The origins of Indians

What our genes are telling us.

BY SRINATH PERUR

In December 2000, Kumarasamy Thangaraj flew to Port Blair, took an overnight ferry to Hut Bay on Little Andaman Island, and then travelled by road and a small motor-boat to Dugong Creek. "It was thrilling," he says, mentioning crocodiles in the water and the rough sailing on the last leg of the journey, where a rivulet meets the sea.

At Dugong Creek, he was amazed to see the Onge people spearing fish with sharpened wooden sticks. This semi-nomadic aboriginal people, of whom fewer than a hundred were left, alternate between foraging and living in government-run camps. They were the reason K. Thangaraj was here. Armed with permissions that had taken much time and effort to secure, he planned to collect blood samples from them.

K. Thangaraj made himself understood with the help of a social worker who spoke the Onge language. "They were looking at us and laughing," he recalls, but they indulged him, and he returned with blood drawn from around 40 members of the group. DNA was isolated from the samples and added to the DNA bank at the Centre for Cellular and Molecular Biology (CCMB), Hyderabad, where he has been a researcher for two decades.

CCMB's DNA bank has been put together with the help of researchers like K. Thangaraj, and students who return from vacations carrying blood samples and cheek swabs from their communities. It is the largest such repository in India. Samples from around 25,000 people, including some from India's most far-flung

reaches, are stored in trays of tiny barcoded vials kept refrigerated at -70°C. Associated with each sample is the donor's information: name, geographical coordinates, age, sex, language, caste or tribe or other ethnic grouping, and a signed (or often thumb-printed) consent form. Today, in addition to proving useful for research on medicine and health, this database is casting light on a contentious period of Indian history to reveal who we are and where we come from.

The human body—hair, skin, muscle, organs, blood, bone—is made of trillions of cells (37.2 trillion, according to one recent estimate). The genetic recipe for an individual is contained in the nucleus of most cells. This recipe is in strands of DNA packed into 23 pairs of chromosomes. The DNA from a single microscopically small cell's nucleus would extend to several feet in length if—as science writers are always threatening—it is unwound and laid out end to end.

This paired filament would be thousands of times thinner than a human hair and shaped like a twisted ladder. Molecules that form the rungs of this ladder are the letters in which DNA's information is written. There are four of these—named adenine, guanine, cytosine and thymine—and so DNA can be transcribed using only four characters: A, G, C, and T. For example, a small portion of the genetic code for haemoglobin, which carries oxygen and makes our blood red, reads:
ACTCCTGAGGAGAAGTCT.

Written like this, it would take more than three billion characters for all the DNA from a cell's nucleus to be transcribed. A small amount of DNA—a little over 16,500 characters—is present inside cells but outside the nucleus, in bodies called mitochondria.

Mitochondrial DNA or mtDNA has a special property: it is passed on unchanged from mother to child, and so is an ancient inheritance down the maternal line. But occasionally there is a random mutation in mtDNA—a misprint in the recipe, say, an A turning to a T—that gets passed on and marks all of a woman's descendants.

Since the rate at which mtDNA undergoes mutation is known, it also acts as a time-keeper of sorts. By finding out when and where and in which order mutations took place, studies of mtDNA have allowed the creation of entire family trees of human populations, along with a reconstruction of the geographical paths they took as they peopled the world.

***

The history of all humankind begins approximately 2,00,000 years ago when the first anatomically-modern humans are thought to have appeared in Africa. Then, around 60,000 years ago, a band of people ventured out of Africa, into the Middle East, branched out into India and Europe, and ultimately settled all over the planet, replacing other early human populations. These new settlers changed as they adapted to different conditions, as they migrated and interbred in complex ways. Eventually, they gave rise to the wide variety of humans found today across the earth: dark-skinned, light-skinned, red-haired, blue-eyed, able to digest milk as adults, able to resist some diseases but prone to others, and so on.

