

# Interpretation of Genomic Data in Human Population Genetics

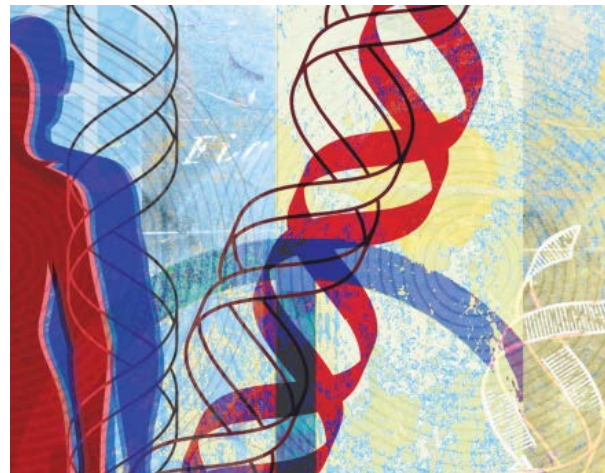
Gyaneshwer Chaubey  
Banaras Hindu University, Varanasi



# Fun fact

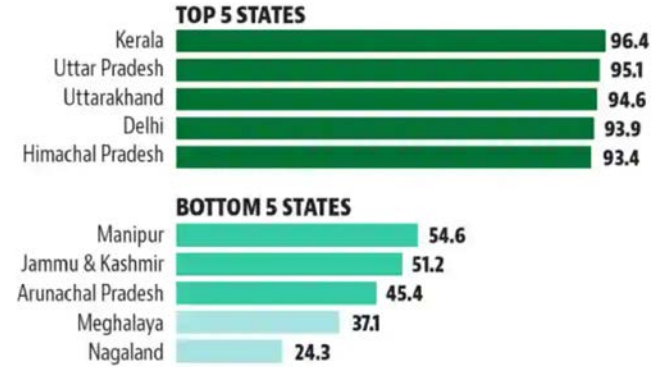


- If two random Indians meet, there is only a 1.6%- 95% chance that they can talk to and understand each other (HT Survey 2011).
- Any two random Indian share >30%- 100% of common genetic ancestry.

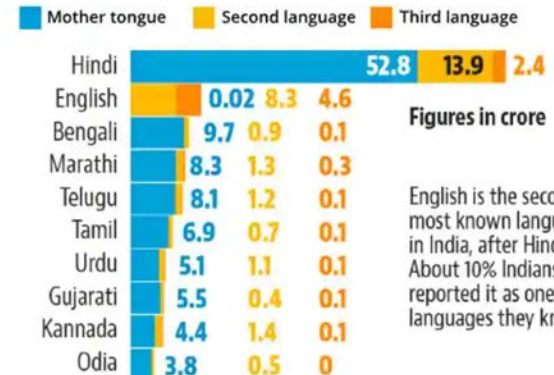


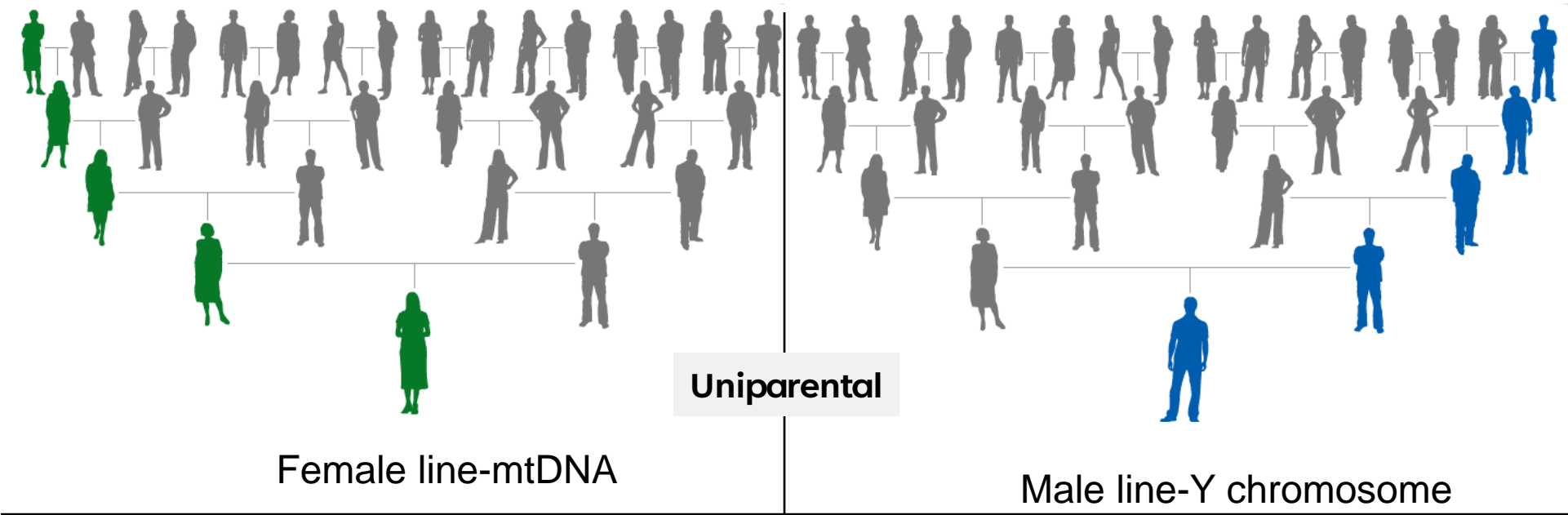
## It varies even within state

The chance that two random persons from the same state can converse is the highest in Kerala and the lowest in Nagaland

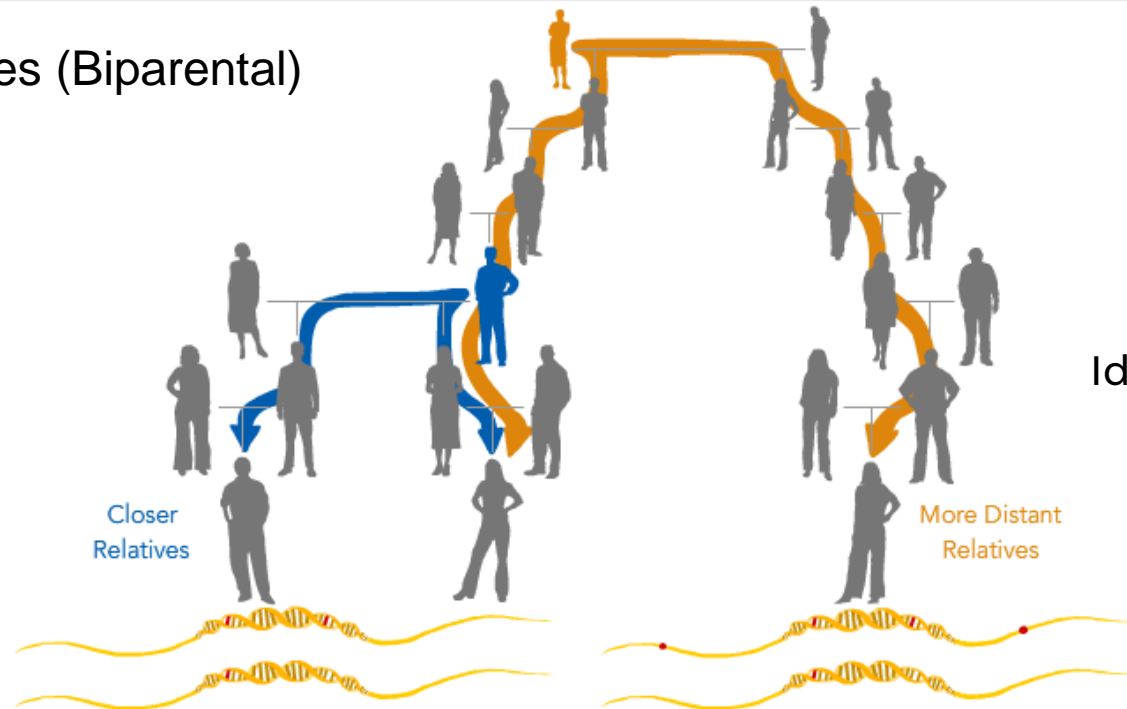


## English second most known language in India





Autosomes (Biparental)

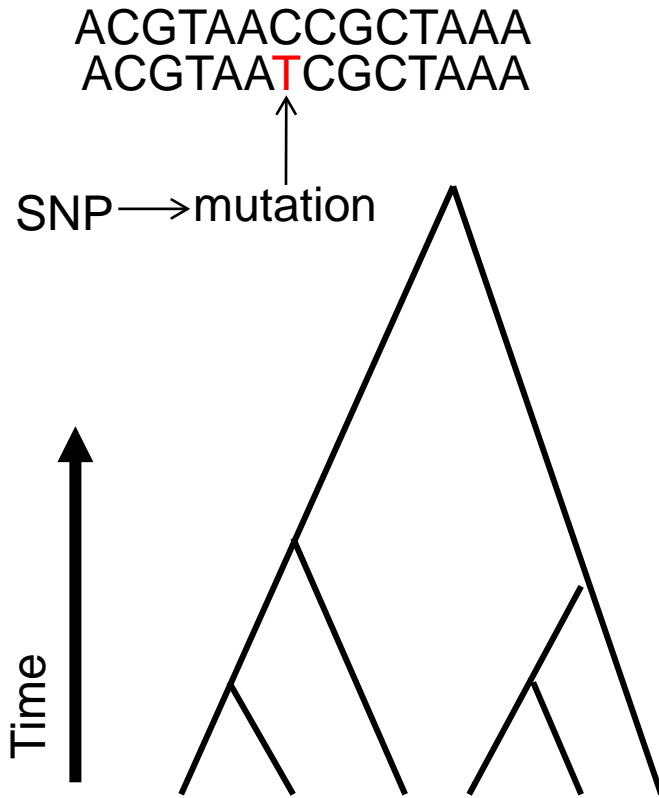


Identity by Descent

People who are more distantly related have more differences.

# Using DNA markers to know the population(s) origin and migration

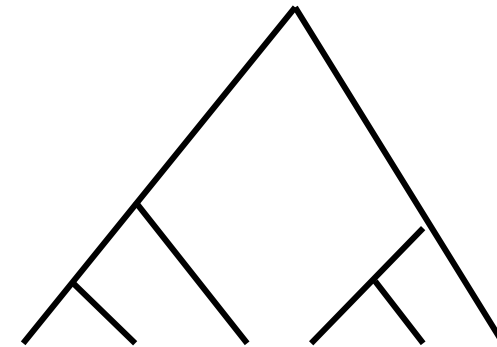
Haplotype ACGTAACCGCTAAA



More diversity  
Large sum of branch lengths  
More time for mutations to accumulate

Haplogroup YY  
ACGTAATCGCTAAG  
ACGTAATCGCTAAG  
ACGTAATCGCTAAG  
ACGTAATCGCTAAG

Haplogroup XX  
ATGTAACCGCTGAA  
ATGTAACCGCTGAA  
ATGTAACCGCTGAA  
ATGTAACCGCTGAA



Less diversity  
Small sum of branch lengths  
Less time for mutations to accumulate



