GENETIC VARIATION REVEALS MIGRATIONS INTO THE INDIAN SUBCONTINENT AND ITS INFLUENCE ON THE INDIAN SOCIETY

Aritra Bose[1,2], Daniel E Platt[2], Laxmi Parida[2], Peristera Paschou[3,4], Petros Drineas[1]

[1] Department of Computer Science, Purdue University, West Lafayette, IN, USA;

[2] Computational Genomics, IBM T.J. Watson Research Centre, Yorktown Heights, NY, USA;

[3] Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis, Evros, Greece;

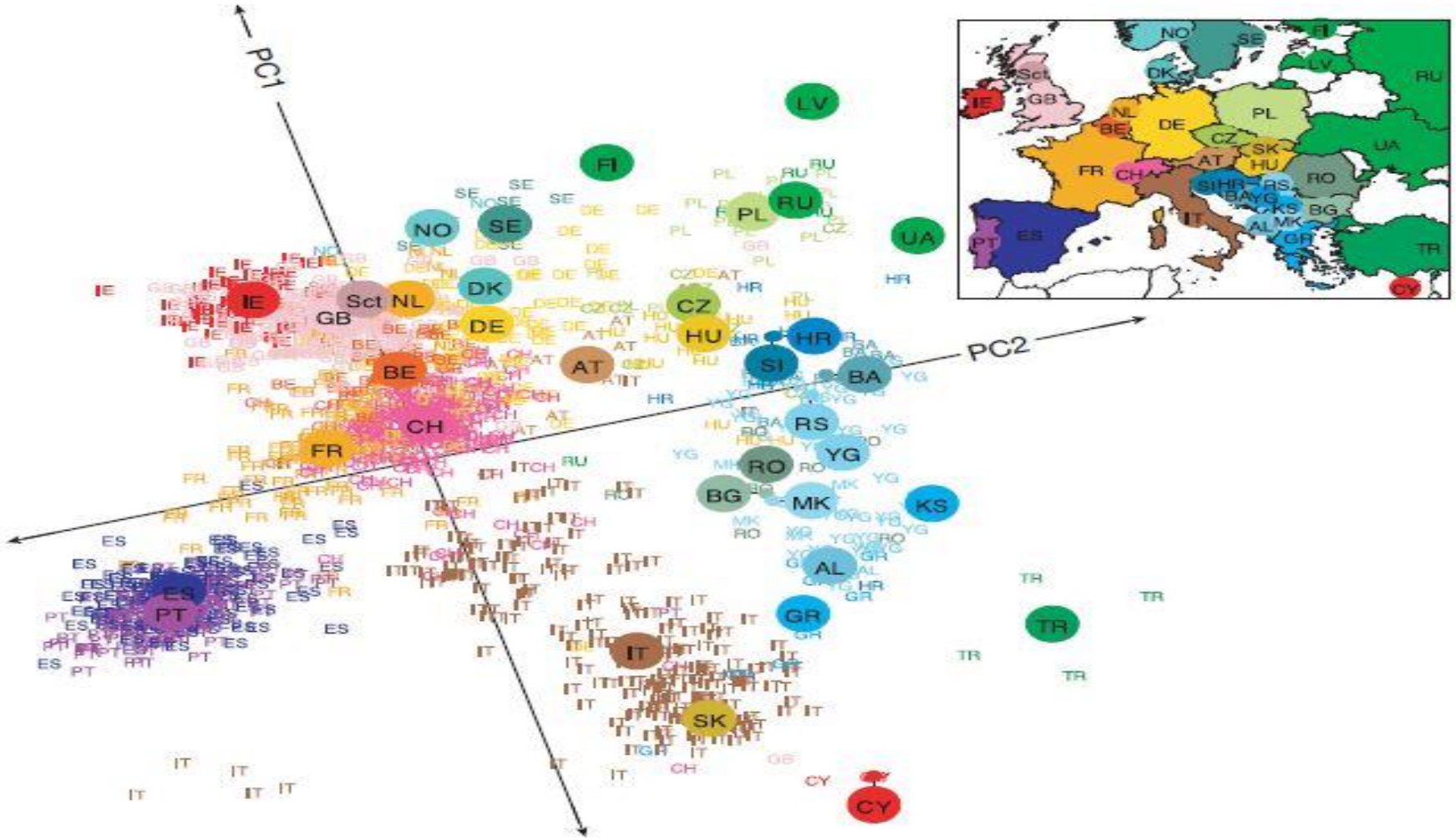
[4] Department of Biological Sciences, Purdue University, West Lafayette, IN.

PURDUE UNIVERSITY

IBM

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ

EUROPE: GENES MIRRORING GEOGRAPHY



Paschou et al. (2010), Drineas et al. (2010), Lao et al. (2008) and Novembre et al (2008) showed the Pearson correlation coefficient, r^2 between the geographical coordinates and the principal components for 197,146 SNPs in 1,387 samples (POPRES project) collected across Europe to be:

0.71 for PC1 v/s Latitude and 0.72 for PC2 v/s Longitude

DATA

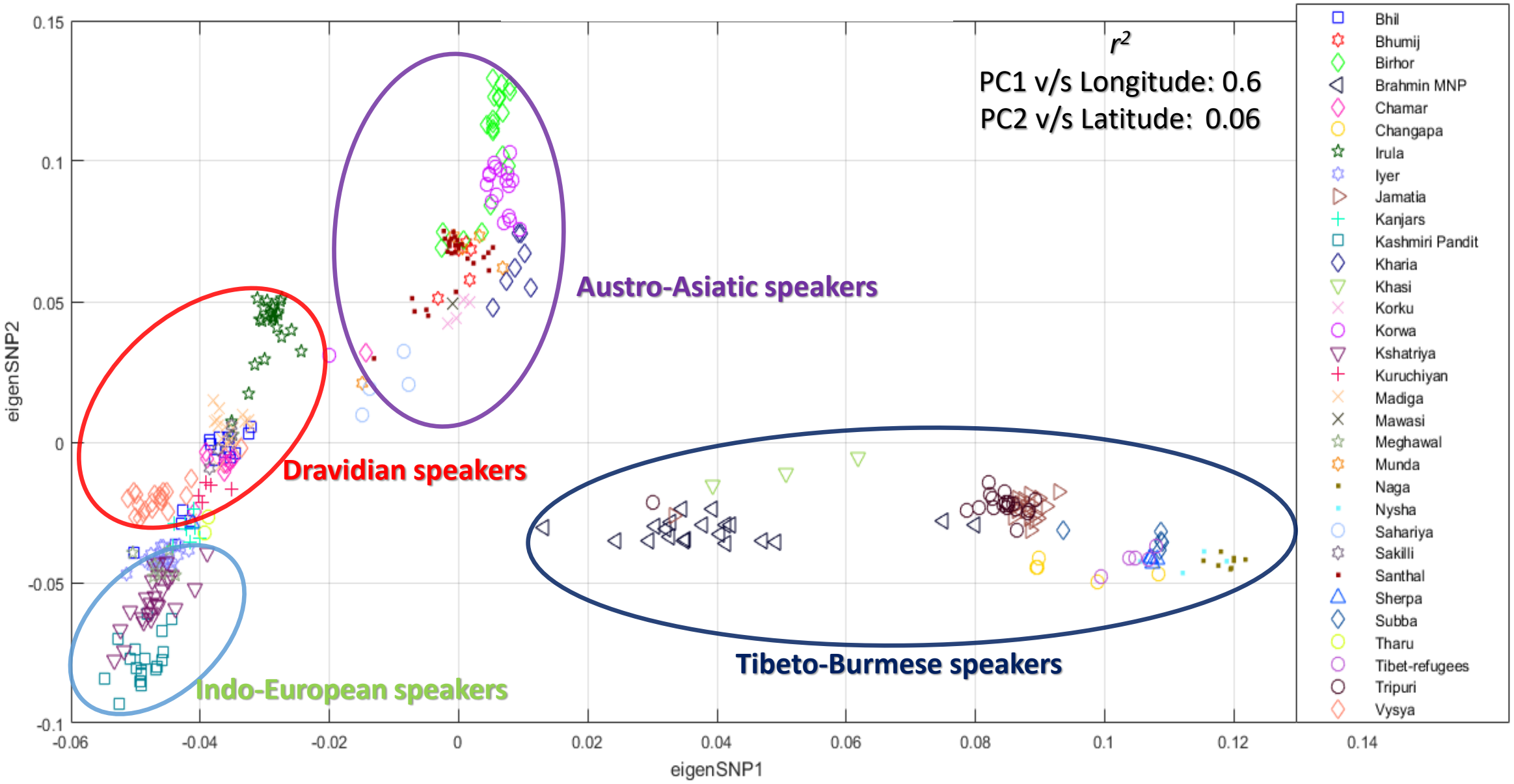
Combining data from various sources:

