Genetic Variation reveals migrations into the Indian subcontinent and its influence on the Indian society[1]

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]IBM T.J. Watson Research Center, 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.

Archaeological excavations revealed artefacts used by homo Erectus as long as 500-200ky. The moistening at the end of the last glacial period brought expanded subsistence; drying then spread agriculture from 8-5kya, marking some of the earliest migrations and expansions. Around 5ky, the Indus Valley civilization began with the much matured Harappan civilization, whose de-urbanization led to the initiation of the Vedic period. Following this, displacements followed as foreign rulers established dominance in the Indian subcontinent: from Greeks and Scythians, to the first seeds of Muslim invasions, followed by the Mughal Empire. In this phase, India had diverse rulers (including Afghans, Turks, and Mongols).

The migrations led to widespread admixture of the Indian population, influencing language, culture, caste endogamy, and metallurgical technologies, and more, resulting in a complex and differentiated structure. We set out to explore modern genetics correlating with migration routes into the subcontinent, and to study genomic variation in 48,570 SNPs genotyped in 1484 individuals, across 104 population groups. We propose, COGG (Correlation Optimization of Genetics and Geodemographics), a novel optimization method to model genetic relationships with social factors such as castes, languages, occupation, and maximize the correlation with geography. We calculated the shared ancestry between different caste groups in the subcontinent with other reference populations from Eurasia. We tested different migration theories into the subcontinent using a Linear Discriminant Analysis of redescription clusters and study recombination events shaping the gene pool.

Our results demonstrate that COGG gives us significantly higher correlations, with p-values lower than 10^{-8}. Identification of significant components among caste, language and genetics simplifies the complex structure. We identify varnas (Brahmins and Kshatriyas) to be closely related to reference Eurasian populations, whereas tribal groups show no shared ancestry with them and conclude that they resided in India before migration from Eurasia happened. We identify probable migration routes from Mongolia through Central Asia, and another via Anatolia into the subcontinent. Tibeto-Burman speaking populations share some ancestry with populations from East Asia; on the other hand, Austro-Asiatic speakers did not share ancestry with other Mon-Khmer language speaking populations.

[1] Bose A., Platt D.E., Parida L., Paschou P., Drineas P., *Genetic Variation reveals migrations into the Indian subcontinent and its influence on the Indian society*; (Abstract/Program #27, Platform Session #70, October 20, 2016, 9:15 am,). Presented at the 66th Annual Meeting of the American Society of Human Genetics, Vancouver, Canada.