENSOFT package 6.1.4 [15] as well as our own MatLab implementation of PCA [44, 45]. We also implemented our own version of Linear Discriminant Analysis.

## 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 and the Geodemographic matrix, containing geographical coordinates, caste, tribe and language information encoded as indicator variables. We restrict our encoding into three castes: Forward castes, Backward castes and Tribal or nomadic hunter gatherers. $\mathbf{u}$ is the vector containing either one of the top two principal components, computed by EIGENSTRAT [15]; the Geodemographic matrix is denoted by $\mathbf{G}$. The caste (Forward, Backward and Tribals) and language (AA, DR, IE, TB) encoding was performed as follows:

$$\textbf{Castes (or Languages)}= \begin{cases} 1, & \text{if the sample belongs to that caste (or language)} \\ 0, & \text{otherwise} \end{cases}$$

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):

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

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

We used a greedy feature selection algorithm described in [27] to select features of the Geodemographic matrix $\mathbf{G}$. We obtain two sets of the three most significant 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 words, 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 features in $S_1$ and $S_2$ recover almost nothing. We also checked all $\binom{9}{3}$ possible sets of three features exhaustively and concluded that (for both PC1 and PC2) $S_1$ and $S_2$ return the maximum correlation.

## Estimating population admixture

We used the ADMIXTURE v1.22 software [46] for all admixture analyses and used our in house script to plot the admixture estimates. Before running ADMIXTURE, we pruned for LD using PLINK [33, 34] 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 $K$. We also performed a quantitative analysis of ADMIXTURE's output using a method described and implemented in [17]. To visualize the results of this quantitative analysis, we designed a color-coding scheme, 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, network analysis, and TreeMix

We used ADMIXTOOLS [31] to compute $f_3$ statistics for our data sets to find signs of admixture using the qp3Pop program. To better visualize and understand the connection between the populations included in our study, we performed a network analysis on the results of ADMIXTURE, using a method presented by a previous study [30]. Finally, TreeMix [32] was used to analyze the population divergence, mainly for the IE language dispersal into the Indian subcontinent. We used migration values from zero to eight to infer language dispersal routes.

# References

[1] Sahoo, S. et al. A prehistory of Indian Y chromosomes: Evaluating demic diffusion scenarios. *Proc. Natl. Acad. Sci.* **103**(4), 843–848 (2006).

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

[3] ArunKumar, G. et al. Population Differentiation of Southern Indian Male Lineages Correlates with Agricultural Expansions Predating the Caste System. *PLoS One* **7**(11) (2012).

[4] Novembre, J. et al. Genes mirror geography within Europe. *Nature* **456**(7218), 98–101 (2008).

[5] Wang, C. and Zöllner, S. & Rosenberg, N. A. A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations. *PLoS Genet.* **8**(8) (2012).

[6] Metspalu, M. et al. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. *Am. J. Hum. Genet.* **89**(6), 731–744 (2011).

[7] Chaubey, G. et al. Population genetic structure in indian austroasiatic speakers: The role of landscape barriers and sex-specific admixture. *Mol. Biol. Evol.* **28**(2), 1013–1024 (2011).

[8] Moorjani, P. et al. Genetic evidence for recent population mixture in India. *Am. J. Hum. Genet.* **93**(3), 422–438 (2013).

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

[10] Dubey, A. Caste in 21st Century India: Competing Narratives. *Economic and Political Weekly* **Vol. 46**(Issue No. 11) 12 (2011).

[11] Pati, R.N. & Dash, J. *Tribal and Indigenous People of India: Problems and Prospects.* A.P.H. Publishing Corporation, New Delhi, (2002).

[12] Sengupta, S. et al. Polarity and Temporality of High-Resolution Y-Chromosome Distributions in India Identify Both Indigenous and Exogenous Expansions and Reveal Minor Genetic Influence of Central Asian Pastoralists. *Am. J. Hum. Genet.* **78**(2), 202–221 (2006).

[13] Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nat. Genet.* **40**(5), 646–649 (2008).

[14] Chisholm, B., Cavalli-Sforza, L., Menozzi, P., and Piazza, A. The History and Geography of Human Genes. *J Asian Stud.* **54**(2), 490 (1995).

[15] Price, A. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**(8), 904–909 (2006).