Number of Samples	Number of Populations	Source
142	30	Metspalu et al. (2011)
26	10	Chaubey et al. (2010)
19	4	Behar et al. (2010)
132	25	Reich et al. (2009)
188	21	Moorjani et al. (2013)
367	20	Basu et al. (2016)
874	**110**	

We analyzed many subsets of the above data sets to create an equal representation of caste, languages and geographical locations of India.

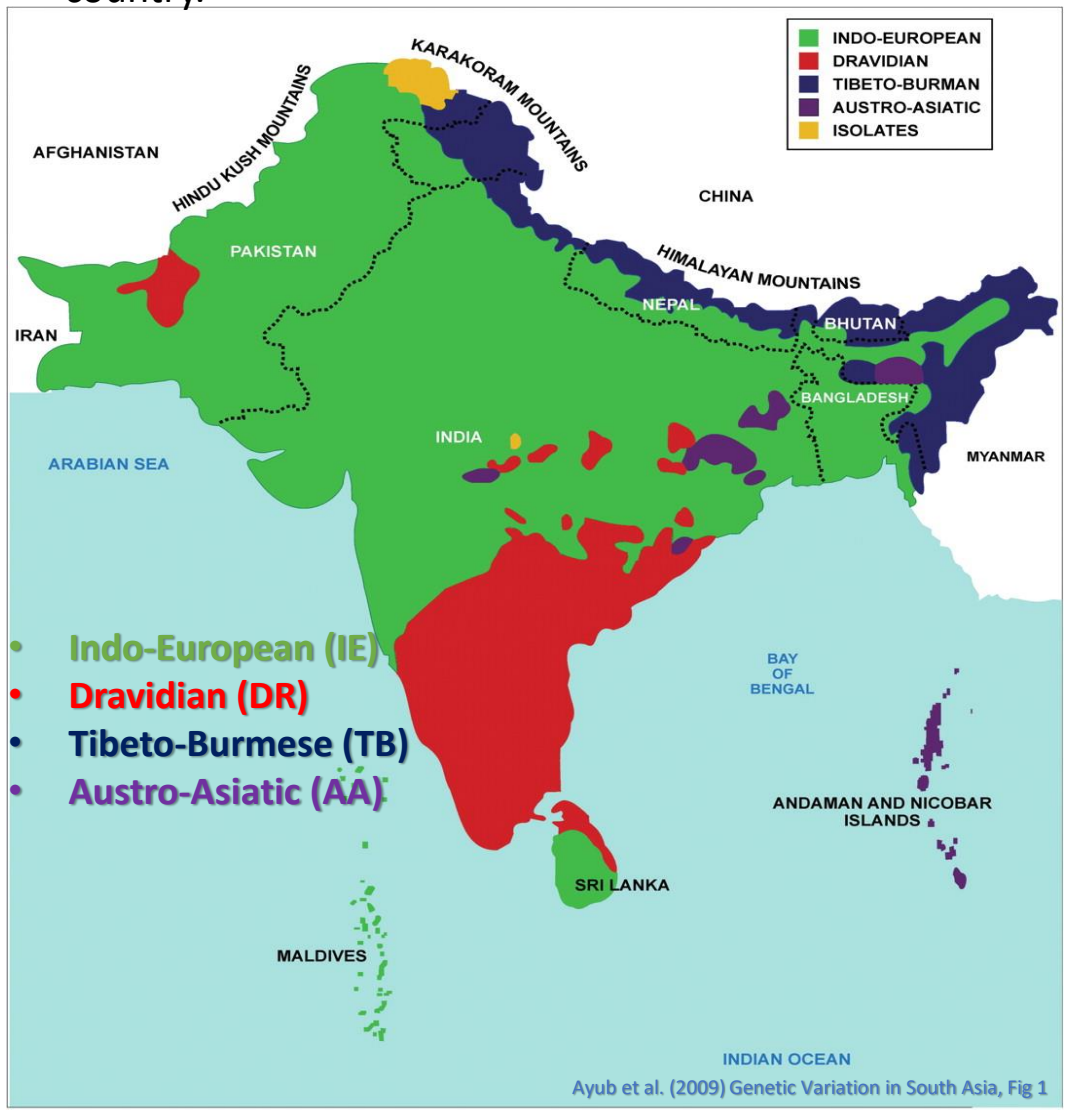
A typical data set has approximately 50,000 SNPs in 370 samples across 33 populations.

INDIA: Genes Mirroring geography and languages?