In 2005, K. Thangaraj and his colleagues at CCMB published their findings about the origin of Andaman islanders in the journal *Science*. The Onge turned out to have surprisingly unmixed origins. They had likely lived isolated in the islands since the arrival here of the first group of humans out of Africa. There were mutations in their mtDNA that were found nowhere else in the world. These mutations must have originated here and not spread. The Onge were an untouched link to the earliest humans who settled the planet.

Among those who noticed CCMB's work was David Reich, a geneticist at Harvard Medical School with an interest in studying how human populations had mixed in the past. He approached CCMB with a view to working together on the genetic history of indigenous Andamanese.

"This has been a wonderful collaboration," Reich writes by email when asked about working with CCMB. Their comparison of people such as the Onge with diverse ethnic groups on the Indian mainland has now led to important insights into the ancient history of Indians.

All humans carry DNA from both parents in their body cells. The nucleus in a cell contains two sets of 23 chromosomes, one from our mother and one from our father. When a man's body produces sperm, or when a woman's produces eggs, bits of DNA from that person's parents are spliced together to create a single recombined version. Their child will then have DNA assembled from the DNA of his four grandparents, and thus, generation to generation, we carry our entire family tree in our genes.

It is relatively straightforward to draw conclusions about our origins by looking at mtDNA because it passes unchanged from mother to daughter. Every change observed is a significant marker. This is also the case with another part of DNA called Y-DNA, that passes unchanged from father to son. But the rest of DNA recombines every generation, and it is far trickier to make sense of ancestry from the wealth of jumbled information that is the entire human genome.

Two people from different places or ethnic groups will possess characteristically different markers in their DNA. "You can think of their DNAs as being two long strips of paper, one red and one green," K. Thangaraj tells me on the phone from Hyderabad. If those two

When a man's body produces sperm, or when a woman's produces eggs, bits of DNA from that person's parents are spliced together to create a single recombined version. eir child will then have DNA assembled from the DNA of his four grandparents, and thus, generation to generation, we carry our entire family tree in our genes

people have a child together, the recombined DNA can be thought of as a single strip with alternating stretches of red and green, a sort of bar code of history and ancestry.

By looking at the lengths of stretches of the two colours and how often they occur, it can be estimated how much ancestry each colour contributed, and how many generations ago those colours interbred. "Of course in reality it's a little more complicated than that," says the CCMB researcher. "And that's where David Reich comes in."

***

Reich has been working for over a decade on developing statistical methods and tools to analyse population mixtures. When the genome of Neanderthal man—a close, early relative of humans—was sequenced, Reich helped compare it with DNA from modern humans and found that all non-African DNA contained a small amount of Neanderthal DNA. The conclusion was startling: humans must have inter-bred with Neanderthals on their way out of Africa. Earlier, he had looked into the genome of African-Americans and found a significant European component in their genetic history with implications for their susceptibility to certain diseases.

For their work on India's population history, Reich and his CCMB collaborators tracked hundreds of thousands of markers in all the DNA samples they studied, a level of detail several times greater than previous genetic studies of Indian populations. This allowed for a more fine-grained measurement of genetic differences and similarities between groups of people. Using samples from the CCMB DNA bank (as well as some data from other researchers and projects), they found that samples from individuals who came from the same ethnic group tended to share more markers and cluster together.

Further, when they compared different groups to a European reference they noticed an interesting pattern. Almost all Indian groups had inherited varying portions of their ancestry from a population related to western Eurasians.

Most Indians alive today are descended from a mixture of two very different populations, Reich and colleagues reported in *Nature* in 2009 based on a study of 25 ethnic groups. These two populations—the red and green of the earlier analogy—were given the names Ancestral North Indians (ANI) and Ancestral South Indians (ASI).

The ASI, likely aboriginal inhabitants of India since no trace of them is found outside the subcontinent, were a sister population of the Onge. The two must have diverged after being separated, one on the mainland, one on the islands.