[16] Drineas, P., Lewis, J., and Paschou, P. Inferring geographic coordinates of origin for Europeans using small panels of ancestry informative markers. *PLoS One* **5**(8) (2010).

[17] Stamatoyannopoulos, G. et al. Genetics of the peloponnesean populations and the theory of extinction of the medieval peloponnesean Greeks. *Eur. J. Hum. Genet.* **25**(5), 637–645 (2017).

[18] Bamshad, M. et al. Genetic evidence on the origins of Indian caste populations. *Genome Research* **11**(6), 994–1004 (2001).

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

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

[21] Basu, A. et al. Ethnic India: A genomic view, with special reference to peopling and structure. *Genome Res.* **13**(10), 2277–2290 (2003).

[22] Brahmachari, S. K. et al. The Indian Genome Variation database (IGVdb): A project overview. *Hum. Genet.* **118**(1), 1–11 (2005).

[23] Roychoudhury, S. et al. Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum. Genet.* **109**(3), 339–350 (2001).

[24] Rosenberg, N. A. et al. Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet.* **2**(12), 2052–2061 (2006).

[25] Tamang, R. & Thangaraj, K. Genomic view on the peopling of India. *Investig Genet* **3**(1), 20 (2012).

[26] Sirajuddin, S. M., Duggirala, R., and Crawford, M. H. Population structure of the Chenchu and other south Indian tribal groups: relationships between genetic, anthropometric, dermatoglyphic, geographic, and linguistic distances. *Hum Biol* **66**(5), 865–884 (1994).

[27] Natarajan, B. K. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing* **24**(2), 227–234 (1995).

[28] Silva, M. et al. A genetic chronology for the Indian Subcontinent points to heavily sex-biased dispersals. *BMC Evol. Biol.* **17**(1), 88 (2017).

[29] Kumar, V. & Reddy, B. Status of Austro-Asiatic groups in the peopling of India: An exploratory study based on the available prehistoric, linguistic and biological evidences. *J Biosci.* **28**(4), 507–522 (2003).

[30] Paschou, P. et al. Maritime route of colonization of Europe. *Proc. Natl. Acad. Sci.* **111**(25), 9211–9216 (2014).

[31] Patterson, N. et al. Ancient admixture in human history. *Genetics* **192**(3), 1065–1093 (2012).

[32] Pickrell, J. K. & Pritchard, J. K. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLoS Genet.* **8**(11), 1–17 11 (2012).

[33] Purcell, S. et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**(3), 559–575 (2007).

[34] Chang, C. C. et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**(1), 7 (2015).

[35] Cann, H. et al. A Human Genome Diversity Cell Line Panel. *Science* **296**(5566), 261–262 (2002).

[36] Di Cristofaro, J. et al. Afghan Hindu Kush: Where Eurasian Sub-Continent Gene Flows Converge. *PLoS One* **8**(10) (2013).

[37] Yunusbayev, B. et al. The caucasus as an asymmetric semipermeable barrier to ancient human migrations. *Mol. Biol. Evol.* **29**(1), 359–365 (2012).

[38] Kovacevic, L. et al. 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).

[39] Yunusbayev, B. et al. The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLoS Genet.* **11**(4), 1–24 (2015).

[40] Behar, D. et al. The genome-wide structure of the Jewish people. *Nature* **466**(7303), 238–242 Jul (2010).

[41] Fedorova, S. 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).

[42] Raghavan, M. et al. Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**(7481), 87–91 Jan (2014). Letter.

[43] Rajeevan, H. et al. ALFRED: The ALelle FREquency Database. Update. *Nucleic Acids Res.* **31**(1), 270–271 (2003).

[44] Paschou, P. et al. PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* **3**(9), 1672–1686 (2007).

[45] Paschou, P. et al. Intra- and interpopulation genotype reconstruction from tagging SNPs. *Genome Res.* **17**(1), 96–107 (2007).

[46] Alexander, D. H. and Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* (2009).

[47] Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34**(90001), D590–D598 (2006).

[48] Auton, A. et al. A global reference for human genetic variation. *Nature* **526**(7571), 68–74 (2015).

[49] Excoffier, L. and Smouse, P E & Quattro, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**(2), 479–491 (1992).