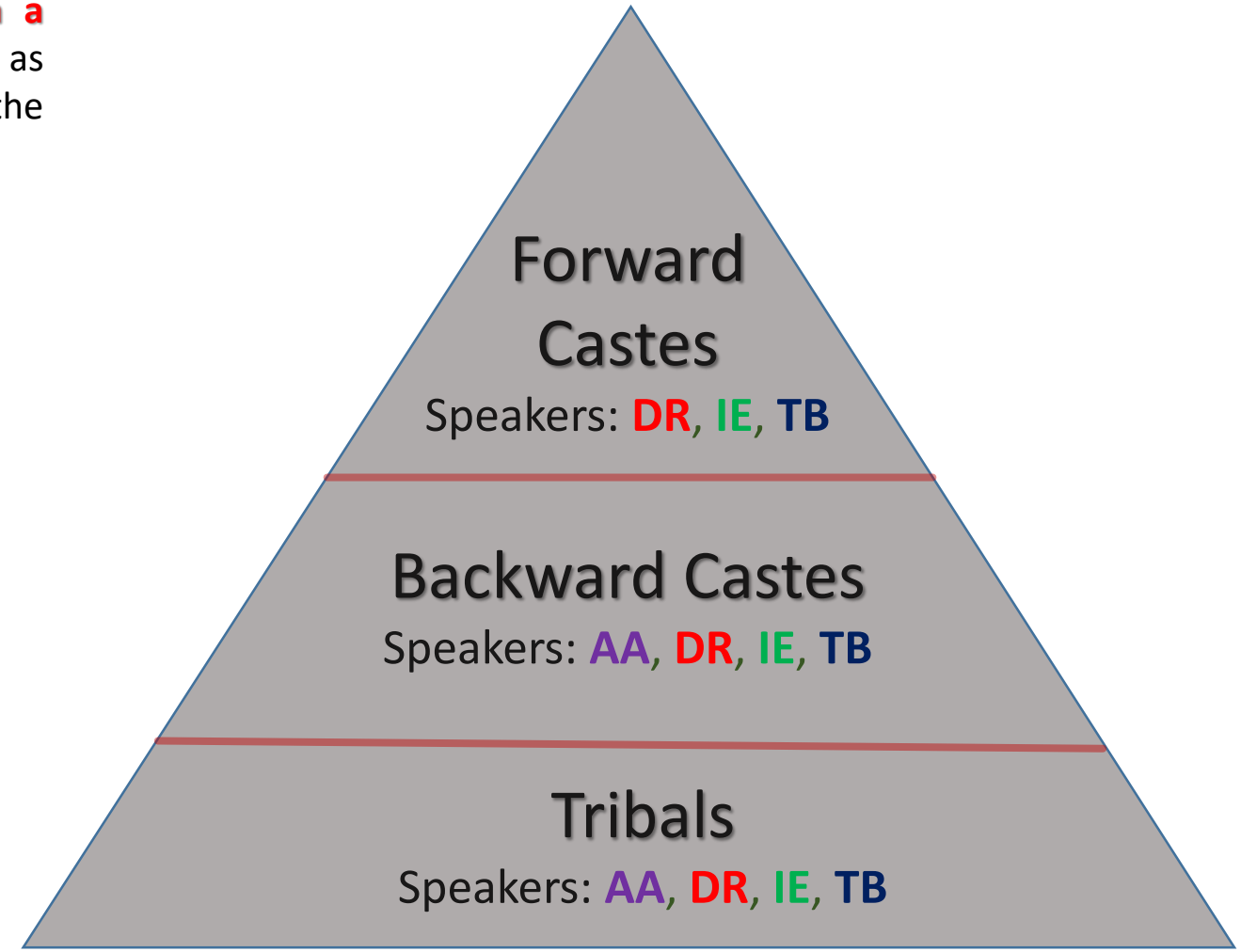


INDIA: SOCIOLINGUISTICS

- According to 2001 census, 29 languages have **more than a million** native speakers, of which 22 languages are recognized as official, with a total of 1,652 mother tongues spoken across the country.



- Indo-European (IE)
- Dravidian (DR)
- Tibeto-Burmese (TB)
- Austro-Asiatic (AA)



- Social stratification in terms of Caste System was documented first around 300 BC.
- There are 4,635 well-defined endogamous populations in India with 532 tribal communities constituting ~8% (2001 Census, Govt. of India) of the total population.

COGG

COGG stands for **Correlation Optimization of Genetics and Geodemographics**. A novel method to correlate genetic information with geographical axes with social factors (castes and languages).

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

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.

COGG: Results

We analytically solved for α and found that we get the maximum correlation for

$$\alpha = [\mathbf{Var}\,[U] \cdot \mathbf{Cov}\,[G_i, G_j]]^{-1} \cdot \mathbf{Cov}\,[U, G_i]$$

$U \in \mathbb{R}^n$, is the vector corresponding to the eigenSNPs.
$G \in \mathbb{R}^{n \times k}$, is the Geodemographic matrix.

Plugging in the value of α we get:

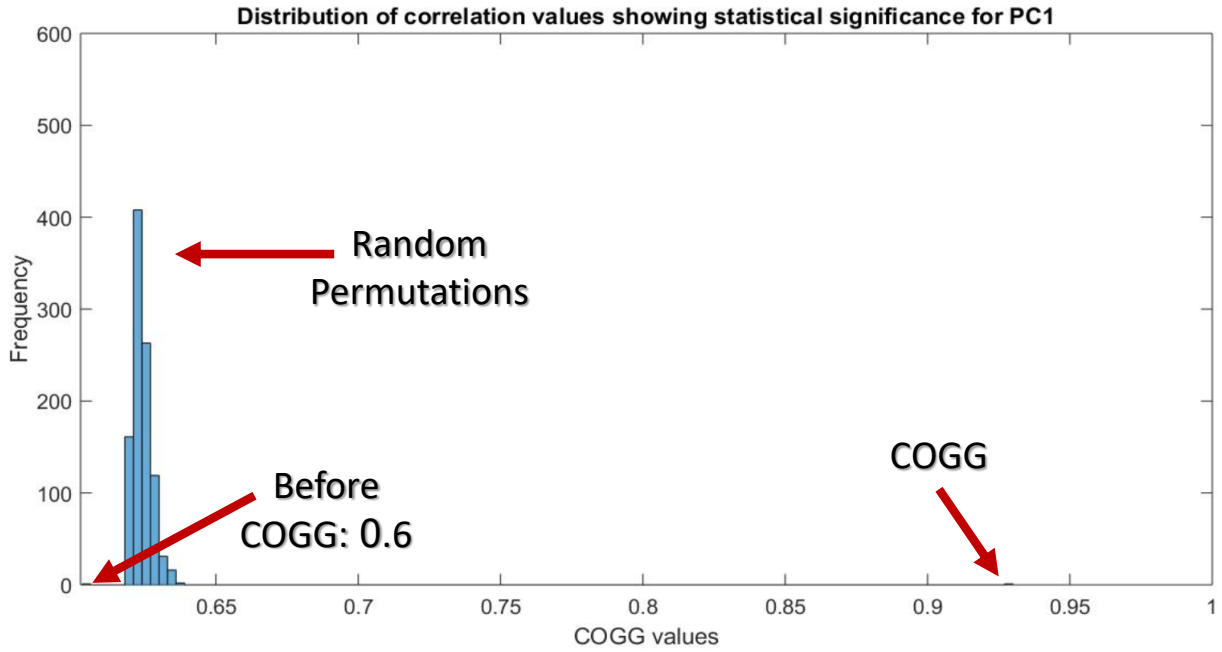
0.93 for **eigenSNP1 v/s G**

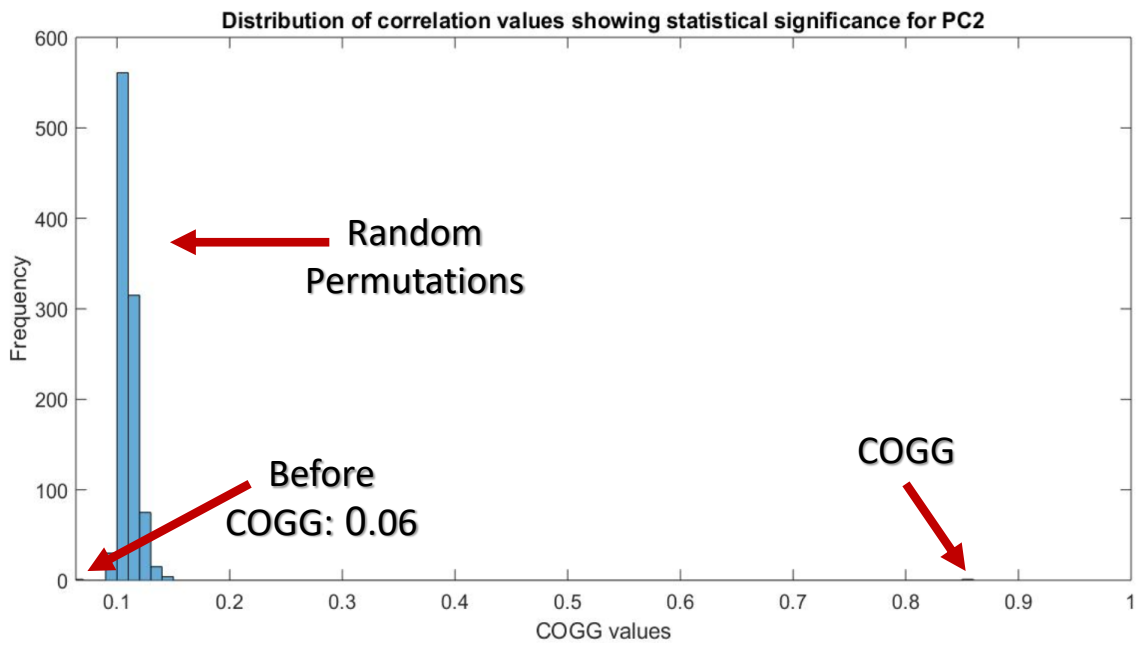
0.85 for **eigenSNP2 v/s G**

COGG: Statistical Significance

We checked for statistical significance by randomly permuting Caste and Language information (we ran 1,000 iterations).

r^2 values / *Stages*	PC1	PC2
Before COGG	0.6005	0.0655
COGG	0.9280	0.8529
Random Permutations (max over 1000 iterations)	0.6461	0.1431





Visualizing COGG

To check whether COGG can trace back samples to their geographical origins, we devised an unsupervised algorithm which colors the points depending on the coefficients returned by COGG.

Algorithm:

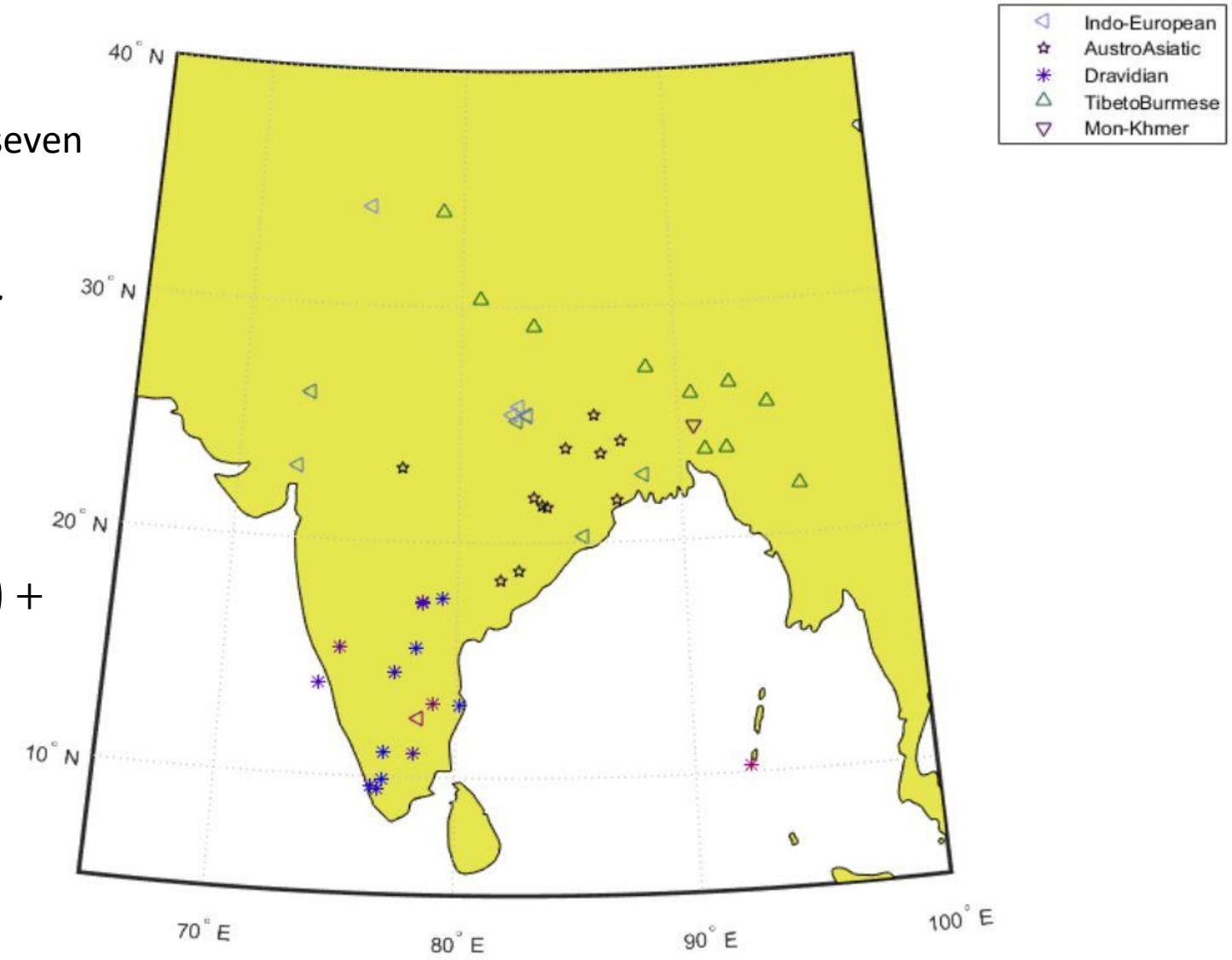
1. Compute $\sum_i^k \alpha_i \cdot G_i$, normalize and scale by seven (representing VIBGYOR color palette).

2. Define a color map, *cmap,* corresponding to *k.*

3. Separate integers and floating points from the scaled values into vectors, *ints* and *floats*.

4. for (i = 1: number of samples)
   ColCogg = $\left(1 - floats(i) \cdot cmap\left(ints(i,*)\right)\right) + floats(i) \cdot cmap(ints(i+1,*))$

