GenomeIndia

The Epistemologies and Politics of Stitching together Genetic Variation in Indians

Mahendra Shahare

Abstract *Since colonial times, attempts at racial classification and naming the diversity and potpourri of India's population have remained controversial. However, the postgenomic era has created an impetus within the life sciences to study human biological diversity. The ongoing GenomeIndia project (GIP) is a significant development in this direction. The GIP aims to create a catalog of genetic variations in Indians and construct a reference haplotype structure for Indians. But what epistemological assumptions and concepts inform the GIP? And how do such epistemic conceptualizations influence the discourse on identity politics? By focusing on empirical research practices, I highlight here how the GIP relies on taken for granted epistemologies for characterizing human differences. I posit that while population genetics and medical genomics continue to be guided by the normative logic of genetic nationalism, in the case of India, the logic remains closely tied to the diffuse politics of identity and caste. As such, the GIP's notions of a reference haplotype or catalog of genetic diversity cannot fully dissociate science epistemologies from politics.*

Introduction

In January 2020, the Government of India initiated an ambitious mission-mode project, "GenomeIndia: Cataloguing the Genetic Variation in Indians," with a substantial grant of ₹ 238 crores (approximately USD 30 million). The GenomeIndia Project (GIP), an ongoing collaborative effort between twenty academic and research institutes[1] in India, aims to identify genetic variation in the Indian population through whole genome sequencing of 10,000 representative individuals, with a vision to facilitate the development of precision

1 IIIT Allahabad (Uttar Pradesh), IBSD Imphal (Manipur), MZU Aizawl (Mizoram), NIBMG Kalyani (West Bengal), ILS Bhubaneswar (Odisha), CSIR-CCMB Hyderabad (Telangana), CDED Hyderabad (Telangana), IIT Madras (Tamil Nadu), RGCB Thiru-vananthapuram (Kerala), CBR Bengaluru (Karnataka), IISc Bengaluru (Karnataka), NIMHANS Bengaluru (Karnataka), NCBS Bengaluru (Karnataka), IISER Pune (Maha-rashtra), GBRC Gandhinagar (Gujrat), AIIMS Jodhpur (Rajasthan), IIT Jodhpur (Raja-sthan), SKIMS Srinagar (Jammu and Kashmir), IGIB Delhi (Delhi), IIT Delhi (Delhi).

healthcare and major diseases diagnostics at affordable costs (PIB 2020a). Given the scale and ambition of the GIP and potential long-term consequences on India's sociopolitical spheres, including medicine and law, as well as group, ethnic, and individual identity, it becomes pertinent to critically analyze definitions, concepts, and assumptions employed by the project in framing and understanding human biological diversity.

In engaging with scientific discourse on classifying human diversity in the life sciences, this volume opens up some key questions, including: How do assumptions about the preexistence of human races shape scientific practices? In what ways do such epistemic conceptualizations influence the discourse on identity politics? What are the differences and continuities between historical and present population naming practices in the life sciences? These concerns also resonate with the GIP. The GIP aims to create a catalog of genetic variations in Indians and construct a reference haplotype structure for Indians. But what epistemological assumptions and concepts inform the GIP? Does the cataloging of genetic variation reproduce problematic social constructs, and if so, how? How is diffuse politics around caste and identity implicated in the practices of GIP? And can the notion of a reference haplotype dissociate science epistemologies from politics? In this chapter, I analyze these questions by examining and situating the GIP in the context of scientific discourse around medical genomics and population genetics in India.

Contemporary social sciences scholarship that closely engages with genomic science offers three relevant strands for problematizing the GIP. First, notwithstanding the GIP's well-rehearsed future imaginaries that have been in circulation since the Human Genome Project (HGP), which promised to pinpoint linkages between common human phenotypic traits with common genomic differences, genomic science is facing a conceptual and epistemic quandary. As Reardon (2017, 2) notes, medical genomics is caught up in the problem of meaning, wherein biologists have access to a large number of complete human genomes but are grappling with the question—what does it mean? While at the time of writing, the GIP had completed its target of sequencing 10,000 human genomes, how this massive information will produce meaningful knowledge for "transformative" medicine remains a question (PIB 2024). Second, the GIP is a "national project" that aims to "create a reference haplotype structure for Indians" (CBR 2023). The project therefore can be viewed as a move to assert "genomic sovereignty" through practices of bionationalism that seek to control shared genomes bounded by nation-state borders (Subramaniam 2019, 165). Third and most importantly, the GIP aims to understand India's genetic makeup, putatively constituted by more than 4,600 population groups inhabiting the country. As various scholars have illustrated, within India the discourse on genetic diversity is imbued with questions of origins and ancestry, deeply entangled with complex caste and tribe identities, and is a site of contestation for the politics of "Hindu nationalism" (Sur and Sur 2008, 210). In this chapter, I broadly use these strands to question the epistemic assumptions underlying the GIP. I further show how these assumptions uncritically reproduce the diffuse politics in India around caste and identity. In this article, I draw on my ongoing ethnographic research and documentary analysis of published resources related to the GIP available in the public domain. In particular, I draw on thematically coded semi-structured interviews of three principal investigators working on the GIP, as well as discussions with researchers and stakeholders at a rare diseases conference held in November 2023 in Bengaluru, India.