The ANI showed genetic similarities with Europeans, Middle Easterners, and Central Asians. Some ANI ancestry was present in almost all Indian groups, but the percentage was found to be greater in the north of India and lesser in the south—for example, Kashmiri Pandits could trace about 70 per cent of their ancestry to the ANI people, and the Mala, a Dalit community from Andhra Pradesh, around 40 per cent.

Broadly, groups that spoke Indo-European languages and were traditionally considered upper-caste had a larger ANI component. No groups in mainland India were seen with only ASI ancestry. The Onge were the only group studied that showed absolutely no trace of ANI ancestry. At some time in the past, these two very different populations had inter-bred, and at some later point the castes, clans, communities and tribes we see now had formed as endogamous groups that only married within themselves.

Most Indians alive today are descended from a mixture of two very di erent populations, Reich and colleagues reported in 'Nature' in 2009 based on a study of 25 ethnic groups. ese two populations—the red and green of the earlier analogy—were given the names Ancestral North Indians (ANI) and Ancestral South Indians (ASI)

It was still unknown when exactly these populations had mixed. Those details came in August this year in the *American Journal of Human Genetics*. K. Thangaraj and Reich's groups had assembled data from 73 different ethnic groups from across India and two from Pakistan: among others, Kashmiri Pandits, Bhils from Gujarat, Gonds from Madhya Pradesh, Srivastavas from Uttar Pradesh, Naidus from Andhra Pradesh, Adi-Dravidars from Tamil Nadu, and the Paniyas from Kerala.

Thangaraj says, "We tried to represent all the states, all language families, and all social classifications."

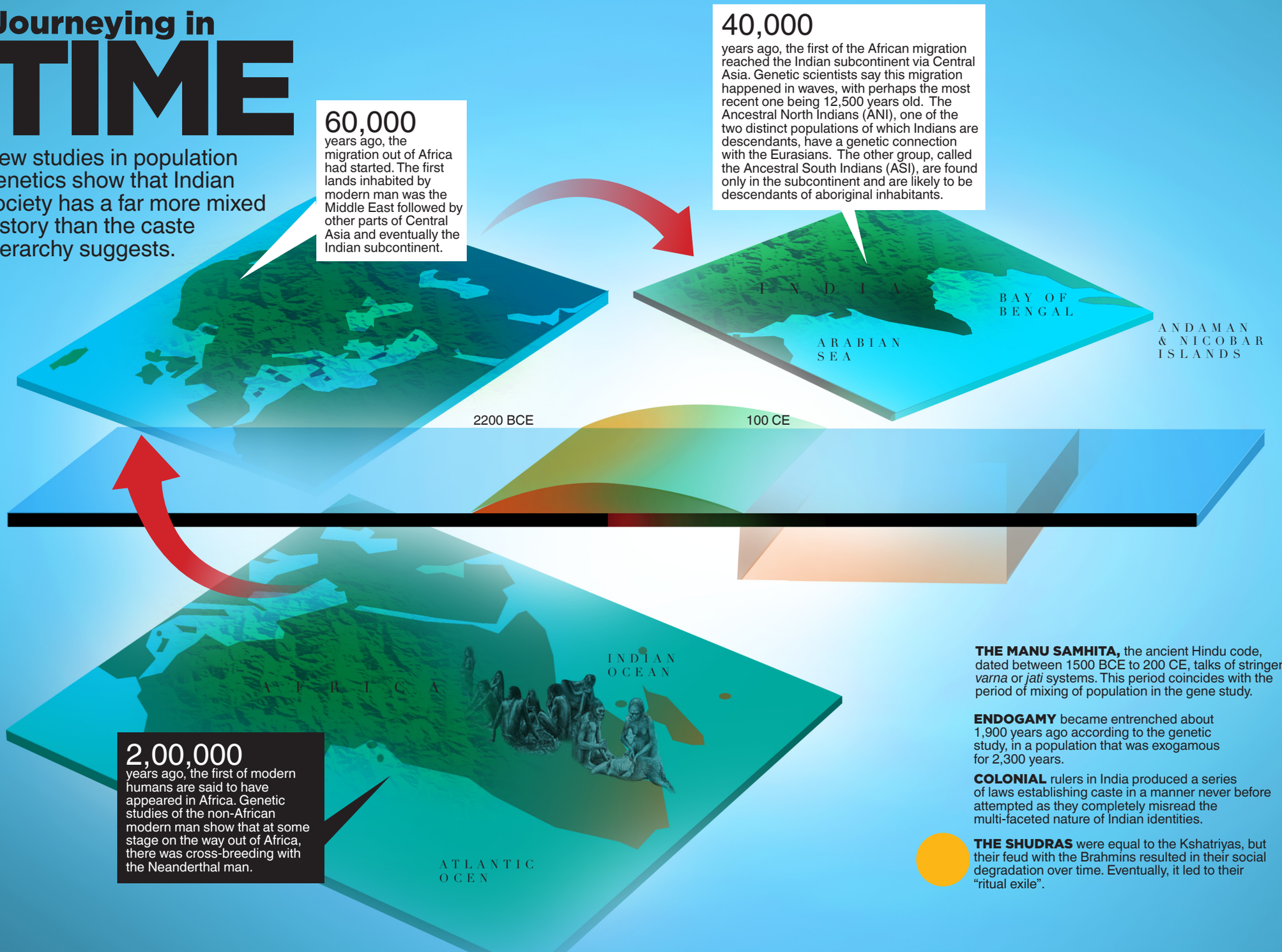
This meant ensuring the inclusion of speakers of both Indo-European languages such as Hindi, Punjabi and Gujarati, and Dravidian languages such as Tamil and Malayalam. (Speakers of Tibeto-Burman languages, spoken in Northeast India, and speakers of Austroasiatic languages, people like the Santhals who live largely in Eastern India, were excluded since they were known to have a different population history: the former show a genetic proximity to the Chinese, and the latter are known to have had significant genetic infusion from Southeast Asia.) In this study, the researchers had managed to find out when ANI admixture took place in various populations.