5. Use ColCogg to color the points on the map to get meaningfully colored clusters.



Extending COGG to CCA

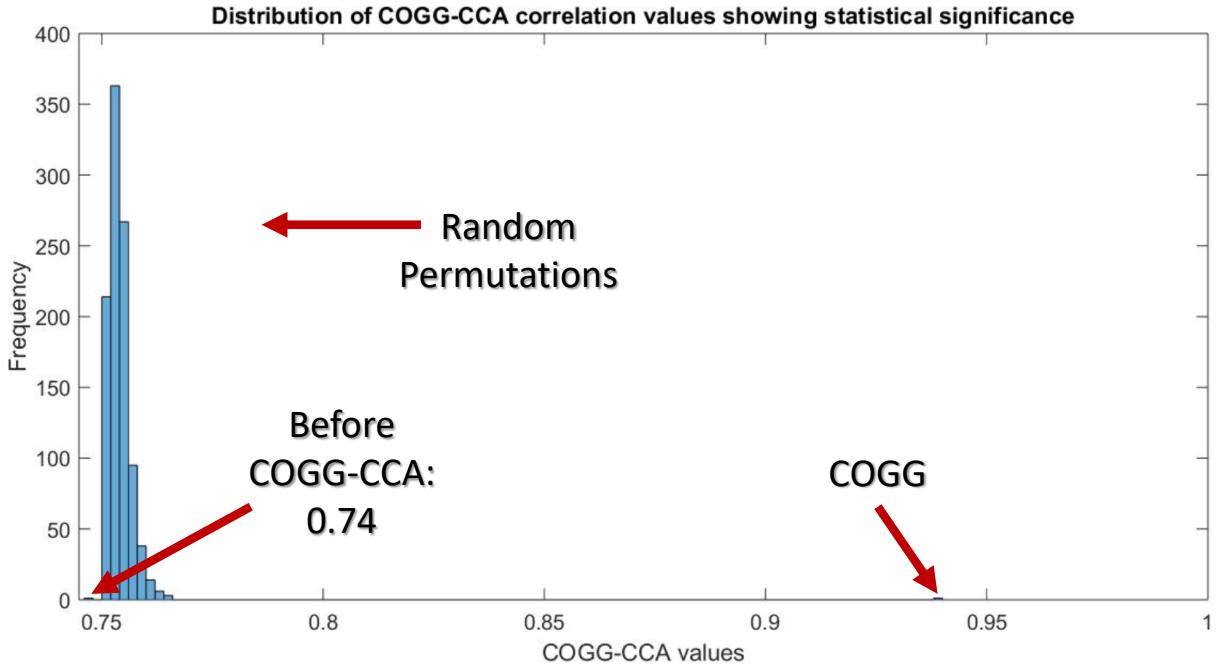
We extended COGG to a Canonical Correlation Analysis, where we compute an overall squared correlation between the top ***p*** eigenSNPs and Geography, caste, and languages.

Formally:

$$\max_{\alpha,\beta} \mathbf{Corr}\left(\sum_{j=1}^{p} \beta_j \cdot U_j, \sum_{i=1}^{k} \alpha_i \cdot G_i\right)$$

where $G \in \mathbb{R}^{n\times k}$, $U \in \mathbb{R}^{n\times p}$, p is the number of eigenSNPs.
$\beta = (\beta_j)$ and $\alpha = (\alpha_i)$ are unknown vectors of the coefficients for each feature and each eigenSNP.

We analytically solved for α, β and in our case a maximum correlation (r^2) of **0.94** was obtained.

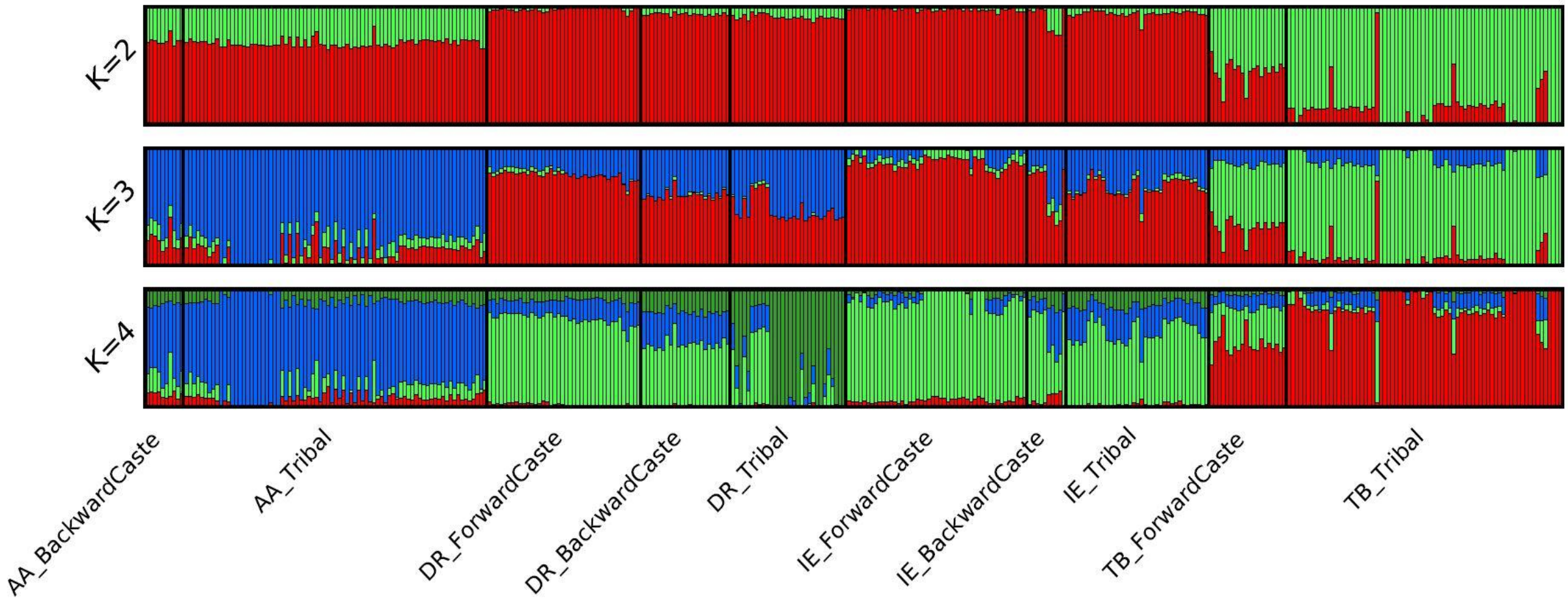


Stages \ r^2 values	r^2 (p=8)
Before COGG-CCA	0.746
COGG-CCA	**0.94**
Random Permutations (max over 1000 iterations)	0.765

ADMIXTURE

We ran ADMIXTURE and observed that endogamy in the hierarchical caste system **and** peopling by language families have caused genetic stratification in the Indian subcontinent.

We observed the caste groups to be well admixed with the tribal groups for each language family.



Shared Ancestry

Our objective is to compute the shared ancestry[1] between populations *'X'* and *'Y'*. We create two matrices, P_x and P_Y of dimensions *n-by-K* containing the estimates from ADMIXTURE and project P_Y onto the subspace of P_X.

We take the top *'k'* eigenvectors of P_X and create a matrix V_x . We perform the projection to find the shared ancestry between populations *'X'* and *'Y'* using the following formula:

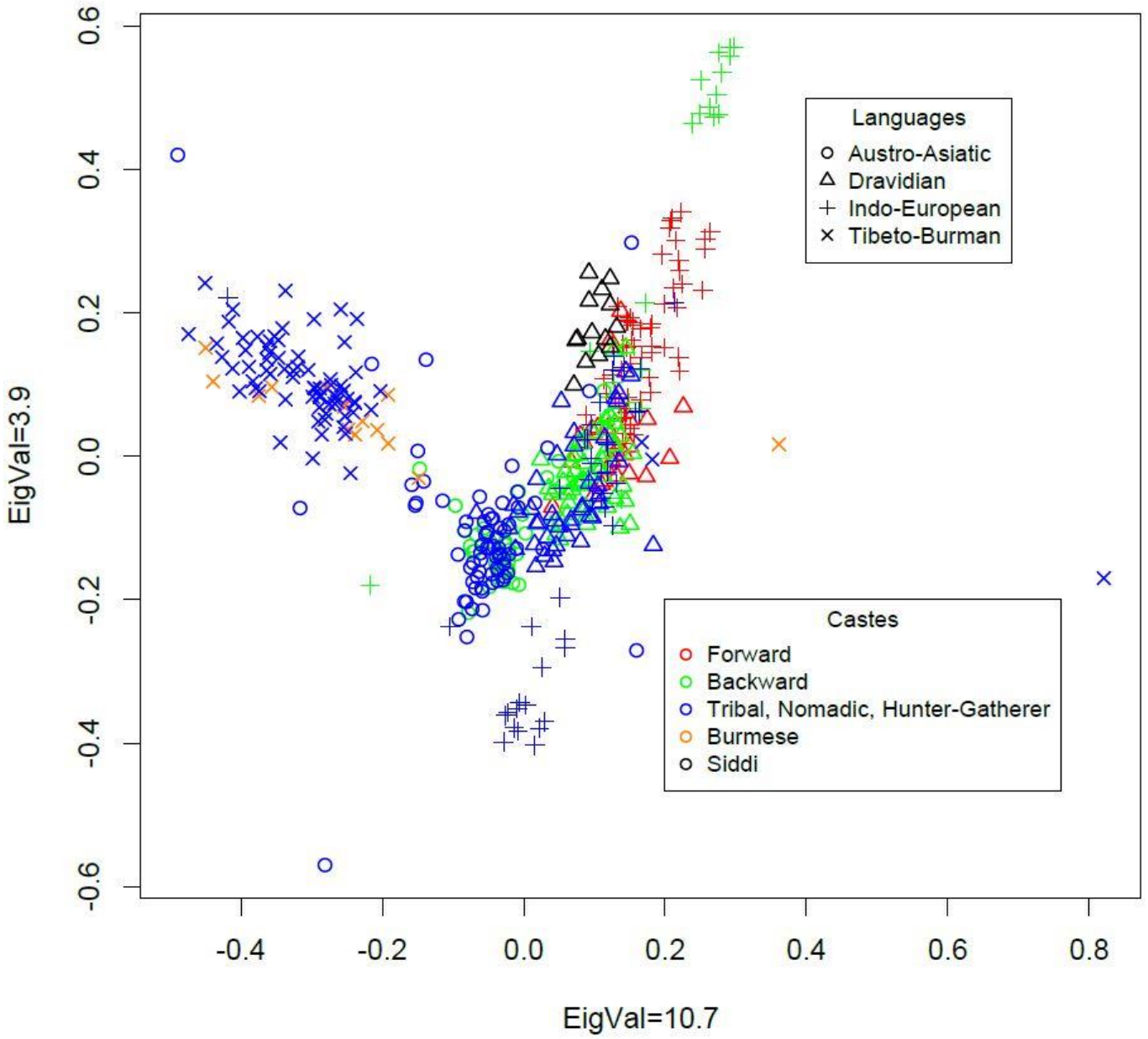
$$\frac{\|P_Y \cdot V_X\|_F^2}{\|P_X\|_F^2} \qquad \|.\|_F = \text{Frobenius Norm}$$

Castes / Tribes	DR_Forward Caste	IE_Forward Caste	TB_Forward Caste	AA_Backward Caste	DR_Backward Caste	IE_Backward Caste
AA_Tribal	14.65	4.22	11.69	94.13	41.12	27.33
DR_Tribal	71.05	51.95	26.15	79.09	93.02	73.93
IE_Tribal	96.84	86.94	34.19	44.69	97.27	90.26
TB_Tribal	1.07	1.83	75.35	6.71	1.22	3.37

Table showing shared ancestry between the Tribal and Caste populations

[1]Stamatoyannopoulos G., Bose A., et al. (2016) [under review]

Linear Discriminant Analysis

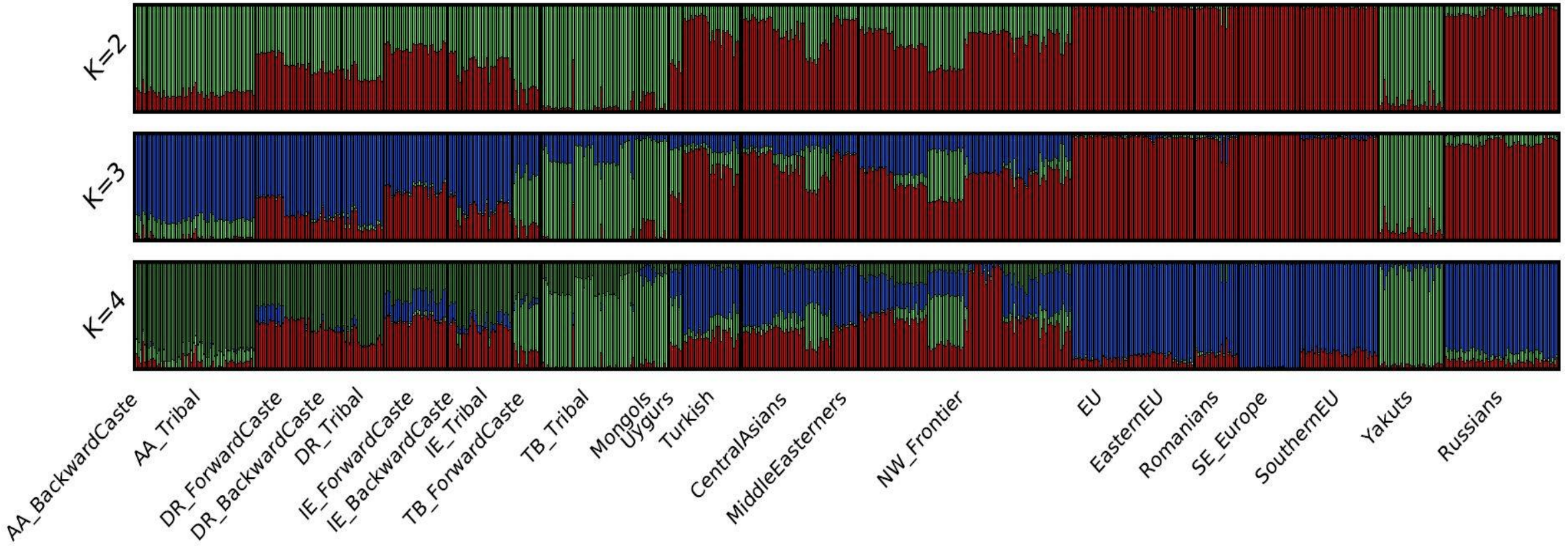


- Linear Discriminant Analysis (LDA) reveals a cline representing a language gradient from DR through IE speakers.
- LDA finds linear combinations of SNP contributions that maximize the variation between specified groups with variations within those groups (in our case, Caste and Languages).
- There are a few "satellite" populations in IE speaking Forward castes, Backward castes and tribals (might be because of their relatedness with Eurasian populations).

Indo-Aryan migrations

IE Forward Castes share a very small amount of ancestry with Europeans, unlike IE Backward Castes and Tribals.

All of IE and DR speaking populations share some ancestry with Central Asians and Middle Easterners.



DR Forward Castes have Central Asian components, with a small or no European component.

TB Tribals share significant amount of ancestry with Mongols, Uygurs, and Yakuts.