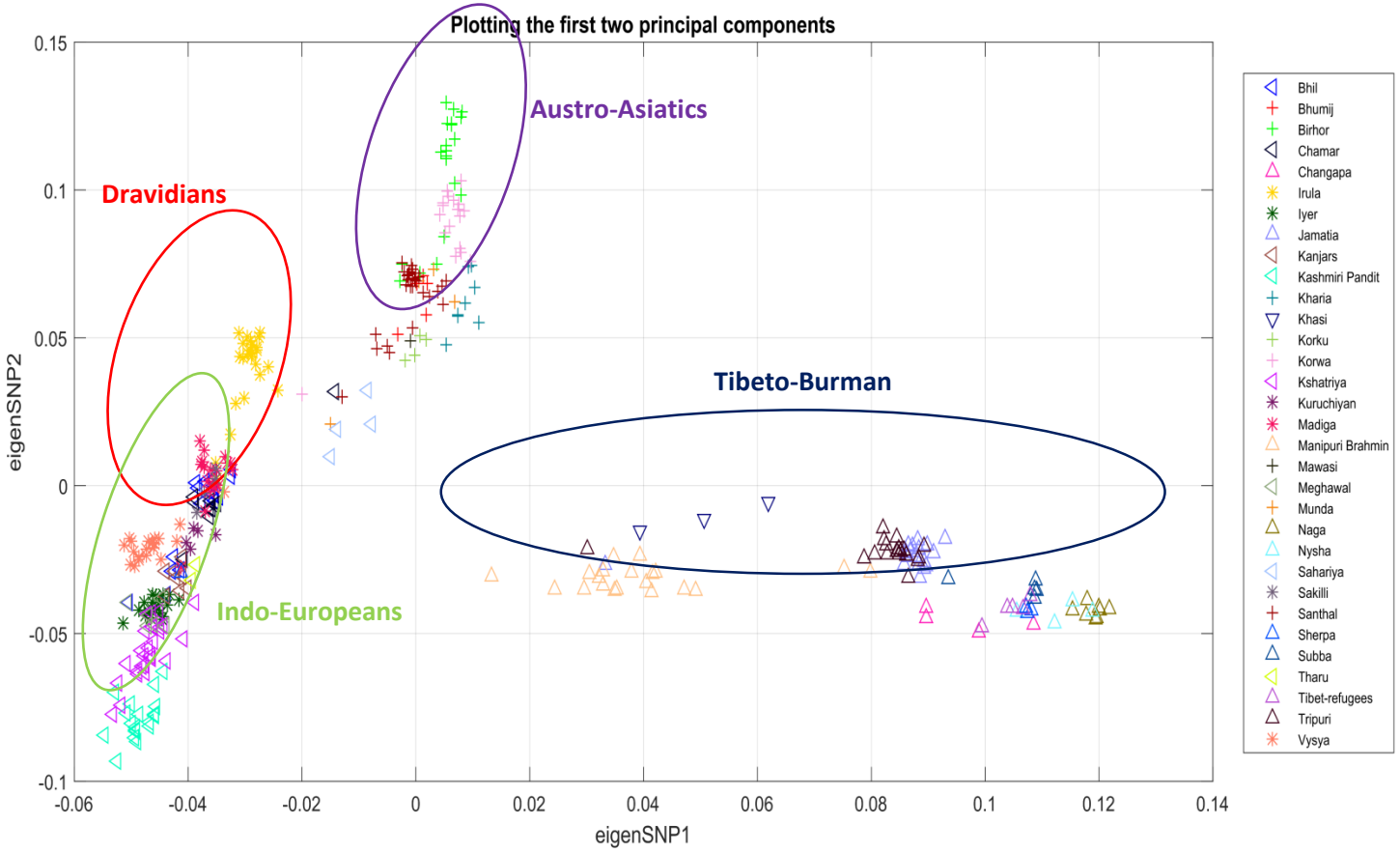
[50] Rao, C. R. The Utilization of Multiple Measurements in Problems of Biological Classification. *Journal of the Royal Statistical Society. Series B (Methodological)* **10**(2), 159–203 (1948).