Different ethnic groups showed different ANI admixture dates, all between 4,200 and 1,900 years ago (2200 BCE to 100 CE). Within this period, speakers of Dravidian languages, largely South Indians, tended to show earlier dates of admixture and a smaller proportion of ANI ancestry when compared to Indo-European speakers. There was evidence to suggest that the ANI ancestry in Indo-European speakers had come in multiple waves.

In summary: about 4,200 years ago, there would have been people in the Indian subcontinent who were completely ANI in their genetic makeup, and others who were completely ASI. About 1,900 years ago, there were likely no pure populations of either ANI or ASI left. So, there began about 4,200 years ago

Journeying in TIME

New studies in population genetics show that Indian society has a far more mixed history than the caste hierarchy suggests.



THE MANU SAMHITA, the ancient Hindu code, dated between 1500 BCE to 200 CE, talks of stringent *varna* or *jati* systems. This period coincides with the period of mixing of population in the gene study.

ENDOGAMY became entrenched about 1,900 years ago according to the genetic study, in a population that was exogamous for 2,300 years.

COLONIAL rulers in India produced a series of laws establishing caste in a manner never before attempted as they completely misread the multi-faceted nature of Indian identities.

THE SHUDRAS were equal to the Kshatriyas, but their feud with the Brahmins resulted in their social degradation over time. Eventually, it led to their "ritual exile".

a period of demographic change due to inter-breeding among two dramatically different populations. Then, after about 1,900 years ago, there was no significant inter-breeding, pointing to cultural changes that brought in a strong form of endogamy, the practice of marrying within one's group.