In 2005

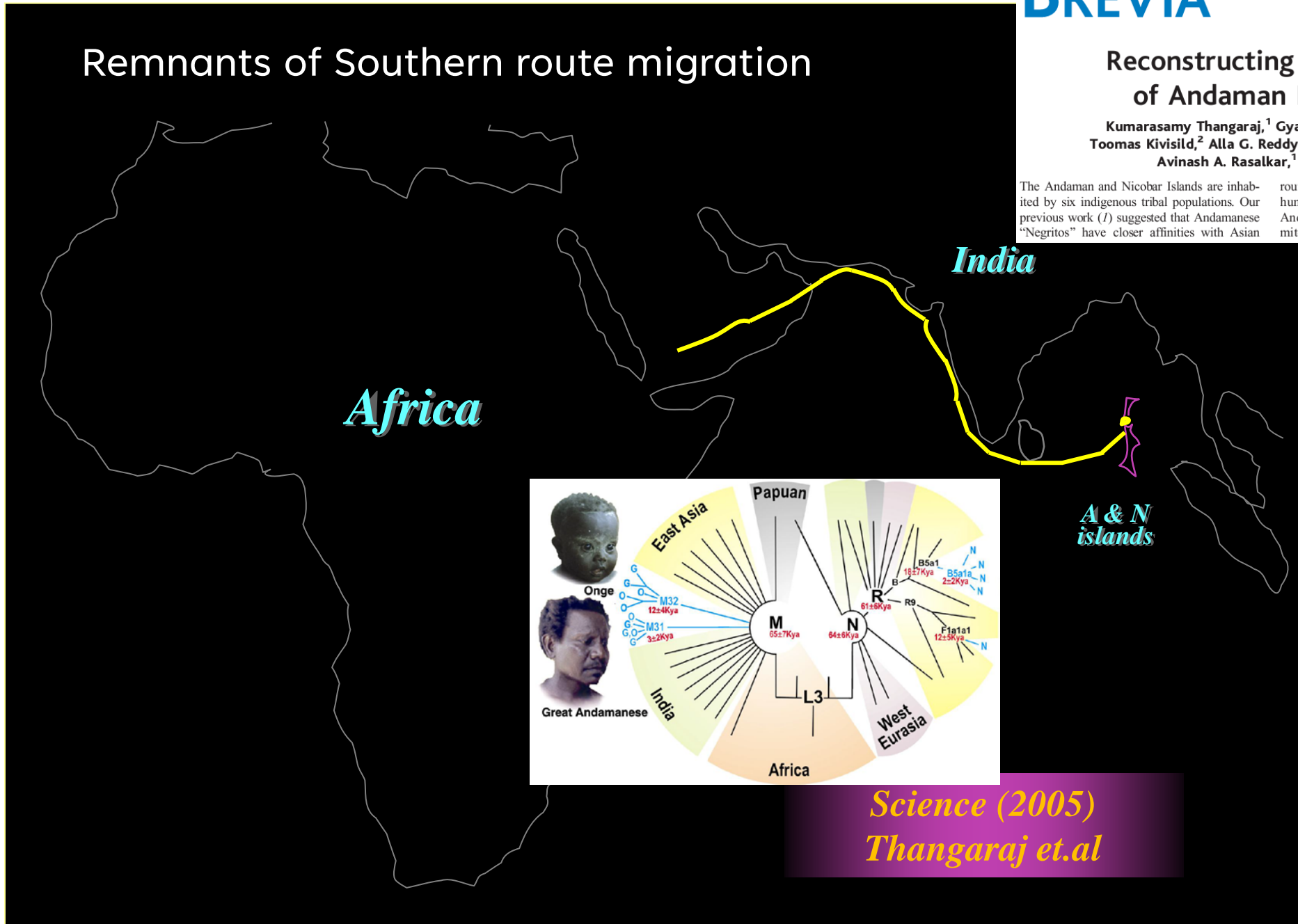
## Remnants of Southern route migration

### Reconstructing the Origin of Andaman Islanders

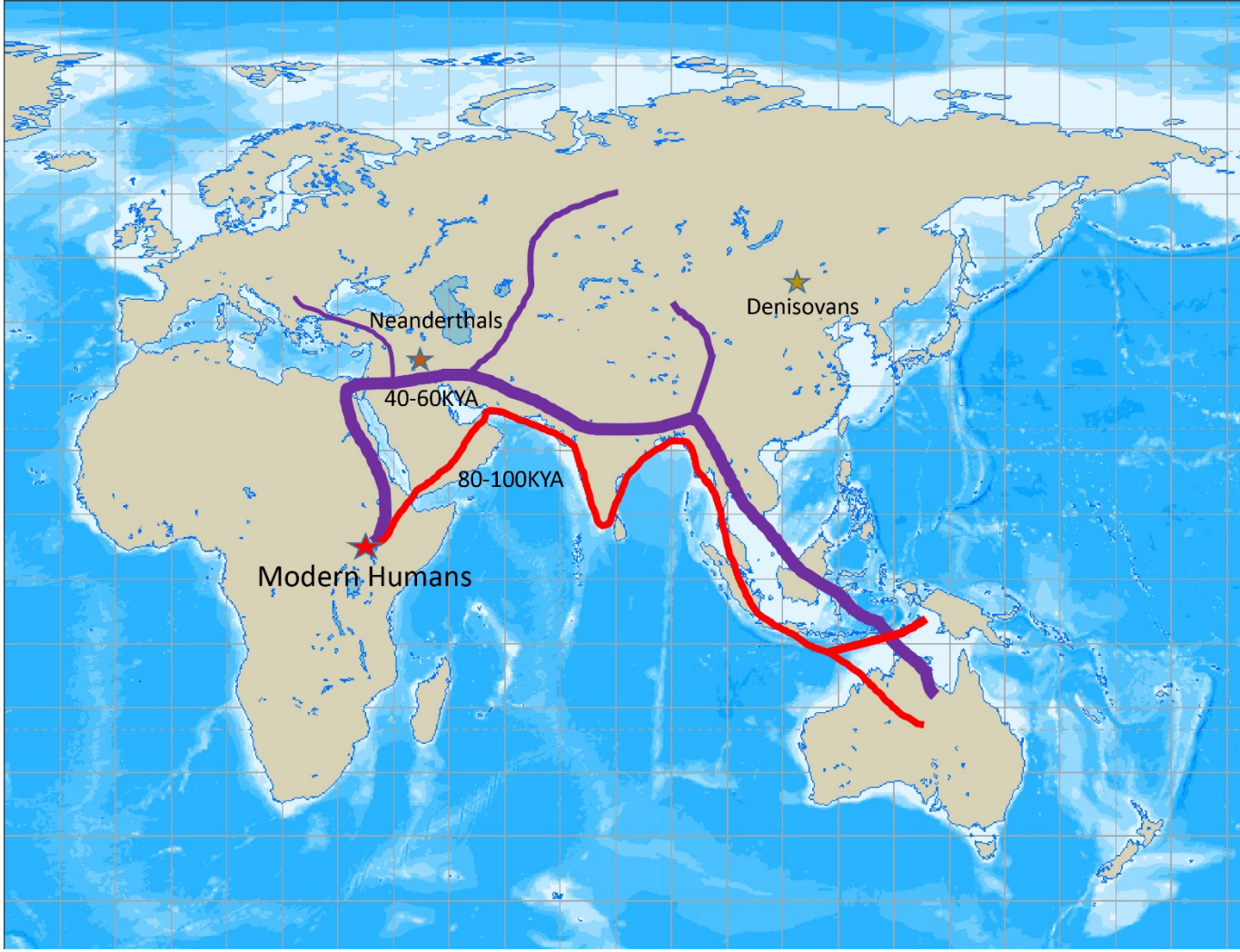
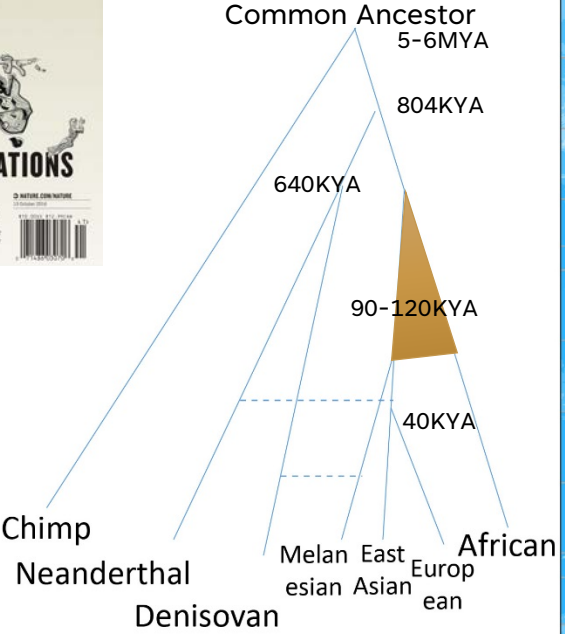
Kumarasamy Thangaraj,<sup>1</sup> Gyaneshwer Chaubey,<sup>1</sup>  
Toomas Kivisild,<sup>2</sup> Alla G. Reddy,<sup>1</sup> Vijay Kumar Singh,<sup>1</sup>  
Avinash A. Rasalkar,<sup>1</sup> Lalji Singh<sup>1\*</sup>

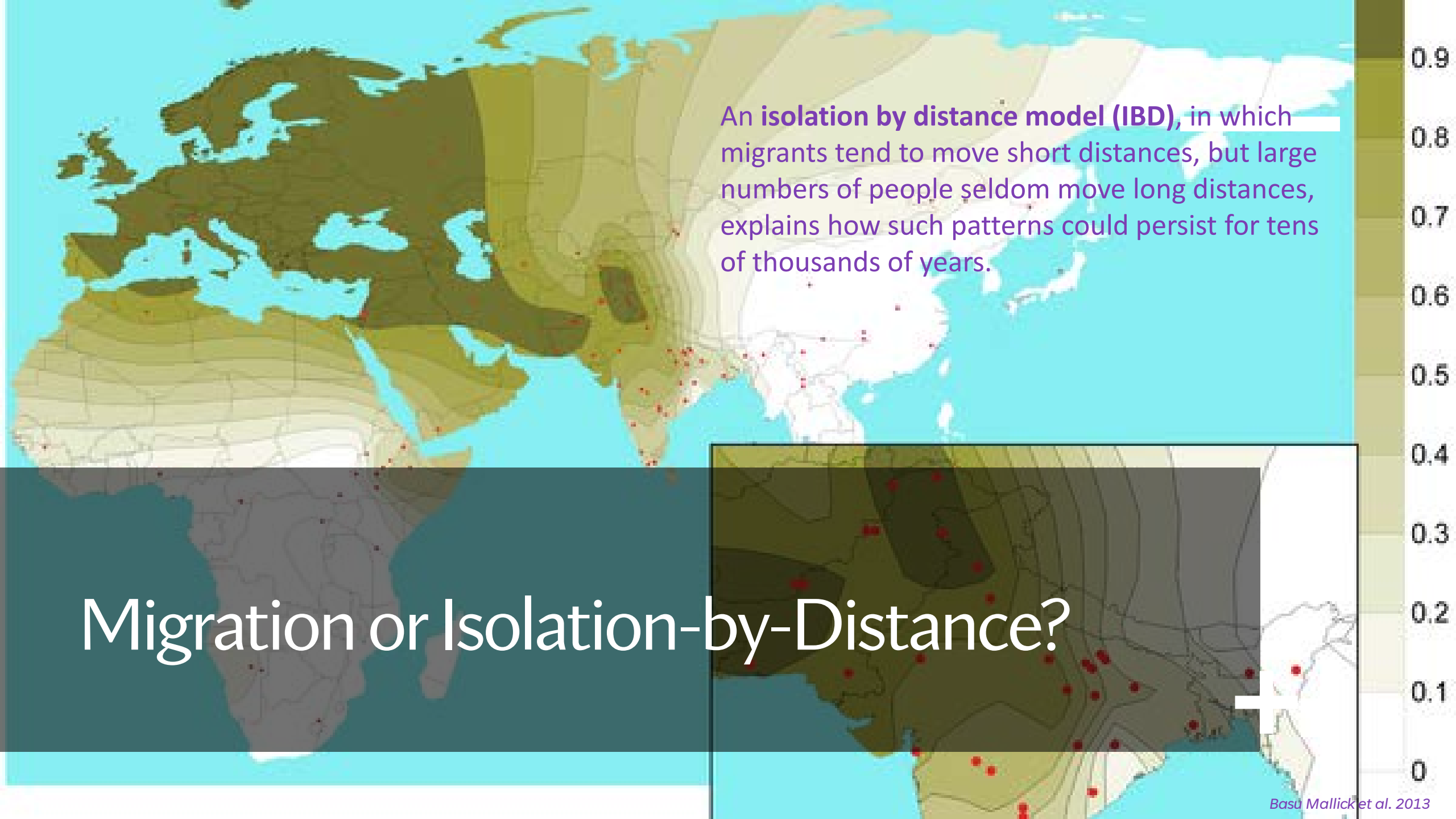
The Andaman and Nicobar Islands are inhabited by six indigenous tribal populations. Our previous work (1) suggested that Andamanese "Negritos" have closer affinities with Asian

route in one wave after the exit of modern humans from Africa (1, 3-5). Because the Andaman Negrito populations carry only one mitochondrial founder haplogroup (M) and



# In 2015- complex making of modern humans



A world map illustrating genetic differentiation by distance. The map uses a color scale from 0 (lightest) to 0.9 (darkest) to represent the degree of genetic differentiation. Darker colors indicate higher differentiation, which is most prominent in Europe and Africa. A vertical color scale legend is located on the right side of the map, ranging from 0 to 0.9 in increments of 0.1. Numerous red dots are scattered across the map, representing sampling locations. A text box in the upper right quadrant explains the Isolation by Distance (IBD) model.

An **isolation by distance model (IBD)**, in which migrants tend to move short distances, but large numbers of people seldom move long distances, explains how such patterns could persist for tens of thousands of years.

# Migration or Isolation-by-Distance?



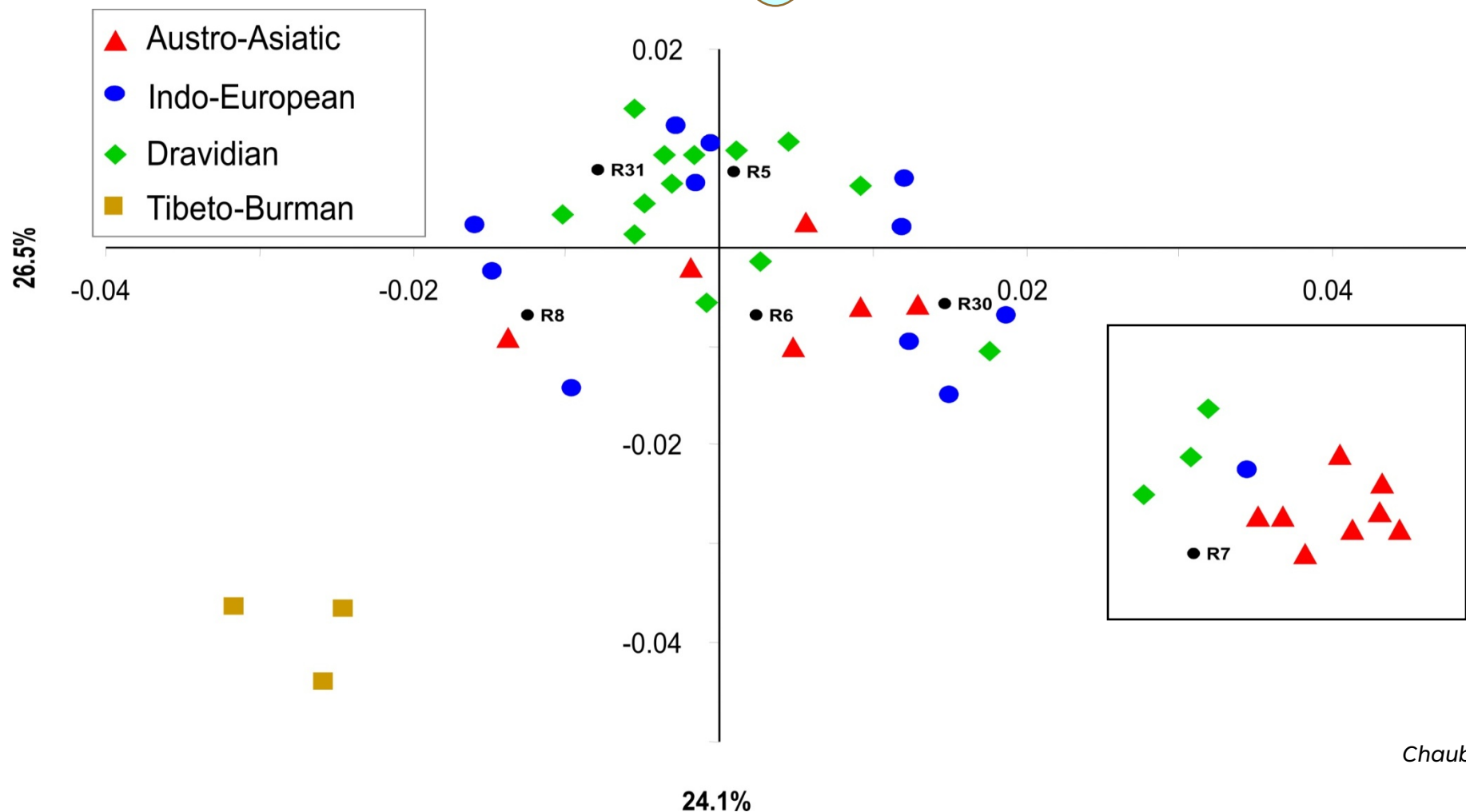
# Data resolution





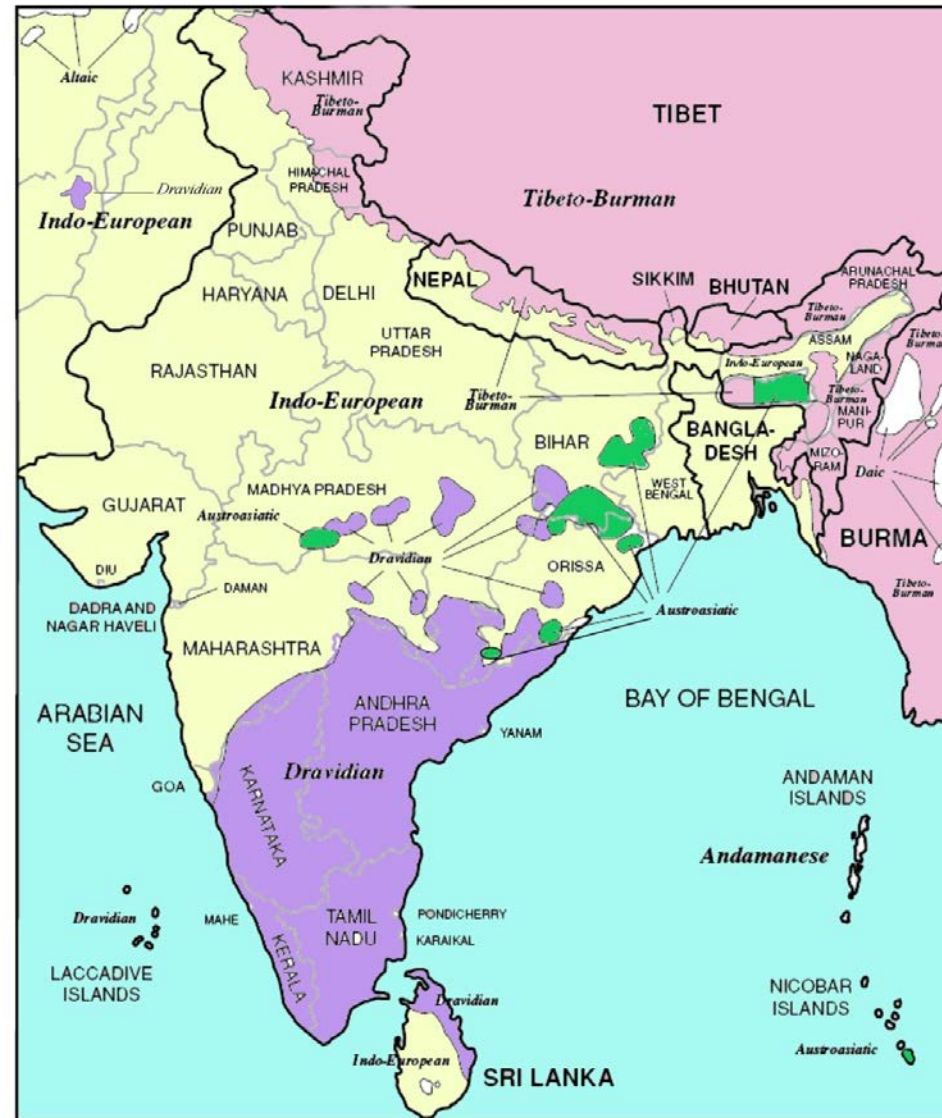
# The mystery of mtDNA haplogroup R7

	R2	R5	R6	R7	R8	R30	R31	Total Samples
Austroasiatic	0.41%	1.12%	4.29%	<b>5.90%</b>	2.64%	0.61%	0.00%	883
Indo-European	0.31%	3.62%	1.70%	<b>0.58%</b>	1.61%	2.63%	0.85%	2240
Dravidian	0.23%	3.65%	1.69%	<b>1.37%</b>	1.64%	2.15%	0.32%	2190
Tibeto-Burman	0.00%	1.74%	0.00%	<b>0.00%</b>	0.58%	0.58%	0.00%	172



