Understanding the Aryan debate: population genetic concepts and frameworks

Partha P. Majumder

A long-standing debate on whether 'Aryans' (central Asians) had entered India has recently gained momentum. The debate is polarized. In the recent set of articles, some authors have strongly criticized inferences drawn using genomic data and population genetic methods. Some criticisms are flawed. These criticisms stem from lack of clear understanding of population genetic concepts and frameworks. Such inappropriate criticisms have led to unnecessary confusion and further polarization among readers. This article attempts to place the ongoing 'Aryan debate' in the perspective of population genetic frameworks.

Keywords: Ancestry, admixture, central Asia, haplogroup, lineage.

SPARKED by a recent scientific publication in *BMC Evolutionary Biology* by Silva *et al.*[1], we have witnessed in the pages of *The Hindu* (Joseph[2]; Danino's critique of Joseph[3]; Joseph's rebuttal[4]) and elsewhere[5] a bitter battle pertaining to the 'Aryan issue'. The battle will persist. The Aryan issue has been debated with emotion, when it does not really have to be. Population genetic inferences have been criticized as subjective, when it should be considered as more objective than inferences of most other empirical sciences, especially because (a) the framework and methodologies of drawing inferences from data are defined with mathematical and statistical rigour in population genetics, and (b) better sampling of populations and generation of larger data sets are making the inferences more reliable and robust. Irrelevant issues, such as whether there was 'out-of-India' migration, are brought into the Aryan debate that adds more haze than transparency. Of course, the 'out-of-India' issue is related to the debate on whether Sanskrit is the mother of all Indo-Aryan languages, and moved from India to Eurasia. The fact that many words of Sanskrit are phonetically similar to those in Greek, Latin, Gothic, Celtic, etc. indicates that these languages share a common origin, possibly developed by a set of people now extinct. Sir William Jones, the founder of the Asiatic Society, had pointed this out long ago. The 'debate' will, of course, live on, unless there is a strong intent to weigh evidence objectively.

In the articles published in *The Hindu* and elsewhere, many methodological and conceptual issues related to analyses of genetic data of populations have been touched upon. However, these issues were not discussed in sufficient detail for the general reader to clearly comprehend the framework and the basis of inferences in population genetics. Even I, a practising population geneticist working on ethnic populations of India for about 40 years, had a hard time understanding the logical veracity of some of the claims made in these articles. Sweeping statements such as 'most studies of population genetics suffer from shortcomings and flaws…'[3] have been made. Here, I am neither deeply focusing on the 'Aryan debate' nor intending to critique any of these recent articles. My intent is only to try and more clearly inform the reader on certain relevant concepts and frameworks that underlie the population genetic studies pertinent to these articles. I am driven to do so because the Joseph and Danino essays are being discussed by students and researchers, often without clarity of their own thought. I must also point out that population geneticists have not always been able to peg their inferences in the context of social history of a region; sometimes this is not possible; sometimes geneticists are not adequately informed of social history. This inability has contributed to a lot of interpretative confusion.

I will try to provide some clarity to four population genetic concepts that underlie the debate between Joseph and Danino, and also present sketches of relevant inferential methods for a broad understanding. These include, (a) Evolution of populations, (b) Sampling for unbiased and representative genetic information, (c) Inferences from fossil or other palaeontological or archaeological records, and (d) Admixture detection and estimation.

Evolution of populations

The kind of population genetic studies referred to in the recent articles pertain to evolution of populations. Populations evolve over time. In the remote past, when we

Partha P. Majumder is in the National Institute of Biomedical Genomics, Kalyani 741 251, India and Indian Statistical Institute, Kolkata 700 108, India. e-mail: ppm1@nibmg.ac.in

were hunting and gathering to feed ourselves to survive, bands of people had to break away from a core group of people when numbers increased by reproduction. Natural resources in the area became scarce for an increased number of mouths to feed. The breakaway band had to move to a new geographical location and set up a new colony, simply to assure their survival. This must have happened for many evolutionarily ancient tribal groups of Africa, India, etc. These ancient populations were ancestral to many descendant populations.

