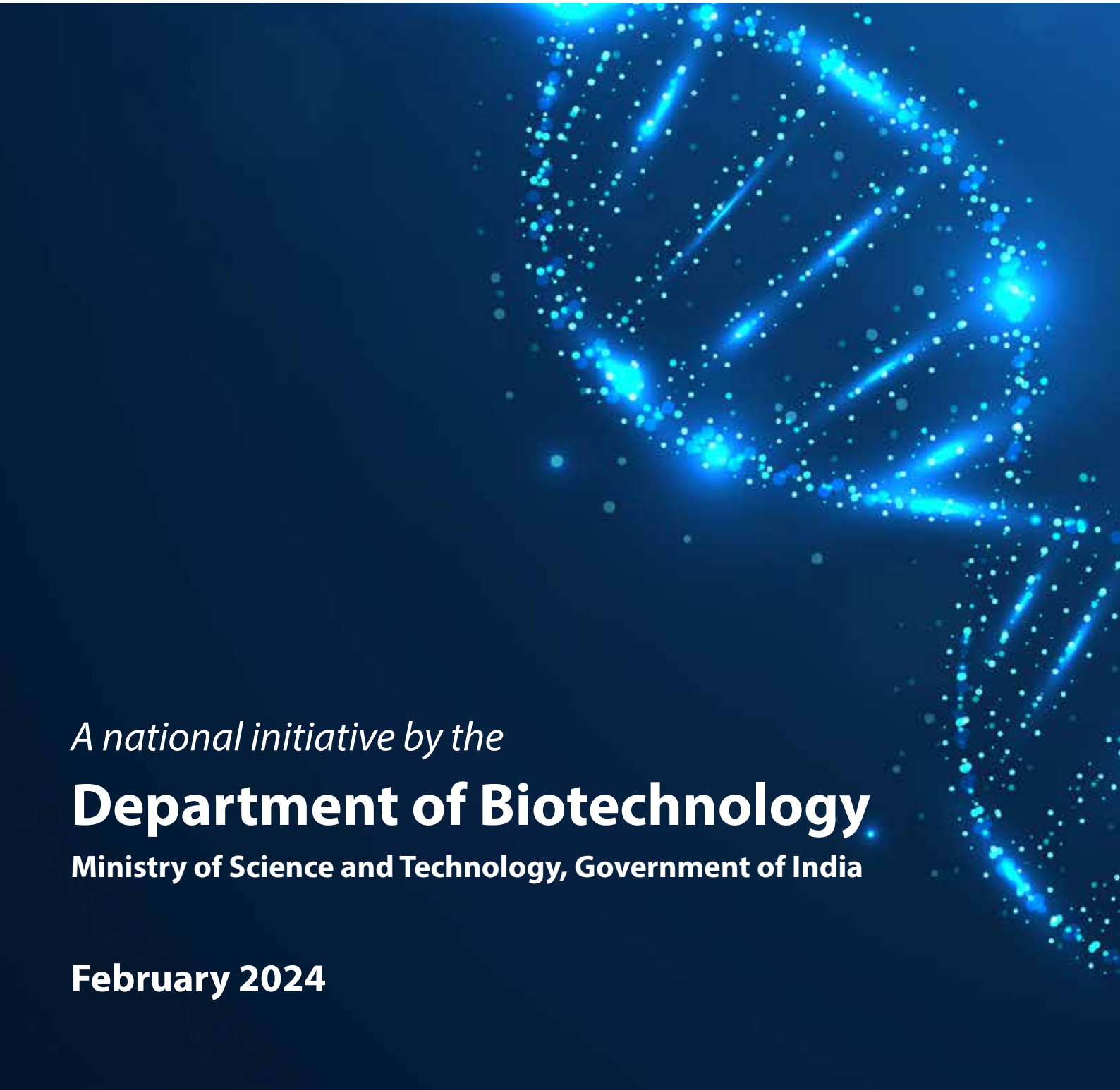
DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology
Government of India

GenomeINDIA

Population Genomics for Public Health



A national initiative by the

Department of Biotechnology

Ministry of Science and Technology, Government of India

February 2024

DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology
Government of India

GenomeINDIA

Population Genomics for Public Health

A national initiative by the

Department of Biotechnology

Ministry of Science and Technology, Government of India

February 2024

Published by

The Department of Biotechnology
Ministry of Science and Technology
Government of India

DBT Coordination

Dr. Suchita Ninawe
Adviser

Dr. Richi V. Mahajan
Scientist D

GenomeIndia Project Coordinators

Prof. Y. Narahari
Indian Institute of Science and Centre for Brain Research, Bengaluru

Prof. K. Thangaraj
CSIR-Centre for Cellular and Molecular Biology, Hyderabad

GenomeIndia Project Founding Coordinator

Prof. Vijayalakshmi Ravindranath
Former Director, Centre for Brain Research, Bengaluru

डॉ0 जितेन्द्र सिंह
राज्य मंत्री (स्वतंत्र प्रभार)
विज्ञान एवं प्रौधोगिकी मंत्रालय;
राज्य मंत्री, प्रधान मंत्री कार्यालय;
राज्य मंत्री कार्मिक, लोक शिकायत एवं पेंशन मंत्रालय;
राज्य मंत्री परमाणु ऊर्जा विभाग तथा
राज्य मंत्री अंतरिक्ष विभाग
भारत सरकार

सत्यमेव जयते

Dr. JITENDRA SINGH
Minister of State (Independent Charge)
Ministry of Science and Technology;
Minister of State in the Prime Minister's Office;
Minister of State in the Ministry of Personnel,
Public Grievances and Pensions;
Minister of State in the Department of Atomic Energy and
Minister of State in the Department of Space
Government of India

Message

The 'GenomeIndia Project' is one of pioneer initiatives of Government of India funded through the Department of Biotechnology, Ministry of Science and Technology. Recognizing the exceptional genetic landscape of the Indian population, this initiative set the ambitious goal to identify and catalogue the genetic variations of diverse Indian populations by sequencing the whole genome of 10,000 healthy individuals representing all major ethnic groups across the country.

I am happy to announce the completion of whole genome sequencing of 10,074 individuals from 99 communities, representing all major linguistic and social groups of the Indian population. By unraveling the genetic intricacies of the Indian population, the project lays the foundation of Genomic Hub as an opportunity for India to enter the emerging field of personalized medicine. A "Reference Genome for Indian Population" created under the project will lead to a better understanding of the nature of diseases and specific interventions essential for various ethnic groups. The GenomeIndia will place India on the world map of genome research and will collectively facilitate future large-scale human genetic studies for researchers across the globe.

GenomeIndia led by a consortium of 20 national institutes exemplifies the significance of collaborative, nation-wide, mission-oriented scientific partnerships, and visionary funding by the Department of Biotechnology, Government of India. I congratulate all the contributors for this successful endeavour.

