My research interest is nested. There are two major components of my research. One component is to understand human genome diversity and evolution and systematically connecting that to genomic underpinning of diseases. This endeavor leads to the understanding of human genetic diversity and variation and provides us a window into human pre-history as well as corroborate findings from genetic architecture of populations to documented events in human past. It also provides the basic framework to look at the natural history of diseases, the multifactorial nature of disease causation considering host and exposure factors together. This facilitates identification of populations to design genetic epidemiological studies to identification of genomic regions responsible for differential expression of heritable traits, i.e. connecting genotype with the phenotype.

The Science Council of the World Health Organization (WHO) has identified genomics as a tour de force and its goal has been to extend the substantial and extensive benefits of genomics for personal and public health. It categoricaly emphasize in its release "Current uses and future applications of genomic technologies are critical for improving the health and livelihood of people in all parts of the world, regardless of economic status". The major imperative has been to extend genomics of the people, by the people and for the people of the developing countries. My endeavor and achievements perfectly resonates with that vision of WHO.

I identify myself as a Human Geneticist.I have developed novel statistical methods of gene-mapping using admixed populations. I have taken leadership in largescale studies of gene-mapping for metabolic diseases which I started in United States during my post-doctoral research tenure and I am continuing successfully after my return to India.

By the turn of the last millennium, Molecular biology and genetics have completely redefined the modus operandi, the gold standards and the grammar of many scientific endeavors. The study of human past, its evolutionary and demographic history, its connection with the natural history of diseases and identification of its genetic underpinnings is one of them. The technology that rapidly enables the reading of DNA and RNA, the fundamental molecules of life, has been a game-changer. My PhD thesis with Professor Partha Pratim Majumder, was able to answer many of these unanswered questions, settled plenty of debates and initiated some more. Our study of analyzing the population diversity of India – encompassing mitochondrial DNA, Y-chromosomal DNA and autosomal DNA variations –was the first comprehensive reconstruction of the trails and dynamics of how anatomically modern humans(AMH) entered and populated the Indian sub-continent. Our work provided deep and fundamental insights on the population histories and affinities among Indian population groups. This paper in *Genome Research* (2003)[1] is one of the most cited papers on Indian population genetics (cited 425 times) and still remains one of the most referred publications in the entire discourse. In addition to providing support to aphorisms about Indian population diversity and structure, we showed that India has been populated by a small number of 'founding mothers'. We showed that the Austro-Asiatic speaking tribal populations are the autochthones of the India. Besides the widely studied population migrations through North-

Western India, this study reinstated the importance of the North-East corridor for ancient migrations of people into mainland India. We also postulated that Dravidian speakers were widespread throughout India and were possibly pushed to the southern region by the entry of Indo-European speakers from South-Central Asia. The major findings of this study have stood the test of time and many of them were later substantiated by larger population genetic studies by ours and other groups. More recently our contribution in defining the ancestral 'components' of Indian population has been a revelation. We have corrected the conclusion of an earlier major study – published in 2009 (Reich et al *Nature* 461, 489-494)[2] – that modeled the population history of India and concluded that the present-day Indians are derived from two ancestral groups of people, one of whom is ancestral primarily to all north Indians and the other ancestral south Indians. I quote from the abstract of the above mentioned paper "*India has been underrepresented in genome-wide surveys of human variation. We analyse 25 diverse groups in India to provide strong evidence for two ancient populations, genetically divergent, that are ancestral to most Indians today. One, the 'Ancestral North Indians' (ANI), is genetically close to Middle Easterners, Central Asians, and Europeans, whereas the other, the 'Ancestral South Indians' (ASI), is as distinct from ANI and East Asians as they are from each other.*" This understanding, although profound and is one of first systematic dissection of the Indian population, is grossly incomplete particularly for a "genomics-aware" biomedicine. This would be extremely misleading and incomplete when we consider the genomes of individuals and populations where a large proportion of components might be inherited from other ancestries if they remained undiscovered.