Often, a breakaway band would have no contact with the core group. This process of colonization of new territories continued during the entire period of our evolution. The core groups from which the bands broke out successively would be their ancestors. Ancestral groups, therefore, may not be the same for all populations that are present today. Lack of contact with the ancestral group would mean that genetic differences would rapidly accrue over time, mainly by a natural process called mutation – a process in which DNA alphabets (nucleotides) unpredictably change from a pre-existing one to a new one, often with no physiological or biological effect but sometimes with large effects. The ancestral and descendant groups would evolve independently and genetically diverge from one another. Yet, groups of people who broke out from a common ancestor would be genetically more similar than those with different ancestors. Migration and consequent reproductive contact between individuals of populations that results in genetic mixing (admixture) tends to reduce genetic differences. When social structures were developed, in addition to geographical distance as a deterrent to genetic mixing, new social factors impacting genetic mixing were introduced. If two populations arising from two different ancestral groups admix significantly, then they may appear to have arisen from the same common ancestor since the genetic characteristics between them would be shared to a large extent, as one would expect of two groups arising from the same ancestral group. By not taking admixture history into account, one could draw a flawed inference about ancestry from genetic data alone. Admixture history is often not known in granular detail. Therefore, inferring admixture using simple methods may not always provide full details of ancestry. A detailed discussion of admixture detection and estimation of the amount of admixture is presented in a later section. We also note that populations that are ancestral to some existing populations would themselves have arisen from some other ancestral groups. Further, ancestral groups themselves are genetically heterogeneous. That is, individuals belonging to an ancestral group differ in their genetic composition. Natural biological processes – mutation, natural selection (a process by which one or more genetic variants rapidly increase (or decrease) in relation to environmental changes that favour (or disfavour) carriers of these genetic variants to become reproductively more successful and produce a larger number of children) and admixture – contribute to the genomic diversity of a population. In view of the process of evolution of populations, diversity within an ancestral population and other considerations, there is really no 'true' ancestor; not even conceptually. I underscore this because 'true ancestral populations of India' have been discussed in the articles cited above. Despite these limitations, it is possible to identify ancestral populations, provided that admixture is contained within limits and natural selection has not swept through the genes of individuals. The confidence that underlies this assertion comes from a variety of simulation studies in which populations are allowed to evolve under known mutation–selection–admixture scenarios and then the evolutionary paths are reconstructed by studying genetic variations in extant populations using methods that have been refined for over 50 years and still continue to be refined. The above is a reasonably simplified paradigm under which population geneticists operate. Many of the phenomena described above can either be treated as tautologies or as assumptions. There are limitations in every branch of empirical science, and inferences get more robust as more data are collected and analysed using more refined methods. There is really no last word in most empirical sciences.

Sampling for unbiased and representative genetic information

When we carry out genetic studies, we sample individuals from extant populations, obtain genetic information from them and look backwards in time to reconstruct their evolutionary history and identify common ancestors. In reality, populations diverge from a common ancestor. When we look backwards in time, populations converge or coalesce to a common ancestor. Given genetic data on a set of populations, statistical methods, that are relatively unaffected by small violations of assumptions, have been developed to identify which populations have diverged from a common ancestor and whether all the populations under consideration have a single common ancestor. (Of course, if we go very far back in time, all populations will converge to a single common ancestor. For modern humans, that single ancestral population will be rooted in Africa.) Under reasonable assumptions, such as the rate at which alterations take place in DNA per unit time, we can even estimate the times of divergence of populations from a common ancestor. The estimated times are, of course, sensitive to the rate of DNA alterations. (As I have mentioned before, the natural process of mutation introduces alterations in the DNA over time. These alterations occur at a specific rate over time. For example, in a given stretch of 1000 alphabets, a change may take place once in every 100 years or every 2000 years. The rate of change/alteration is variable in different regions of the human DNA.) Therefore, the estimated times of divergence often have a wide range of variability, since the rate of DNA alteration is not easy to estimate precisely. For how long have we evolved from our most recent common ancestor in Africa? The estimates will vary between 60,000 and 100,000 years; that is a wide range. But, that is the nature of inferences in population genetics in the face of uncertainty of rates of DNA alterations. As we gather more data, uncertainty diminishes and the range of estimates of the time of divergence becomes narrower and more precise. This is the nature of all sciences. More data usually results in improved inference. *The Hindu* articles under consideration have cited past studies and have strongly criticized past conclusions because of wide ranges of estimates. Readers also need to understand that estimates presented in past studies have not been based on equal quantities or comparable qualities of data; the more recent studies are usually based on larger quantities of data.

