New study ties India's genetic diversity to language, not geography

January 26 2021

Credit: Pixabay/CC0 Public Domain

The popularity of genetic and ancestry services like Ancestry.com and 23andMe attests that people care about where their ancestors originated. The underlying assumption is that the geography of one's forebears

affects one's genes today.

Historically, scientists have found that geography is the biggest driver behind the genetic diversity of a [population](). Now, new research from Purdue University indicates that while that may be true for European countries, it is not true for all other parts of the world—especially places like India, where language and social systems have strongly affected how and where people live. The [model]() the researchers developed to analyze India's population genetics will allow other researchers to analyze populations where genetics are not as closely tied to geography. Understanding the genetics of human populations helps scientists understand the history of human movement and cultures, and paves the way to understanding human health and susceptibility to disease.

Peristera Paschou, a population geneticist and associate professor of biological sciences at Purdue, studies [human genetic variation]() all around the world and led the study with Petros Drineas, associate head of Purdue's Department of Computer Science.

"Our genome carries the signature of our ancestors, and the genetic structure of modern populations has been shaped by the forces of evolution. What we are looking for is what led different groups of people to come together and what drove them apart." Paschou said. "To understand the genetics of human populations, we created a model that allows us to consider jointly many different factors that may have shaped genetics. Interdisciplinary research bringing together genetics and computer science was key to our work, as well as analyzing a comprehensive dataset that represents the diversity of the Indian subcontinent."

Many population analyses mostly rely on datasets from European-ancestry individuals living in Europe or North America; [genomic data]() for populations from other parts of the world is lacking. The data from

European samples showed that genetics correlates very closely with geography: If you know someone's genetics, you can guess where they are from, to within a few kilometers in some cases, and if you know where someone's ancestors came from, you have a close approximation of their genetic makeup.

Aritra Bose earned his doctorate at Purdue in both data science and genetics. Reading studies about how European genomes map onto geography, Bose, who was born and raised in Calcutta, thought, "Huh. That wouldn't work in India." India is home to more than 800 languages as well as a millennia-old caste system that regulates who can marry—and have children with—whom.

"I read these papers, and I thought, "How can I use this concept in a stratified population like India?'" Bose said. "I grew up there, I have an understanding of the castes and the languages, and the intricacies of the society that can affect genetics."

Former studies of the Indian population had shown that the European model of population genetics and geography failed in trying to explain Indian population genetics. Bose wondered if he could come up with a model that would take into account other factors affecting the Indian population, including the caste system, culture and language.

The model, and the conclusions the team of geneticists and data scientists reached using it, were just published in a study in the journal *Molecular Biology and Evolution*. Their study revealed that shared language, not geography, is the most powerful force in shaping gene flow in India.

Developing the model was not easy. Early on, Bose hit a roadblock with some of his equations and mentioned the problem to his mentor at IBM Research, where he was an intern at the time. Working with both his

doctoral advisors and several computer scientists from IBM Research, the team was able to craft a robust, flexible model.

Drineas, one of Bose's doctoral advisers, said: "I was intrigued by the interplay between genetics and socio-demographic factors in shaping the population structure of the Indian continent. It was exciting to see that our model detected spoken language as a major force in bringing people together in India, across geographic and social barriers. We were fortunate to have Aritra Bose, our former doctoral student (jointly advised with professor Paschou) work on this project, since he has extensive background in both the algorithmic and the human genetic sides of our research, as well the expertise to interpret our findings in the context of human genetic diversity within India."

The resulting model, the first to be able to take into account so many different variables, has been highly successful at analyzing the genetics of the Indian population, giving scientists a lens into how the Indian people moved into India and how various groups of people commingled. People who speak the same language—or even similar languages—tended to be much more closely related, even if they lived far apart geographically.

"It sheds light on how genetics work in our society," Bose said. "This is the first model that can take into account social, cultural, environmental and linguistic factors that shape the gene flow of populations. It helps us to understand what factors contribute to the genetic puzzle that is India. It disentangles the puzzle."

The data helps place India in context with the rest of the globe genetically. Indians who spoke Indo-European and Dravidian languages were more closely tied to Europeans, while Indians who speak Tibeto-Burman languages were more closely related to East Asians.

This type of interdisciplinary research, pairing data science with population genetics, and this model in particular, will help researchers understand the genetics of the human world, especially non-European countries with rich histories of diversity and migrations.

More information: Aritra Bose et al. Integrating Linguistics, Social Structure, and Geography to Model Genetic Diversity within India, *Molecular Biology and Evolution* (2021). DOI: 10.1093/molbev/msaa321

Provided by Purdue University

Citation: New study ties India's genetic diversity to language, not geography (2021, January 26) retrieved 26 April 2024 from https://phys.org/news/2021-01-ties-india-genetic-diversity-language.html