Shared Ancestry

Eurasians / Indians	Eastern Europe	Central Europe	Southern Europe	Uygurs	Turkish	North West Frontier (Afghanistan and Pakistan)	Middle Easterners	Central Asians
IE_ForwardCaste	20.558	15.857	23.715	26.008	47.253	81.649	54.882	52.779
IE_BackwardCaste	7.799	5.754	9.371	16.297	22.531	51.254	26.951	26.229
IE_Tribals	4.521	3.348	5.917	11.405	16.873	46.281	21.453	20.328
DR_ForwardCaste	5.583	4.198	7.207	12.536	20.368	54.357	26.324	24.207
DR_BackwardCaste	1.141	0.707	1.848	7.266	9.018	35.643	12.573	11.695
DR_Tribals	0.648	0.455	0.955	3.742	4.081	17.201	5.683	5.272

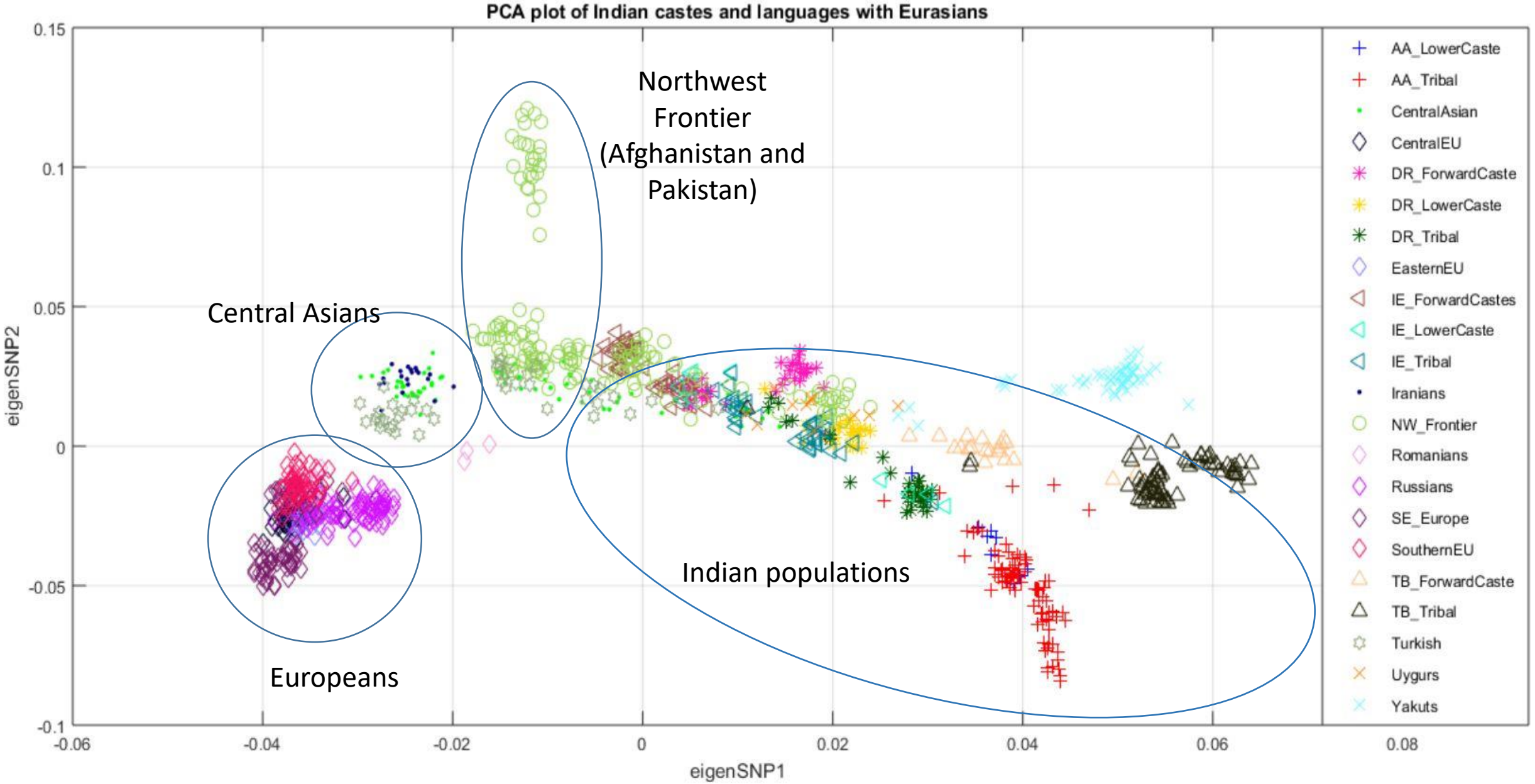
f_3 statistics

f_3 tests (target; source$_1$, source$_2$)	f_3 values	Z
f_3 (IE_ForwardCaste; Mongols, Southern Europe)	-0.0018	-4.960
f_3 (IE_ForwardCaste; Mongols, Southwestern Europe)	-0.0012	-2.923

Note: All other f_3 tests gave positive f_3 values showing no evidence of admixture from the source populations.

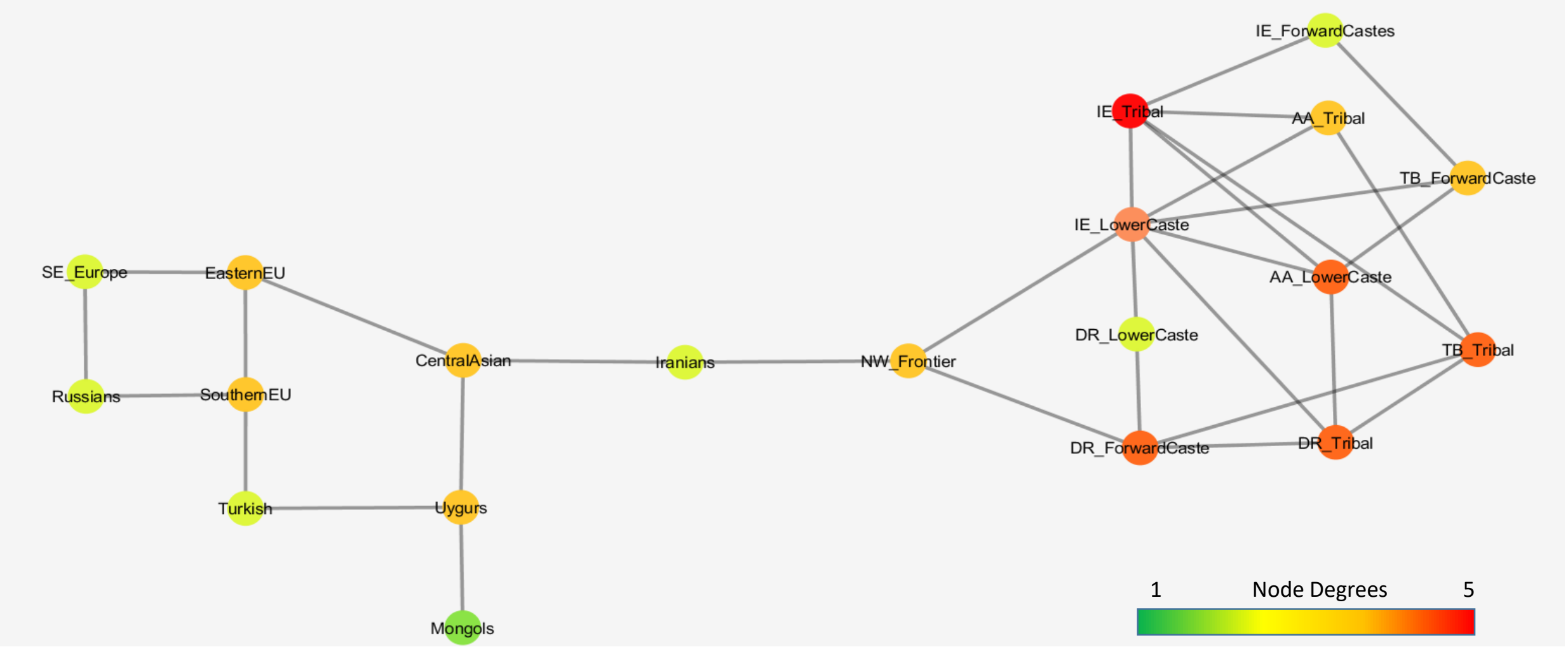
Indo-European populations share little ancestry with Mongols (~2%). f_3 shows evidence that Indo European Forward Castes are admixed populations between Mongols and Southern Europeans.