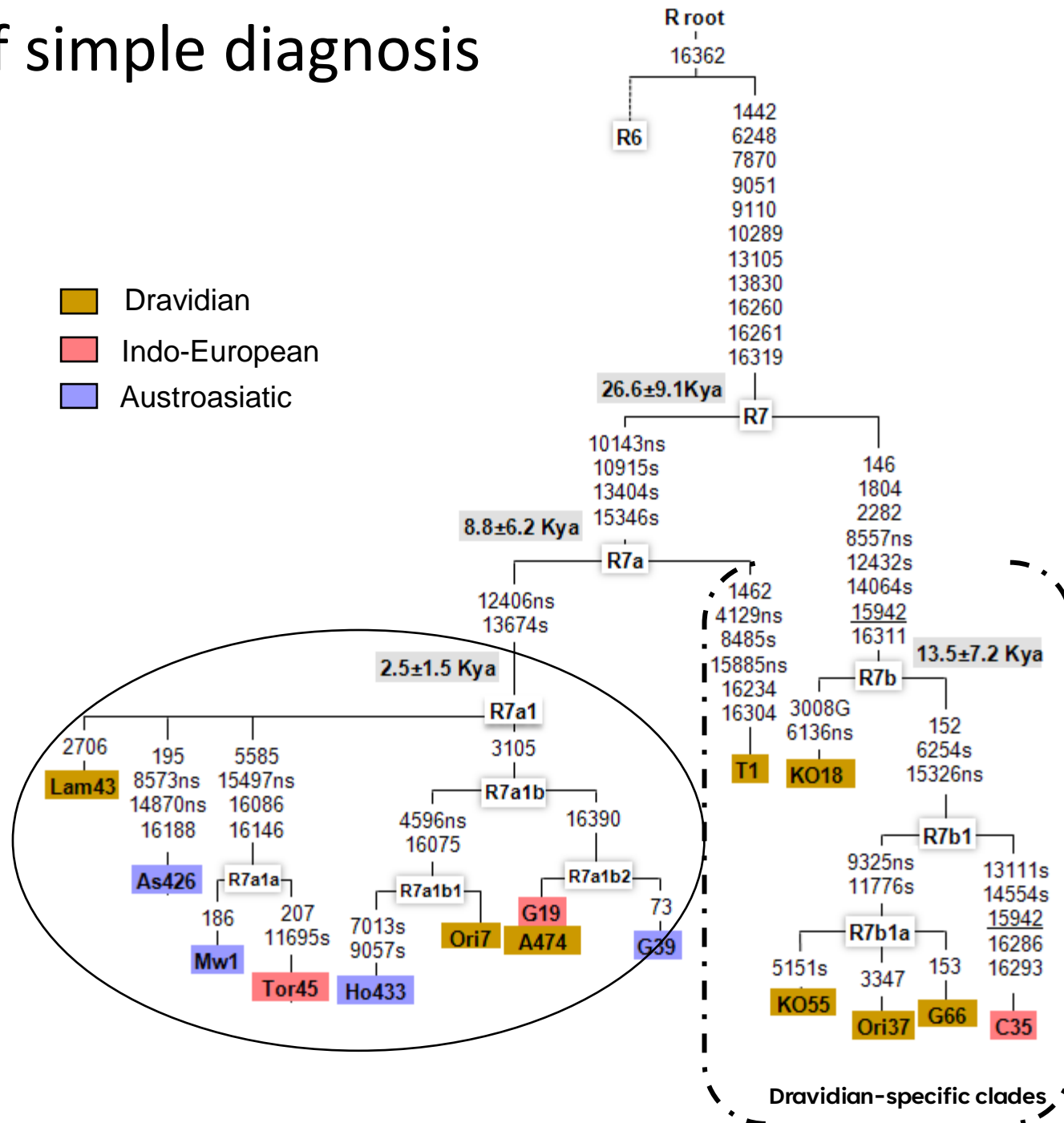
# Language groups in India

There are 121 languages spoken by 10,000 or more people in India

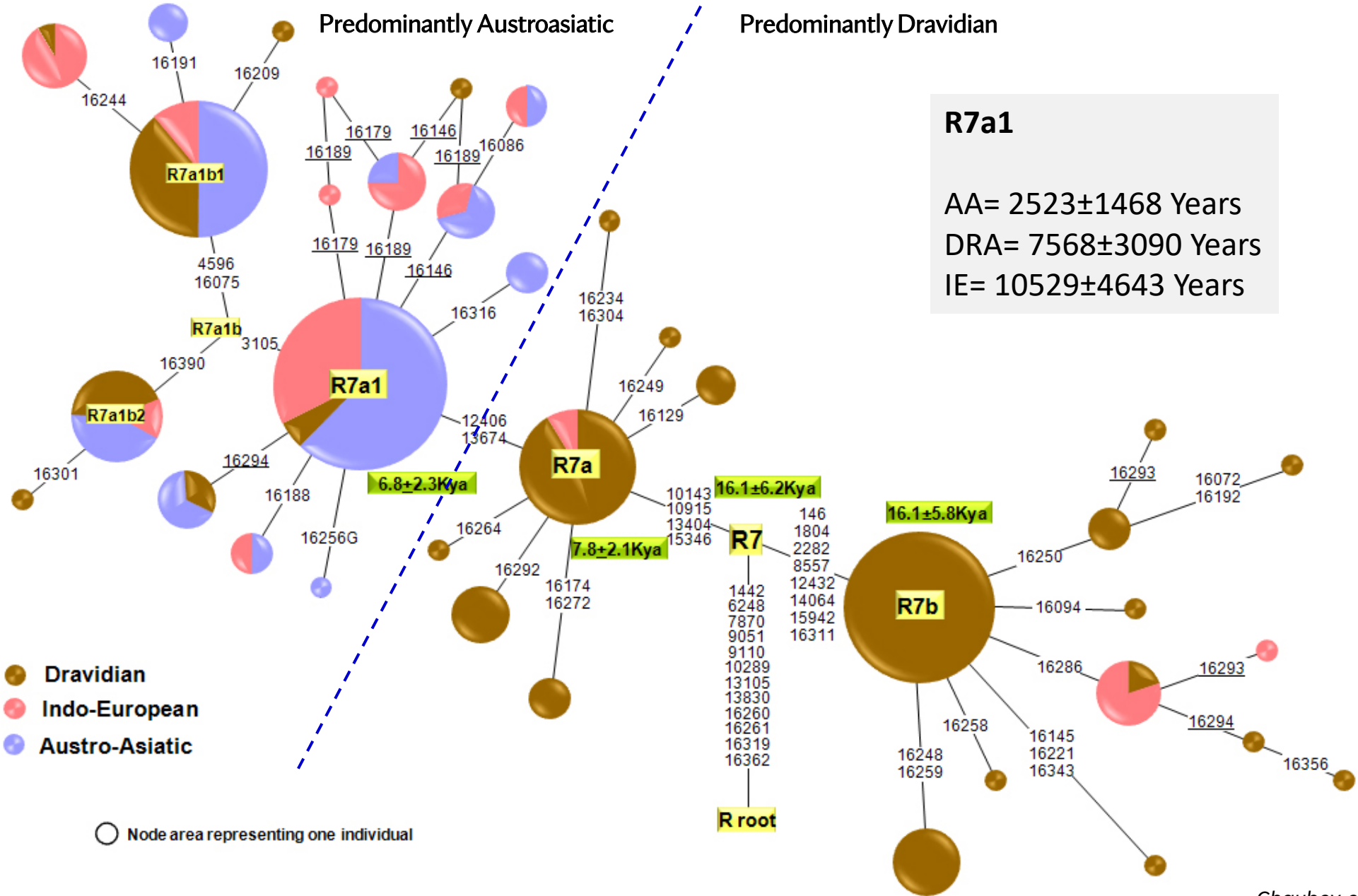


# The demise of simple diagnosis

- Dravidian
- Indo-European
- Austroasiatic

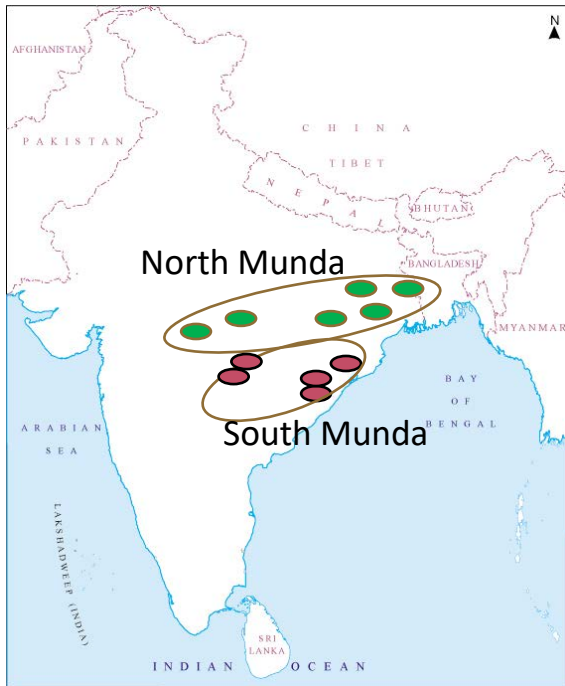
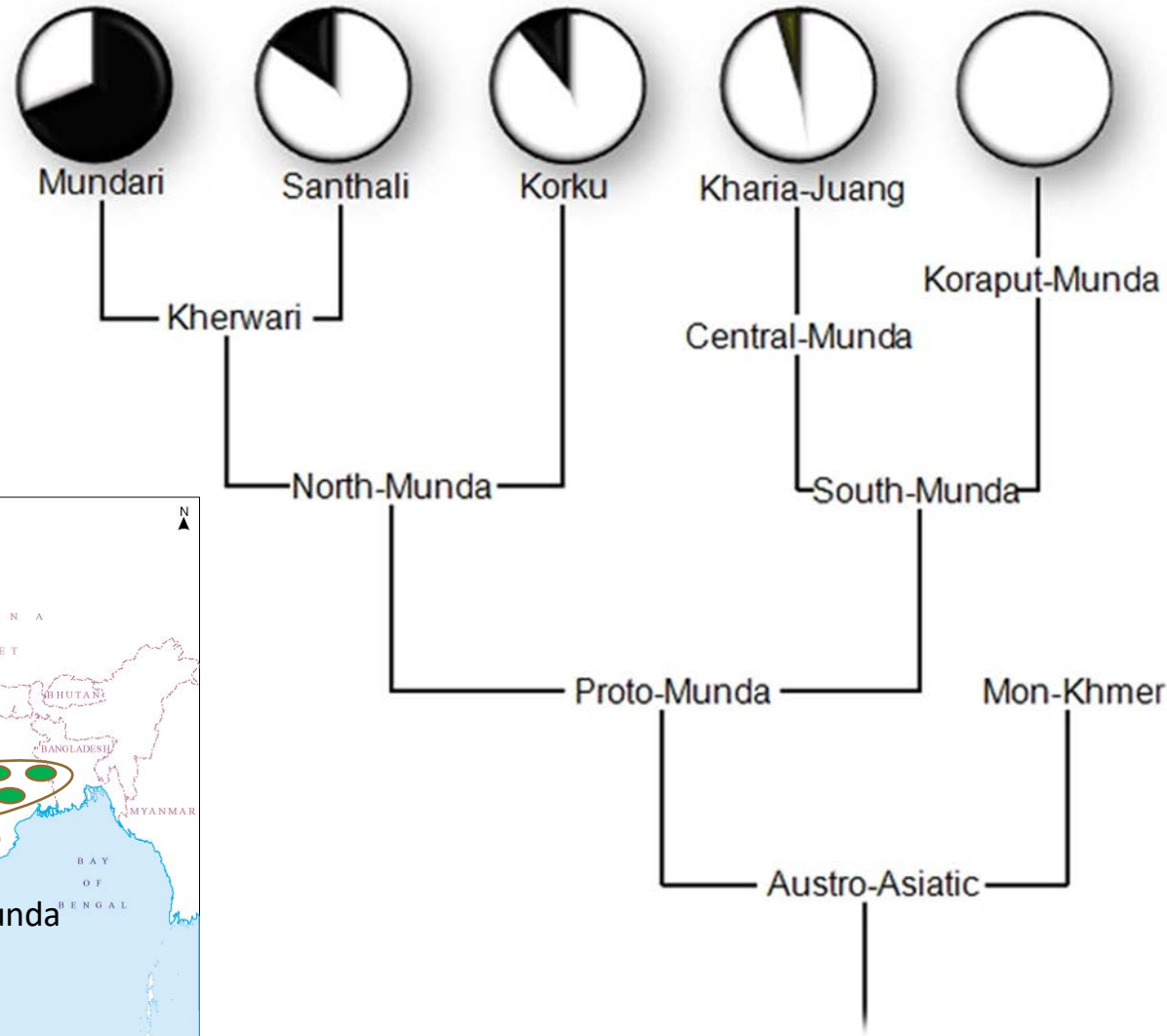


# Linguistic distribution of haplotypes

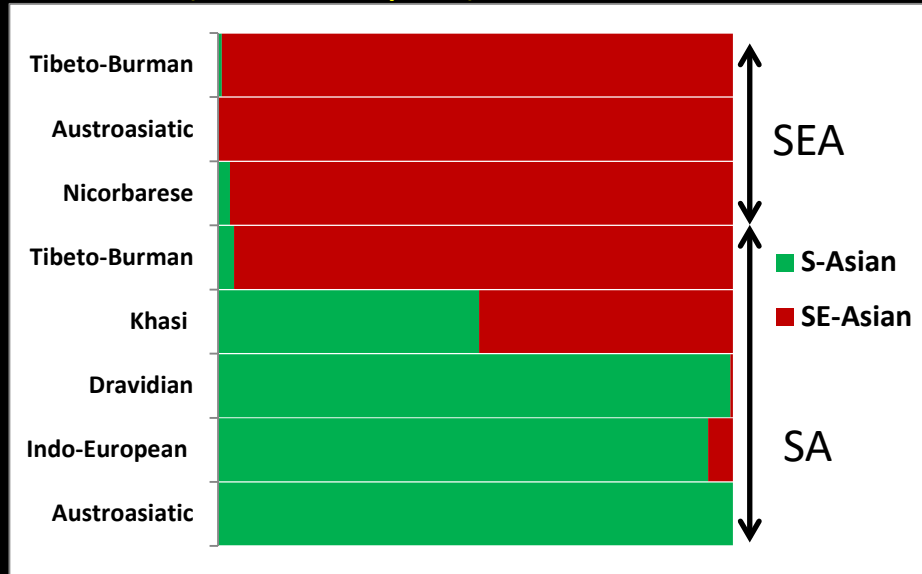




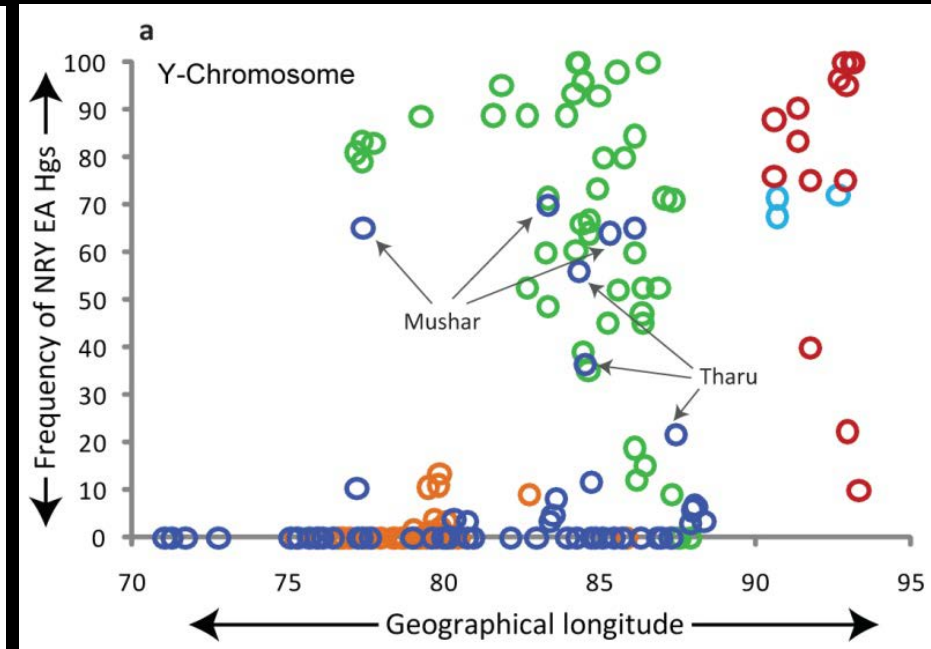
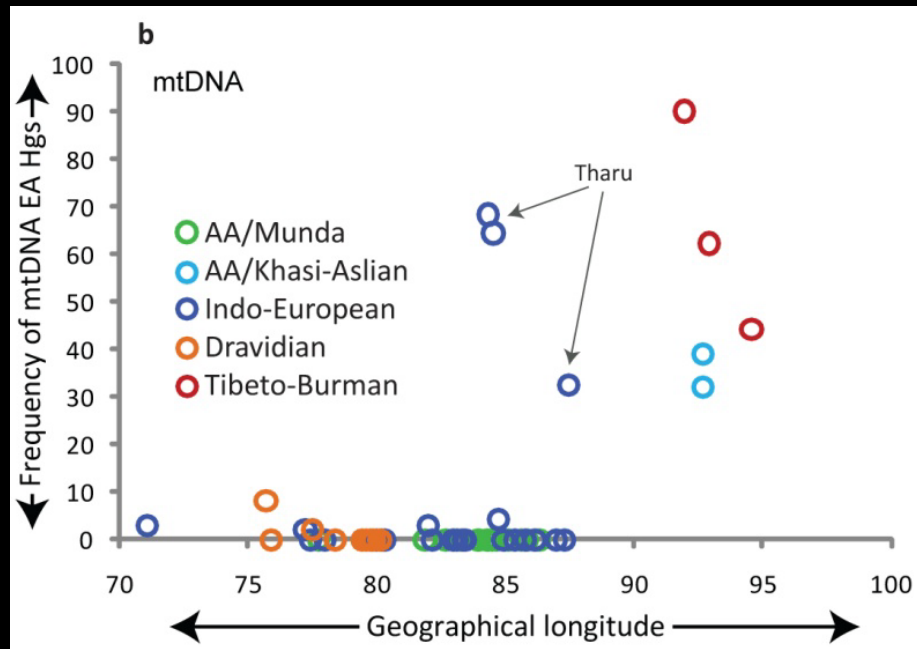
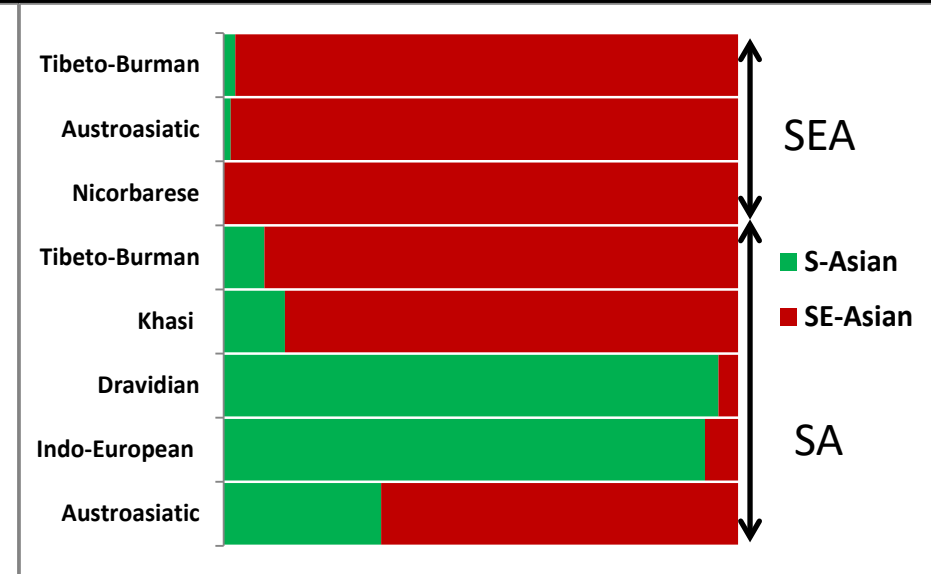
# Internal distribution of R7 among Munda speakers