(Dr. JITENDRA SINGH)
MBBS (Stanley, Chennai)
MD Medicine, Fellowship (AIIMS, NDL)
MNAMS Diabetes & Endocrinology

Anusandhan Bhawan, 2, Rafi Marg,
New Delhi-110 001
Tel. : 011-23316766, 23714230,
Fax. : 011-23316745

South Block, New Delhi-110011
Tel. : 011-23010191, Fax : 011-23017931
North Block, New Delhi-110011
Tel. : 011-23092475, Fax : 011-23092716

सत्यमेव जयते

डॉ. राजेश सु. गोखले
Dr. RAJESH S. GOKHALE

आज़ादी का अमृत महोत्सव

सचिव
भारत सरकार
विज्ञान और प्रौद्योगिकी मंत्रालय
जैव प्रौद्योगिकी विभाग
ब्लॉक-2, 7वां तल, सी.जी.ओ कॉम्पलेक्स
लोधी रोड़, नई दिल्ली-110003

SECRETARY
GOVERNMENT OF INDIA
MINISTRY OF SCIENCE & TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY
Block-2, 7th Floor, CGO Complex
Lodhi Road, New Delhi-110003

Foreword

Genomics has an enormous potential to transform our future. Genome-based predictions and targeted value-based alterations have permeated several sectors including health, agriculture, livestock and bioprocessing industries. Today, new sequencing technologies and advanced computational capabilities allow us to implement genomics for futuristic research goals. India is a land of diversity with over 4,600 distinct population groups. This necessitates a systematic pan-India genomics led initiative to explore the country's genetic diversity.

Towards this, 'GenomeIndia Initiative' is a pioneering step. The flagship project funded by the Department of Biotechnology, Ministry of Science and Technology, Government of India is an exemplar of nation-wide scientific collaboration and innovation involving 20 national institutes. The target of the project is to develop a reference genome for the population of India that will help in designing genome-wide and disease-specific genetic chips for low-cost diagnostics and research.

As on 5th January 2024, the consortium has completed whole genome sequencing of more than 10,000 individuals. The data being been archived at the Indian Biological Data Centre (IBDC), set up by Department of Biotechnology, Government of India at the Regional Centre for Biotechnology (RCB), Faridabad, will become a valuable national resource. The database will be instrumental for conducting next-generation basic and clinical research in India. It will help to foster large-scale human genetic studies, thus empowering both national and international researchers.

GenomeIndia denotes India's commitment towards improved public health interventions, drug development, and tailored treatments using advanced genetic research.

I commend the Coordinators and Scientists who have greatly contributed to this coveted project - 'GenomeIndia'.

(Dr. Rajesh S. Gokhale)

Phone : 24362950/24362881, Fax : 011-24360747, E-mail : secy.dbt@nic.in

Dr. Suchita Ninawe
Scientist 'G'/ Advisor
Phone: 011-24363722
Email: suchita.ninawe@dbt.nic.in

भारत सरकार
विज्ञान और प्रौद्योगिकी मंत्रालय
बायोटेक्नोलॉजी विभाग
ब्लाक–2,7 वां तल, सी० जी० ओ० कम्पलेक्स
लोदी रोड, नई दिल्ली–110003
GOVERNMENT OF INDIA
MINISTRY OF SCIENCE & TECHNOLOGY
DEPARTMENT OF BIOTECHNOLOGY
Block-2, 7th Floor C.G.O. Complex
Lodi Road, New Delhi-110003

PROLOGUE

Embarking on the ambitious quest to unravel the intricate tapestry of Indian human genetic diversity, the GenomeIndia initiative emerges as an epitome of scientific endeavour and national commitment. Spearheaded by the Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India, this visionary project, launched in 2020, represents a transformative leap into the genetic exploration of India's diverse population. The profound unwavering support by Dr. Jitendra Singh, the Hon'ble Minister of State (Independent Charge), Science and Technology has been instrumental in achieving the purpose of this endeavor. The constant guidance by Dr. Rajesh S. Gokhale, the Secretary, Department of Biotechnology has been remarkable.

The genetic diversity within India, with its 4,600 distinct population groups, emphasizes the project's scientific worth. Led by the Centre for Brain Research at the Indian Institute of Science, the consortium of dedicated scientists across the 20 national institutes of GenomeIndia has embarked on the task of whole-genome sequencing of 10,074 representative individuals across 99 communities, covering all major ethnic populations of the country. The revelations from ongoing analyses promise not only a deeper understanding of our collective genetic heritage but also lay the groundwork for targeted clinical interventions and the future of personalized healthcare.

Beyond the sheer scale of sequencing and establishing a Reference Genome, the creation of a biobank housing 20,000 blood samples at the Centre for Brain Research, coupled with data archiving at the Indian Biological Data Centre exemplify the project's commitment to transparency, collaboration, and

future research endeavors. GenomeIndia forms a foundation for precision medicine, acknowledging the need for tailored interventions aligned with the unique genetic makeup of the nation.

As we delve into the ensuing pages, due acclamation is extended to the coordinators, Dr Y. Narahari, Centre for Brain Research at Indian Institute of Science, Bengaluru and Dr. K. Thangaraj, Centre for Cellular and Molecular Biology, Hyderabad, and all the scientists from 20 national institutes who are primary contributors in this massive project on scientific front.

The foresight of the Department of Biotechnology in funding this initiative is evident in GenomeIndia's accomplishments and its trajectory toward constructing a reference genome structure, developing genetic chips, and fostering a comprehensive understanding of India's genetic landscape. The determined efforts and coordination by the DBT team ensuring smooth implementation of GenomeIndia from its inception to its current milestones are creditable. Special acknowledgment is extended to Dr Onkar Tiwari, Dr. Kakali Dey Dasgupta, Dr Amit Kumar Tripathi for their contributions while the efforts and zeal of Dr. Richi V Mahajan during the project's concluding phase are duly recognized.

Having achieved significant milestones in decoding over 10,000 genomes and providing profound insights into India's unparalleled genetic diversity, the GemoneIndia opens the door for deepened genomic research by our scientists. I am confident that the outcomes of this sought-after project will serve as a foundational resource for next-generation basic and clinical research in India. GenomeIndia not only positions India at the forefront of global genome research but also indicates a future where precision medicine transforms healthcare. This compilation will serve as a portal into the collaborative scientific journey that contributes to the ongoing genomic revolution.

Suchita

(Dr. Suchita Ninawe)

Preface

The human genome comprises the basic structural unit of DNA, represented by four letters (nucleotides) A, C, G, and T, stretched over three billion such letters. Any pair of unrelated individuals will have millions of variations in their genomes. These genetic variations among individuals are crucial for understanding our disease predispositions, including rare inherited disorders. They could also determine our response to drugs and help track migration and evolutionary patterns of population groups.

Populations across the world differ in their genetic makeup due to many factors including environment. The Indian population of 1.43 billion consists of more than 4600 population groups, and several thousand of them are endogamous. These factors have contributed to the genetic diversity of the contemporary Indian population. 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. This is where the GenomeIndia project, so generously funded with great vision by the Department of Biotechnology, Ministry of Science and Technology, assumes paramount importance.

We are delighted to report that, as of January 2024, the GenomeIndia consortium has completed the whole genome sequencing (WGS) of more than 10000 individuals and about 8000 WGS files are already archived at Indian Biological data Centre (IBDC). In about two months time, the joint genotyping of 10000 samples will be completed and we are hoping to release a flagship paper from the consortium by June 2024. We have formed 16 specialized working groups to conduct investigations to produce deep scientific insights.