The period is known to be a particularly eventful one for the Indian subcontinent: large-scale changes were occurring in river systems and climate; the Harappan civilisation was fragmenting; and, according to many linguists and historians, the Sanskrit language and Vedic culture were making an appearance.

The findings are also significant for what they say about the history of caste. With all the groups sampled—from Chamars in Uttar Pradesh, to Bhils in Gujarat, to Kashmiri Pandits, to Paniyas in Kerala—showing ancestry from both ASI and ANI populations, it establishes that ethnic groupings such as castes and tribes are a structure imposed on an already mixed population.

"The date of admixture 2,000 to 4,000 years ago is an upper-bound on the beginning of the caste system," explains David Reich. "Specifically, that date marks the time when the genetic data show that mixture between very divergent populations was very common in the ancestry of essentially all present-day Indian groups and thus the caste system in its present form must not have been fully formed. The endogamy that characterises many present-day Indian groups must have set in some time after the dates of admixture."

When B. R. Ambedkar wrote that "the superposition of endogamy on exogamy means the creation of caste" he was of course referring to the idea of marriage being allowed only between people outside one circle of relations (*gotra*, for example) and inside another (caste). But that description might, in a sequential sense, apply to the historical formation of caste as well.

***

It may not be easy for recent findings from population genetics to be reconciled with history given that the social and natural sciences often cannot find common ground. "It is typically the case," says Nicholas Dirks, anthropologist and author of *Castes of Mind: Colonialism and the Making of Modern India*, "that scientists have as much trouble understanding the findings of social scientists as the other way round."

Dirks writes in *Castes of Mind*: "Under colonialism, caste was thus made out to be far more—far more pervasive, far more totalizing, and far more uniform—than it had ever been before, at the same time that it was defined as a fundamentally religious social order . . . In pre-colonial India, the units of social identity had been multiple, and their respective relations and trajectories were part of a complex, conjunctural, constantly changing, political world. The referents of social identity were not only heterogeneous; they were also determined by context."

ere began about 4,200 years ago a period of demographic change due to inter-breeding among two dramatically di erent populations. en, after about 1,900 years ago, there was no signi cant inter-breeding, pointing to cultural changes that brought in a strong form of endogamy, the practice of marrying within one's group

SCIENCE IN HISTORY

A lot of us think of history as something that is forever fixed and timeless. Part of the reason for this impression is the lists of kings one is made to memorise. As these lists rarely change we get the impression of the entire discipline as something outside time.

In the last two centuries, the history of the planet has been rewritten entirely as compelling new evidence from geology and biology became available. In the 19th century the new discipline of historiography provided the tools for a critical analysis of history and this too led to a great deal of revision of local histories. The most spectacular effect of this revision is the way the Bible is viewed today.

Until the scientists got down to the job it was considered a true account of creation. The serious geologist soon realised that there could be no reconciliation between its findings and the Bible's account of Creation. The evidence from historiography provided conclusive proof that the story of Creation was no more than a myth, but such is the power of belief that even today a lot of Christians refuse to believe anything else.

The Aryan invasion theories and the eternal nature of caste are for India similar to the Biblical conundrum. It's hard to convert even the learned though there is increasing evidence that the Aryan invasions never took place. Most of it has come from archaeology and the hard sciences but it is only recently that historians have grudgingly started to accept the need to revise their accounts. The evidence comes from geology, hydrology, archaeology, remotely sensed data from satellite imagery, analysis of palaeo-waters, all of which call for rigour. Each study by itself may be inconclusive but if the conclusions are unimpeachable the cumulative evidence could provide a radically different picture.

The myths that surround the caste system Are of a different order altogether, but it possible given time and accurate data that the truth will finally be told.

- G K Rao

David Reich and K. Thangaraj, in their 2009 paper in *Nature*, cited Dirks's book as an example of a view held by some historians that the idea of caste "became more rigid under colonial rule". "However," they wrote, "our results indicate that many current distinctions among groups are ancient and that strong endogamy must have shaped marriage patterns in India for thousands of years."