## mtDNA (3454 samples)



## Y-Chromosome (6110 samples)




Status of Austro-Asiatic  
An exploratory study based on  
linguistic and biological

## HUMAN GENETICS

# Tracing India's invisible threads

Aravinda Chakravarti

**One measure of the extraordinary level of human diversity found in India is the use of 15 languages on its banknotes. The genetic underpinnings of that population diversity are yielding to whole-genome analysis.**

Vikrant Kumar & B. Mohan Reddy 

*Journal of Biosciences* 28, 507–522 (2003) |

201 Accesses | 32 Citations | 3 Altmetric

### Abstract

Among the most contentious currently in the Indian subcontinent. It has been Austro-Asiatic linguistic family are perhaps the oldest palaeoanthropological evidences suggest that the Indian population is a mixture of BP. Recent speculations, based on both linguistic and biological data, corroborate the aforesaid view. However, the representation of the constituent groups

### Austro-Asiatic Tribes of Northeast India: A Missing Genetic Link between

B. Mohan Reddy  B. T. Langstieh, Vikrant Kumar, T. Nalini, Lalji Singh

Published: November 7, 2007 • [https://doi.org/10.1371/jou](https://doi.org/10.1371/journal.pone.0011111)

Article	Authors	Metric
		

### Abstract

Introduction  
Results  
Discussion  
Materials and Methods  
Supporting Information  
Acknowledgments  
Author Contributions  
References

### Abstract

Northeast India, the only subcontinent and South peopling of East Asia. The oldest and spoken by the earliest expect that populations genetic link between India we analyzed mtDNA and Austro-Asiatic Khasi from North relevant Asian population India represent a genetic thereby advocating that populations from India t

The idea and shape of modern India was an invention of its twentieth-century political leaders, who crafted citizenship defined by civic and universalist, rather than ethnic or religious, criteria precisely because that citizenship is so diverse<sup>1</sup>. As Jawaharlal Nehru, the nation's first prime minister, wrote<sup>2</sup>: “[India] is four hundred million separate individual men and women, each differing from the other ... a bundle of contradictions held together by strong but invisible threads.” Who are these diverse peoples separated by caste, customs and language? Where did they come from, and when? What are the “invisible threads”, beyond claims on the state, that bind them? Studies of biological kinship, which search for the stories of ancestry marked indelibly in a

person's genome, help to provide answers to these questions because they illuminate that unwritten past<sup>3</sup>. The latest addition to our attempts to understand India through genes comes from Reich, Singh and their colleagues (page 489 of this issue)<sup>4</sup>, who arrive at some bold conclusions about its past population history from genome-variation studies.

The earliest occupation of the subcontinent was by Austro-Asiatic people about 60,000 years ago. They were dispersed and driven into smaller enclaves with the arrival of the Dravidian speakers around 3000 BCE (Before the Common Era, the Common Era marking the same divide as BC and AD). The latter people were themselves driven south with the arrival of the Indo-European speakers in about 1500 BCE.

## View, With Special and Structure

<sup>4</sup> Sangita Roy,<sup>2,4</sup> Sanghamitra Sengupta,<sup>1,4</sup> Badal Dey,<sup>1</sup> Monami Roy,<sup>1</sup> Bidyut Roy,<sup>1</sup> Soudhury,<sup>2</sup> and Partha P. Majumder<sup>1,5</sup>

*Institute, Calcutta 700 108, India; <sup>2</sup>Human Genetics & Genomics, India; <sup>3</sup>Crystallography & Molecular Biology Division, Saha Institute*

ona on 58 DNA markers (mitochondrial [mt], Y-chromosomal, and a large number of ethnically diverse populations of India. Our underlying unity of female lineages in India, indicating that the population is small; (2) the tribal and the caste populations are highly related; (3) the earliest settlers in India, providing support to one theory; (4) a major wave of humans entered India through the considerable genetic commonalities with the Austro-Asiatic tribals, shared a common habitat in southern China, but the two groups of autosomal haplotypes; (6) the Dravidian tribals were possibly related to the Indo-European-speaking nomads, but retreated to southern India by fission that resulted in founder and drift effects have left distinct population genetic signatures; (8) the upper castes show closer genetic relationships to those of southern India are more distant than those of northern India, attributed to a considerable obliteration of genetic histories of the region. We present no clear congruence of genetic and geographical or

genome.org. The following individuals kindly provided reagents, in the paper: C.S. Chakraborty, R. Lalthantluanga, M. Mitra, A. Thakur, M.V. Usha Rani, L. Jorde, K. Kidd, A. Merriwether, A.

as seen from an evolutionary

[View article](#)  
[Metrics](#)

Humans, *Homo sapiens sapiens*, evolved in Africa, and (ii) modern humans in other parts of the world, and (iii) the population came into India. India, therefore, served as a genetic melting pot. By studying variation at DNA level in the Indian population, we have provided evidence that mitochondrial

# The Principal Component (PCA) & ADMIXTURE Analysis

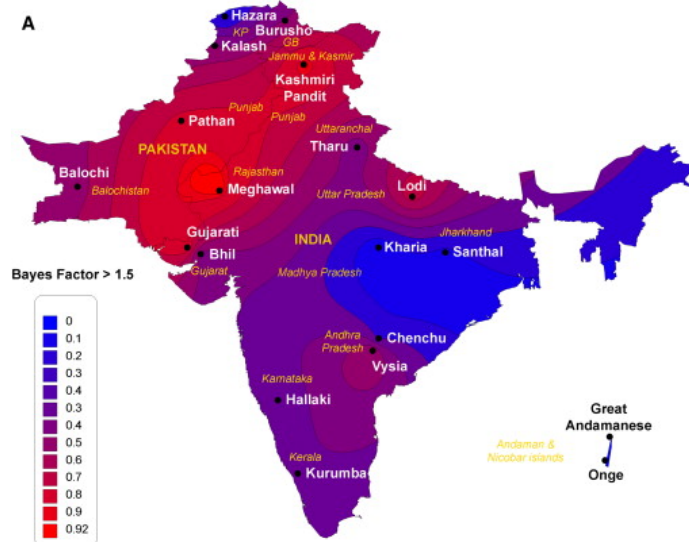
PCA is a statistical method for exploring and making sense of datasets with many measurements (which can be considered dimensions) by reducing the dimensions to the few principal components (PCs) that explain the main patterns.

Thus, the first PC is the mathematical combination of measurements that accounts for the largest variability in the data.

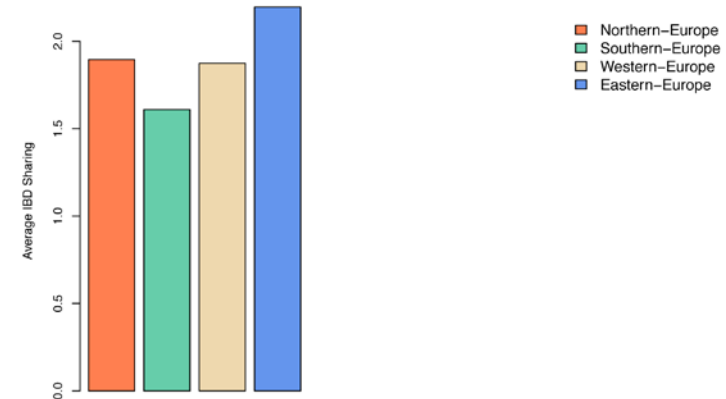
**ADMIXTURE** is a software tool for maximum likelihood estimation of individual ancestries from multilocus SNP genotype datasets. It uses the same statistical model as **STRUCTURE** but calculates estimates much more rapidly using a fast numerical optimization algorithm.



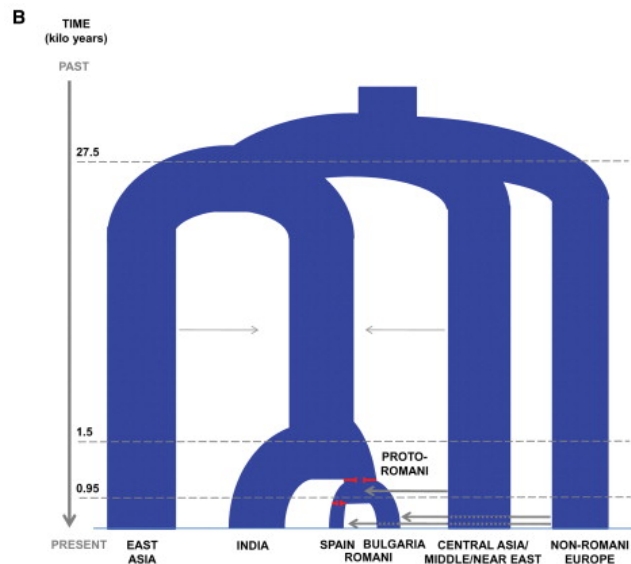
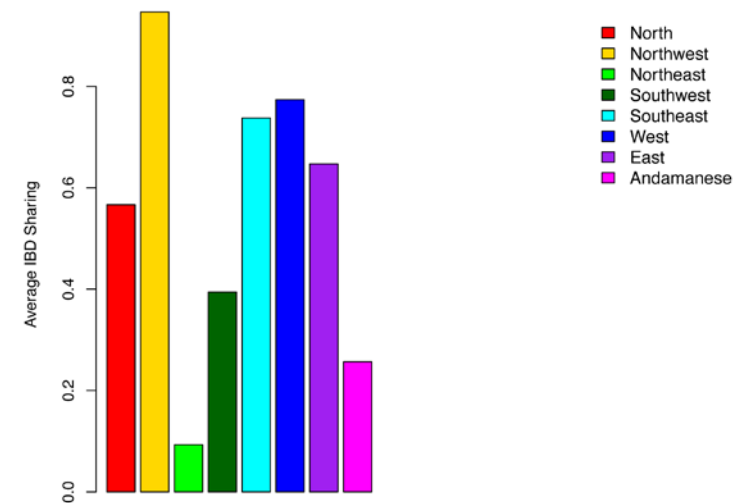
# Roma: from India to Europe



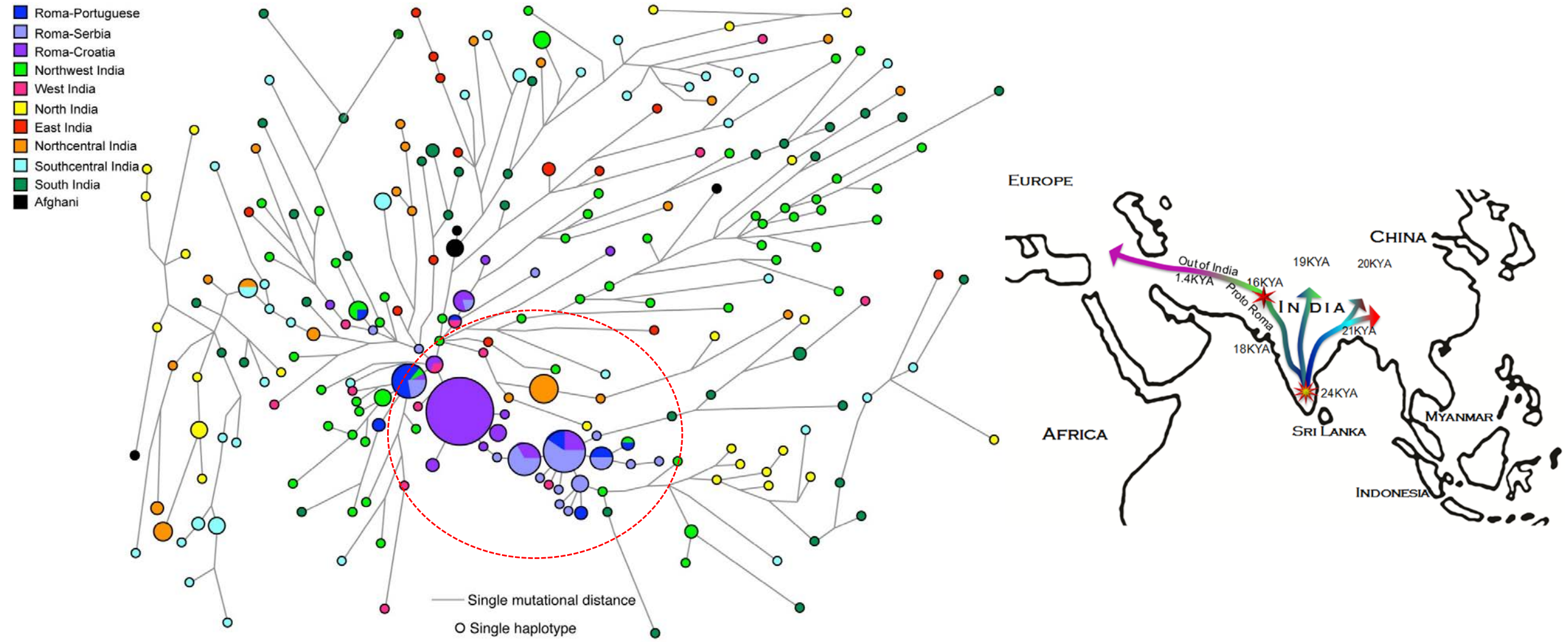
(a) Average pairwise IBD sharing with Europeans