[51] Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**(2), 179–188 (1936).

a


**Fig 1 (a)** Map showing the locations of the 835 Indian samples (from 84 well-defined population groups) that were used as the starting point in our study. **(b)** PCA plot of a normalized dataset consisting of 368 individuals (genotyped on 48,373 SNPs) that guarantees an approximately equal representation of endogamous populations, geographical locations, and language groups shown in 1A. Language groups are clearly quite important in the PCA plot and correlate well with the principal components.

b



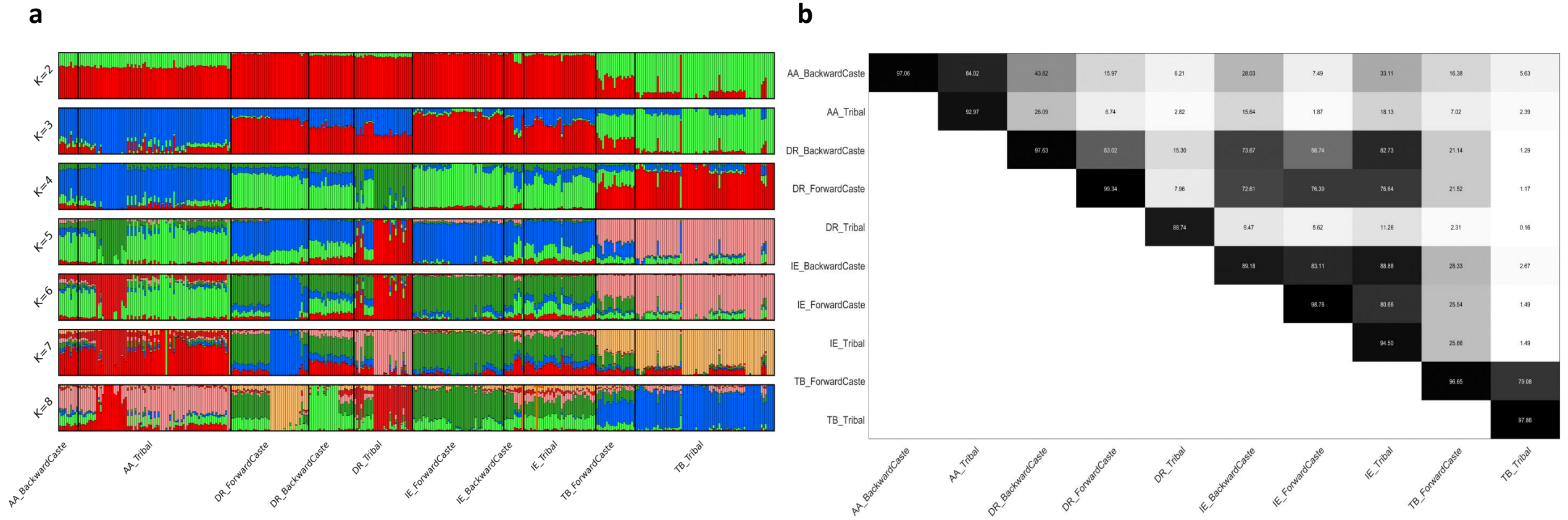



**Fig 2 (a)** An ADMIXTURE plot (for values of K between two and eight) of the normalized data set (368 individuals 48,373 SNPs) clearly shows the four main components related to language groups (Dravidian, Indo-European, Tibeto-Burman, and Austro-Asiatic); see, for example, the plot for K equal to five or six. The plot also shows the divergence of the Dravidian Tribal samples (DR_Tribal). **(b)** We perfomed a meta-analysis of the results of the ADMIXTURE plot (see Methods for details) to visually and numerically quantify the amount of shared ancestry (as revealed by ADMIXTURE) between any pair of populations. Darker colors indicate larger amounts of shared ancestry; we observe a higher amount of shared ancestry between the Indo-European and Dravidian populations, across all castes, indicating the existence of significant admixture between the two linguistic groups. The isolation of the Dravidian Tribal samples is primarily due to the isolation of hill tribes (such as Irula, Kadar, Paniyas, etc.)