Asked to comment, Dirks says, "I wrote little (if at all) about endogamy as a factor in the formation of a 'system' of caste . . . so the Reich study, important though it might be, is not of any particular relevance to my argument."

Sumit Guha, professor of history at the University of Texas and author of *Beyond Caste: Identity and Power in South Asia, Past and Present,* stresses that he has not read all the existing

work on population genetics. "In what I have seen," he says, "conclusions are tendentiously drawn with little or no understanding of actual social conditions in the past. Inferences are hastily made from modern observation to events that supposedly happened thousands of years ago." He identifies small and non-random population samples as major problems with such studies.

Reich responds, "I do not think that the criticism of small sample sizes is valid. We typically have five to 10 samples per ethno-linguistic group. While this may sound small, in fact we have the whole genome for each of these individuals—and a single genome contains multiple of ancestors—which means that the effective sample size is much larger than five to 10. To be concrete, consider for example just a single genome sequence. This contains information segments of DNA inherited from two parents, four grandparents, eight great-grandparents, 16, 32, etc. Thus in fact, each genome represents and conveys information about multitudes of people, and indeed, some studies have convincingly reconstructed whole population histories using single genome sequences." Counting all of a person's ancestors for 10 generations, or roughly 300 years, gives us over 2,000 ancestors.

"I agree with the critique that our sampling is non-random. Nevertheless, it is striking that of the more than 50 Dravidian speaking and Indo-European speaking groups we analysed all were consistent with the patterns we documented in our paper. Since our selection of groups was specifically chosen to be even more diverse than is typical for India, it seems likely that more systematic samplings would detect similar patterns."

The eminent historian of ancient India, Romila Thapar, when asked about the usefulness of population genetics research in arriving at histories, says, "The DNA results from various sources have been so confused and contradictory that it is difficult for me to accept what any of them say. None of them are social historians nor do they consult historians and sociologists before they make their categories, hence the confusion."

This happens to be a not an uncommon view among historians. A 2009 paper by the anthropologist Yulia Egorova asked, in part, how historians received population genetics studies (particularly in the context of caste). Most historians and social scientists interviewed were sceptical of the idea that population genetics studies could contribute to their area of work. Two-thirds felt that geneticists were "bound to be asking the wrong kinds of questions". The paper also summarises several population genetics studies of the subcontinent, which do at times appear to contradict each other (as Romila Thapar claims).

Some papers published in the last 15 years seem to find no evidence for recent European ancestry in the sub-continent, while others do. "[G]eneticists argue that they are able to help historians by providing them with 'hard' evidence," the paper concludes. "However this geneticisation of history is almost completely resisted by historians and social scientists doing research on caste who

A 2009 paper by the anthropologist Yulia Egorova asked, in part, how historians received population genetics studies (particularly in the context of caste). Most historians and social scientists interviewed were sceptical of the idea that population genetics studies could contribute to their area of work. Two-thirds felt that geneticists were 'bound to be asking the wrong kinds of questions'

insist on the primacy of socio-historical analysis in their field."

"We do have access to data that historians and archaeologists have not had access to previously, and that turns out to be useful for addressing some previously contentious questions," says Reich.

He continues: "Part of the problem is that geneticists are amateurs in the presentation of historical evidence." He points out that apparently contradictory results may not be so. For instance, studies using Y-DNA, which passes from father to son, have found recent European ancestry in Indians, while studies that used mtDNA, which passes down the female line have not. This is consistent with European ancestry having come in largely through males. Reich says, "I do not think the right approach is to ignore the data because our community does not always communicate well."