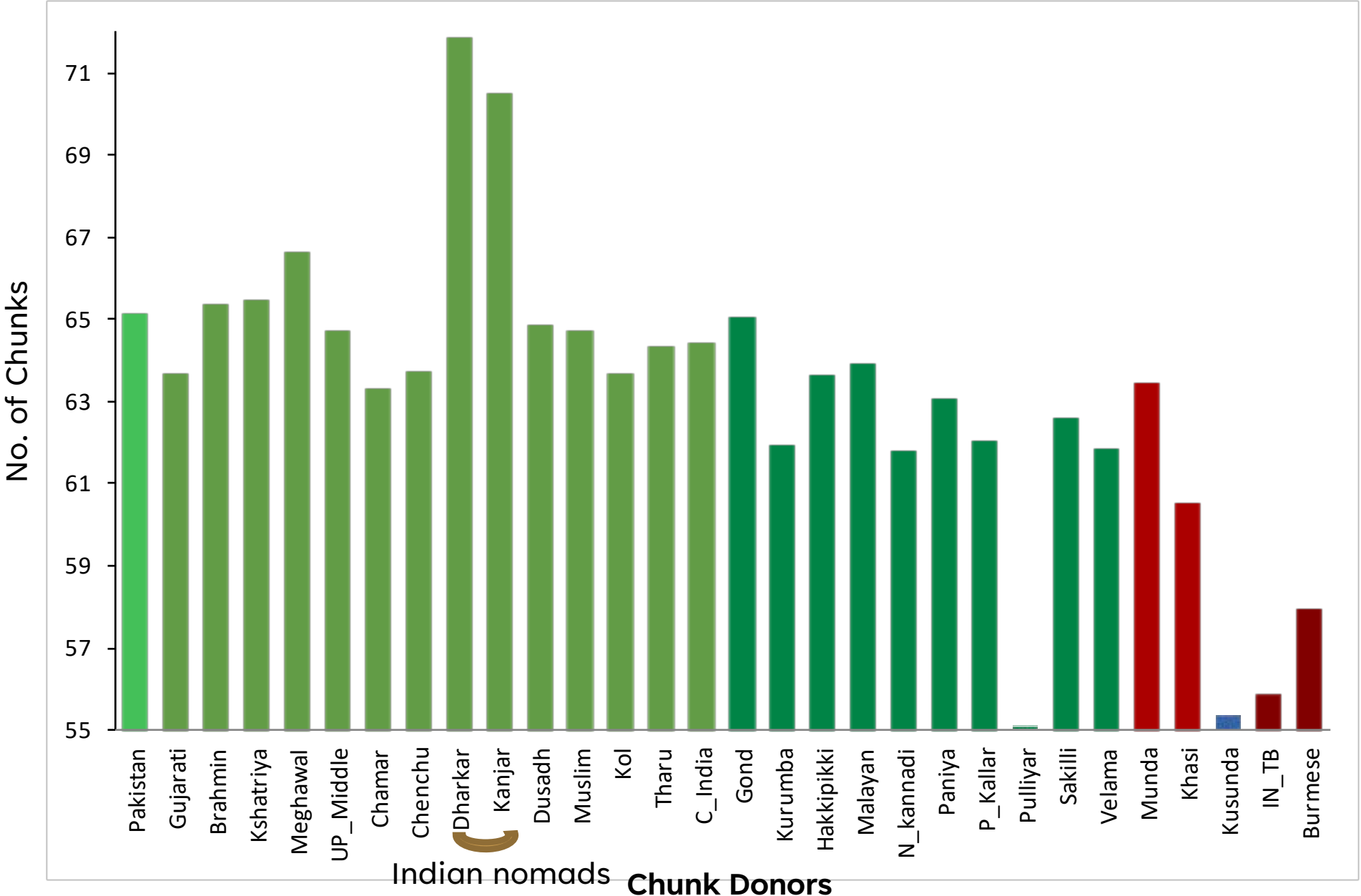
(b) Average pairwise IBD sharing with Indian groups



# Roma: from India to Europe (1.4KYA)



# Closest Indian population to Roma (fineStructure analysis)



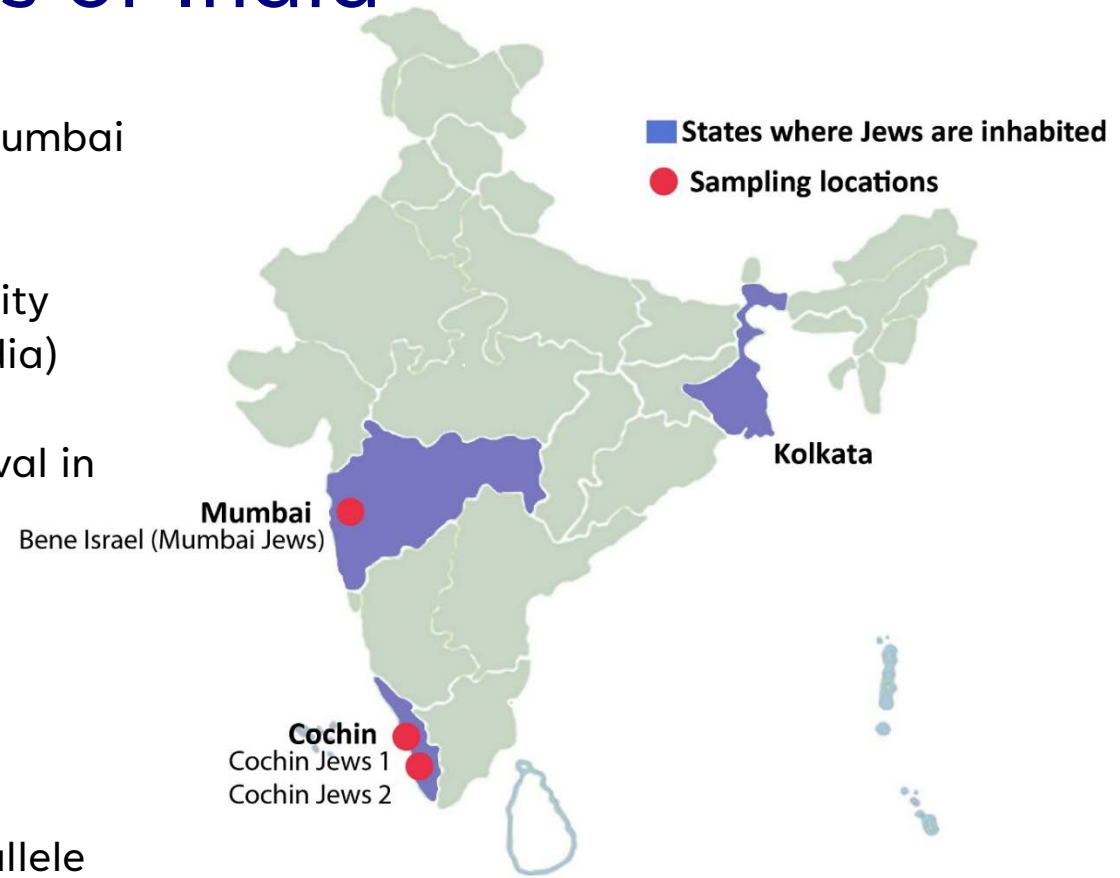
# Jews of India

## Population details

- Total ~5 - 6000 - half of them live in Mumbai (Bombay)
- June 2016- recognized as official minority religious group in Maharashtra state (India)
- Lack of written records about their arrival in India

## Earlier Genetic studies

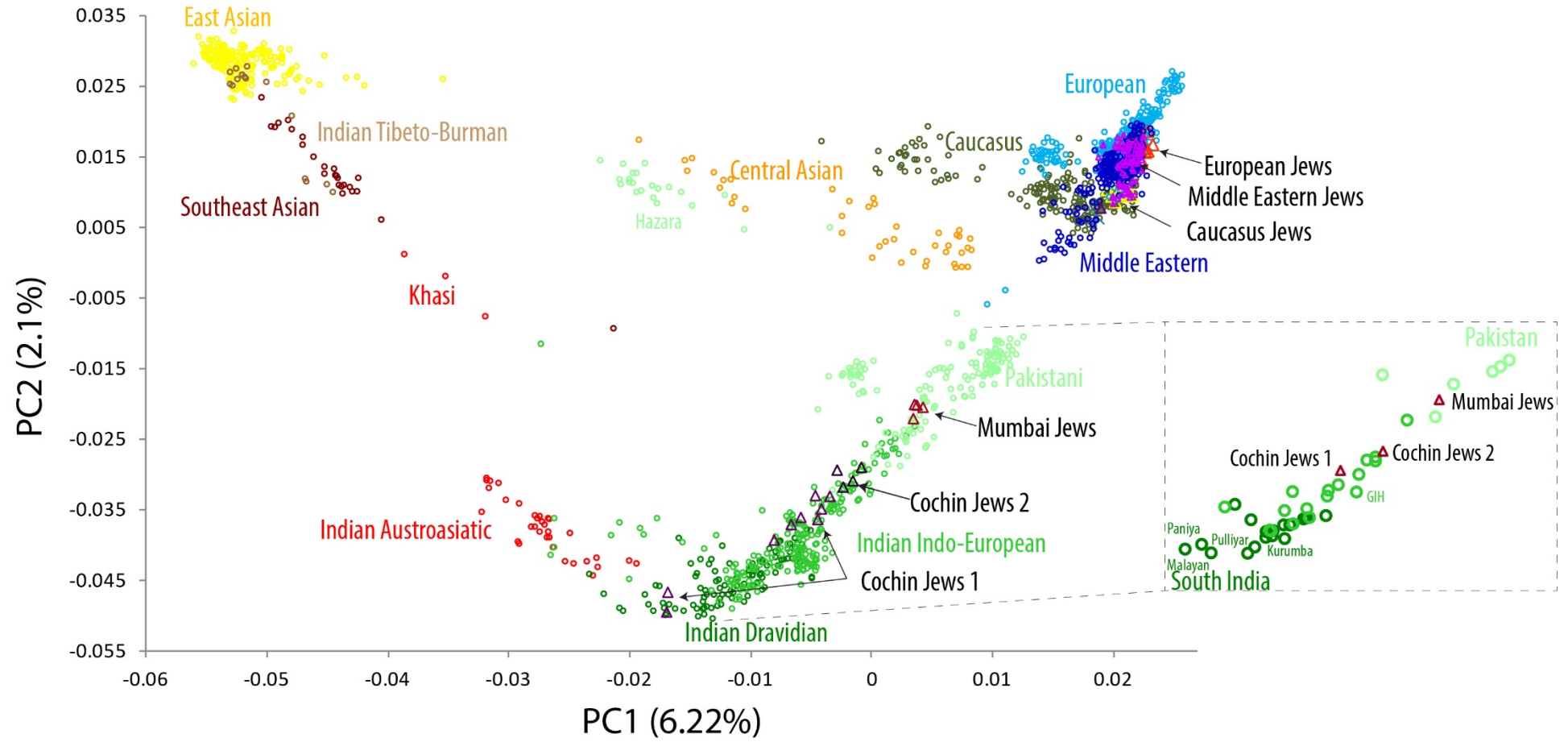
- ✓ Classical marker study suggests their allele sharing with Yemenite Jews and South Indian populations (Cohen et al 1980).
- ✓ mtDNA- South Asian; Y chromosome- Indian and Middle Eastern (Behar et al. 2008, 2010).
- ✓ Autosomal study has reported minor Middle Eastern specific ancestry (Behar et al. 2010).



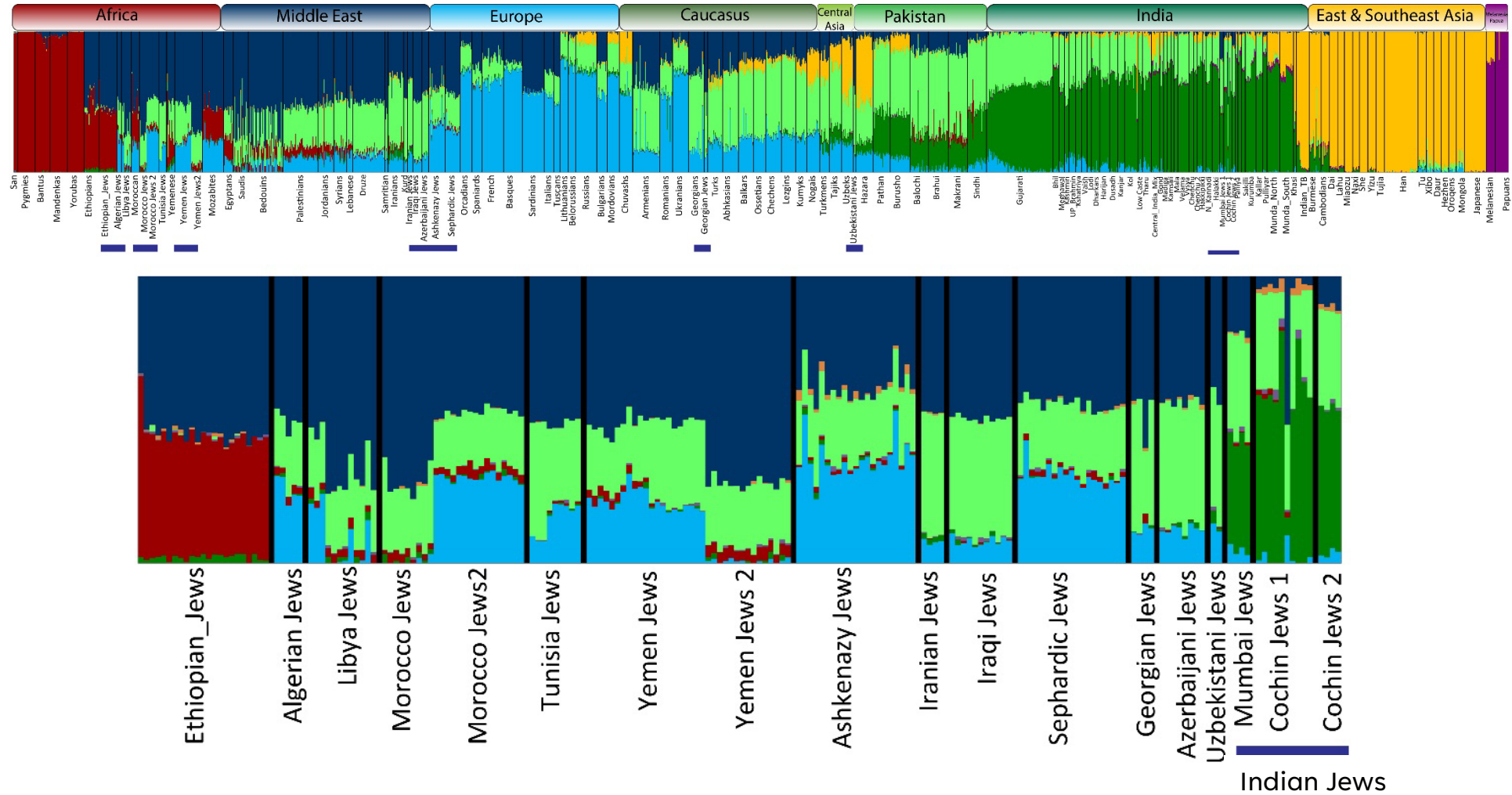
Present study  
mtDNA  
Y chromosome  
Autosome