We have been able to provide robust evidence that four – not two – ancestral groups contributed to the genetic diversity of present-day mainland Indians [*Proceedings of the National Academy of Sciences* 113: 2016: Number of citations: 197; Altmetric Attention score: 293; In the top 5% of all research outputs and top 1% in all research outputs of the same age][3]. These ancestral groups are roughly identifiable with the four language families in India – Indo-European (north India), Dravidian (south India), Tibeto-Burman (north-east India) and Austro-Asiatic (fragmented in east and central India; spoken exclusively by the tribals). We identified the ancestry predominant among the Indo-European speakers to be co-ancestral to South-Central Asia. This is the ancestry which was identified as the ANI ancestry in [2]. We identified the ancestry predominant among the Tibeto-Burman speakers to be co-ancestral to East and South-East Asia; but the ancestor groups or origin of the Austro-Asiatics and the Dravidian tribals remain unidentified. Besides unravelling the history, this discourse has underscored the importance of representation of under-represented populations in studies of genetic variation. It laid the foundation to construct the framework to undertake genomics driven health solutions both at the personal as well as the public health context.

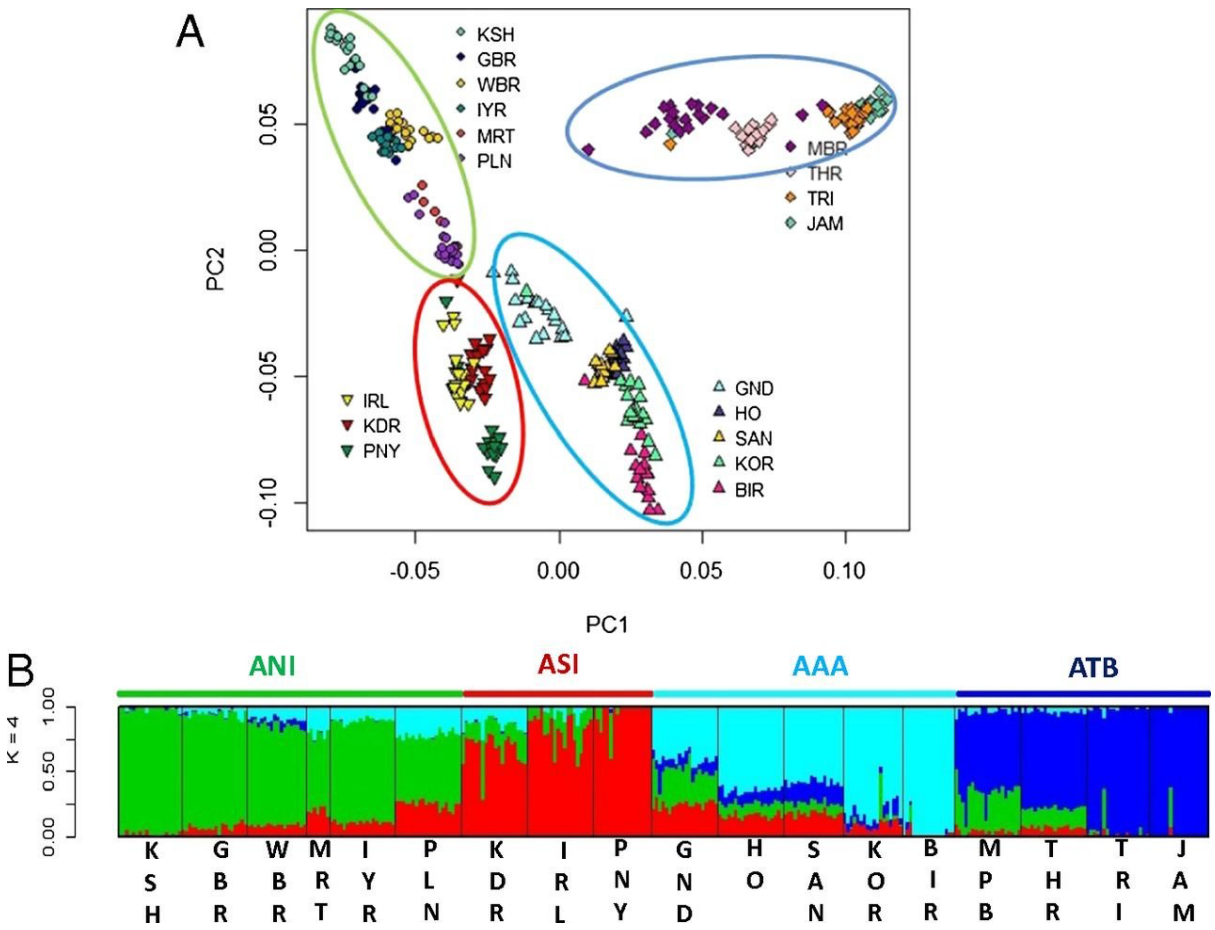

Figure 1: (A) Scatterplot of 331 individuals from 18 mainland Indian populations by the first two PCs extracted from genome-wide genotype data. Four distinct clines and clusters were noted; these are encircled using four colors. (B) Estimates of ancestral components of 331 individuals from 18 mainland Indian populations. A model with four ancestral components (K = 4) was the most parsimonious to explain the variation and similarities of the genome-wide genotype data on the 331 individuals. Each individual is represented by a vertical line partitioned into colored segments whose lengths are proportional to the contributions of the ancestral components to the genome of the individual. Population labels were added only after each individual's ancestry had been estimated. We have used green and red to represent ANI and ASI ancestries; and cyan and blue with the inferred AAA and ATB ancestries. These colors correspond to the colors used to encircle clusters of individuals in A.

We also identified a fifth ancestral lineage that is dominant among the hunter-gatherer tribals of Andaman and Nicobar Islands (Jarawa and Onge). We also found evidence that this lineage is also ancestral to the present-day Pacific Islanders.