**a**

$$\max_{\alpha} \textbf{Corr}\left( \overset{U}{\begin{bmatrix} \vdots \\ eigenSNP \\ \vdots \end{bmatrix}_{[n\times 1]}}, \sum_{i=1}^{9} \overset{\begin{matrix}\alpha_1\cdot G_1 & \alpha_2\cdot G_2 & \alpha_3\cdot G_3 & \alpha_4\cdot G_4 & \alpha_5\cdot G_5 & \alpha_6\cdot G_6 & \alpha_7\cdot G_7 & \alpha_8\cdot G_8 & \alpha_9\cdot G_9\end{matrix}}{\begin{bmatrix} \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Latitude & Longitude & Forward\ Castes & Backward\ Castes & Tribals & AA & DR & IE & TB \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}_{[n\times 9]}} \right)$$

Formally: $\max_{\alpha} \textbf{Corr}\left(U, \sum_{i=1}^{k} \alpha_i \cdot G_i\right)$

where $U \in \mathbb{R}^n$, is the vector corresponding to the eigenSNPs.
$G \in \mathbb{R}^{n\times k}$, is the Geodemographic matrix.
$\alpha = (\alpha_i)$ is the unknown vector of coefficients for each feature.

**b**

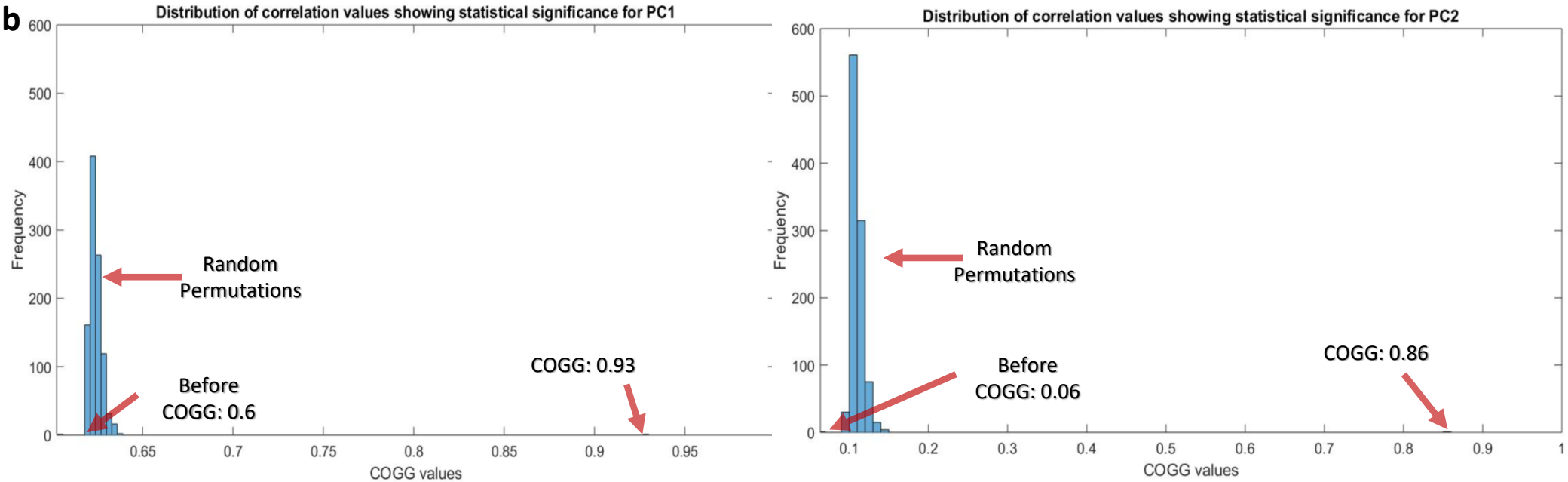
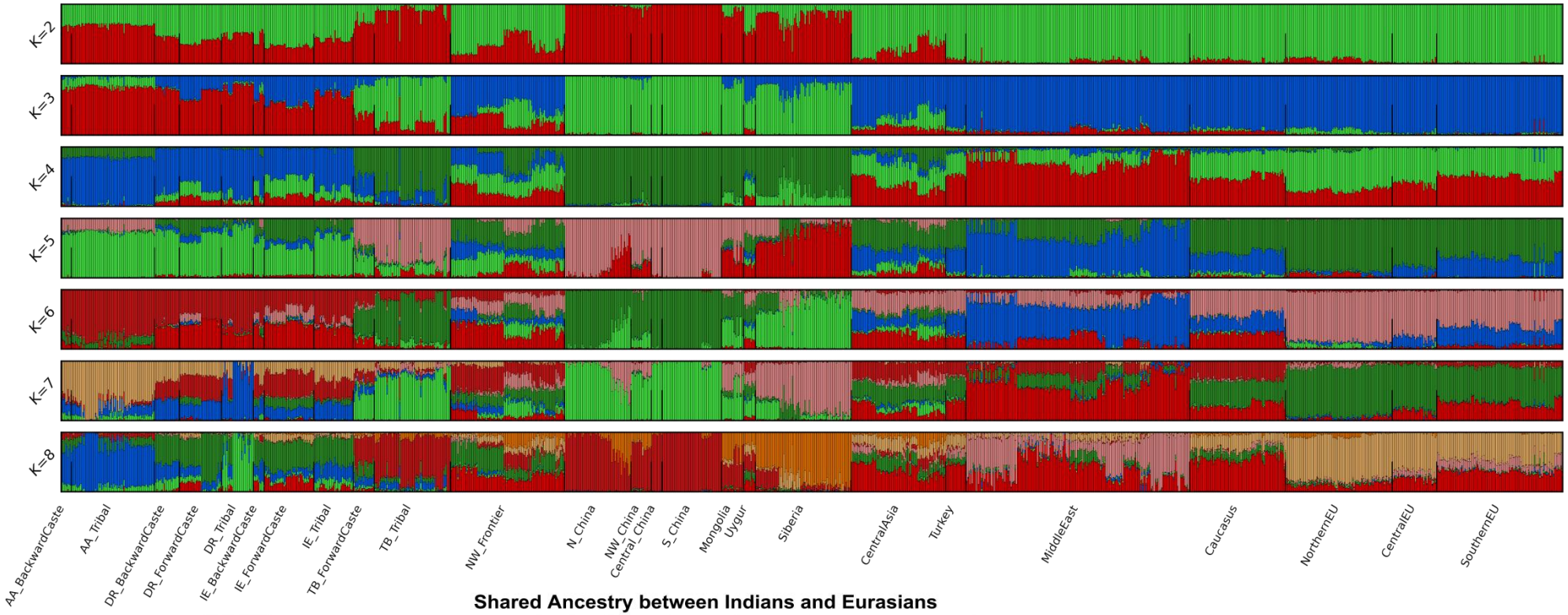