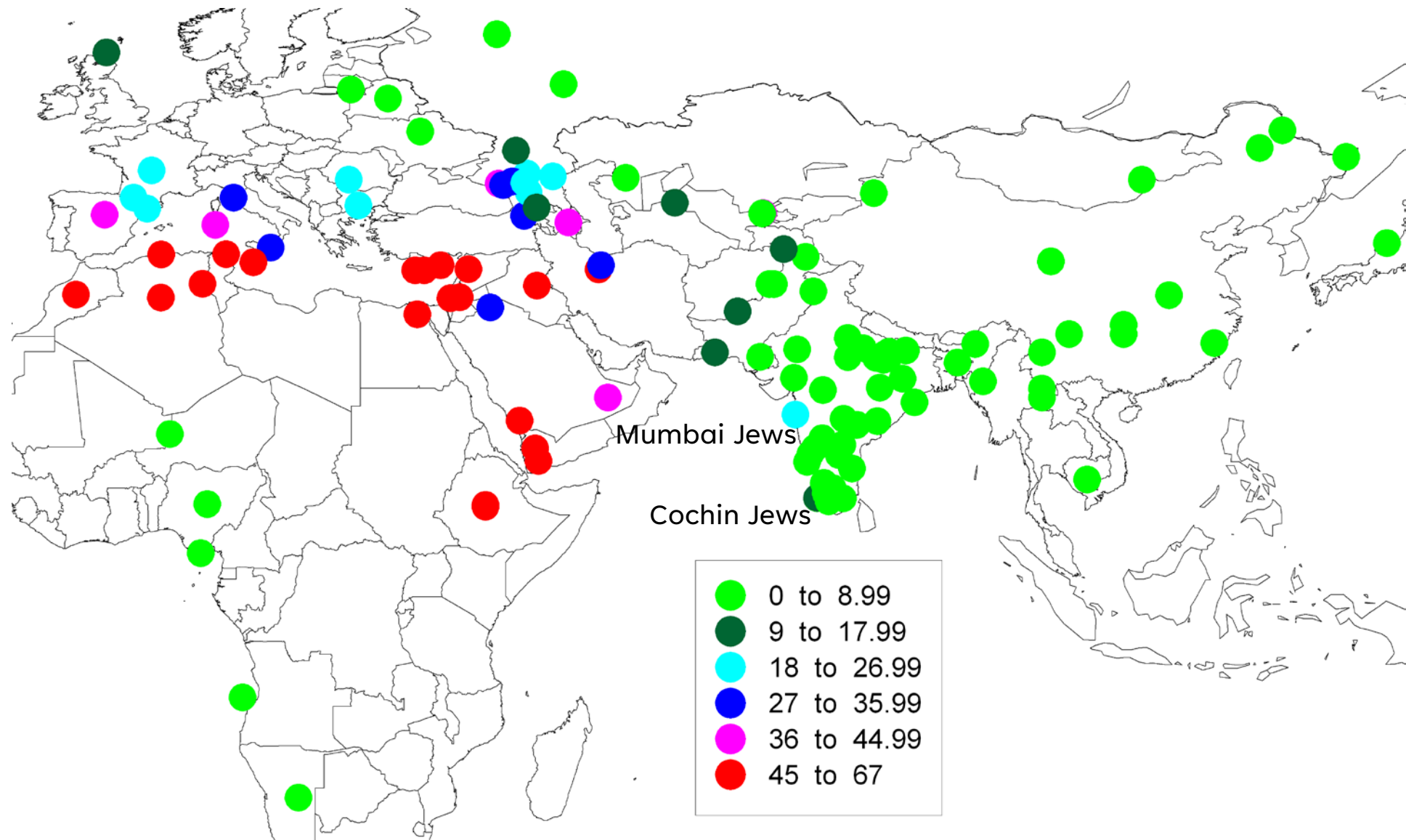
# Principal Component (PC) analysis



# ADMIXTURE analysis

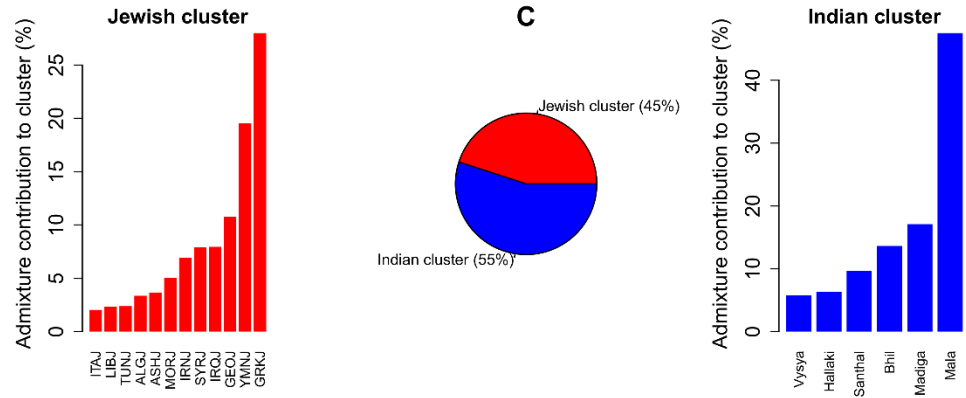
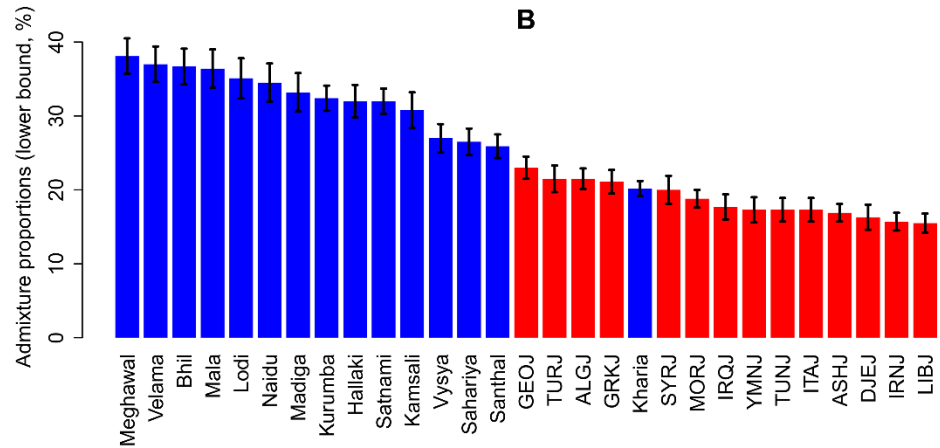


# Geospatial distribution of Middle Eastern specific ancestry



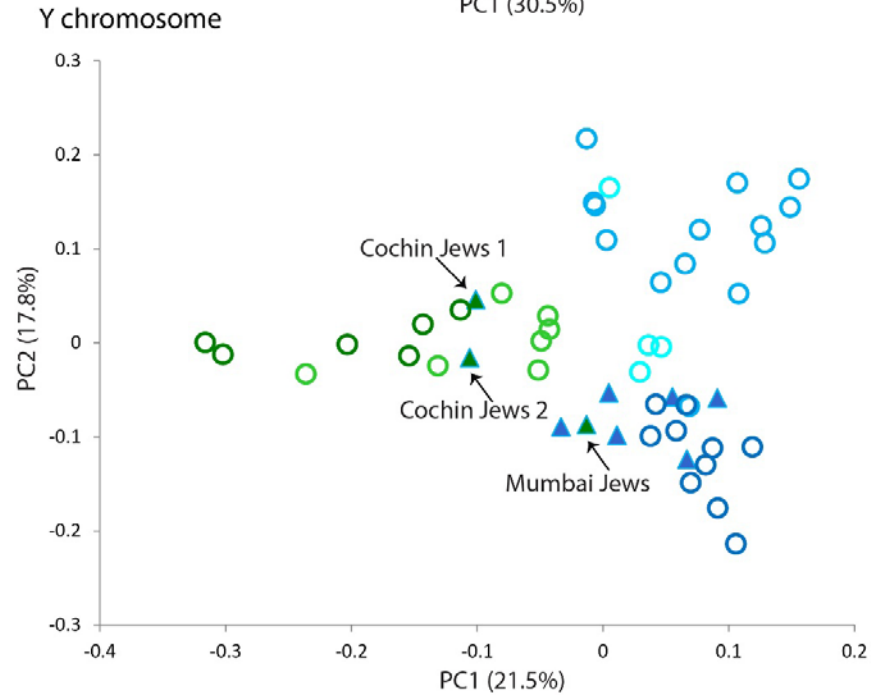
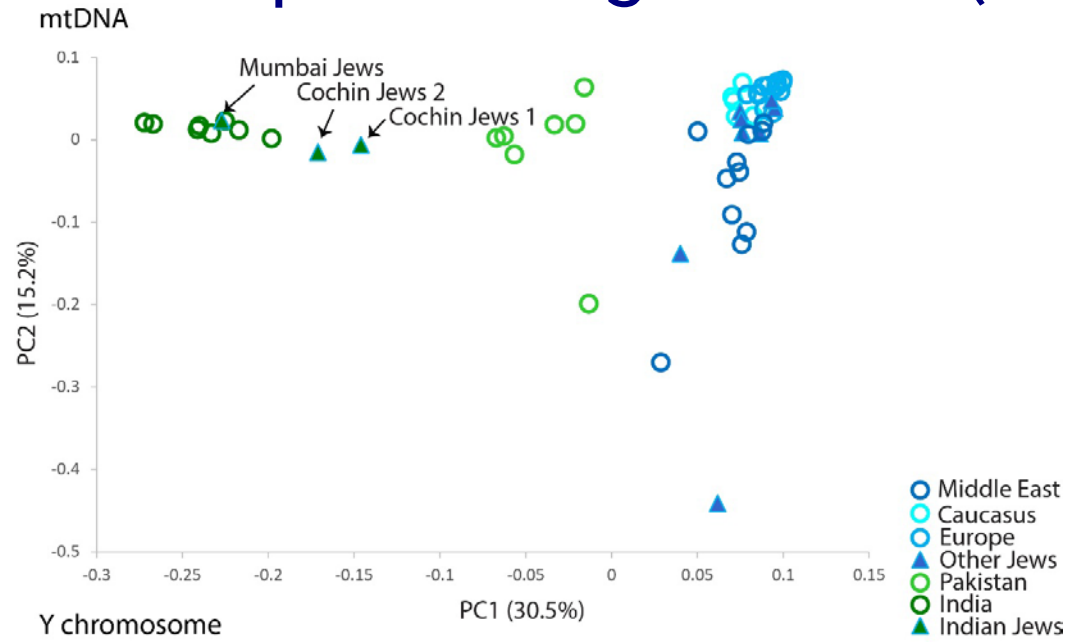
# Time of Admixture (Jews)

Admixed population	Surrogate population 1	Surrogate population 2	Generations from Admixture (SD)	p value	Z score
Mumbai Jews	GIH	Druze	37 (9)	6.9 x10 <sup>-5</sup>	3.98
Cochin Jews 1	Kurumba	-	37 (11)	1.6 x10 <sup>-4</sup>	3.23
Cochin Jews 2	Kurumba	Druze	53 (11)	3.1 x 10 <sup>-5</sup>	4.17

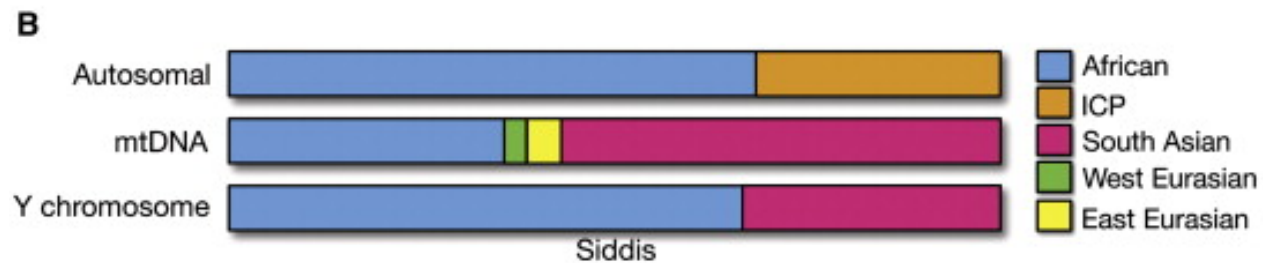
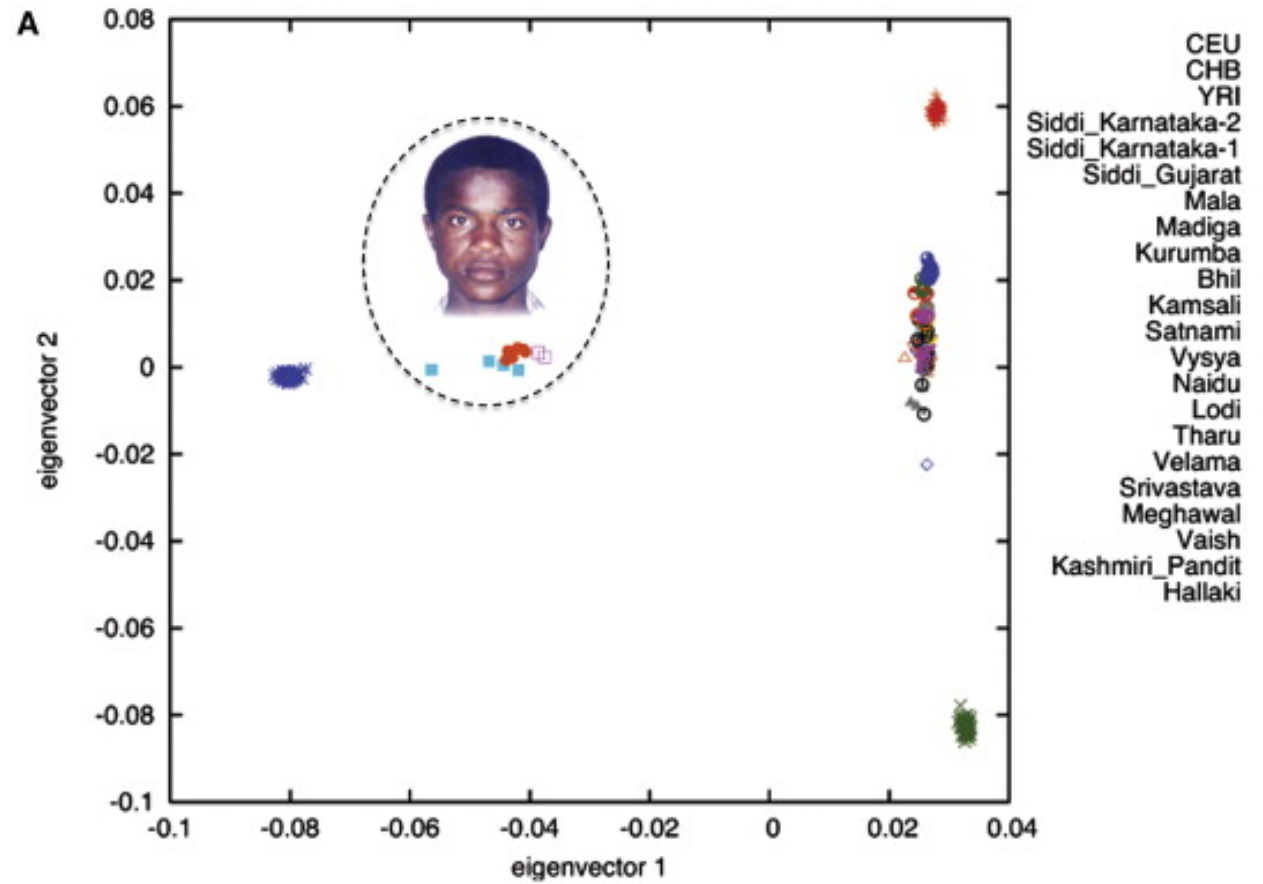




# Uniparental geneflow (Jews)



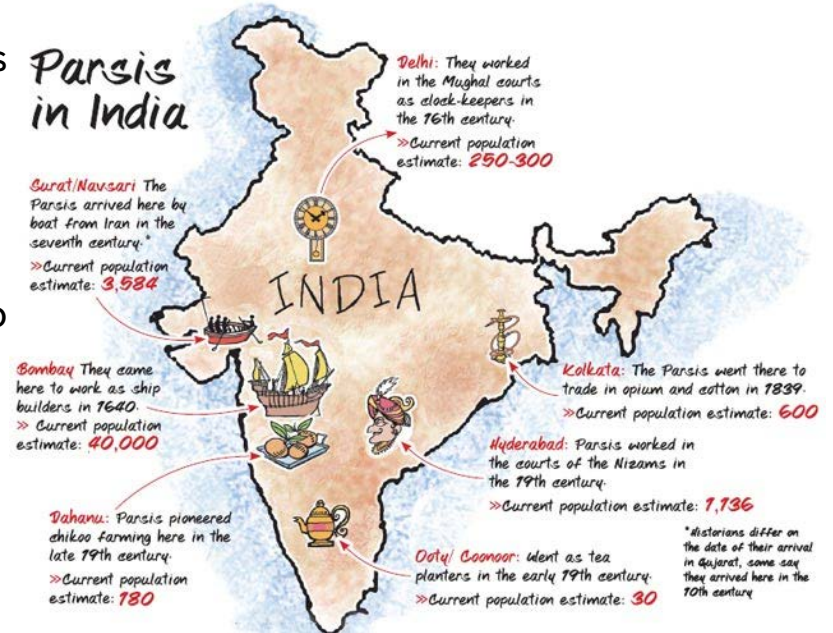
# Siddi: Slave trade from Africa (15<sup>th</sup>-19<sup>th</sup> Century)



# Like Sugar in Milk: Genetic history of Parsi population

## Population details

- Parsi is one of the smallest ethnoreligious communities of the world, 57,264 members- 2011 Indian census (down from 69,601 in 2001 and 114,000 in 1940).
- In the 7th century, the Zoroastrian Sassanian dynasty was threatened by Islamic conquest and a small group of Zoroastrians fled to Gujarat in present-day India, where they were called 'Parsi' (literally meaning 'people from Paras or Fars', the local term for Persia).
- *Qissa-e-Sanjaan*- narrative of the events.