We have compared the genomic spectrum of populations of India and showed its uniqueness, under-representation and importance in understanding global diversity [*Genome Biol Evol.* 8 (11): 2016, *Indian Journal of History of Science: 2016, Journal of Biosciences: 2019*][4][5][6]. This discourse culminated in efforts to understand and catalogue the genetic

variation in the (a) Indian-subcontinent, or broadly in the (b) Asian continent. In order to achieve (a), the Department of Biotechnology (DBT), has initiated a multi-institutional, multi-centric effort, known as "Genome India"; while to achieve (b), there is an international public-private effort: "GenomeAsia 100K", where I as an investigator have played a lead role [*Nature* 576: 2020][7]. The GenomeAsia 100K Pilot study already have a profound impact on understanding of disease genomics and a tremendous impact on the genomics awareness of Asian populations. It has added enormously to our understanding of clinically relevant mutations. It has catalogued founder effects in Asian populations which can greatly facilitate new studies on genetic discoveries. It has identified mutations pertaining to rare diseases and cancer. It has also documented variants of pharmacogenetic relevance which have adverse effects on response to different drugs often have large frequencies in many Asian populations.

[1] BASU A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP (2003) Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Research* 13(10):2277-90

[2] Reich, D., Thangaraj, K., Patterson, N. *et al.* (2009) Reconstructing Indian population history. *Nature* **461**, 489–494. https://doi.org/10.1038/nature08365

[3] BASU A, Sarkar-Roy N, Majumder PP (2016) Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. *Proc Natl Acad Sci U S A.* Feb 9;113(6):1594-9.

[4] Sengupta D, Choudhury A, BASU A, Ramsay M (2016) Population stratification and underrepresentation of Indian subcontinent genetic diversity in the 1000 Genomes Project dataset. *Genome Biol Evol.* 8 (11): 3460-3470.

[5] BASU A (2016) The Dazzling Diversity and the Fundamental Unity: Peopling and the Genomic Structure of Ethnic India. *Indian Journal of History of Science* 51.2.2 406-416

[6] Chakraborty S, BASU A (2019) Reconstruction of ancestral footfalls in South Asia using genomic data. *Journal of Biosciences.* 44: 74. https://doi.org/10.1007/s12038-019-9875-5

[7] GenomeAsia 100K Consortium (2020) The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576, 106–111

Signature of the Applicant

Date: 2nd September 2023