==========

THE FIRST SIX PAPERS ARE RELATED TO GENOMICS OF INDIAN POPULATIONS, IT'S UNDER-REPRESENTATION IN GLOBAL STUDIES OF GENOMIC VARIATION AND IT'S IMPLICATION IN STUDIES OF BIOMEDICAL GENOMICS

THE OTHER FOUR ARE A COMBINATION OF PAPERS WHERE THERE HAS BEEN A STRONG COMPONENT OF METHOD DEVELOPMENT AND APPLICATION OF THOSE METHODS FOR A DEEPER UNDERSTANDING OF PROBLEMS RELATED TO BIOMEDICAL GENOMICS AND PUBLIC HEALTH

---

1) **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. **(IF:12.779)**

[Citations: 197, Altmetric Score: 293, this article is in the 99th percentile and it is in the top 5% of all research outputs ever tracked by Altmetric.]

This paper unravels the mosaic of complex population admixture and the deep intricacies of migration in south and south-east Asia. This knowledge is fundamental for fine-mapping of genetic architectures of diseases common in India as well as any progress towards precision medicine. This study underscores the importance of studying small isolated populations for the purpose of biomedical genomics. It shows that admixture *was* wide-spread in Indian populations and hence populations which have small census sizes today can have a large genetic presence in all populations through ancient admixture. We have identified four ancestral genetic components that have contributed to the formation of Indian population groups, which was a significant advance from the then prevailing notion of two ancestors. We estimated proportions of these ancestral components in individual ethnic populations and dissected the genetics of traits of biomedical significance.

2) **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 **(IF:9.438)**

[Citations: 425]

This early paper of ours was quoted in the popular press as a "Unity in Diversity" paper. In this paper we showed that there is a fundamental unity among the mitochondrial haplogroups in India across population groups, indicating a small number of female founders. This has an immense long-standing impact on the genomic underpinning of diseases. We also showed that population groups in India are genetically largely differentiated and

formation of populations by fission that resulted in founder and drift effects have left their imprints on the genetic structures of contemporary populations. This warrants a comprehensive planning for genomics driven health approaches to be diversity aware. We also showed that linguistic groups can be good proxies for genetic similarity. We predicted the migration routes of populations and inferred that the ancestors to Austro-Asiatic speaking tribal populations as the autochthons of the Indian subcontinent.

[Many findings of this paper have later been validated and substantiated by the work of multiple researchers, including ours]

3) Tagore D, Aghakhanian F, Naidu R, Phipps ME, **BASU A*** (2021) Insights into the demographic history of Asia from common ancestry and admixture in the genomic landscape of present-day Austroasiatic speakers. *BMC Biology* 19(1):61 **(IF: 7.37)**

[Citations: 7]

In this recent paper we have shown the deep connection between the Austro-Asiatic speaking tribal people who live in scattered isolated populations across South and SouthEast Asia. The Austro-Asiatic speaking tribal people have been the earliest inhabitants of both South and South-East Asia. Although, their population sizes are small, because of their widespread presence across the vast region, their genomes are integrated and admixed with all populations which inhabit the region. It is therefore crucial that these underrepresented populations be studied in detail, not only to have a better understanding of history and prehistory but also for a better understanding of the health and well-being of all populations.

4) GenomeAsia 100K Consortium (2020) The GenomeAsia 100K Project enables genetic discoveries across Asia. *Nature* 576, 106–111 **(IF:69.5)**

[Citations: 238, Altmetric Score: 536, this article is in the 99th percentile and it is in the top 10% of all articles of similar age in Nature.]

This study is a culmination of the series of studies through which we emphasized the underrepresentation of non-Europeans in human genetic studies so far has limited the diversity of individuals in genomic datasets and led to reduced medical relevance for a large proportion of the world's population. Population-specific reference genome datasets as well as genome-wide association studies in diverse populations are needed to address this issue. This study includes a whole-genome sequencing reference dataset from 1,739 individuals of 219 population groups and 64 countries across Asia. We catalogue genetic variation, population structure, disease associations and

founder effects. We also explore the use of this dataset in imputation, to facilitate genetic studies in populations across Asia and worldwide.

5) 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. **(IF:4.065)**

[Citations: 36]

Genomic variation in Indian populations is of great interest due to the diversity of ancestral components, social stratification, endogamy and complex admixture patterns. With an expanding population of 1.4 billion, India is also a treasure trove to catalogue innocuous as well as clinically relevant rare mutations. Acknowledging this importance, the 1000 Genomes Project (KGP) Phase-3 data include about 500 genomes from five linguistically defined Indian-Subcontinent (IS) populations (Punjabi, Gujrati, Bengali, Telugu and Tamil) some of whom are recent migrants to USA or UK. This is a corrective measure compared to previous studies of global variation where the populations from the Indian sub-continent were either absent or grossly under-represented. Comparative analyses show that despite the distinct geographic origins of the KGP-IS populations, only one among the four ancestral components of the Indian population is predominantly represented in this dataset. These substructured populations have characteristic/significant differences in heterozygosity and inbreeding coefficients. Moreover, we demonstrate that the substructure is better explained by factors like differences in proportion of ancestral components, and endogamy driven social structure rather than invoking a novel ancestral component to explain it. Therefore, using language and/or geography as a proxy for an ethnic unit is inadequate for many of the IS populations. This highlights the necessity for more nuanced sampling strategies or corrective statistical approaches, particularly for biomedical and population genetics research in India.