As coordinators of this unique initiative, we thank Dr. Rajesh Gokhale, Secretary, Department of Biotechnology, for his unwavering support throughout the duration of this project. We are grateful to Dr. Suchita Ninawe, Adviser, for her foresight and constant support throughout the duration of the project. We thank Dr. Richi V. Mahajan, Dr. Kakali De Dasgupta, Dr. Onkar N. Tiwari and Dr. Amit K. Tripathi for their fine administrative support. We are indebted to Prof. Vijayalakshmi Ravindranath, who was the principal architect for this initiative. We gratefully acknowledge the valuable guidance and precious suggestions provided by the members of the GenomeIndia Technical Monitoring and Assessment Committee (TMAC) chaired successively by Prof. G Padmanaban, Late Prof. MRS Rao and currently Prof. Partha Pratim Majumder. We thank all the principal investigators and co-investigators at the 20 partner institutions, particularly, the main contributors from the four sequencing centres: Dr. Bratati Kahali (CBR), Dr. Tej Sowpati (CCMB), Dr. Mohammed Faruq (IGIB), and Dr. Analabha Basu (NIBMG). We wish to thank Dr. Prathima Arvind (CBR), who as project manager for GenomeIndia, diligently coordinated the numerous operational aspects of the project. We also thank Mr. Jothibasu (CBR) and his IT team for taking charge of the onerous responsibility of data sharing among the sequencing centres and IBDC. We express our thanks to Prof. Ganganath Jha, Vinoba Bhave University, Hazaribagh, for his splendid help in sample collection.

Y. Narahari & **K. Thangaraj**

Joint Coordinators, GenomeIndia

Table of Contents

GenomeIndia:
A Bird's Eye View

What are Genomes?

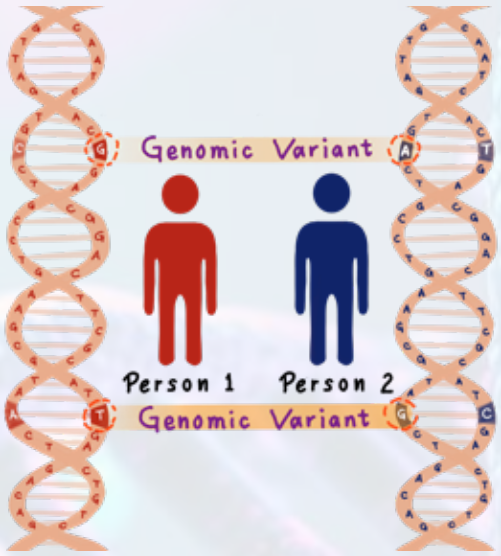
The human genome is the instruction manual of our life. It is made up of DNA, represented by four letters (nucleotides) A, C, G, and T. This genetic script, making up the human genome, extends across a staggering three billion such letters.

Housed within the cells of our body, the genome of an individual is embedded in the 23 pairs of chromosomes contained in each cell of the individual. In the intricate phenomenon of heredity, we inherit our genomes from our parents. Half our DNA is from our mothers, and half is from our fathers. This genetic inheritance orchestrates the very essence of our being.

Our genomes are a blueprint for the function of our bodies. We have copies of our genomes in all our cells.

Why Study Genetic Variations?

About 1 in every 1000 positions in the DNA differ between any two individuals on average, resulting in diverse genetic backgrounds. These genetic variations among individuals affect our disease risks, and also determine a person's responses to drugs. Understanding these genetic nuances is key to deciphering people's predispositions to certain diseases and designing effective treatments.



All humans share 99.9% of our DNA. But the 0.1% difference--called genetic variations--add up to millions of positions in the genome.

Genetic variations accumulate with time. The Indian population is one of the oldest populations in the world. Most importantly, the Indian population is an amalgamation of ethnic subgroups, well-defined by physical, linguistic, cultural and other attributes. Characterisation of the genomic diversity of the ethnic subgroups is valuable to understand the history, natural selection and adaptation of these subgroups. A "reference genome" tailored to the Indian population will lead to a better understanding of the nature of diseases. It will open doors to specific interventions essential for diverse ethnic groups, thereby paving the way for more targeted and effective healthcare solutions.

What is GenomeIndia?

GenomeIndia, a visionary national project funded by the Department of Biotechnology, Ministry of Science and Technology, Government of India, was launched in January 2020. Its ambitious goal is to sequence genome of 10000 individuals spanning the length and breadth of the country.

The primary aim of GenomeIndia is to construct a comprehensive catalogue of genetic variations of Indian population that will better capture our unique diversity. This initiative is not just about decoding our genes; it is about creating a detailed reference that encapsulates the Indian population's genetic makeup and enables a deeper understanding of its diversity, health and disease.

Creating a Precious Resource for India's Public Health

While millions of genomes worldwide have been sequenced, the glaring gap lies in the severe under-representation of Indian populations in these global studies. Our population of 1.43 billion consists of more than 4600 groups, many of which are endogamous (that is, marriages within the group). All these population groups have genetic variations unique to themselves, thus contributing to our unparalleled genetic diversity.

Distinct population groups differ in their genetic makeup and exhibit different risk factors for diseases. Consequently, findings from human genetics research from other populations of the world cannot be extrapolated to Indians. It is crucial to undertake studies that rigorously account for our own variations for better clinical practice. Moreover, sequencing our population is a first step towards bringing genetics to the world of personalized medicine. A national resource like GenomeIndia is an excellent opportunity for clinicians to find a genetic diagnosis for patients with complex and rare disorders.

This effort has the potential to revolutionise healthcare, empowering clinicians and basic researchers, leading to transformative precision interventions. The impact of GenomeIndia, therefore, extends far beyond the lab, promising a healthier nation in the future.

Who is Involved?

GenomeIndia comprises dedicated scientists and researchers from 20 partner institutions, with some of the institutions playing multiple roles.

Sample Collection, Sequencing and Analysis Centres

Centre for Brain Research (CBR), IISc Campus, Bengaluru
CSIR - Centre for Cellular and Molecular Biology (CSIR-CCMB), Hyderabad
CSIR - Institute of Genomics & Integrative Biology (CSIR-IGIB), New Delhi
iBRIC - National Institute of Biomedical Genomics (iBRIC-NIBMG), Kolkata

Sample Collection Centres

All India Institute of Medical Sciences (AIIMSJ), Jodhpur
Gujarat Biotechnology Research Centre (GBRC), Gandhinagar
iBRIC - Institute of Bioresources & Sustainable Development (iBRIC-IBSD), Imphal
Indian Institute of Science Education and Research (IISER), Pune
iBRIC - Institute of Life Sciences (iBRIC-ILS), Bhubaneshwar
Mizoram University (MZU), Aizawl
National Institute of Mental Health & Neurosciences (NIMHANS), Bengaluru
iBRIC - Rajiv Gandhi Centre for Biotechnology (iBRIC-RGCB), Thiruvananthapuram
Sher-i-Kashmir Institute of Medical Sciences (SKIMS), Srinagar

Method Development Centres

iBRIC - Centre for DNA Fingerprinting and Diagnostics (iBRIC-CDFD), Hyderabad
Indian Institute of Information Technology (IIITA), Allahabad
Indian Institute of Science (IISc), Bengaluru
Indian Institute of Technology Delhi (IITD), New Delhi
Indian Institute of Technology Jodhpur (IITJ), Jodhpur
Indian Institute of Technology Madras (IITM), Chennai
National Centre for Biological Sciences (NCBS), Bengaluru

Biobanking and Data Archival Centres

Biobank at the Centre for Brain Research (CBR)
Data Archival at Indian Biological Data Centre (IBDC)

Goals and Impact of GenomeIndia

- Develop a reference set of genetic variations for the Indian population by carrying out whole genome sequencing of 10000 individuals from 99 ethnic groups.
- Create a biobank of 20000 blood samples for future genomic studies.
- Make genome data available for public access (digital public goods) for academic / research purposes through IBDC.
- Design genome-wide and disease-specific genetic chips for low-cost diagnostics and research activitiess.
- First big step towards developing genome-based precision medicine in India.
- An inspiration for India's young minds and young researchers to explore the exciting area of genomics research and innovation for the health of Indian population.