**Fig 3 (a)** Our approach for Correlation Optimization of Genetics and Geodemographics (COGG). The inputs to COGG are an eigenSNP (e.g., a singular vector from the covariance matrix based on the sample genotypes) and the geodemographic information (longitude, latitude, membership to a language group, caste membership, etc.). The output is the correlation between the eigenSNP and the geodemographics. Interestingly, a closed form solution exists for the COGG optimization problem. **(b)** Statistical significance of the COGG output (using random permutations). Clearly, COGG is statistically significant for both the first and the second principal components.

**a**




**b**

**Shared Ancestry between Indians and Eurasians**

| | NW_Frontier | N_China | NW_China | Central_China | S_China | Mongolia | Uygur | Siberia | CentralAsia | Turkey | MiddleEast | Caucasus | SouthernEU | CentralEU | NorthernEU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| AA_BackwardCaste | 19.44 | 4.79 | 4.97 | 4.88 | 4.83 | 3.82 | 8.81 | 1.31 | 5.42 | 1.37 | 0.83 | 1.26 | 0.31 | 0.21 | 0.27 |
| AA_Tribal | 15.12 | 3.54 | 3.67 | 3.63 | 3.60 | 2.76 | 6.35 | 1.04 | 3.60 | 0.68 | 0.42 | 0.61 | 0.14 | 0.11 | 0.15 |
| DR_ForwardCaste | 55.47 | 0.03 | 0.15 | 0.03 | 0.03 | 0.34 | 13.42 | 0.37 | 26.83 | 14.87 | 5.64 | 17.57 | 7.78 | 6.79 | 6.45 |
| DR_BackwardCaste | 39.26 | 0.07 | 0.18 | 0.07 | 0.07 | 0.28 | 8.31 | 0.30 | 14.53 | 6.19 | 2.71 | 7.20 | 2.13 | 1.72 | 1.73 |
| DR_Tribal | 22.58 | 0.12 | 0.22 | 0.12 | 0.12 | 0.27 | 4.79 | 0.29 | 7.39 | 2.99 | 1.56 | 3.30 | 1.19 | 0.97 | 0.99 |
| IE_ForwardCaste | 79.60 | 0.12 | 0.44 | 0.09 | 0.09 | 1.18 | 26.08 | 1.37 | 52.48 | 35.12 | 11.35 | 41.24 | 24.27 | 22.71 | 21.92 |
| IE_BackwardCaste | 54.11 | 1.97 | 2.29 | 1.96 | 1.94 | 2.25 | 17.45 | 1.18 | 28.45 | 16.47 | 6.11 | 18.90 | 9.48 | 8.37 | 8.03 |
| IE_Tribal | 49.66 | 0.14 | 0.30 | 0.13 | 0.13 | 0.50 | 12.58 | 0.49 | 22.96 | 11.94 | 4.14 | 14.09 | 6.24 | 5.65 | 5.66 |
| TB_ForwardCaste | 34.10 | 67.92 | 68.11 | 67.95 | 67.77 | 54.91 | 59.34 | 15.83 | 19.62 | 7.13 | 2.20 | 6.42 | 3.04 | 2.82 | 3.15 |
| TB_Tribal | 10.10 | 94.65 | 92.87 | 94.61 | 94.39 | 73.75 | 52.91 | 21.47 | 6.43 | 1.19 | 0.31 | 0.65 | 0.26 | 0.23 | 0.34 |