Inferences from fossil and other palaeontological or archaeological records

Inferences from fossil and other palaeontological or archaeological records that are testimony to the temporally forward process of evolution may appear to be less flawed than 'backward-looking' population genetic methods. However, it is important to realize that such records are often scanty and fragmented (a broken skull or a finger bone, an arrowhead, etc.). Further, methods of ascertaining ages of fossil remains are also not free of assumptions. Newer methods are more robust and more reliable, as is usually true for all branches of science. Interestingly, it is now possible to isolate DNA from many ancient DNA remains, generate DNA data and draw inferences using population genetic methods. Such methods have now yielded valuable information that Neanderthals admixed with humans over wide geographical regions outside of Africa, even though the extent of admixture was relatively infrequent.

Admixture detection and estimation

The most popular method of admixture detection and estimation of the amount of admixture depends on identification of genetic signatures in specific geographical regions or specific hypothesized ancestral populations. These genetic signatures are combinations of specific DNA alphabets that appear on a chromosome and get passed on from parents to offspring, usually unaltered. Normally, the region or the population in which a genetic signature arises has a high frequency of the signature, because the signature gets passed on from a parent to many children. A signature can move to a new region when an individual carrying the signature moves to a new region, marries someone from the new region and produces children who remain in the new region. Recipient regions or populations are, therefore, expected to have lower frequencies. The farther one moves away from the geographical location where the signature arose, its frequency diminishes, almost exactly like the epicentre of a cyclone, with diminishing ferocity from the epicentre. Thus, if one can identify the epicentre of a genetic signature and for a region/population away from the epicentre, if one seeks to find whether this region received any migrants from the epicentre, it is easy to do so. If one seeks to find how many migrants arrived from the epicentre, it is more difficult; often, hardly possible. With respect to the Aryan issue, the debate is possibly not whether there was any migration from central Asia (Aryan homeland) to India, but whether there was a tsunami of migrants who entered India. If the framework of identification of the geographical location where a genomic signature arose and its diminishing impact via migration on regions with increasing distance is accepted, then genetic evidence is unequivocal that there was significant migration from central Asia into India. The amount of migration, in terms of the number of migrants, is very hard to estimate from genetic data. However, the presence of footprints of some genetic signatures in India that arose in Central Asia (i.e. occurring with a very high frequency in that region) is overwhelmingly evident. The migration may have taken place over a longish period of time, and involved a large number of people. With more data, deeper insights were obtained. A DNA signature termed as haplogroup U of the female-lineage mitochondrial DNA is widely found in India. The epicentre of this signature is in Western Eurasia. It was initially felt that Haplogroup U was brought into India exclusively by migrants, which would mean that the number of migrants was large. However, more extensive data showed that this signature U comprises two sub-signatures U2i and U2e. (Sub-signatures are similar, but not exactly the same, derived from the original signature, identifiable by a deeper examination of DNA alphabets that comprise the signature.) We had shown that tribals of India possess *exclusively* the U2i sub-signature and that the age of this sub-signature predates the postulated time period (4000–5000 years ago) when migrants from Eurasia began to enter India[6]. The caste populations possess U2i and U2e in the proportions, 88% and 12% respectively. This means that the present-day caste populations are largely indigenous, but have also admixed with Eurasians (Indo-European speakers). Before the discovery of the India-specific sub-signature U2i, one had assumed that a very large number of migrants had arrived carrying the signature U. This example shows that the Aryan debate – if at all there is a debate – can only be resolved by collecting and analysing more scientific data; not by verbal accusations and duels. Another paternal-lineage signature, called R1a1a on the Y-chromosome, has also provided similar evidence and resulted in a similar inference. The most likely period of entry of this signature into India is between 4500 and 5000 years ago.