What we have Achieved so far

- Collected more than 19200 blood samples from 99 ethnic groups, against a target of 20000. These valuable samples are available in the GenomeIndia biobank, as a reservoir for future research breakthroughs.
- Achieved the milestone of completing whole-genome sequencing of more than 10000 individuals from 99 ethnic groups.
- Sequencing data for more than 7800 samples is securely archived at IBDC and all sequencing data for 10000 samples will soon be available for academic/ research purposes from IBDC.
- As part of Phase 1 analysis, we have analyzed (technically called joint genotyping) 5750 samples, unravelling unique facets of the genomic structure of Indians. Within this genomic treasure trove, we have uncovered a wealth of rare variations unique to our populations or specific subsets, reflecting our unique population history and diversity.
- Beyond the scientific revelations, many of these variations will have clinical significance, leading to targeted clinical interventions for specific sub-groups. The genetic roadmap holds the promise of precision medicine and will potentially transform the landscape of clinical care for the benefit of common people.

Roadmap

- Joint calling of 10000 genomes (joint genotyping) and comprehensive analysis of the resulting data to obtain India-specific insights.
- Development of a reference genome for the Indian population.
- Archival of all the data at IBDC and making the digital public goods available for academic/ research purposes.
- Design of genome-wide and disease-specific genetic chips for low-cost diagnostics and research activities.
- Production of a flagship manuscript from the GenomeIndia Consortium for worldwide dissemination of the research findings.
- Completion of a multitude of scientific investigations through specialized working groups, leveraging the dataset of genetic variations.

Summary

- GenomeIndia exemplifies the significance of collaborative, nation-wide, mission-oriented scientific partnerships, and visionary funding by the Department of Biotechnology, Ministry of Science and Technology, Government of India.
- By unraveling the genetic intricacies of the Indian population, the project lays the foundation for advancements in public health interventions, drug development, and personalized medicine.
- As the scientific community eagerly awaits the forthcoming insights, GenomeIndia stands as a beacon of India's commitment to deep science by advancing genetic research for the benefit of every citizen of our nation.

Insights from Preliminary Analysis

In phase-1, joint variant calling has been executed for 5750 samples, comprising 2587 samples from CBR, 1055 from CCMB, 572 from IGIB, and 1536 from NIBMG. These 5750 individuals represent 69 distinct population groups across India. The joint variant calling has been independently done at both NIBMG and CBR. These analyses have shed light on new knowledge about the genetic makeup of the Indian population, underscoring the importance of a population genomic project like GenomeIndia.

Huge Number of Common and Rare Variants

In the callset of 5750 samples, we have identified more than 135 million genetic variations, mostly comprising biallelic single nucleotide polymorphisms (SNVs) and short insertions-deletions (INDELs), and a small proportion of multi-allelic variants. SNVs largely outnumber INDELs. As expected, the majority of the variants (~65%) are ultra-rare, with a minor allele frequency (MAF) of less than 0.1% in the overall population.

Although most of the identified variants were ultra-rare or rare, from the joint analysis of the data we identified a large number (> 6.9 million, representing 11%) of common variants, detected after quality check of the genetic varients. These common variants, which are largely shared among Indian population groups, are candidates for association study design to identify genetic factors underlying common traits. They can also be used for optimization of gene-chips to be used for Indian populations since they are a better representation of Indian genetic variation than known before. Moreover, many of these common variants are rare or non-existent in global variant databases. These variants will allow us to create a larger set of benign innocuous variants that can be eliminated from candidate variant lists in clinical cases.

Unprecedented Diversity Leading to Deeper Understanding of Population History

Our sequence data reveals novel insights into Indian population history. The sampled populations capture the linguistic diversity of India, with representation from the four major linguistic groups (Indo-European, Dravidian, Austro-Asiatic, and Tibeto-Burman) as shown in Figure 1a. The top principal components of the genetic data capture this diversity. Principal component 1 (PC1) primarily separates the populations from North-East India from the rest. This primarily includes Tibeto-Burman speaking populations.The Austro-Asiatic speaking tribal populations are clustered among themselves and are also proximal to each other (Figure 1b). There is an approximate correspondence between the geographical location of the population and their genetic position on the PC1-PC2 scatter plot, with a correlation of 0.7 between PC1 and longitude, and a correlation of 0.45 between PC2 and latitude.

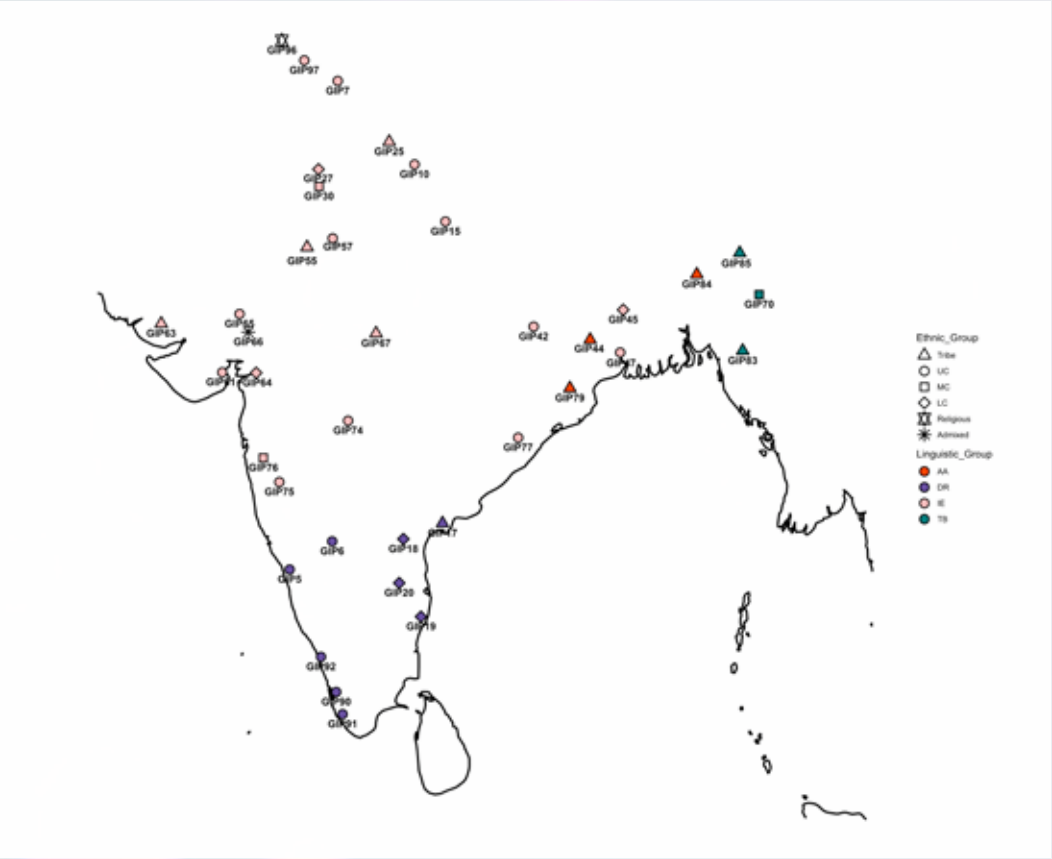


Figure 1a. The approximate geographical position of the populations (co-ordinates of the sample collection sites) on the map of India. The colours stand for the broad linguistic identity (pink-peach for Indo-European (IE) speaking populations, blue for Dravidian (DR) speakers, red for Austro-Asiatics (AA), green for Tibeto-Burman (TB) speakers) while the shapes stand for the social hierarchy of the populations.