The chapter is organized as follows. In section two, I provide a brief sketch of the role of science in constituting race discourse in India and its interpolation with the social categories of caste. In section three, I situate the GIP in relation to similar research efforts in the last two decades. Focusing on the idea of cataloging differences, section four details how the GIP is reproducing the problematic population classification of the past and its imbued caste and identity politics. In the fifth section, I analyze the work of stitching together Indian reference haplotype as a way to nationalize ancestry. In the concluding section, I discuss the entanglements between population genetics and medical genomics and argue that the GIP cannot fully dissociate science epistemologies from politics.

Science, Race, and Caste Discourse

Human biological diversity, when principally understood in terms of ancestral collectives, enables a biological concept of human differences. How the life sciences frame and conceptualize human variation shapes and constrains knowledge production processes. Over the last two decades, with technological capacity for DNA sequencing of entire genomes, genomics has become the key mode of knowledge production for anthropological, biomedical, and genealogical studies of human variation. Nonetheless, genomic science operates on a highly politicized epistemic terrain, where the problematic legacy of racial and eugenic theories emanating from the eighteenth- and nineteenth-century paradigms of anthropology, ethnology, and human genetics continues to linger. Through empirically rich case studies, several scholars have argued that research practices in population genetics, which studies human variation, continue to use concepts, classification, and categorization employed by erstwhile race science (Reardon 2005; Burton 2021; Lipphardt 2014). Since the GIP intends to construct a catalog of genetic variations in Indians, it is important to probe and understand the historical ruptures and continuities that inform the conceptualization of human diversity mobilized by the project. As I further elaborate in the next section, the GIP uses caste as its proper unit of analysis. Therefore, it is crucial to contextualize how colonial practices of ordering Indian populations were informed by various contemporary scientific theories and paradigms and how the social construct of caste anchored and mediated these explorations.

Since colonial times, attempts at racial classification and naming the diversity and potpourri of India's population have remained controversial. But long before the colonial push for dividing and labeling natives into ancestral collectives, India had its own preexisting sociocultural system of stratifying populations, i.e., the caste system. The religiously sanctioned Hindu varna system ranks people into four groups—Brahmins (priests), Kshatriyas (warriors), Vaishyas (traders), and Sudras (peasants and artisans). While tribal people are not part of the caste society, the Dalits, formerly known as "untouchables," are outcastes. In general, each varna acts as a broader caste category that comprises multiple subcastes and accords graded social and economic privileges, with Brahmins enjoying the greatest freedoms and Sudras placed at the lower rung of the four groups. The tribal and Dalit populations constitute the most oppressed lot of the society. These historical discriminatory practices are at the heart of India's caste politics. Notwithstanding regional variation and complexity, the social construct of the *jati*/caste

system has firmly entrenched a hereditary, hierarchical, and rank-based way of categorizing individuals and communities. The *jatis* commonly signify occupational groups and guilds, comprise myriad subcastes, are governed by complex kinship norms, and generally observe endogamy. Therefore, predictably, questions regarding the genesis, proliferation, and stability of the institution of caste continue to occupy a central place in scholarship attempting to understand the social stratification of Indian society.

The riddle of caste became more complex and pronounced as scholars in the colonial period sought to index and map caste onto race. Science perforce played a pivotal role in efforts to determine racial affinities and label populations in the colonial period. The "science of man" or anthropology, aided by the production of anthropometric data, was the governing paradigm for colonizers to accumulate knowledge about the natives (Philip 2004, 107). These attempts at racial classification were systematized early on through the incorporation of the category of race in the Census of India (Bhagat 2006). In the late 1880s, British colonial administrators initiated ethnographic surveys of the people of India to measure human differences using anthropometric parameters, viz. the cephalic index or nasal index (Bates 1995). Herbert Risley was the key figure of this colonial project. Using anthropometric measurements, he classified Indian populations into seven racial types, namely—Aryo-Dravidian, Dravidian, Indo-Aryan, Mongolo-Dravidian, Mongoloid, Scytho-Dravidian and the Turko-Iranian (Malhotra and Vasulu 2019). Furthermore, the classification of India's tribal populations into distinct racial categories such as the Negrito, the Proto-Australoid, and the Mongoloid was a product of colonial efforts (ibid.). These theories held that caste status and ranking are associated with racial differences.

In parallel to the comparative ethnology work in the nineteenth century, the work of philologists demonstrated close linguistic linkages between Sanskrit and other European languages. Max Müller, an orientalist philologist, argued that linguistic affinity indicates racial kinship and proposed his two-race theory that classified Indian populations into the Hamites (Aboriginals) and the Japhites (immigrant Aryans) (Sur and Sur 2008). Müller's work contributed to the discourse on Aryan and Dravidian racial identity. The two-race theory had a profound impact on India's academic and political spheres and even today acts as a meta category. For example, Govind Ghurye, a pioneering native scholar of Indian sociology, accepted and advanced the racial categorization of the Indian populations. These debates gave rise to what is referred to as the Aryan invasion/migration theory and created its own long protracted identity politics (see Barbosa in this volume). Nevertheless, this discourse directly or indirectly utilized caste as a marker.

During the interwar period, as Projit Mukharji illustrates in his cogent historical account of the development of sero-anthropology, scientists took up the notion of blood groups as a more objective basis to explain human differences. Through the works of Hirszfeld, Macfarlane, and other scientists, Mukharji notes, "from being a metonym for all of India in 1918," the blood group B became a metonym "for the lower-castes in the mid-1930s" (Mukharji 2014, 165). The transnational serological discussion about race therefore got translated into a debate about caste and its origin. Although some of the foremost scholars, including Dhirendranath Majumdar, Iravati Karve, and L. D. Sanghavi, combined anthropometric and serological approaches to study Indian populations, by the late 1950s the field witnessed a decline (ibid.). Though a small number

of studies on blood group systems (viz. ABO, MN, Rh) from India were reported till the 1980s (Papiha 1996), these academic debates remained subdued.