Sometimes, as in the case of the hypothetical ANI (Ancestral North Indian) and ASI (Ancestral South Indian) that have been fiercely debated in the articles in *The Hindu*, it may not always be possible to identify an epicentre of a signature. It may not even be possible to identify a specific signature attributable to a putative ancestral region. New statistical methods have been developed to analyse the diversity of genetic data within and between populations to estimate how many putative ancestral populations may have contributed their genes to these populations. These methods also allow us to estimate for every individual in any population, how much the putative ancestral populations may have contributed to her or his genome. Based on a limited sampling of populations, Reich *et al.*[7] identified that genomes of north-Indians showed a high level of admixture with an ancestral population; this level diminished as one travelled south. Similarly, genomes of south Indians showed a high level of admixture with another ancestral population; this level also diminished as one travelled north. Neither of these ancestral populations can be identified with any specific extant population. Thus, an ancestral population – currently unidentifiable with an extant population – contributed highly to the genomes of north Indians and another similarly unidentifiable ancestral population contributed highly to the genomes of south Indians. This is the story of postulation of ANI and ASI; not ‘advanced simply because they were expected to conform to predetermined results, such as ANI entering as Indo-Aryan speakers’[3]. However, it also does not ‘prove that most groups in India can be approximated as a mixture’ of ANI and ASI (as claimed by Joseph[4]). Such empirical data do not ‘prove’ anything; such data can only support a model or a hypothesis more strongly than competing ones. Danino[3] has pointed out that ‘we might as well put forth constructs of ‘Ancestral Eastern Indians’ and ‘Ancestral Western Indians’ and demonstrate that most Indian populations can be approximated by a mixture of these two’. Indeed, we agree with him, albeit partially. The geographical and social spread of the populations studied by Reich and his team[7] were somewhat limited. Their analysis did not include data from tribal groups. No population of north-east India was included in their study. Based on a much better sampling of populations across diverse geographies and social strata, we have indeed provided significant genetic evidence that there were four ancestral populations of mainland India. In addition to ANI and ASI, there were two other ancestral populations; one from which most populations of north-east India evolved and another from which some Austro-Asiatic speaking eastern and central Indian tribal groups evolved. This underscores the importance of representative sampling. Do we then need to sample ‘thousands of communities from all over the subcontinent,’ as Danino[3] has contended, to infer the ancestral history of Indian populations? I think not. Groups who occupy habitats in geographical and social proximity may be reasonably assumed to have similar evolutionary histories. Therefore, the quality of inferences that can be drawn by sampling one or a few of these groups will not improve significantly if a large number of them are included. Sampling small number of populations is not necessarily ‘biased,’ provided that they are representative.

Finally, Danino[3] has expressed dissatisfaction that many genetic studies have attributed the ‘spread of agriculture’ into the subcontinent to migrations. Just to be sure that no one thinks that the practice of agriculture is in the genes, I would like to point out that the spread of the technology of agriculture was associated with the movement of people; agriculturists who took the technology to new regions and taught it to the locals in the new region. Movement of people implies movement of genes. Some migrants ‘export’ their genes to a new region by taking spouses from the new region and producing children with them who stay in the new region. We can never be sure that the attribution of agriculture having been introduced to the Indian subcontinent by migrants is fully true. However, genetic data do support this model, especially of the spread of modern, organized agriculture. Having said this, I must also emphasize, once again, that collection of more extensive data is always more helpful in understanding our past and of the spread of our inventions and innovations. A Y-chromosomal signature, haplogroup J, was shown to be associated with the spread of modern agriculture. This signature has its highest frequency in the Fertile Crescent region – the region comprising the present-day countries of Syria, Lebanon, Turkey – where the technology of modern agriculture was invented about 7,000–10,000 years ago. Collection of deeper data showed that this signature is quite heterogeneous and is composed of at least four sub-signatures, one of which – haplogroup J2b2 – is confined to the India–Pakistan region (Figure 1). This sub-signature arose over 13,000 years ago and hence its introduction into India could not have been by migrants who introduced modern agriculture into India. We showed that the haplogroup J2b2 possibly arose in India, because the highest frequency of this haplogroup is found in India[8]. We discovered multiple epicentres of this haplogroup in India and interestingly these epicentres nearly coincided with the seats of introduction of early forms of agriculture in India (as evidenced by the study of fossilized pollen grains by Fuller and his team[9]. It is unlikely that haplogroup J arose independently multiple times in geographically separated places. It probably arose in an ancient population who had spread themselves in geographically separated regions and they invented rudimentary forms of agriculture independently in multiple geographical regions. However, it is notable that these early forms of agriculture remained largely confined to India and Pakistan region.