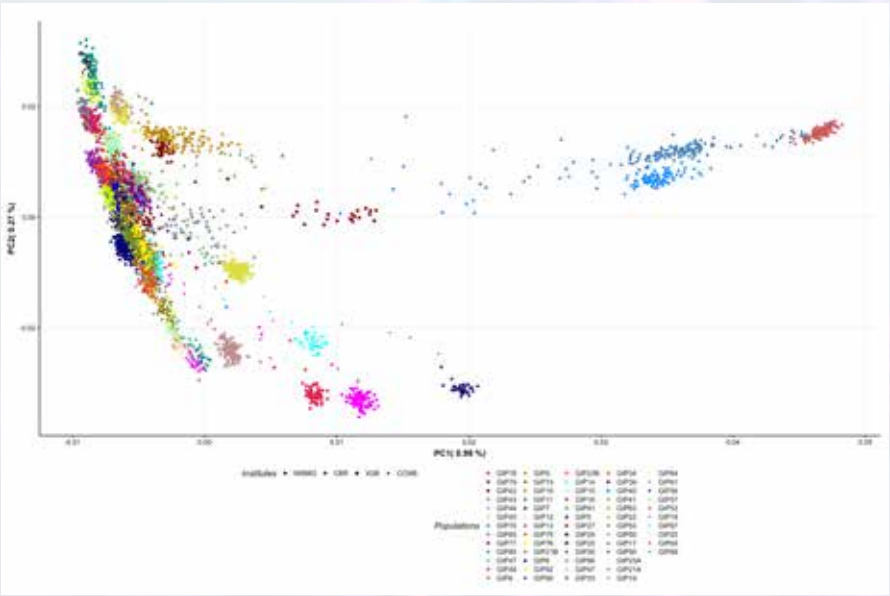


Figure 1b. Scatterplot of the first two principal components obtained from the genetic distances between the individuals (covariance matrix). The correlation of the PC1 co-ordinate of a population and the longitude from where it is collected is 0.7 while the correlation of the PC2 co-ordinate and the latitude is 0.45.

Medical relevance of the identified variants

Functional relevance of the identified variants

We have followed up our genetic variant identification with detailed genomic annotation, including measures of deleteriousness at specific genomic regions. Expectedly, most SNVs and INDELs are in intergenic and intronic regions.

However, we observe many putatively functional variants: more than 1.4 million variants are missense, frameshift or splice variants, or affect the untranslated regions of genes, potentially leading to phenotypic alterations. We also get a glimpse of medically relevant findings, for example, in the *LDLR* gene (implicated in familial hypercholesterolemia, a disease with high prevalence in India), there are at least 10 novel missense variants in our callset, which could allude to crucial functional consequences for the Indian population.

We also found an additional 27 million variants present at low frequency in the overall dataset but present at significantly higher allele frequencies in one or more of the populations. About 7 million of these 27 million are novel variants and not found in the global catalogue of all variants. In rare disease genetics, databases are used to filter based on allele frequency with the idea that common alleles are unlikely to be responsible for rare highly penetrant disorders. We explored whether the novel variants can improve the ability to identify disease-relevant variants. We annotated these 7 million variants that were identified in the GenomeIndia dataset against the Human Gene Mutation Database (HGMD) disease-causing pathological, InterVar and ClinVar pathogenic variants. This analysis identified 213 variants in InterVar which were identified by American College of Medical Genetics (ACMG) as deleterious; 396 ClinVar variants which were identified by an expert panel as deleterious and 75,476 variants which were identified as deleterious by SIFT. This exercise will eventually result in reclassification of many variants which were classified as deleterious even after screening them against all existing genome sequence databases.

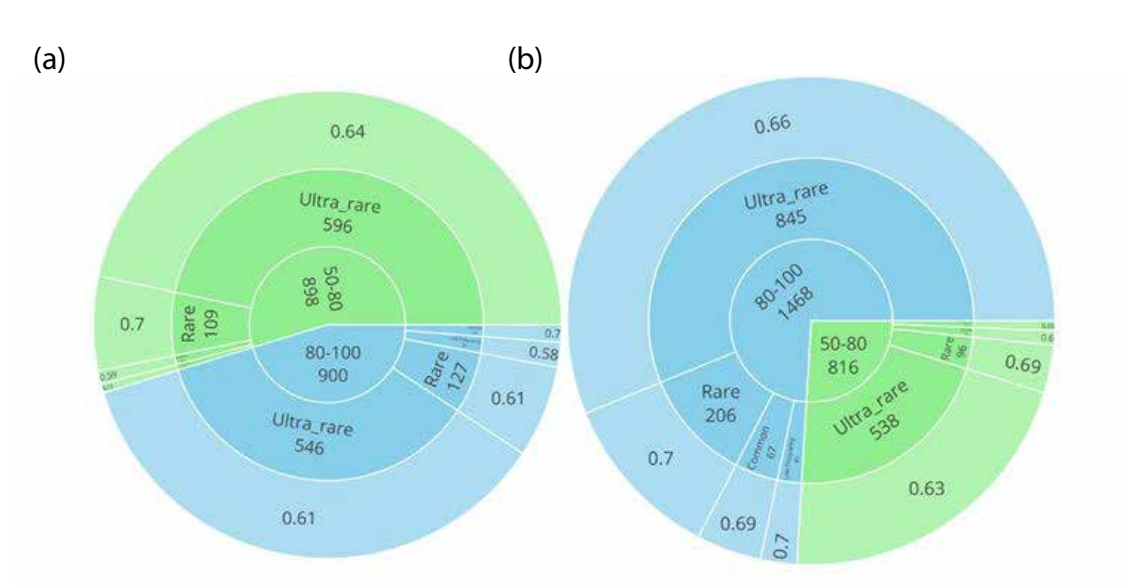


Figure 2: The loss of function metric for genes affected by SNVs (a) and INDELs (b) are shown. For genes having high propensity of deleteriousness, there are at least 100 of them (the thinnest slices) which are affected by variants present commonly in our population. This can have serious implications for protein structure and function and thereby disease susceptibilities. Similarly, for overall genomic annotation, tens of thousands of variants were detected in our population that can cause disruptive changes in the protein coding regions. Such information is critical for understanding disease susceptibilities. This information will be invaluable in planning and designing genomics informed public health solutions as well as personalized health care in the post-genomics era.