Although the notion of race is now understood as a biologically dubious concept, it acquired a new meaning in the nineteenth century as a tool and method for colonial domination and was an active product of "scientific racism." Nonetheless, after independence, India ended the practice of racial classification, and the Decadal Census of India (1951 onwards) does not enumerate or recognize any racial groups in the country. Notwithstanding various social movements between the 1960s and late 1980s, by and large, the scientific discourse in India around human variation, race, caste, and origins remained muted. At the turn of the millennium, a study titled "Genetic Evidence on the Origins of Indian Caste Populations" (Bamshad et al. 2001) became the moment that revived the race and caste discourse in India, bringing genomic science into conversation with the diffuse politics of identity around Hindu nationalism and caste politics. The authors emphasized that their data analysis suggests "upper castes being more similar to Europeans, whereas lower castes are more similar to Asians" (ibid.). These findings revived the "two-race theory of Aryan migration from Central Asia into India" (Sur and Sur 2008, 206), ascribed linkages between race and caste, and threw open to debate "the basis for the majoritarian, communal identity as forged by Hindu nationalism" (Sur and Sur 2008, 215). Furthermore, citing these findings, when a few Dalit organizations demanded that caste oppression be discussed at the United Nations conference against racism at Durban, the Indian government disapproved the move and sought to control the entanglement of the social and the biological through bionationalism (Subramaniam 2019, 155). Debates on ethnicity, ancestry, and indigeneity of the diverse Indian population (Joseph 2021), which were put on ice after the emergence of democratic India in 1947, were thus reopened by advances in population genetics that promised to answer the question—Who are Indians? These developments came with claims of better scientific evidentiality afforded by the DNA and genomic sciences, and placed sciences back at the core of the politically vexed question of origins (Figueira 2015). As a response to the political anxieties stirred up by entanglements of legacy racial frameworks and scientific explorations of the peopling of India, Harvard geneticist David Reich, along with his Indian collaborators published a very influential paper in 2009, which invented two new population labels—Ancestral North Indians (ANI), and Ancestral South Indians (ASI) (Reich et al. 2009). As Barbosa argues (see this volume), these new categorical denominations not only nationalized ancestry but also sought to conceal questions surrounding the foreignness of ancestry, caste inequalities, and religious nationalism. Nevertheless, most models for racial/ethnic classification of the people of India continue to utilize caste as the unit for denoting differences and similarities. In the following sections, I trace and show how the GIP is not disjointed from such historical notions of human variation.

Situating the GenomeIndia Project (GIP)

In the late 1980s, with the availability of commercial DNA sequencers which could facilitate sequencing of the complete genome, the epistemic frame in life sciences witnessed a shift. It was against this backdrop that the Human Genome Project was launched in

1990. With the promise of discovering the "blueprint of life," transforming medicine, and curing cancer, the HGP became a vehicle and marker of exponential growth in genomics. Nonetheless, as Reardon points out, researchers were soon faced with the question: what does genomic data mean? She posits, "this turn to the question of meaning—the question of the uses, significance, and value of the human genome sequence" as the "postgenomic condition" (Reardon 2017, 2). A recent editorial in the leading Indian newspaper echoes the same question about the GIP:

> When the Human Genome Project published its reference "human genome" in 2003 … it rang with a "brave-new-world" promise of … mapping every awry gene to a disease and a future of "personalized medicine." Much of Genome India's sales pitch reflects similar promises. However, the subsequent decades have tempered such expectations … In other words, genome sequencing only opened up new realms of complexity. (Hindu 2024)

But despite the uncertainty and complexity involved, over the last two decades, India has extended public funding to build genomic infrastructure, and the GIP is the culmination of these efforts. While India lacked resources in the 1990s to partner with big science initiatives such as the HGP, it decided not to participate in the International Haplotype Map Project (HapMap) and instead initiated the Indian Genome Variation (IGV) Project in 2003. The Human Genome Diversity Project (HGDP), precursor to the HapMap project, which intended to survey and catalog human genetic diversity, faced public censure in the early 1990s. The criticism of the HGDP as a "vampire project" practicing "biocolonialism" (Burton 2021, 244) had not only alerted countries in the Global South to how genomic resources might be exploited and capitalized by the powerful Global North but also conveyed the political import of safeguarding nations' shared genomes. Already by November 1997, the Indian Council of Medical Research had issued the "Guidelines for Exchange of Human Biological Material for Biomedical Research Purposes," mandating government permission for shipping outside the country any human tissue samples that were collected in India (Hardy et al. 2008).

The IGV project was implemented through six laboratories of the Council of Scientific and Industrial Research (CSIR) in collaboration with the Indian Statistical Institute Kolkata and the Anthropological Survey of India (AnSI). The IGV project was granted USD 5 million to develop markers for predictive medicine and "used SNP [i.e., single nucleotide polymorphism]- based genotyping of 900 genes from over 1800 individuals across fifty-five subpopulations to underscore the heterogeneity of the Indian population" (Jain et al. 2020). Though the IGV project did not perform whole genome sequencing (WGS), in 2009, India announced that a team of scientists at the CSIR Institute of Genomic and Integrative Biology Delhi had completed the entire human genome sequencing of a healthy Indian male (PIB 2009). As Subramaniam has suggested, India thus sought to assert "genomic sovereignty" through practices of bionationalism "that seeks to secure and control their national genomics" (Subramaniam 2019, 165).