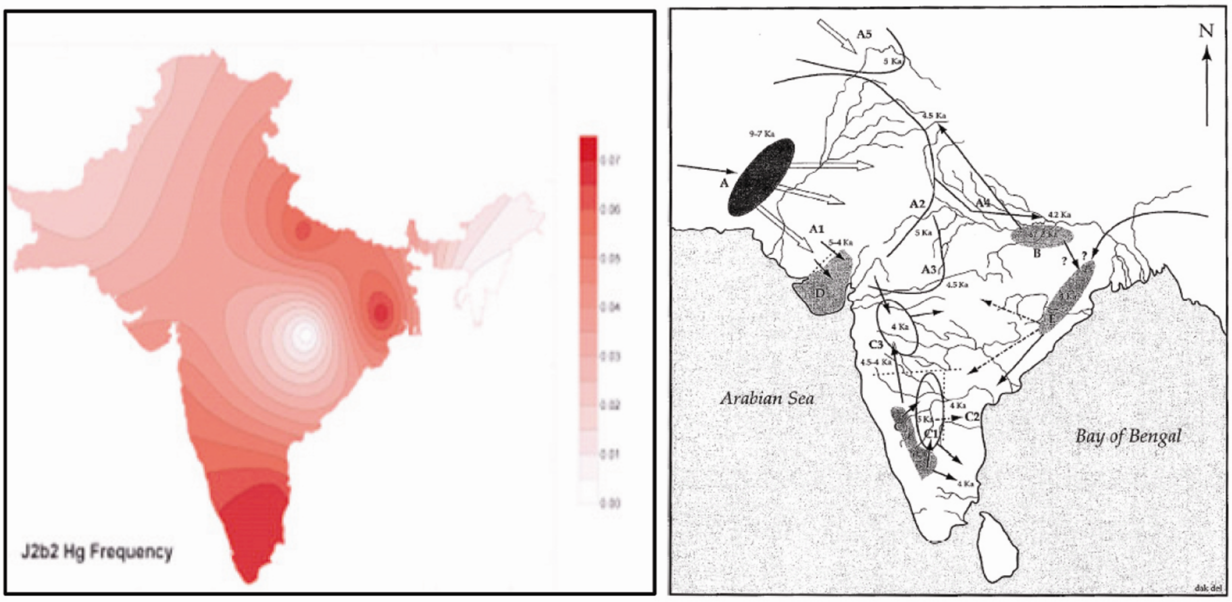

Figure 1. Epicentres of frequency of the Y-chromosomal haplogroup J2b2 and its spread (figure on the left) nearly coincides with the regions of introduction of early forms of agriculture and its spread (figure on the right; drawn from Fuller's paper). Arrows on the figure indicate the directions of spread of agriculture.

Overall, therefore, undoubtedly there are uncertainties in population genetic inferences. This is true of all empirical sciences. At least there are well-defined frameworks and methodologies of inference in population genetics, which is possibly the most quantitative of all biological sciences. Hence, given a body of data, the inferred results are replicable and inferences are generally robust and not as subjective as many would have us believe. Uncertainties can be quantified. Sample sizes can be large and choice of populations can be made by statistical designs, unlike many sciences in which generation of evidence is based on chance finds and not sought out by statistical design.

1. Silva, M. *et al.*, A genetic chronology for the IndianSubcontinent points to heavilysex-biased dispersals. *BMC Evol. Biol.*, 2017, **17**, 88; doi:10.1186/s12862-017-0936-9.
2. Joseph, T., How genetics is settling the Aryan migration debate. *The Hindu*, 16 June 2017.
3. Danino, M., The problematics of genetics and the Aryan issue. *The Hindu*, 29 June 2017.
4. Joseph, T., *The Hindu*, 29 June 2017.
5. Elst, K., Genetics and the Aryan invasion debate. http://www.pragyata.com/mag/genetics-and-the-aryan-invasion-debate-367#
6. Basu, A. *et al.*, Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.*, 2003, **13**, 2277; doi/10.1101/gr.1413403.
7. Reich, D. *et al.*, Reconstructing Indian population history. *Nature*, 2009, **461**, 489; doi:10.1038/nature08365.
8. Sengupta, S. *et al.*, Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of central Asian pastoralists. *Am. J. Hum. Genet.*, 2006, **78**, 202.
9. Fuller, D., Agricultural origins and frontiers in South Asia: a working synthesis. *J. World Prehist*, 2006, **20**, 1.

Received 13 September 2017; accepted 15 September 2017

doi: 10.18520/cs/v114/i05/971-975