6) Majumder PP, **BASU A** (2014) A Genomic view of peopling and population structure of India

Cold Spring Harb Perspect Biol. pii: a008540. **(IF:9.708)**

[Citations: 40]

This is a review article which collates evidence from a large body of our work as well as work done in this field. It starts with the genetic evidence of evolution of anatomically modern humans in Africa about 150,000 years ago and, consistent with paleontological evidence, its evidence of recent migration out of Africa. And through a series of settlements, demographic expansions, and further migrations, they populated the entire world. One of the first waves of migration from Africa was into India. Subsequent, more recent, waves of migration from other parts of the world have resulted in India being a genetic

melting pot. Contemporary India has a rich tapestry of cultures and ecologies. There are about 400 tribal groups and more than 4000 groups of castes and subcastes, speaking dialects of 22 recognized languages belonging to four major language families. The contemporary social structure of Indian populations is characterized by endogamy with different degrees of porosity. The social structure, possibly coupled with large ecological heterogeneity, has resulted in considerable genetic diversity and local genetic differences within India. In this essay, we provide genetic evidence of how India may have been peopled, the nature and extent of its genetic diversity, and genetic structure among the extant populations of India. We also emphasize how these variations need to be considered in designing studies of biomedicine in India.

7) Das S, Biswas N, **BASU A*** (2023) Mapinsights: deep exploration of quality issues and error profiles in high-throughput sequence data. *Nucleic Acid Research* 51(14):e75 **(IF:19.16)**

This paper investigates errors in variant calling that results from High-throughput sequencing (HTS). A false variant discovery can have disastrous implications in application of genomic technologies to health and diseases. It can mislead genomics driven decision making in precision medicine, like but not limited to genomics driven treatments in human cancers. HTS has revolutionized science by enabling super-fast detection of genomic variants at base-pair resolution. Consequently, it poses the challenging problem of identification of technical artifacts, i.e. hidden non-random error patterns. Understanding the properties of sequencing artifacts holds the key in separating true variants from false positives. Here, we develop Mapinsights, a toolkit that performs quality control (QC) analysis of sequence alignment files, capable of detecting outliers based on sequencing artifacts of HTS data at a deeper resolution compared with existing methods. Mapinsights performs a cluster analysis based on novel and existing QC features derived from the sequence alignment for outlier detection. We applied Mapinsights on community standard open-source datasets and identified various quality issues including technical errors related to sequencing cycles, sequencing chemistry, sequencing libraries and across various orthogonal sequencing platforms. Quantitative estimates and probabilistic arguments provided by Mapinsights can be utilized in identifying errors, bias and outlier samples, and also aid in improving the authenticity of variant calls.

8) Bhattacharyya C, Barman D, Tripathi D, Dutta S, Bhattacharya C, Alam M, Choudhury P, Devi U, Mahanta J, Rasaily R, **BASU A***, Paine SK (2023) The influence of maternal Breast milk and Vaginal microbiome on Neonatal gut microbiome: A longitudinal study over the first year of birth. *Microbiology Spectrum* Apr 17;e0496722. doi: 10.1128/spectrum.04967-22. Online ahead of print. **(IF: 9.04)**

Recent research has identified the crucial role of Gut microbiome in maintaining healthy life. Several beneficial bacteria and fungi have been identified to be symbiotically associated with human life, by protecting the host from infectious as well as metabolic diseases. They also play a pivotal role in Immune homeostasis. It has also been observed that the gut microbiome is also involved in human developmental biology including brain development and cognition. Our paper explores the dynamic nature of the neonatal gut microbiome starting from inception and its establishment during the breastfeeding tenure to its shift to adult-like architecture after initiation of solidified food. As the gut microbiome of the offspring is likely to be a resultant of the environment and the exposure, the study explored the mothers' birth canal microbiome on onset of labor. We followed up with the mothers and mapped the maternal breastmilk for 6 months; from the initiation of lactation to 6 months after childbirth at an interval of 3 months. The study observed that the gut of infants who are born through vaginal delivery acquired beneficial bacteria like Lactobacillus, Bifidobacterium and Megashera in higher proportion compared to babies born through C-Section, although the architectural difference becomes less pronounced after initiation of solidified food. C-Section delivered neonates, who are deprived from birth canal exposure, developed less diversified and unstable bacteria colonies at early life and had a lower proportion of beneficial bacteria in their gut even after initiation of solidified food. This less diversified unstable gut microbiome is associated with long term adverse consequences on metabolism and immune maturation and may be a susceptibility factor for various metabolic and chronic diseases. The study also postulates that usage of antibiotics as prophylaxis in C-Section mothers also reflects in breast milk microbiome by reducing its diversity compared to breast milk of mothers who deliver vaginally. This is the first study which simultaneously explore the maternal and neonatal microbiome dynamics and is likely to open a window of opportunity for future research on the formulation probiotic supplement to C-Section neonates who are deprived of the birth canal microbiome exposures and also for milk microbiome supplementation for the babies from non-lactating mothers.