Distribution of the variants in ACMG actionable genes

In clinical exome and genome sequencing, there is a potential for the recognition and reporting of incidental or secondary findings unrelated to the indication for ordering the sequencing but of medical value for patient care. The American College of Medical Genetics and Genomics (ACMG) recently published a policy statement on clinical sequencing that emphasized the importance of alerting the patient to the possibility of such results in pretest patient discussions, clinical testing, and reporting of results.

We analyzed the distribution of identified variants from the GenomeIndia callset in the list of ACMG actionable genes to identify any potential incidental findings. 237,414 variants lie in the 73 ACMG identified actionable genes. Of these, 131 variants are pathogenic or likely pathogenic. These variants are also clustered in sets of samples, where 245 samples share at least one of these 131 variants.

Important diseases linked to these variants include Familial hypercholesterolemia, *BRCA* mutations for inherited breast and ovarian cancers as well as hypertrophic cardiomyopathy. We hypothesize that even a dataset of relatively healthy controls from the Indian population can enable the identification of potentially pathogenic variants, and indicate an increased risk of disease for these carriers.

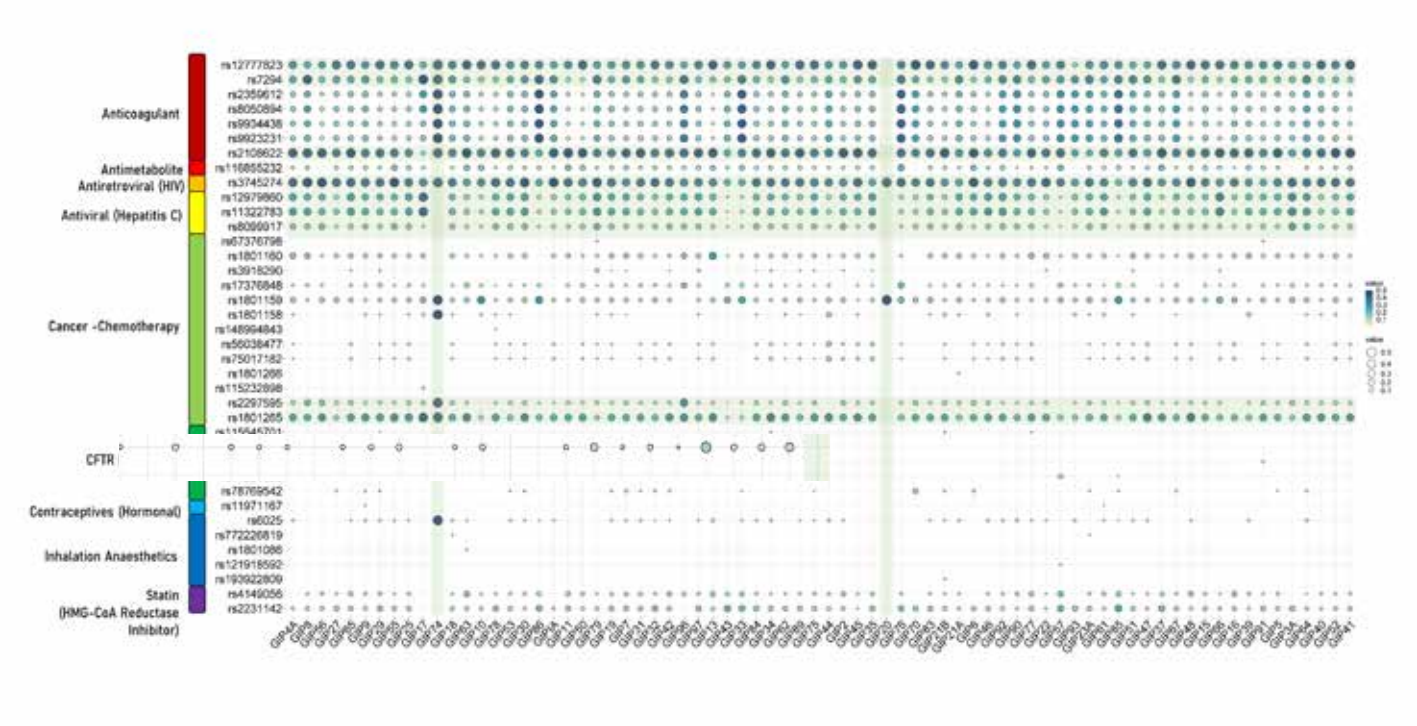


Figure 3: The protocol followed to identify the variants which have high frequency in Indian populations and also are from the ACMG actionable genes. We graphically represent them in two categories (1) pathogenic and (2) likely pathogenic.

Pharmacogenomics

Our genes affect our inherent ability to metabolize drugs and modulate drug response. Pharmacogenomics is the discipline that looks at how genetic variations can affect an individual's biological responses to drugs. Previous studies have identified and catalogued a set of such genes and variants in the PharmGKB database. Of these, 118 variants--categorized as Level 1A and 1B--most adversely impact an individual's response to drugs.

In the GenomeIndia variant callset, 38 of these 118 variants are carried by individuals from different populations of India. A high frequency of one of these variants in one population would imply ineffectiveness of certain drugs for individuals of the population and is hence a public health issue. We observe that many populations in India carry a substantial proportion of these variants which will reduce efficiency and efficacy of anticoagulant, anti-retroviral, and anti-viral drugs.