However, the commencement of a large-scale study of human variation utilizing WGS happened only around 2017. The Centre for Brain Research (CBR) Bengaluru, which is also spearheading the GIP, inaugurated the Genome India initiative (GII) in

collaboration with the Institute of Bioresources and Sustainable Development (IBSD) Imphal. The GII was a precursor to the GIP but with a focus on a specific region of India (and particularly tribal populations). The GII intended to carry out WGS of 2000 individuals from the northeastern states of India to understand "the genetic disease burden which would help in the development of personalized medicine" (IANS 2017). Soon, after a hiatus of a decade, the CSIR also launched the IndiGen Program in April 2019 with the motto—"Genomics for public health in India." Within a span of six months, the IndiGen consortium announced the completion of whole genome sequencing of 1029 healthy Indians and emphasized "the need for an India centric population genomic initiative" (PIB 2020b). Consequently, in early 2020, the Department of Biotechnology (DBT) launched the GIP.

The ambitious "national project" (CBR 2023) since then has acquired the genetic material from a cohort of Indians to catalog genetic variations. To ensure that genetic variation in Indians is well captured, individuals chosen for genome sequencing were seemingly selected in a way so as to represent the country's diverse population. The GIP collected genetic material from individuals belonging to ninety-nine distinct populations residing in different regions of the country (PIB 2024). The CBR at the Indian Institute of Science, Bengaluru, is leading this ambitious collaborative effort of scientists from twenty research institutes across India. The CBR website lists four major aims for the GIP:

> 1. Create an exhaustive catalog of genetic variations (common, low frequency, rare, single nucleotide polymorphisms or SNPs and structural variations) in Indians.
> 2. Create a reference haplotype structure for Indians. This reference panel can be used for imputing missing genetic variation in future GWA [genome-wide association] studies.
> 3. Design genome wide [*sic*] arrays for research and diagnostics at an affordable cost.
> 4. Establish a biobank for DNA and plasma collected for future use in research. (CBR 2023)

The GIP thus seems to be an attempt at identifying and characterizing genetic differences within Indian populations and constructing a knowledge base for healthcare and medical purposes. On 27 February 2024, the GIP consortium announced that it has achieved the milestone of whole genome sequencing of 10,000 Indian individuals (PIB 2024). Accordingly, the GIP is another addition to various worldwide genomics resource creation efforts, embracing a postgenomic epistemological framework. However, the cataloging of variations and biobanking of DNA would also be a valuable resource for population genetics studies, including reconstructing human migration history. How such aggregation of genetic material will simultaneously characterize differences amongst the Indian populations and generate a reference haplotype that encodes similarities between Indians needs to be problematized. In the following section, I probe epistemological assumptions and concepts that inform the GIP to open up this debate.

Catalog of Genetic Variation or Differences

The HGP made a strong assertion about human DNA similarity, i.e., the DNA sequences of two randomly selected humans would be 99.9 percent identical. As a corollary, the epistemological base of human genetics and genomics entirely lies in that 0.1 percent fraction. Even so, this fraction translates into differences of about two to three million base pairs, as the human genome comprises more than three billion nucleotide base pairs. Hence medical genomics and population genetics research essentially relies on human (genetic) differences. The degree of genetic variation between two populations then becomes a marker of their differences or similarities. However, as Troy Duster has argued, "[i]t is possible to make arbitrary groupings of populations (geographic, linguistic, self-identified by faith, identified by others by physiognomy, etc.) and still find statistically significant allelic variations between those groupings" (Duster 2005). Nevertheless, the differences population genetics valorizes in practice often are not between any arbitrary groupings of individuals or populations but rather between discrete populations identified as relevant using prevalent norms, viz. race/caste. Consequently, epistemic assumptions that inform scientific theories and practices shape how knowledge claims are contested and legitimated and how knowledge is used and applied in multiple contexts. As Lipphardt has argued, for empirical work, researchers continue to rely on race concepts through their "sampling practices, group labels, narratives and the concept of the isolate" (Lipphardt 2014, 51). In the following paragraphs, I illustrate similar continuities within the GIP.

Since the GIP aims to construct an exhaustive catalog of genetic variations in Indians, it is vital to probe how the notion of variation or difference is operationalized. In practice, human biological diversity is conceptualized not in terms of differences between individuals but between populations. As a consequence, the science of genomics inevitably deifies social constructs (viz. ethnicity, geography, race), which is akin to presuming (pre)existence of human races or caste. The GIP is no exception. In fact, an important justification for the GIP is that the majority of genomic studies are based on individuals of European descent. The CBR website puts the rationale for the GIP as follows:

> The Indian population of 1.3 billion consists of >4,600 population groups, and several thousand of them are endogamous … Thus, the Indian population harbors distinct variations and often many disease-causing mutations are amplified within some of these groups. Therefore, findings from population based or disease based human genetics research from other populations of the world cannot be extrapolated to Indians. (CBR 2023)

Therefore, the non-European population, specifically the Indian population, is a distinctive lens for the GIP. Furthermore, the Indian population itself is conceptualized as an aggregate of more than 4,600 separate populations, which are presumed to significantly differ biologically from each other. So, two-level distinctions are operationalized—one, the Indian population is genetically distinct from the rest of the global human population; and second, multiple Indian population groups exhibit distinct gene pools. The GIP then simultaneously intends to look at the distinctiveness of genetic diversity amongst

Indian population groups while paradoxically constructing sameness across thousands of population groups. But what is the ground for the GIP to assume that there are more than 4,600 population groups in India that need to be cataloged? The GIP has adopted the categorization of the Indian population postulated by the AnSI. Between 1985 and 1992, AnSI undertook an ambitious project, "The People of India" (POI), to generate an anthropological profile of all communities living in independent India (Singh 1993). The POI project primarily utilized caste groups as a marker and basis for recording the sociocultural diversity of India. It identified 4,635 stratified communities across India (castes, tribes, minorities, etc.), many of which practice endogamy (Joshi, Gadgil, and Patil 1993). If anthropologically defined discrete groups are the criteria, then the imagined biological Indian identity and the sameness thereof is then necessarily a mosaic of small and large population groups rather than a single large homogeneous population.