9) Paine SK, Rout U, Bhattacharyya C, Parai D, Alam M, Nanda R, Tripathi D, Choudhury P, Kundu C, Pati S, Bhattacharya D, **BASU A*** (2022) Temporal Dynamics of Oropharyngeal Microbiome among SARS-CoV-2 patients reveals continued dysbiosis even after Viral Clearance. *NPJ Biofilms and Microbiomes* 67 (8) **(IF: 7.55)**

This study was conducted in rapid response to the outbreak of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) pandemic. Clinical features and sequela of SARS-CoV-2 infection include long-term and short-term complications often clinically indistinguishable from bacterial sepsis and acute lung infection. Post-hoc studies of previous SARS outbreaks postulate secondary bacterial infections with microbial dysbiosis. Oral microbial dysbiosis, particularly the altered proportion of Firmicutes and Proteobacteria, observed in other respiratory virus infection, like influenza, has shown to be associated with increased morbidity and mortality. Oropharynx and lung share similar kinds of bacterial species. We hypothesized that alteration in the Human Oropharyngeal Microbiome in SARS-CoV-2 patients can be a clinical indicator of bacterial infection related complications. We made a longitudinal comparison of oropharyngeal microbiome of 20 SARS-CoV-2 patients over a period of 30 days; at three time points, with a 15 days interval; contrasting them with a matched group of 10 healthy controls. Present observation indicates that posterior segment of the oropharyngeal microbiome is a key reservoir for bacteria causing pneumonia and chronic lung infection on SARS-CoV-2 infection. Oropharyngeal microbiome is indeed altered and its α-diversity decreases, indicating reduced stability, in all SARS-CoV-2 positive individuals right at Day-1; i.e. within ~24 h of post clinical diagnosis. The dysbiosis persists long-term (30 days) irrespective of viral clearance and/or administration of antibiotics. There is a severe depletion of commensal bacteria phyla like Firmicutes among the patients and that depletion is compensated by higher proportion of bacteria associated with sepsis and severe lung infection from phyla Proteobacteria. We also found elevated proportions of certain genus that have previously been shown to be causal for lung pneumonia in studies of model organisms and human autopsies' including Stenotrophomonas, Acenetobactor, Enterobactor, Klebsiella and Chryseobacterium that were to be elevated among the cases. We also show that responses to the antibiotics (Azithromycin and Doxycycline) are not uniform for all individuals.

10) **BASU A**, Tang H, Arnett D, Gu CC, Mosley TH, Kardia S, Luke A, Tayo B, Cooper R, Zhu X, Risch N (2009) Admixture Mapping of Quantitative Trait Loci for BMI in African Americans: Evidence for Loci on Chromosomes 3q, 5q and 15q. *Obesity* 17(6): 1226-31. **(IF:9.298)**

[Citations: 41]

This paper is one among few papers where we have studied admixed populations in the United States, who have extreme burden of non-communicable diseases but are considerably underappreciated in

genomewide studies. We developed novel methods for analysis of admixture and used a modified admixture mapping technique to identify regions of the genome which have high probability of harboring a disease causing allele which is in high frequency in one of the ancestral populations. In this paper we study obesity which is a heritable trait and a major risk factor for highly prevalent common diseases such as hypertension, cardiac diseases and type 2 diabetes. Obesity is also a major public health concern worldwide. Using Individual Ancestry (IA) estimates at 284 marker locations across the genome, we now present a Quantitative Admixture Mapping (QAM) analysis of body mass index (BMI) in the same population. We used a set of unrelated individuals from Nigeria to represent the African ancestral population and the European Americans in the Family Blood Pressure Program as the European ancestral population. The analysis was based on a common set of 284 microsatellite markers genotyped in all three groups. We considered the quantitative trait, BMI, as the response variable in a regression analysis with the marker location specific excess European ancestry as the explanatory variable. After suitably adjusting for different covariates such as sex, age and study center, we found strong evidence for a positive association with European ancestry at chromosome locations 3q29 and 5q14 and a negative association on chromosome 15q26.