***

According to Reich, there are conclusions that genetics leads us to with certainty: "All or nearly all Indian groups today that speak Dravidian or Indo-European languages are descended from an ancient mixture of two very divergent ancestral populations, one genetically closely related to West Eurasians"; then, "The mixture was a gender-biased process, whereby most of the West Eurasian ancestry that is present in India today came into the population through males"; and finally, "Prior to 4,200 years ago, there were unmixed groups in India. Sometime between 1,900 to 4,200 years ago, profound, pervasive convulsive mixture occurred, affecting every Indo-European and Dravidian group in India without exception. The pervasive mixture was then followed by a switch to endogamy, which in many groups in India we can show has been strong and persistent for thousands of years."

Reich also points out what their work does not show: "The fact that we document major mixture in the period between 4,200 to 1,900 provides no evidence that there was substantial migration from West Eurasia into India during this time." This, of course, is important to state since the history of the period has been marked for over a century by politically-charged debate about "Aryan" invasion or migration theories that suggest a large and possibly violent influx into India from the northwest. Different groups—western colonisers, lower castes, Hindu nationalists—have invoked the idea, or their strong

opposition to it, to argue about who the "original" inhabitants of India are, usually with a view to justifying their own ascendancy.

"Nobody thinks there was a mass immigration at this time," says Mait Metspalu, speaking of the community of population geneticists. Metspalu is research director at the Estonian Biocenter, and has collaborated with several researchers investigating Indian population history.

According to Metspalu, the population of the subcontinent was already large during the time in question, and it is hard to find a West Eurasian source large enough to contribute so much to the Indian genetic makeup. In addition, the West Eurasian component in Indians appears to come from a population that diverged genetically from people actually living in Eurasia, and this separation happened at least 12,500 years ago. K. Thangaraj believes it was much longer ago, and that the ANI came to India in a second wave of migration that happened perhaps 40,000 years ago.

Metspalu summarises: "So the scenario at present seems to show that there were two populations colonising South Asia, one close to West Eurasian populations but not derived of them recently. These two populations lived in broad South Asia with little mixing for a long time before admixing quite abruptly relatively recently." (If that ends up being confirmed, it would mean that both proponents and opponents of the Aryan invasion/migration theories are in a sense simultaneously right and wrong—yes, foreigners entered an already inhabited India; but they did so so long ago that they might as well be thought of as original inhabitants too. It would provide a strangely satisfying end to an acrimonious debate.)

We might find out for sure very soon. Metspalu points out that conclusions from population genetics are becoming less tentative as it becomes technically feasible to work with increasingly large portions of DNA. "We are now entering a new era in these studies," he explains. "We are entering the complete genome sequences era and I would expect more definitive answers in the coming year or two."

An important piece of the puzzle when it comes to the history of India two to four millennia ago is the Harappan civilisation, about which very little is known. Jonathan Kenoyer, archaeologist and professor of anthropology at the University of Wisconsin-Madison says, "Any work that is being done on genomics in South Asia will be useful in understanding the legacy of the Indus Civilisation. The main problem now is to be able to get some DNA from ancient Harappans or other bones from Indus burials." In the last decade or so, it has been increasingly possible to extract DNA from ancient remains, sometimes even when they are a few million years old. But the DNA of Harappans, though relatively recent, has proved hard to extract because the arid climate of their erstwhile land does not preserve genetic material well. Tantalisingly, even if a single well-preserved Harappan tooth turned up, it could unlock the history of the period.

Population genetics has answered other questions about the past. It settled

the debate about whether the Polynesian islands were inhabited by people from Southeast Asia or the Americas, adding to other evidence from linguistics and archaeology. In Central Europe, it has revealed, again in conjunction with other methods, that groups of indigenous hunter-gatherer people existed side by side with immigrant farmers in the period between 7,000 and 5,000 years ago, with women from the foragers sometimes marrying into the farmers but not the other way round. (It may be that ancient India went through a similar phase soon after with ANI and ASI people.)

"The new genomic research in general has been a great boon for the deep history of the human race," says Thomas Trautmann, Professor Emeritus of history and anthropology at the University of Michigan, who has written prolifi-