The GIP in practice is therefore ambiguously reproducing the historical/colonial practices of population classification that reinforce problematic social constructs (viz. caste, tribe). The sampling strategy and methodology applied for the collection of physical DNA offers a way to discern the assumptions. Institutes that form the GIP consortium are spread across India (twelve states and two union territories) and functioned as DNA collection centers for each region. The CBR, as the nodal center, provided guidance to these centers on which population groups (read caste/tribe) in the respective region are to be sampled and the number of samples to be drawn. A principal investigator of the GIP whom I interviewed as a part of my fieldwork added that DNA samples from ninety-nine distinct population groups were collected for the WGS. They also noted that, based on the absolute population size of the concerned community, the number of individual DNA samples collected from each group varied from 160 to 350. In general, a greater number of DNA samples were collected from communities with large absolute population sizes and vice versa.

The catalog of genetic variations in Indians, as envisaged by the GIP, is in practice deeply rooted in specific forms of sociomateriality. In particular, differences are of prime concern for the GIP, and how this delineation was operationalized in relation to human bodies and embedded DNA provides insights into underlying epistemic assumptions. In order to obtain DNA, the GIP collected blood samples of volunteering individuals from the relevant population groups. Since many participating institutes are located in urban areas, whereas most relevant population groups reside in small towns, villages, and remote hamlets, sampling posed logistic challenges. Furthermore, ethical considerations limited nonmedical academic institutes in collecting samples themselves. Therefore, some institutes partnered with doctors and medical institutes, who organized special blood sample collection camps at various places. Approximately 10 ml of blood sample was collected from each volunteer on site. Furthermore, to ensure precision, apart from unrelated individuals, samples were also collected in trios (mother-father-child) from each respective population group. According to a PI, "We collected extra samples because it is extremely difficult to go back to the community and get samples." This practice reflects both economic and social constraints. They further added that, in the process, the GIP in its first phase of the study has collected blood samples of approximately 20,000 individuals belonging to various Indian population groups, i.e., twice the number of samples it aims to process. However, only those samples that were

found to be originating from "normal and healthy" individuals were sent for genome sequencing. As the CBR website notes, "The study individuals from diverse socioeconomic backgrounds undergo a plethora of anthropometric tests, blood biochemical analyses informative of complex conditions (such as diabetes, dyslipidemia, kidney function, and liver function)" (CBR 2023). The primary (population group) and secondary (blood screening tests) inclusion criteria point to continuities with colonial practices of population classification. Furthermore, as corroborated by a PI in an interview, anthropometric measurements of each participating individual were also carried out. Thus, as Ray notes, "[e]ven though the experimental system of anthropometry is dead, the experimental reasoning of race science … continues to animate genomic research" (Ray 2018). Such epistemic practices associated with naming and grouping populations thus give rise to their own contradictions and politics (Sur and Sur 2008).

Nonetheless, the DNA of the members of discrete population groups is central to the GIP's conceptualization of human biological diversity and its catalog of variations. In an interview, a PI explained, though it was through self-reporting and on occasion through local community vetting, blood samples for the GIP were collected only from the individuals in whose family no "cross-marriage" had taken place in the last five generations, i.e., caste/tribe endogamy. Such methodological and epistemic moves are not neutral but intrinsically linked with identity practices and politics, wherein researchers frame the questions as to what constitutes a valid and legitimate sample of DNA. Nonetheless, if the individual was found to be normal and healthy through blood biochemical analyses, the collection center transferred the blood sample into multiple tubes and added a unique GIP tag. The centers used one tube to extract DNA and dispatched two sample tubes with GIP tags to one of the four sequencing centers. The sequencing center utilized these tubes to extract and store DNA and perform WGS. In the process, the GIP has constructed a new material basis and genomic infrastructure for studying human biological diversity. It has done so first through genome sequencing and construction of baseline data for Indians, with the eight petabytes of data generated by the GIP now hosted at the Indian Biological Data Centre Faridabad; second, by establishing DNA biobanks at the collection as well as four sequencing centers; and third, by means of a biobank of blood and plasma samples at the collection centers and the CBR (housing 20,000 samples).