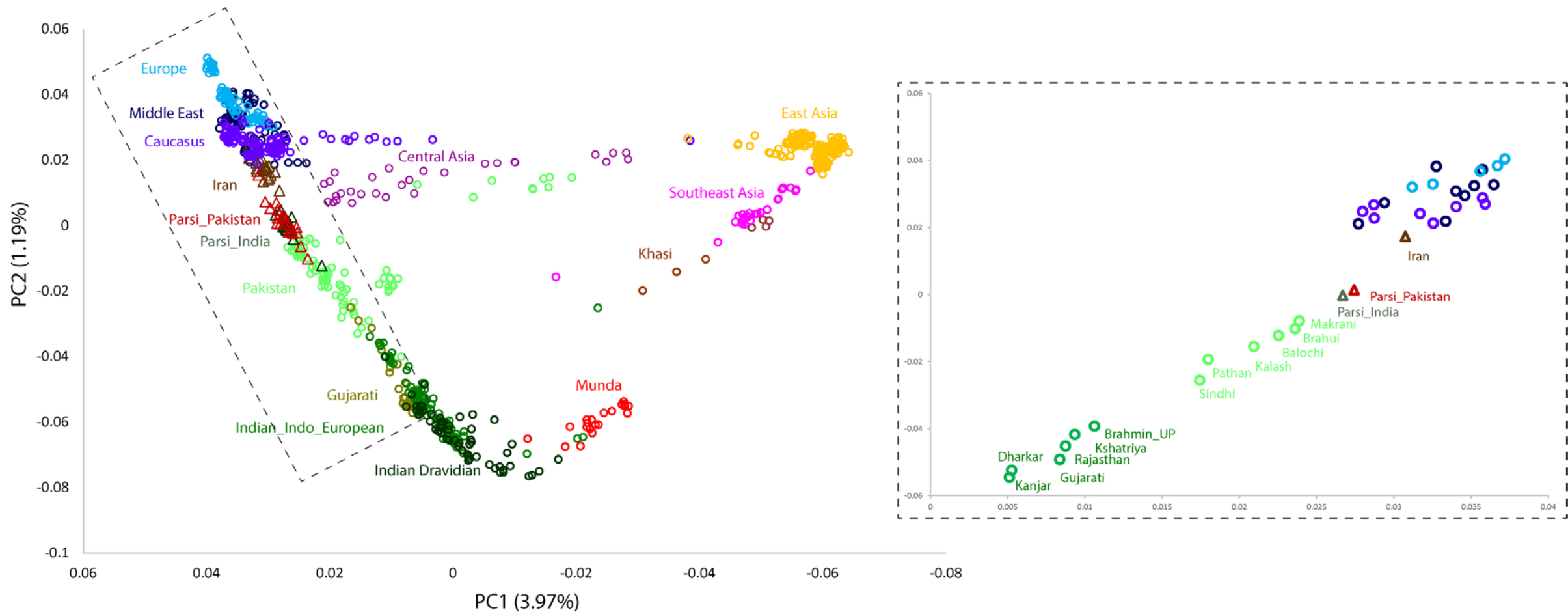


## Earlier Genetic studies

- Microsatellite and HLA study: intermediate position among South Asia and Middle East/Europe (Rosenberg et al. 2006; Mohyuddin and Mehdi 2005).
- Haploid DNA: 60% South Asian maternal and 100% Iranian lineages (Qamar et al. 2002; Quintana-Murci et al. 2004).
- Small sample sizes and low resolution genetic markers. Major Parsi group (living in India) has been often underrepresented, and likewise, no high-resolution autosomal evidence has been considered.

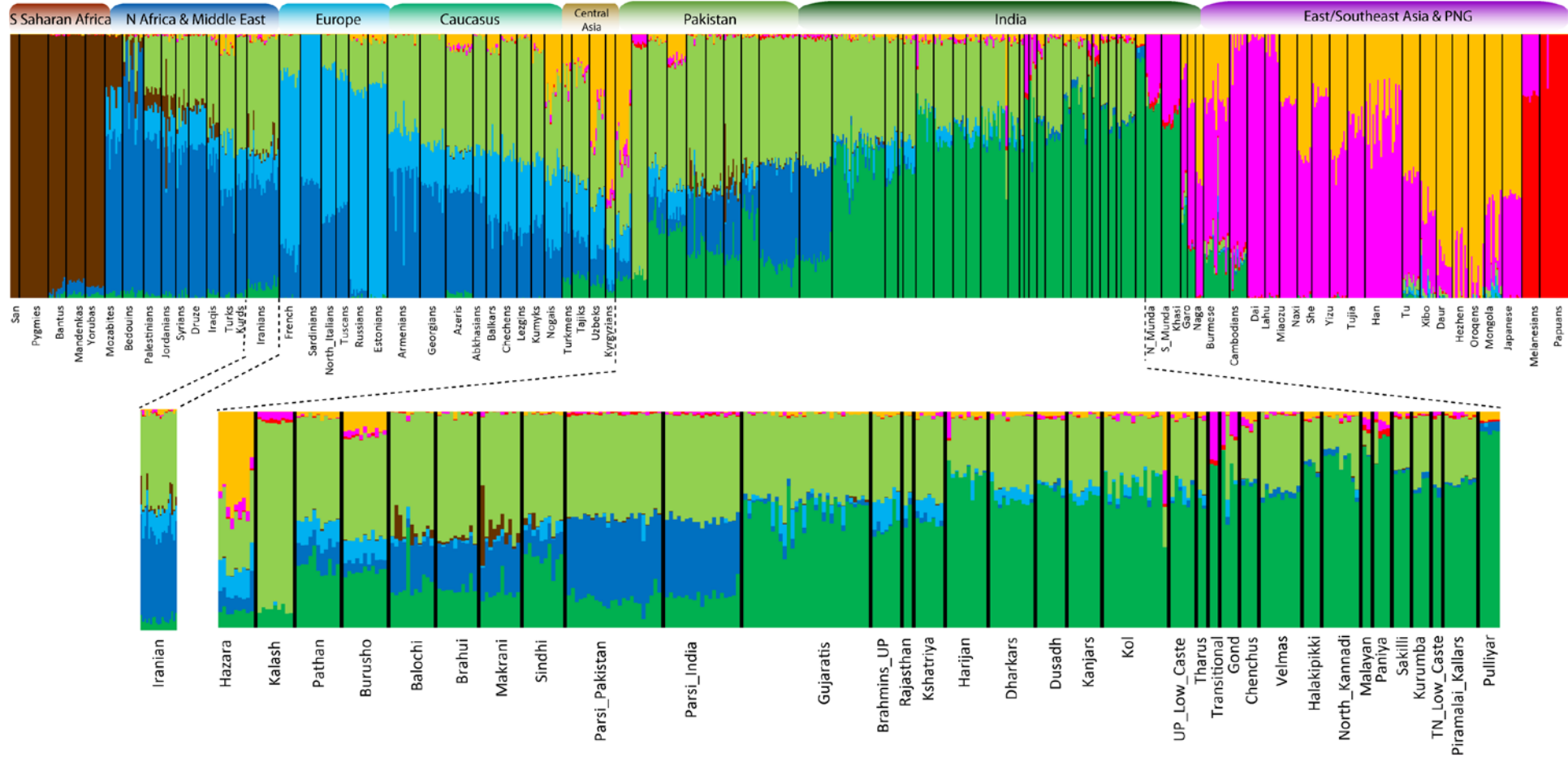
Present study  
mtDNA (aDNA)  
Y chromosome  
Autosome

# Principal Component (PC) analysis





# ADMIXTURE analysis



# Excess of Iranian ancestry among Parsis

The South Indian (SIND), and Iranian ancestry among Parsis and neighbouring populations

<b>Population (X)</b>	<b>SIND (SE)</b>	<b>Iranian (SE)</b>
Pathan	31 (2.4)	56.5 (1.8)
Sindhi	22.9 (2.7)	63.1 (1.9)
Parsi_Pakistan	6.4 (2.4)	76.6 (1.5)
Parsi_India	8.8 (2.5)	74.6 (1.7)
Gujarati	58.1 (2)	34.3 (1.6)

South Indian ancestry= Yoruba,Papua;X,French/Yoruba,Papua;South\_India,French

Iranian ancestry= Yoruba,Papua;X,South\_India/Yoruba,Papua;Iranian,South\_India

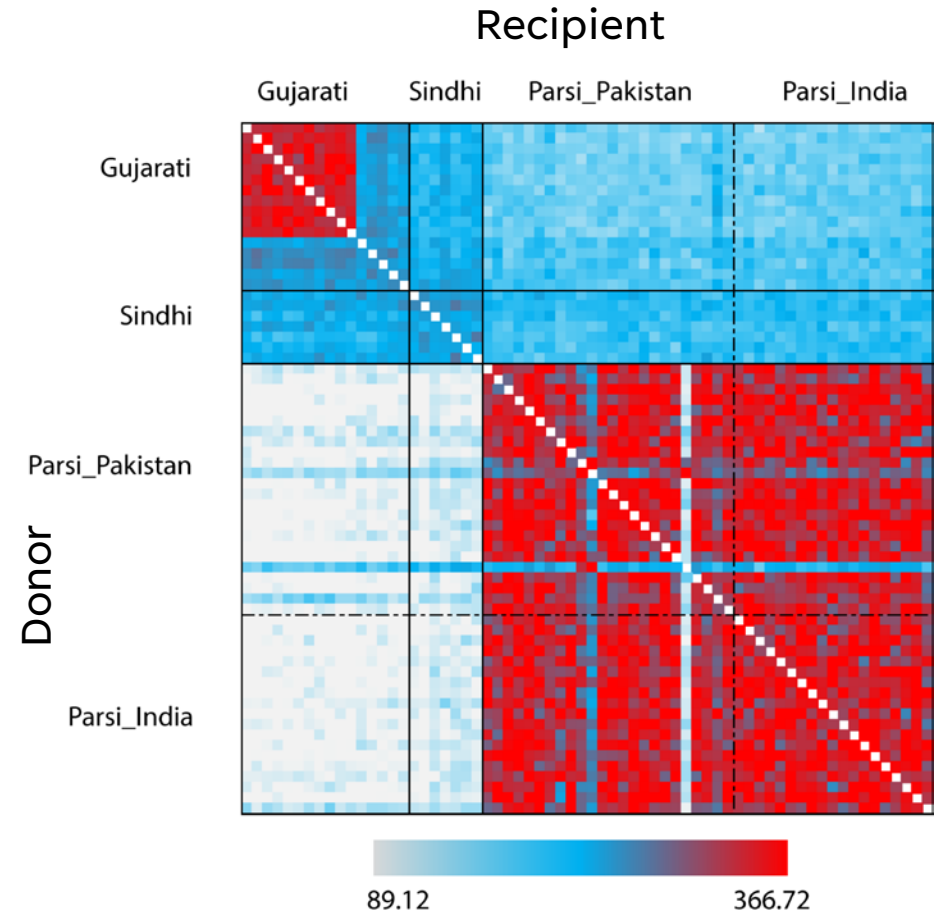
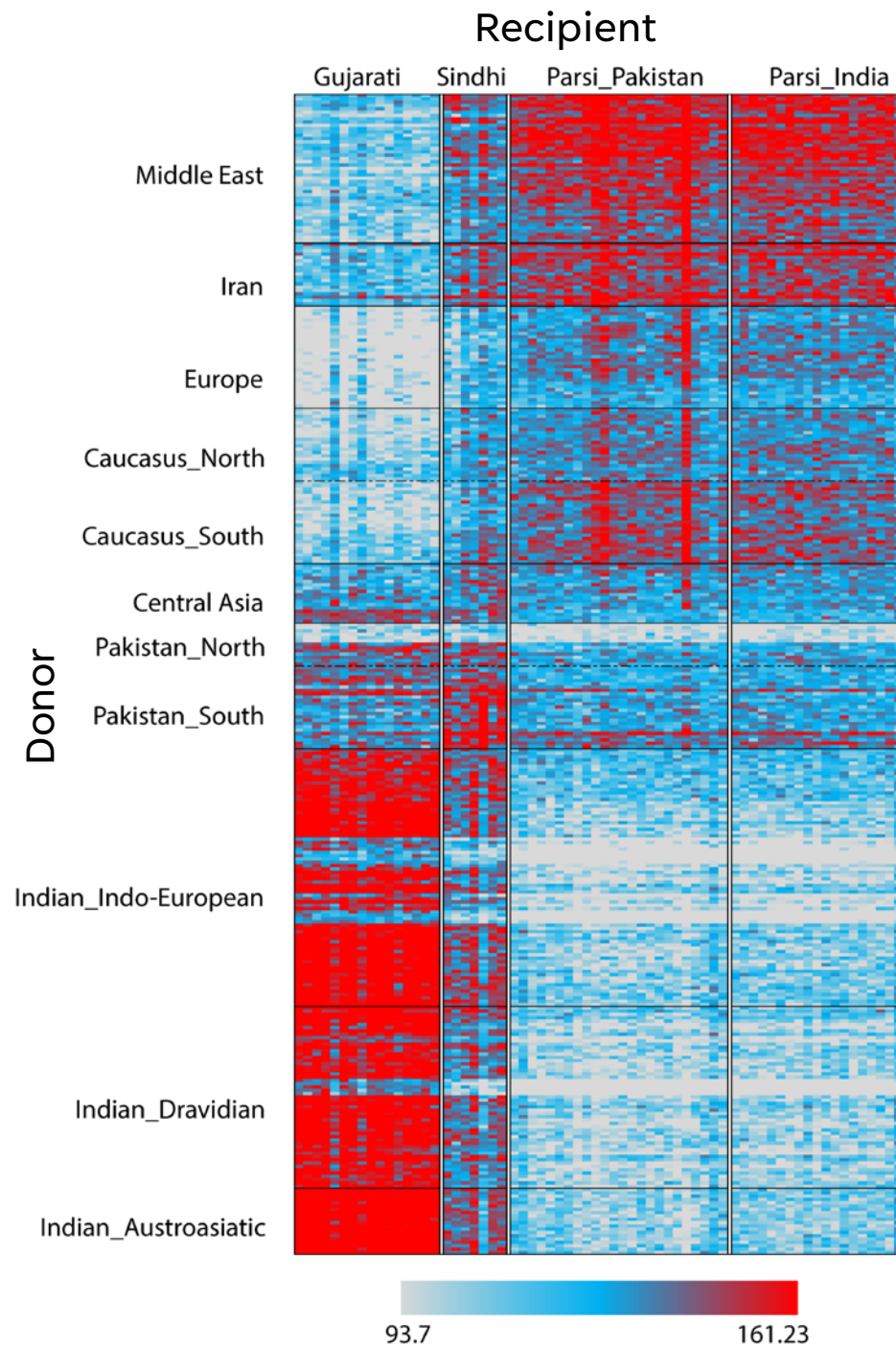
# Highest genflow from Iran Neolithic

The test of geneflow (*D* statistics) between Parsis, modern Iranian, Neolithic Iranian, Sindhi and Gujarati

Gp1	Gp2	Gp3	D value	Z score
Parsi_India	Parsi_Pakistan	Sindhi	0.0309	35.229
Parsi_India	Parsi_Pakistan	Iranians	0.0318	42.175
Parsi_India	Gujaratis	Sindhi	-0.0015	-1.89
Parsi_Pakistan	Parsi_India	Sindhi	0.0309	34.184
Parsi_Pakistan	Parsi_India	Iranians	0.0313	39.061
Parsi_Pakistan	Gujaratis	Sindhi	-0.002	-2.482
Sindhi	Parsi_India	Parsi_Pakistan	0	-0.036
Sindhi	Parsi_India	Gujaratis	-0.0038	-4.609
Sindhi	Parsi_Pakistan	Gujaratis	-0.0038	-4.688
Sindhi	Iranians	Parsi_Pakistan	-0.0124	-16.085
Sindhi	Iranians	Parsi_India	-0.0124	-15.308
Sindhi	Iranians	Gujaratis	-0.016	-17.608
Gujaratis	Parsi_Pakistan	Parsi_India	-0.0005	-0.898
Gujaratis	Iranians	Parsi_Pakistan	-0.0154	-21.245
Gujaratis	Iranians	Sindhi	-0.0211	-22.669
Gujaratis	Parsi_India	Sindhi	-0.0054	-5.982
Gujaratis	Parsi_Pakistan	Sindhi	-0.0058	-6.7
Iran_Neolithic	Sindhi	Iranians	<b>-0.0002</b>	<b>-0.141</b>
Iran_Neolithic	Gujaratis	Iranians	<b>-0.008</b>	<b>-5.075</b>
Iran_Neolithic	Parsi_India	Iranians	<b>0.005</b>	<b>3.2</b>
Iran_Neolithic	Parsi_Pakistan	Iranians	<b>0.0058</b>	<b>4.083</b>

D= (Gp1,Yoruba;Gp2,Gp3)