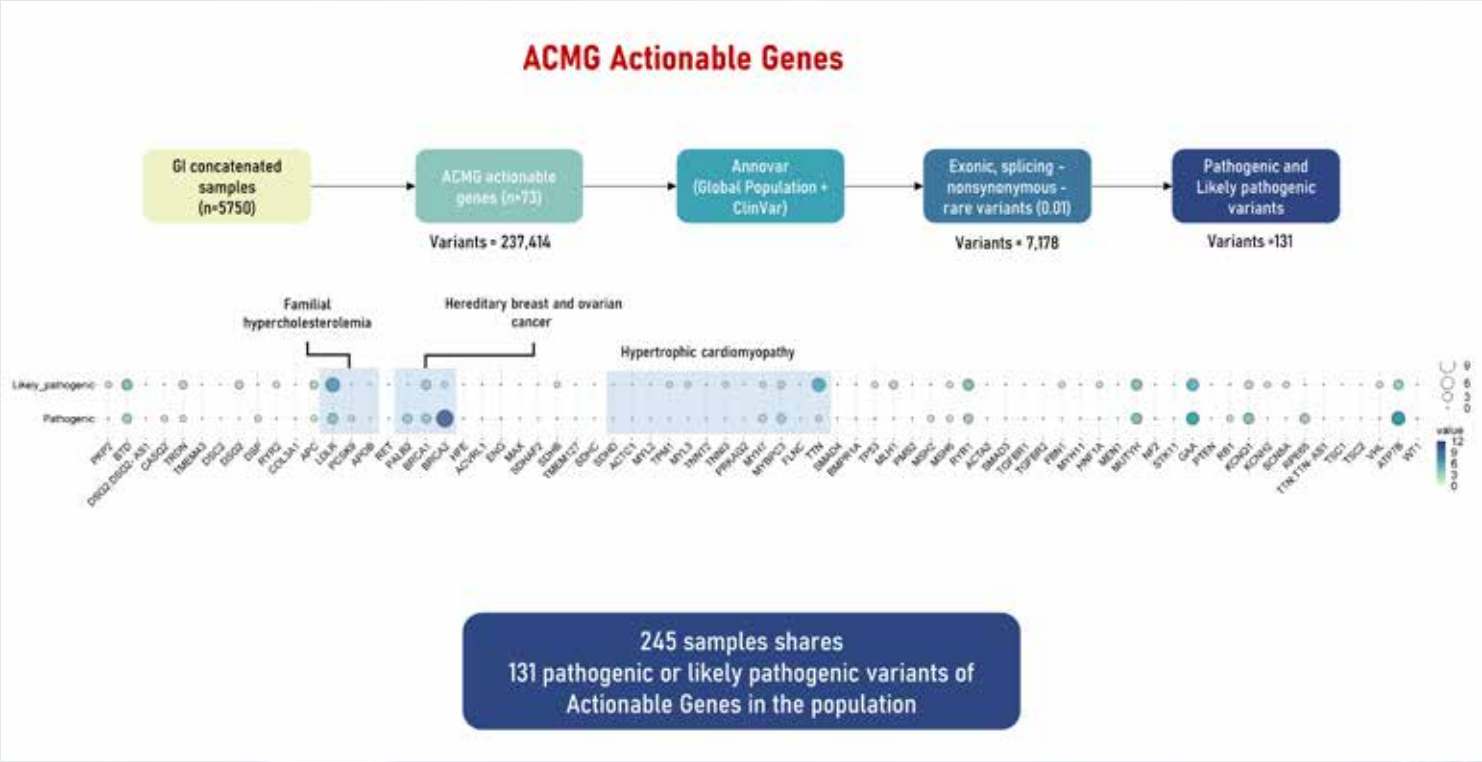


Figure 4: Pharmacogenomics markers in GenomeIndia samples. The 38 out of 118 Level 1A and 1B pharmacogenetics relevant variants are arranged along the Y-axis. They are also classified in accordance to their mode of action and the disease for which the class of drugs are designed. The X-axis contains the names of the different populations of GenomeIndia. The color and size of the circles determine the frequency and effect of the variant. The deeper the color and larger the size, the variant will have a larger effect on the corresponding population. We also identify two populations where there are many variants with high frequency which impacts the response of multiple drugs pertaining to different diseases.

Unprecedented power in imputing genotypes using the GenomeIndia Data

A genome-wide reference imputation panel has been constructed with the variant call dataset, showing improved imputation accuracy and allelic concordance for Indian population genotypes compared to that of TOPMed and Haplotype Reference Consortium panels, even for rare (<1%) and low-frequency variants (minor allele frequency between 1 and 5% in the population) in addition to the variants having common minor allele frequency (>5%).

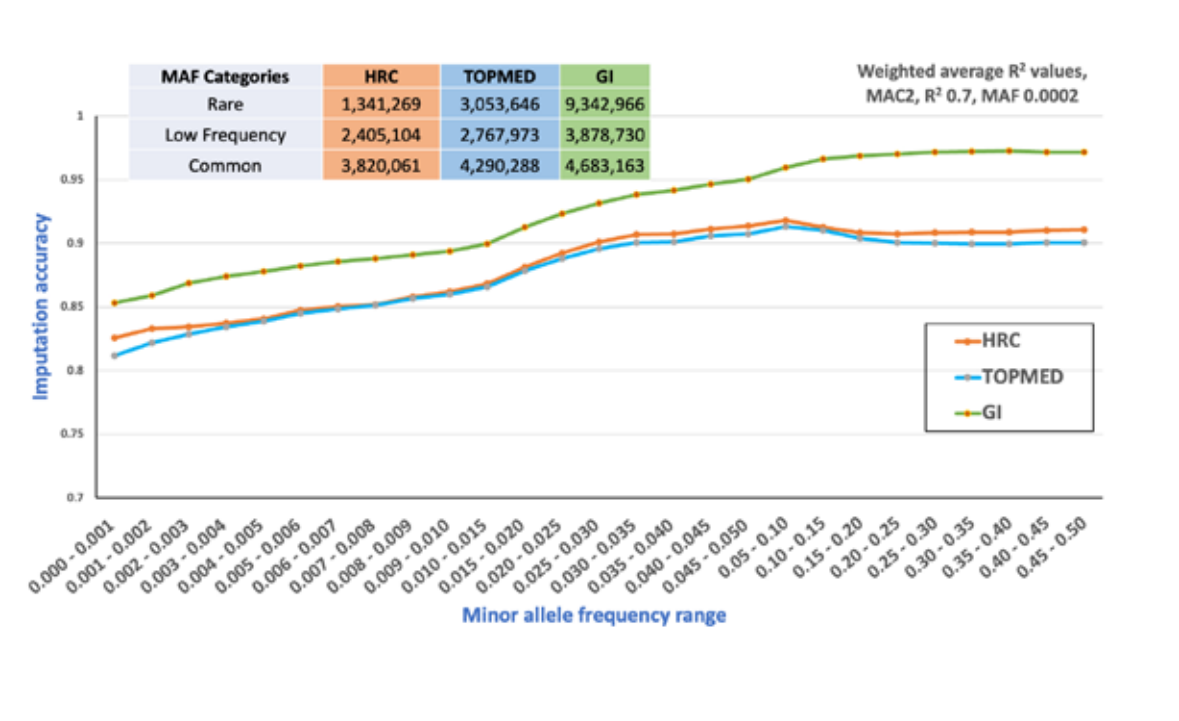
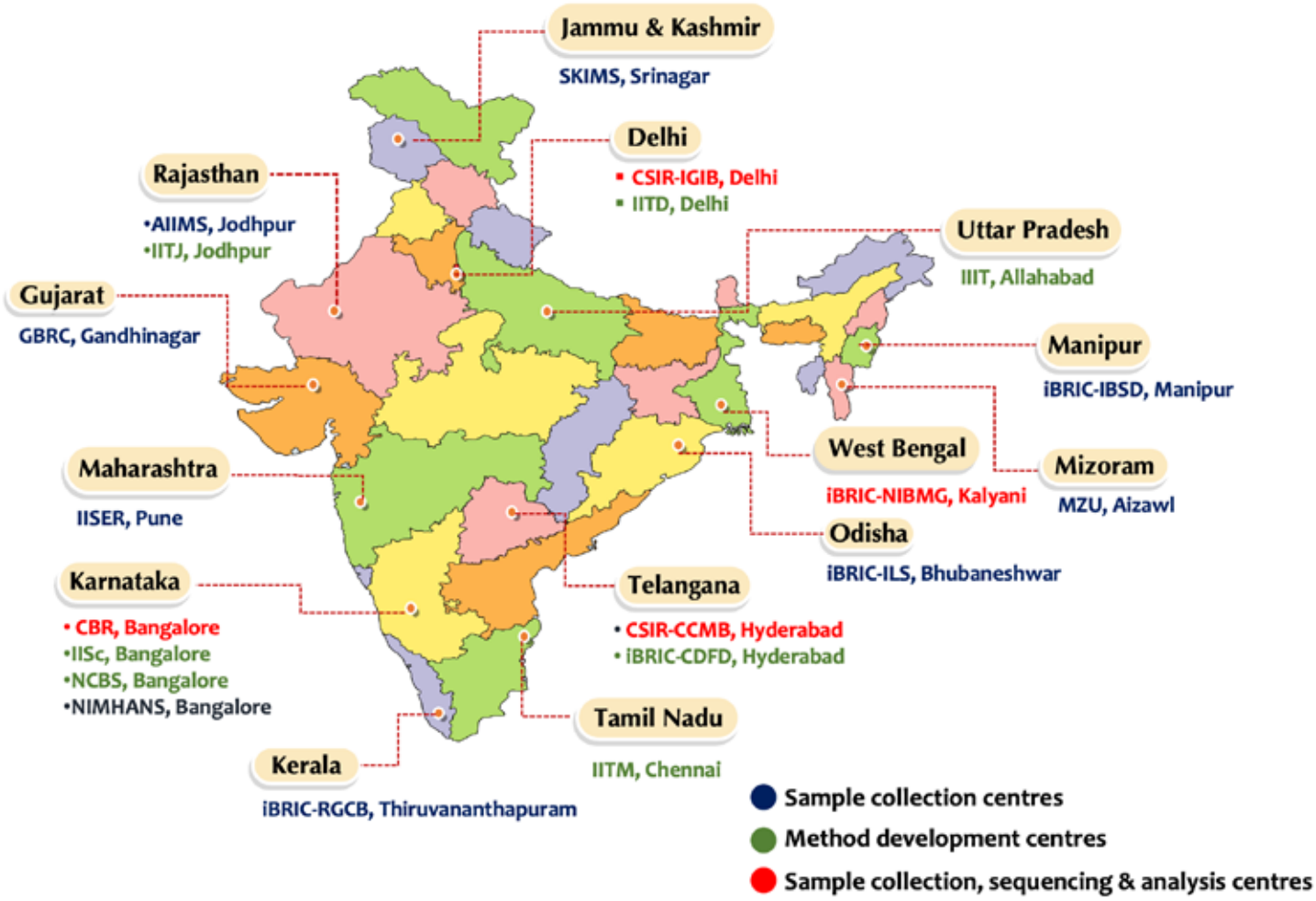