As evident from the discussion above, the GIP sampling is not only based on typological reasoning provided by AnSI but also relies on the assumption that caste/tribe affords a rational unit for the project. Two types of rationales have been employed to justify this historical continuity—1) representing the country's diverse population, and 2) the need to identify disease-causing mutations resulting from prevalent endogamy for medical purposes. It must be noted that India is home to about 17 percent of the world's population, and as the POI project indicated, there is a remarkable resilience of regional identities (Singh 1993). Furthermore, disease-causing mutations could also be detected using regional sampling in combination with an effective disease surveillance network. Nevertheless, the GIP seems to assert that not every difference counts the same. While some differences are valued as significant, others are judged as trivial and to be ignored safely. The ways of ordering and classifying populations employed by the GIP are thus not entirely divorced from historical practices. As a consequence, the GIP would contribute to the naturalization and biological essentialization of caste/tribe constructs (Egorova

2010). In effect, the GIP is essentially an exercise of difference (re)production—not just vis-à-vis global populations but also between people inhabiting India—which relies on taken-for-granted epistemologies and reinforces problematic social constructs. In the following section, I focus on the notion of similarities and the ways in which the GIP intends to stitch together the category of "Indian" through the imaginary of a reference haplotype.

Stitching Together Indian Reference Haplotype

One of the key motivations and assumptions guiding the GIP is that the Indian population has a distinct genetic makeup compared to other populations of the world. Yet, notwithstanding the myriad differences, the project emphasizes unity in diversity through the idea of a single Indian population demarcated by geographical and nation-state boundaries. The ambition of the GIP is to stitch together all such distinct variations and rich genetic diversity of Indian population groups under one single ancestral group. Apropos, one explicitly stated goal of the GIP is to "create a reference haplotype structure for Indians," with the aim to employ the resulting reference panel "for imputing missing genetic variation in future GWA studies" (CBR 2023). I explore below close linkages between concepts of reference haplotype and descent, the production of similarities thereof, and the embedded politics of nationalism.

A haplotype refers to a single or a set of markers (polymorphisms) present along a single chromosome that tend to be conserved or inherited together. Such specific DNA variations at different locations on a single chromosome provide a genetic signature. In practice, such DNA signatures are used to understand disease predispositions of population groups and to trace migration history and evolutionary patterns. International HapMap Project, which constructed a reference panel of 420 haplotypes, established this approach to throw light on human genetic variation (IHC 2003). A reference panel is constituted by genetic sequences of individuals that belong or are attributed to the same ancestral group or collective (read geography, ethnicity, caste, etc.). A reference data or panel made up of individuals sharing genetic ancestry can be used as a baseline for conducting genome-wide association studies to statistically predict variants associated with a specific trait and imputing missing genetic variation. The governing idea here being, "[i]n samples of unrelated individuals, the haplotypes of the individuals over short stretches of sequence will be related to each other by being identical by descent (IBD)" (Marchini and Howie 2010, 500). Thus, notions of similarity such as descent, ancestry, and kinship are closely tied to the creation of a reference haplotype. Furthermore, while the GIP is invested with (re)production of difference amongst Indian population groups, as discussed in the previous section, a reference haplotype paradoxically aims at both the production of similarity within the Indian population and the simultaneous construction of difference between Indian and non-Indian populations. The GIP is thus a Janus-faced enterprise, where scientific knowledge-making is shaped by population labels that both construct and subsume differences, simultaneously invoking different scientific meanings and practices.

Nonetheless, what might it mean to construct a reference haplotype that represents the genetic diversity of the Indian population? The GIP has obtained DNA samples from ninety-nine discrete population groups, which is inadequate to truly represent genetic diversity. The GIP researchers are currently engaged in genotyping and comparing each individual genome with other genomes from the same population groups, as well as comparing genomes of one population group with other population groups. Such comparative analysis of genomic variation within and between sampled Indian population groups will lead to a collection of SNPs and haplotypes, which would be used to construct a reference haplotype structure for Indians. During a field interview, a PI emphasized that the biggest challenge the GIP researchers are facing is that of genome alignment and stitching together genetic information. Preliminary findings released by the GIP, denoting differential prevalence of variations across Indian and other world populations, suggest that 69.02 percent of genetic variations have less than 5 percent differences in prevalence (CBR 2023). On the other hand, approximately 10 percent of genetic variations are unique to the Indian population and not found in other populations (ibid.). But are those variations uniformly observed in all population groups in India? Drawing attention to the GIP data, a PI commented that there is no singular Indian genome but rather multiple genomes, and in particular, northeastern region of the country exhibits the most diverse genetic makeup. Therefore, in practice, multiple contradictions challenge the assumptions of similarity made by the GIP researchers.

Even so, how can we make sense of the politics of stitching together a reference haplotype for Indians? The adjective Indian is an operational trope in this context. By naming the proposed baseline haplotype as Indian, i.e., on a national basis, the GIP is attempting to formulate a form of homogeneous biological identity. Such aggregation is necessary for the making of what Sung terms as "bionation," which engages in the politics of resource-making—be it through people's bodies or information coded in their genes (Sung 2010). The imaginary of Indian reference haplotype then is a way to reinscribe India as a nation on biological terms, distinct from other global populations. This however requires establishing "genomic sovereignty" through the formation and assertion of a genetically unique and homogeneous identity of the Indian population. As Warwick Anderson has suggested, "the clinic and laboratory should be added to those sites where the nation—any nation—may be imagined" (Anderson 2006). At another level, this aggregation of genetic diversity by subsuming differences could also be viewed as a move to create a secular biological identity detached from traditional caste/tribe markers. The reference haplotype might thus occasion a partial collapse of two contemporary population labels, i.e., ANI and ASI. Nevertheless, as Barbosa (see this volume) suggests, it needs to be read as (de)politicization of diffuse caste politics. Furthermore, the GIP is also a means for the state to govern its populations through the promise of science-based healthcare for its citizens, viz., rare diseases (Guardian, Sivasubbu, and Scaria 2019). Yet, as discussed above, Indian reference haplotype draws on the notion of descent. In practice, for mapping disease-causing variations onto an individual's genomic data, ethnic information would play a key role because of the defining assumptions and epistemic conceptualizations of the GIP. As Burton has suggested, these continuities reiterate that "nationalism and human genetics are so thoroughly co-constituted, neither better technology nor better intentions are sufficient … to produce politically neutral data about ancestry" (Burton

2021, 248). The production of similarities through the trope of a reference haplotype and shifting focus from population genetics to medical genomics thus does not fully dissociate science epistemologies from the politics of nationalism.