**Fig 4 (a)** An ADMIXTURE plot (for values of K between two and eight) of the Indian dataset merged with Eurasian populations (1,332 individuals, 42,973 SNPs). **(b)** Our meta-analysis of the ADMIXTURE plot in Figure 4A quantifies the ADMIXTURE results (darker colors indicate higher pairwise shared ancestry). Indian populations show a greater proportion of shared ancestry with the so-called Indian Northwestern Frontier populations, namely the tribal populations spanning Afghanistan and Pakistan. Central Asian populations share higher degrees of ancestry with IE and DR Froward castes. Uygurs share high degrees of ancestry with Indian populations.

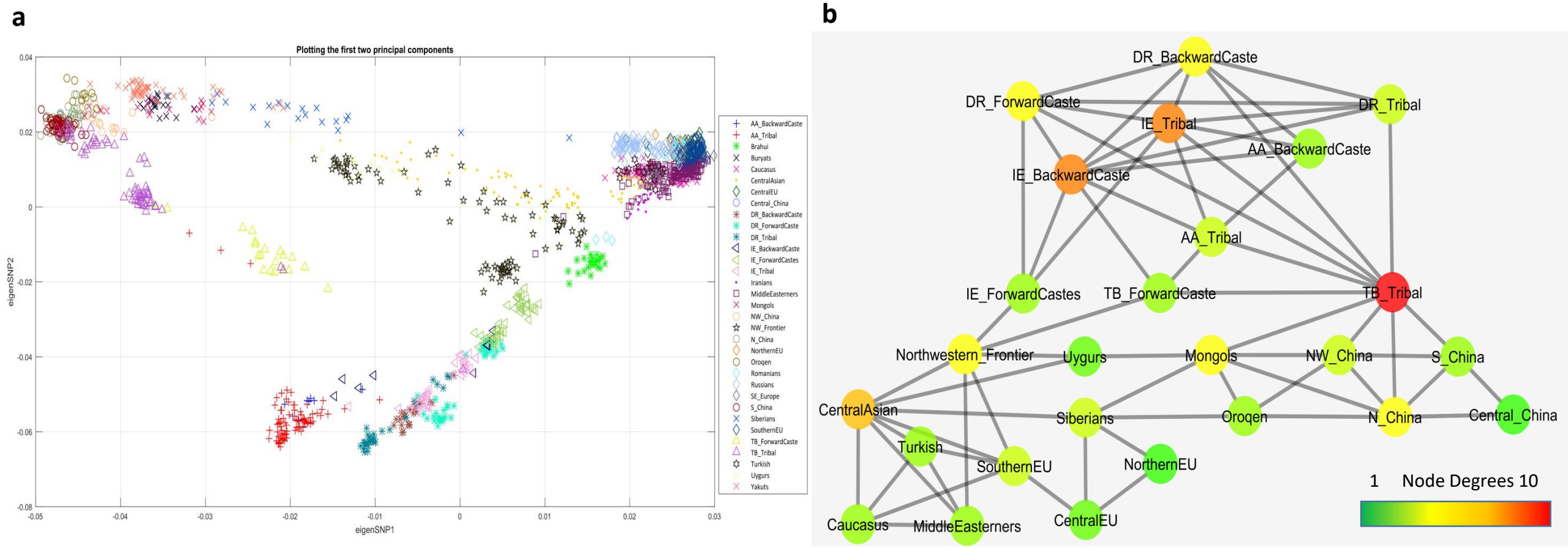



**Fig 5 (a)** PCA plot of Indian and Eurasian populations uncovers a structure that resembles a triangle, with Europeans residing in one corner, the Chinese on another corner and the Dravidian and Austro-asiatic speaking tribal populations of India occupying the third corner. The PCA plot does mirror the geography of Eurasia. **(b)** Population genetic networks formed using the top five principal components (see Methods for the network formation algorithm). We observe three major paths leading to the two entry points of India: from Europe to Central Asia and the Indian Northwestern Frontier, from Northern Europe to Siberia, and then Mongolia, then splitting towards China and Northeast India on one hand or the Uygurs, Central Asia and Northwestern India on the other hand.

**a**


**b**

| A | B | C | $f_3$(C; A,B) | Err | Z |
|---|---|---|---|---|---|
| Central_China | Uygurs | TB_Tribal | -0.00113 | 0.000268 | -4.209 |
| South_China | Uygurs | TB_Tribal | -0.00094 | 0.000212 | -4.424 |
| South_China | Selkups | TB_Tribal | -0.00085 | 0.000224 | -3.807 |
| Yakuts | SouthernEU | IE_ForwardCaste | -0.00146 | 0.000382 | -3.82 |
| Mongols | Caucasus | TB_ForwardCaste | -0.00264 | 0.000384 | -6.868 |
| Mongols | CentralEU | IE_ForwardCaste | -0.00101 | 0.000358 | -2.816 |