PCA: Indians and Eurasians



Connection from India and Eurasia

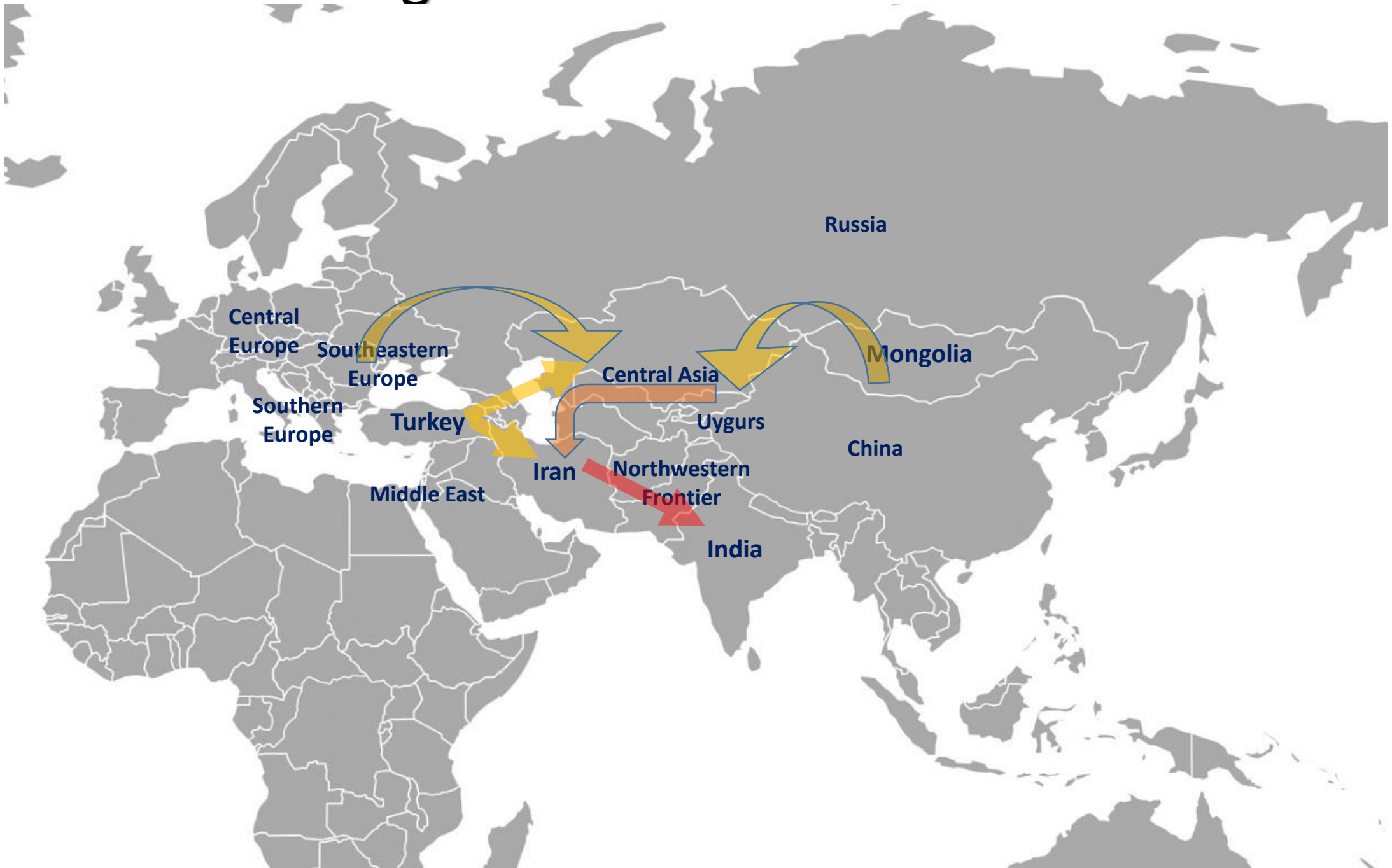
Generating population genetic networks from principal components[2], we observe paths from Mongolia, Turkey and Southeastern Europe, through Central Asia via Iran and the Northwest Frontier provinces of Afghanistan and Pakistan.



We can see that the Himalayas acted as a natural barrier to migrations. Iran was a pivotal region for the migrating populations to the Indian subcontinent.

[2]Paschou P., Drineas P., et al, Maritime route of colonization of Europe, PNAS (2014)

Possible Migration Routes from Eurasia to India



Conclusions

- We proposed a novel method, **COGG** to understand genetic stratification caused by social and linguistic factors. This can be generalized for any set of external factors that influence human population genetics.

- Source code: https://github.com/aritra90/COGG

- **Genetics of Indian subcontinent:**

  - We found evidence to support the hypothesis that the caste system in India evolved from a strong social structure within the tribes and refute the hypothesis that the Indo-Aryan migration led to the creation of the caste system.

  - We found evidence of migration routes from Mongolia and Europe through Iran and Afghanistan to India, showing that the Himalayas acted as a natural barrier to gene flow.

Conclusions

- We proposed a novel method, **COGG** to understand genetic stratification caused by social and linguistic factors. This can be generalized for any set of external factors that influence human population genetics.

- Source code: https://github.com/aritra90/COGG

- **Genetics of Indian subcontinent:**

  - We found evidence to support the hypothesis that the caste system in India evolved from a strong social structure within the tribes and refute the hypothesis that the Indo-Aryan migration led to the creation of the caste system.

  - We found evidence of migration routes from Mongolia and Europe through Iran and Afghanistan to India, showing that the Himalayas acted as a natural barrier to gene flow.

- **Additional Discoveries:**

  - We found evidence of migration routes through the Silk Route for Tibeto-Burmese and East Asian populations, where Tibeto-Burmese tribals are more related to the East Asians than forward castes.

  - We showed that Austro-Asiatic speakers in India are not significantly related to other Mon-Khmer speaking populations of South East Asia.

Acknowledgements

- Part of this work was presented as a poster in the Biology of Genomes meeting (May 10-14, 2016, Cold Spring Harbor Laboratories, NY).

- Part of this work was done while A. Bose was a summer intern at the Computational Biology Center of IBM T.J. Watson Research Center, Yorktown Heights, NY (May – August, 2016).

- Partially funded by NSF grants to Petros Drineas and Aritra Bose; partially funded by European Union grants to Peristera Paschou.

PURDUE
UNIVERSITY