Conclusions

The postgenomic era in the new millennium has created an impetus within the life sciences to study human biological diversity using DNA information. But as Burton suggests, "[a]lthough the capabilities of DNA sequencing far exceed older technologies […] geneticists repeat the process of reifying religious, linguistic, and social differences into biologically detectable ethnic or racial groups" (Burton 2021, 260). In this article, I examined the epistemological assumptions and concepts that inform the GenomeIndia Project and their entanglements with caste and identity politics. The GIP aims to create a catalog of genetic variations in Indians through the whole genome sequencing of 10,000 individuals and construct a reference haplotype structure for Indians. By focusing on empirical research practices employed for sampling and labeling groups, I have highlighted how the GIP relies on taken for granted epistemologies for characterizing difference, which reinforces social constructs of caste/tribe. To be sure, the GIP does promise to remove the personal identifiers of the participating individuals. Although data in its entirety may not be made public, in order to create and maintain a catalog of genetic variations in Indians as well as DNA and blood biobank of samples, caste/tribe labels that trace the provenance of data have been preserved by all the GIP centers. The difficulties in unmooring identities are well reflected in the statement of the DBT secretary, who admitted that "[t]his hidden treasure of genomes also comes with a hidden privacy concern that the Department is striving to address ethically" (Mishra 2024). While there are disagreements amongst researchers on the relationship between the concept of race and caste (Natarajan and Greenough 2009), caste differentiation and identity discourse linked with the question of race and migration continue to remain a politically vexed subject.

Another aim of the GIP is to construct a reference haplotype structure for Indians, with a vision to facilitate the identification of disease-causing mutations. Yet by naming the baseline haplotype as Indian and using a nonbiological criterion in the production of similarities, the GIP seeks to forge a homogeneous biological identity and nationalize ancestry. I suggest that notwithstanding the uncertainty and complexity posited by the postgenomic condition, the GIP must also be understood as a genomic infrastructure and means to assert "genomic sovereignty." Through accumulating a baseline genomic digital database, establishing DNA biobanks, and setting up a blood and plasma biobank, the GIP has constructed a new material basis for studying human biological diversity. Practices of bionationalism steer such aggregation and enable the making of a "bionation." Consequently, human genetics and nationalism are coconstituted. I posit that while population genetics and medical genomics continue to be guided by the normative logic of genetic nationalism, in the case of India, the logic remains closely tied to the diffuse politics of identity and caste. As such, the GIP's notions of a reference hap-

lotype or catalog of genetic diversity cannot fully dissociate science epistemologies from politics.

References

Anderson, Warwick. 2006. The Cultivation of Whiteness: Science, Health, and Racial Destiny in Australia. Durham, NC: Duke University Press.

Bamshad, Michael, Toomas Kivisild, W. Scott Watkins, et al. 2001. "Genetic Evidence on the Origins of Indian Caste Populations." *Genome Research* 11 (6): 994–1004. DOI: https://doi.org/10.1101/gr.173301.

Bates, Crispin. 1995. "Race, Caste and Tribe in Central India: The Early Origins of Anthropometry." In *The Concept of Race in South Asia*, edited by Peter Robb, 219–259. Delhi: Oxford University Press.

Bhagat, Ram B. 2006. "Census and Caste Enumeration: British Legacy and Contemporary Practice in India." *Genus* 62 (2): 119–134.

Burton, Elise K. 2021. *Genetic Crossroads: The Middle East and the Science of Human Heredity.* Stanford, CA: Stanford University Press.

Centre for Brain Research. 2023. GenomeIndia. Accessed March 27, 2024. https://www.cbr.iisc.ac.in/research/flagship-projects/genomeindia.

Duster, Troy. 2005. "Race and Reification in Science." *Science* 307 (5712): 1050–1051. DOI: https://doi.org/10.1126/science.1110303.

Egorova, Yulia. 2010. "Castes of Genes? Representing Human Genetic Diversity in India." *Genomics, Society and Policy* 6 (3): 32–49. DOI: https://doi.org/10.1186/1746-5354-6-3-32.

Figueira, Dorothy M. 2015. *Aryans, Jews, Brahmins: Theorizing Authority through Myths of Identity.* New Delhi: Navayana.

The GUaRDIAN Consortium, Sridhar Sivasubbu, and Vinod Scaria. 2019. "Genomics of rare genetic diseases—experiences from India." *Human Genomics* 13. DOI: https://doi.org/10.1186/s40246-019-0215-5.

Hardy, Billie-Jo, Béatrice Séguin, Peter A. Singer, Mitali Mukerji, Samir K. Brahmachari, and Abdallah S. Daar. 2008. "From Diversity to Delivery: The Case of the Indian Genome Variation Initiative." *Nature Reviews Genetics* 9 (S1): S09–S14. DOI: https://doi.org/10.1038/nrg2440.

The Hindu. 2024. Decoding the Script: On the Genome India Project and Its Sequencing 10,000 Indian Genomes. Accessed March 27, 2024. https://www.thehindu.com/opinion/editorial/decoding-the-script-on-the-genome-india-project-and-its-sequencing-10000-indian-genomes/article67899979.ece.