| Mongols | CentralEU | TB_ForwardCaste | -0.00228 | 0.000398 | -5.713 |
| Mongols | SouthernEU | IE_ForwardCaste | -0.0017 | 0.000361 | -4.705 |
| Mongols | SouthernEU | TB_ForwardCaste | -0.00266 | 0.000403 | -6.592 |
| Mongols | SE_Europe | IE_ForwardCaste | -0.00102 | 0.000359 | -2.839 |
| Mongols | SE_Europe | TB_ForwardCaste | -0.00234 | 0.000405 | -5.772 |
| Mongols | NorthernEU | TB_ForwardCaste | -0.00173 | 0.000407 | -4.244 |
| Buryats | SouthernEU | IE_ForwardCaste | -0.00157 | 0.000353 | -4.437 |
| Caucasus | Mongols | CentralAsian | -0.01138 | 0.000228 | -49.922 |
| Caucasus | Uygurs | CentralAsian | -0.0057 | 0.000175 | -32.559 |
| Caucasus | Yakuts | CentralAsian | -0.0114 | 0.000231 | -49.284 |
| Caucasus | Buryats | CentralAsian | -0.01133 | 0.000225 | -50.34 |
| Caucasus | Selkups | CentralAsian | -0.00793 | 0.000187 | -42.449 |
| Caucasus | North_China | CentralAsian | -0.01328 | 0.000257 | -51.724 |
| Caucasus | NW_China | CentralAsian | -0.01221 | 0.000245 | -49.881 |
| CentralEU | Yakuts | CentralAsian | -0.01241 | 0.000261 | -47.555 |
| CentralEU | Buryats | CentralAsian | -0.0125 | 0.000254 | -49.17 |
| CentralEU | Selkups | CentralAsian | -0.0067 | 0.000219 | -30.636 |
| CentralEU | Mongols | CentralAsian | -0.01252 | 0.000253 | -49.406 |
| CentralEU | N_China | CentralAsian | -0.01497 | 0.000273 | -54.857 |

**Fig 6 (a)** Routes of human migrations through Eurasia and towards India (based on PCA, network and ADMIXTURE analysis). The directions of the arrows are inferred from the f3 statistics shown in Fig 6B. **(b)** f3 statistics (all negative Z-scores are shown) indicate Chinese and Siberian ancestry contributing to the Tibeto-Burman tribal speakers. On the other hand, the Mongols and the Europeans have contributed significant amounts of ancestry to the Indo-European and Tibeto-Burman forward castes. F3 statistics also show that the Central Asians are an admixed population with signs of admixture from Caucasus and other parts of Europe.