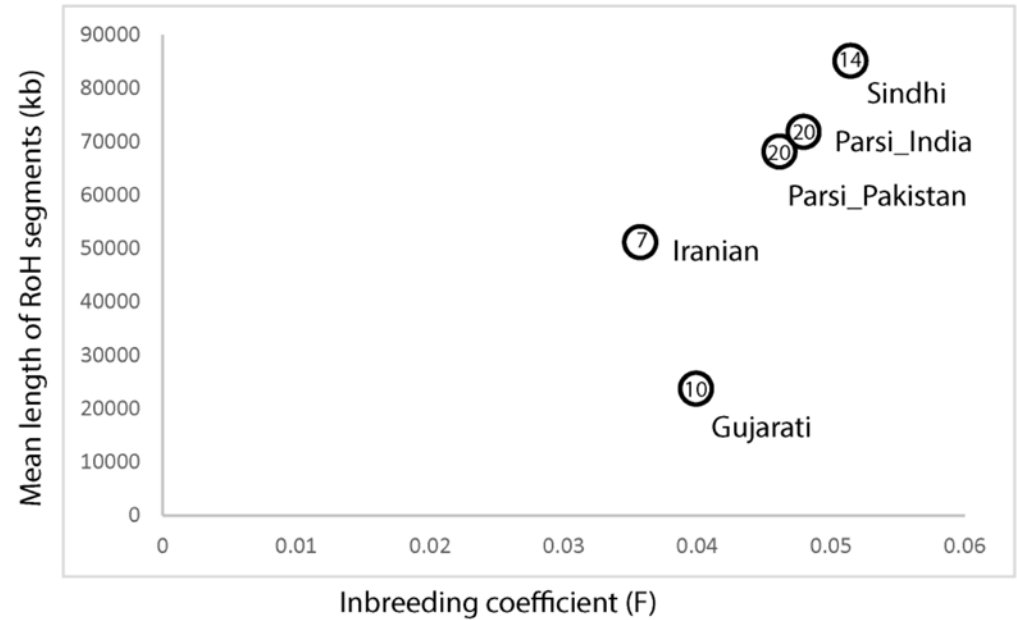
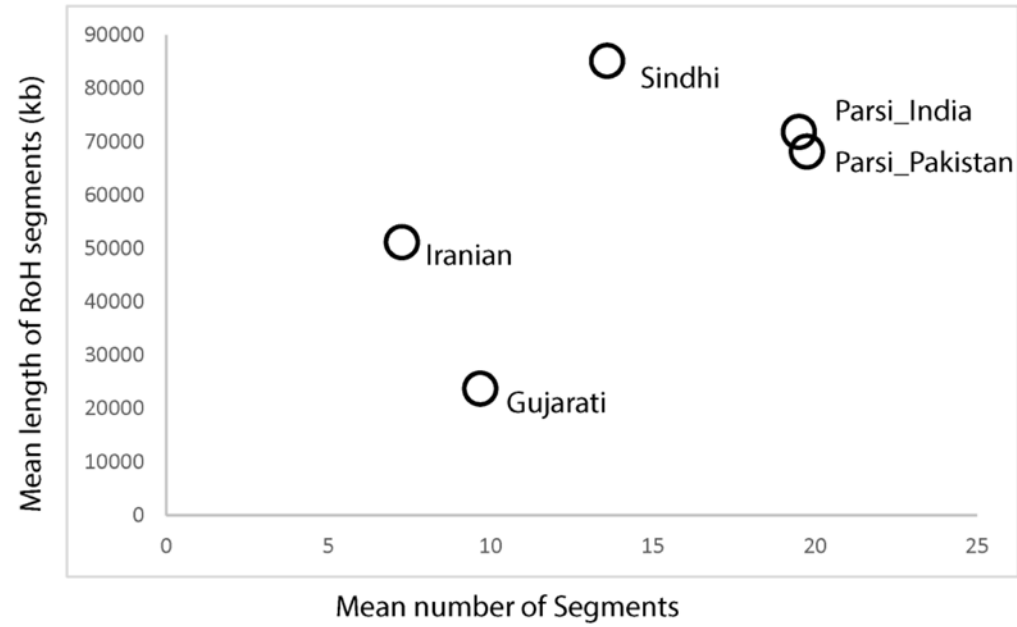
# fineSTRUCTURE analysis





# RoH and Inbreeding

1000kb



# Time of Admixture

The formal text of admixture using ALDER method

Reference 1	Reference 2	Admixed	Gen. Time	p value	Z score
Iranian	Gujarati	Parsi_India	38.26 ± 12.16	<b>0.0017</b>	<b>3.15</b>
Iranian	Sindhi	Parsi_India	32.96 ± 9.42	0.013	2.48
Iranian	Sindhi	Parsi_Pakistan	41.32 ± 8.93	<b>1.7x10<sup>-5</sup></b>	<b>4.3</b>
Iranian	Gujarati	Parsi_Pakistan	30.74 ± 14.04	0.029	2.19

20-38 generations (Lopez et al. 2017)

MALDER: 55 ± 9 Generations, and Single admixture event

# Example of Sri Lanka

iScience

Open access

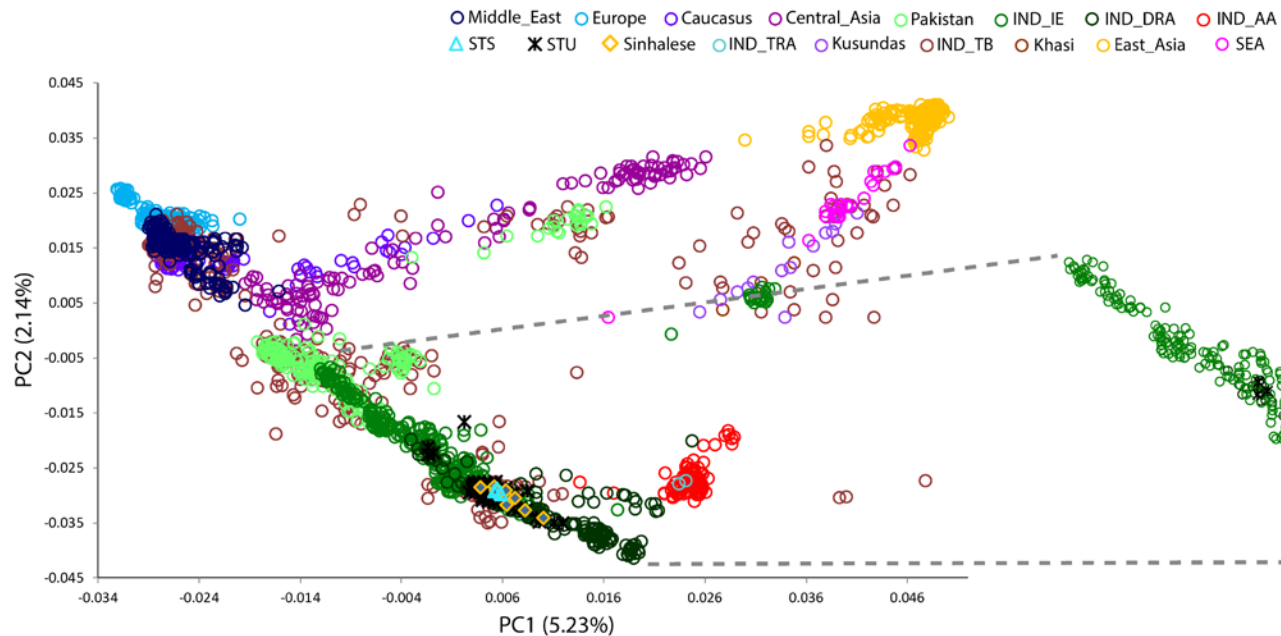
ARTICLE | VOLUME 26, ISSUE 10, 107797, OCTOBER 20, 2023

## Reconstructing the population history of the Sinhalese, the major ethnic group in Śrī Lankā

Prajival Pratap Singh <sup>6</sup> • Sachin Kumar <sup>6</sup> • Nagarjuna Pasupuleti • ... Niraj Rai  

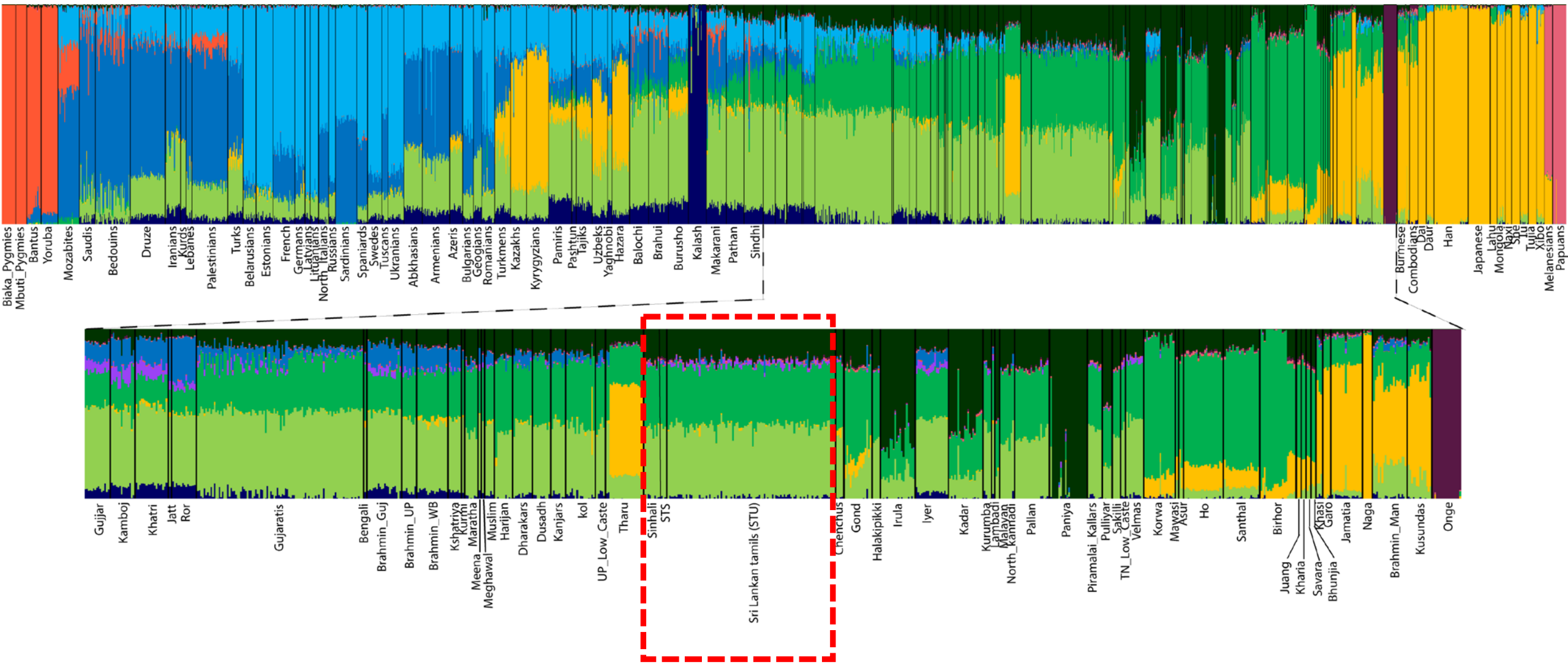
Gyaneshwer Chaubey  <sup>7</sup>  • R. Ranasinghe   • [Show all authors](#) • [Show footnotes](#)

[Open Access](#) • Published: August 31, 2023 • DOI: <https://doi.org/10.1016/j.isci.2023.107797>

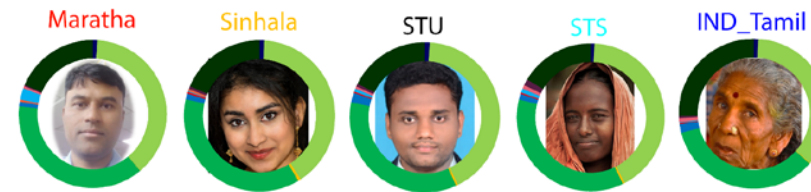
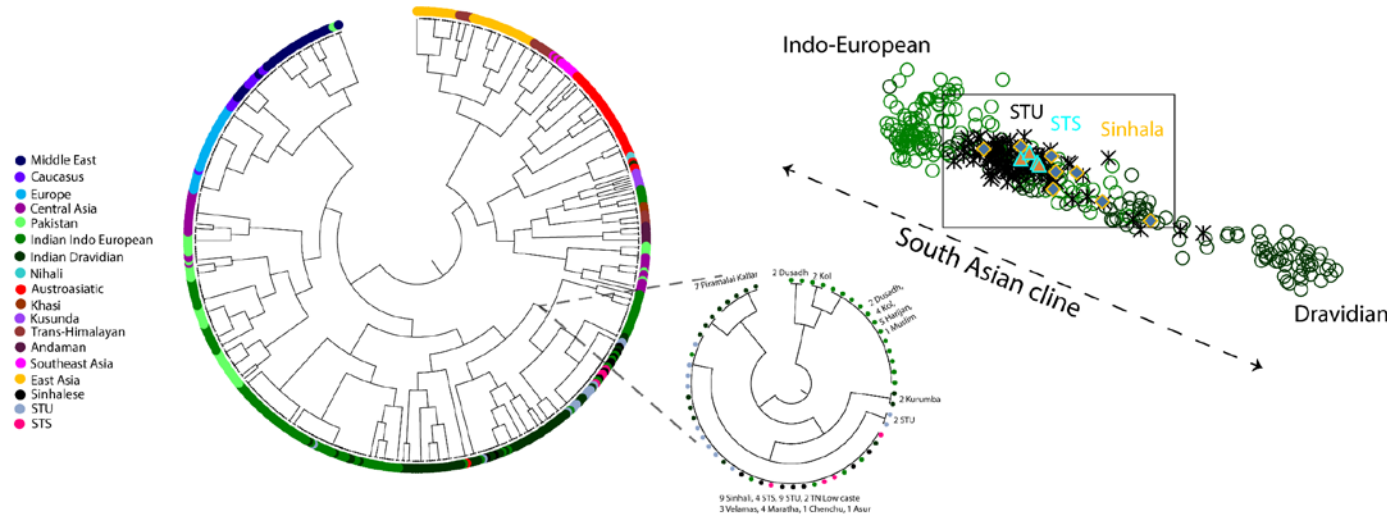
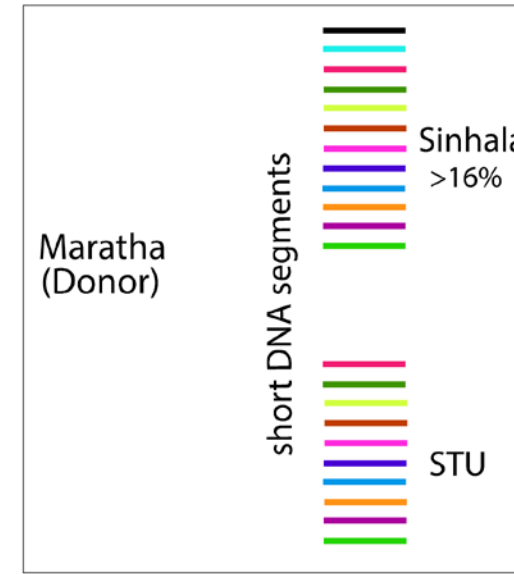
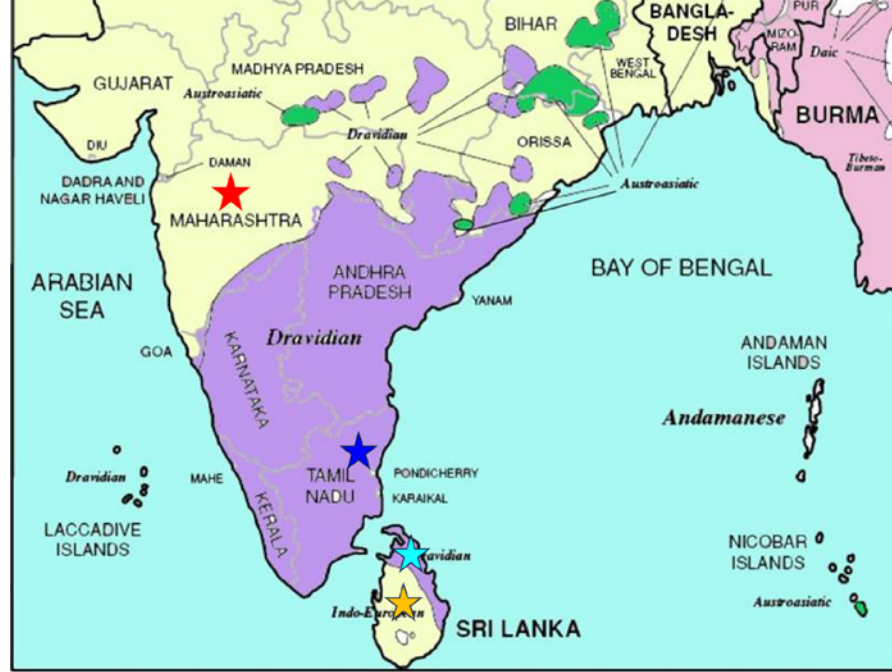


Sinhali: Indo-Aryan  
Tamil: Dravidian

# Genetic components







# Acknowledgements

Institute of Genomics (EBC), and Univ. of Tartu, Estonia

Prof. Richard Villems  
Prof. Toomas Kivisild (KU Leuven)  
Dr. Mait Metspalu  
Dr. Siiri Rootsi  
Dr. Ene Metspalu  
Ajai Kumar Pathak  
Dr Maere Reidla  
Jüri Parik  
Dr. Chandana Basu  
Dr. Monika Karmin



CSIR-CCMB Hyderabad, India

Dr. K. Thangaraj (group leader)  
Dr. Niraj Rai (BSIP)  
Dr. Manvendra Singh  
Satya Prakash  
Dr. Kamayani Singh  
Dr. Ashish Singh  
Dr Deepa Selvi Rani  
AG Reddy  
Dr. Vijay Kumar Singh

Stanford University, USA

Dr. Peter A Underhill

Bern University, Switzerland

Prof. George van Driem

University of Colombo

Prof. Kamani H. Tennekoon  
Dr R. Ranasinghe

University of Calcutta

Dr. Rakesh Tamang



SISEMINISTEERIUM

Eesti Teadusagentuur  
Estonian Research Council