Indio-Asian News Service. 2017. Genome India Project Launched in Northeast. Accessed March 27, 2024. https://economictimes.indiatimes.com/news/science/genome-india-project-launched-in-northeast/articleshow/61203159.cms?from=mdr.

The International HapMap Consortium. 2003. "The International HapMap Project." *Nature* 426 (6968): 789–796. DOI: https://doi.org/10.1038/nature02168.

Jain, Abhinav, Rahul C. Bhoyar, Kavita Pandhare, et al. 2021. "IndiGenomes: A Comprehensive Resource of Genetic Variants from over 1000 Indian Genomes." *Nucleic Acids Research* 49 (D1): D1225–D1232. DOI: https://doi.org/10.1093/nar/gkaa923.

Joseph, Tony. 2021. *Early Indians: The Story of Our Ancestors and Where We Came From*. New Delhi: Juggernaut.

Joshi, N. V., Madhav Gadgil, and Suresh Patil. 1993. "Exploring Cultural Diversity of the People of India." *Current Science* 64 (1): 10–17. https://www.jstor.org/stable/24095541.

Lipphardt, Veronika. 2014. "'Geographical Distribution Patterns of Various Genes': Genetic Studies of Human Variation after 1945." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 47 (A): 50–61. DOI: https://doi.org/10.1016/j.shpsc.2014.05.006.

Malhotra, K. C., and T. S. Vasulu. 2019. "Development of Typological Classification and Its Relationship to Microdifferentiation in Ethnic India." *Journal of Biosciences* 44 (3), article no. 64 DOI: https://doi.org/10.1007/s12038-019-9880-8.

Marchini, Jonathan, and Bryan Howie. 2010. "Genotype Imputation for Genome-Wide Association Studies." *Nature Reviews Genetics* 11 (7): 499–511. DOI: https://doi.org/10.1038/nrg2796.

Mishra, Akanksha. 2024. "India Largest Genetic Lab in the World": What Completion of India Genome Project Means. Accessed March 27, 2024. https://theprint.in/health/india-largest-genetic-lab-in-the-world-what-completion-of-india-genome-project-means/1982036/.

Mukharji, Projit Bihari. 2014. "From Serosocial to Sanguinary Identities: Caste, Transnational Race Science and the Shifting Metonymies of Blood Group B, India c. 1918–1960." *The Indian Economic and Social History Review* 51 (2): 143–176. DOI: https://doi.org/10.1177/0019464614525711.

Natarajan, Balmurli, and Paul Greenough, eds. 2009. *Against Stigma: Studies in Caste, Race and Justice since Durban*. New Delhi: Orient BlackSwan.

Papiha, S. S. 1996. "Genetic Variation in India." *Human Biology* 68 (5): 607–628. https://www.jstor.org/stable/41465511.

Philip, Kavita. 2004. *Civilizing Natures: Race, Resources, and Modernity in Colonial South India*. New Brunswick, NJ: Rutgers University Press.

Press Information Bureau, Government of India. 2009. CSIR Completes First Ever Human Genome Sequencing in India. Accessed March 27, 2024. https://pib.gov.in/newsite/erelcontent.aspx?relid=55470.

Press Information Bureau, Government of India. 2020a. GenomeIndia: Cataloguing the Genetic Variation in Indians' Project. Accessed March 27, 2024. https://pib.gov.in/PressReleseDetailm.aspx?PRID=1605509.

Press Information Bureau, Government of India. 2020b. CSIR IndiGenome Resource of 1029 Indian Genomes Provides a Compendium of Genetic Variants Representing the Contemporary Indian Population. Accessed March 27, 2024. https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1667949.

Press Information Bureau, Government of India. 2024. GenomeIndia. Accessed March 27, 2024. https://www.pib.gov.in/PressReleaseIframePage.aspx?PRID=2009505.

Ray, Subhadeepta. 2018. "Studying Laboratories." In *Towards a New Sociology*, edited by Mahuya Bandyopadhyay and Ritambhara Hebbar, 141–176. New Delhi: Orient Blackswan.

Reardon, Jenny. 2005. *Race to the Finish: Identity and Governance in an Age of Genomics*. Princeton: Princeton University Press.

Reardon, Jenny. 2017. *The Postgenomic Condition: Ethics, Justice, and Knowledge after the Genome*. Chicago: University of Chicago Press.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L. Price, and Lalji Singh. 2009. "Reconstructing Indian Population History." *Nature* 461 (7263): 489–494. DOI: https://doi.org/10.1038/nature08365.

Singh, K. S. 1993. "People of India: The Profile of a National Project (1985–92)." *Current Science* 64 (1): 5–10. https://www.jstor.org/stable/24095540.

Subramaniam, Banu. 2019. *Holy Science: The Biopolitics of Hindu Nationalism*. Seattle: University of Washington Press.

Sung, Wen-Ching. 2010. "Chinese DNA: Genomics and Bionation." In *Asian Biotech: Ethics and Communities of Fate*, edited by Aihwa Ong and Nancy N. Chen, 263–292. Durham, NC: Duke University Press.

Sur, Abha, and Samir Sur. 2008. "In Contradiction Lies the Hope: Human Genome and Identity Politics." In *Tactical Biopolitics: Art, Activism, and Technoscience*, edited by Beatriz da Costa and Kavita Philip, 205–218. Cambridge, MA: MIT Press.