Figure 5: Performance of reference haplotype panel for genotype imputation of Indian ancestry individuals. The green line pertains to the accuracy obtained using the GenomeIndia samples compared to the orange and blue lines which are today's global standards. It is clearly evident that the GenomeIndia samples provide a huge advantage in imputation accuracy.



GenomeIndia collaborating institutions

Brain-storming meeting for conceiving GenomeIndia project, held in the Indian Institute of Science, Bengaluru on 30th June 2017

Meeting of GenomeIndia investigators held in
CBR, Bengaluru on 19 October 2023

Meeting of GenomeIndia investigators held in
CSIR - CCMB, Hyderabad on 21 December 2023

Sample Collection, Sequencing and Analysis Centres

Centre for Brain Research

Indian Institute of Science Campus, Bengaluru

Principal Investigator: Prof. Bratati Kahali

Contributors: Prof. Vijayalakshmi Ravindranath (Founder Director, CBR), Prof. Y. Narahari (National Joint Co-ordinator), Prof. K.V.S. Hari, Dr. Prathima Arvind (Project Manager), Dr. Shweta Ramdas, Dr. Khader Valli Rupanagudi, Dr. Shafeeq KSH, Jothibasu V, Krithika Subramanian, Siddhi Jani, Shreya Chakraborty, Akshaya Rajesh, Raghvendra Agrawal, Debasrija Mondal, Mohammad Hanif K.M, Vinayak H, Dr. Shobha Anilkumar, Mohan C, Rajesh G, Diwakar A, Anand Kumar, Karthik S, Ravindra V.

Role of the Institution in the GenomeIndia Project: CBR is the coordinating centre for the GenomeIndia project. CBR spearheads efforts in strategic planning and execution of the entire project. The institute has been involved in sample collection, whole genome sequencing, data processing and analysis; as well as joint genotyping for all samples and data accumulated over the course of the project. It maintains a biobank of collected DNA samples. CBR, in collaboration with NIBMG, is collating and organising the phenotypic data for all the 20,000 samples.

Accomplishments and Outcomes

1. **Sample collection:** CBR has recruited 3003 individuals from sixteen different communities for the project. They have undergone detailed sociodemographic questionnaire assessment, anthropometric measurements, and blood biochemical investigations. Details are provided in the table below.

Sl. No	Community	No. of samples collected
1	GIP1A	208
2	GIP2A	209
3	GIP3A	159
4	GIP4A	149
5	GIP5	224
6	GIP6	193
7	GIP7	141
8	GIP8	164
9	GIP9	275
10	GIP10	217
11	GIP11	230
12	GIP12	162
13	GIP13	139
14	GIP14	120
15	GIP15	234
16	GIP16	179
Total		3003

* GIP - GenomeIndia Populations

2. **Whole Genome Sequencing and Joint Genotyping:** CBR has formulated the best practices for sequencing and analysis protocols for the consortium for variant calling at individual genome levels, defined the thresholds for various quality check parameters, followed by optimizing the protocol for joint genotyping of >5000 individuals that would be followed for joint genotyping of 10000 individuals' data. The downstream variant checking and analysis protocols were defined by CBR. CBR has also been actively uploading data to IBDC for use by other consortium members. The joint genotyping for 5750 individuals' genomes spanning 69 population subgroups were executed by CBR, and in parallel by NIBMG. Some of the current findings for this work in progress are given below.

   - More than 55 million Single Nucleotide Variations (SNV) and INDELs have been identified.
   - Population subgroup specific variations are enriched in rare and low frequency genetic variants.
   - Genetic variations present in Indians as well as world populations are not equally prevalent, that has implications for uncovering differential genetic architecture of diseases in Indian population, and fine mapping of associated genomic loci for complex traits.
   - More than 20 million rare genetic variants have been identified, that could potentially remodel the rare disease research in the country.
   - Reference haplotype panel of genetic variations for Indians shows markedly greater precision and allelic concordance for genotype imputation, in comparison to other widely used panels, even for rare variants.
   - Current work: Joint genotyping and analyses for 10,000 whole genomes

3. **Phenotyping:** CBR is also the coordinating centre for collating and processing raw blood biochemistry and anthropometric data collected by all sample collection centres. CBR has collated these datasets (collectively called 'phenotype data') for more than 17,000 samples, and along with NIBMG, has built a computational pipeline to clean up the data. These cleaned data will be a central part of downstream genotype-phenotype association analyses.

Sample collection by the CBR team

CSIR - Centre for Cellular and Molecular Biology

Hyderabad

Principal Investigator: Dr. K Thangaraj, Dr. Vinay K Nandicoori

Co-Investigators: Dr. Divya Tej Sowpati, Dr. Karthik Bharadwaj Tallapaka

Contributors: Payel Mukherjee, Pratheusa Maccha, Sofia Banu, Sreelekshmi MS, Malini Nemalikanti, Tulasi Nagabandi, Valli Undamatla, Devavrat Desai, Neha Singh, Priya Pandey, G Mala, Deepak Kashyap, Mahfuj Hassan, Vasanth Kumar

Role of the Institution in the GenomeIndia Project:

- Sample collection and phenotyping from 8 communities
- Whole genome genotyping and sequencing of 28 communities (~2500 samples) from three different sample collection centres
- Raw data processing and QC
- Data analysis at population scale

Accomplishments and Outcomes

We have finished sample collection from all our target communities, including critical tribes in Andhra Pradesh and Uttarakhand. Over 3200 samples, including those collected by GBRC and AIIMS-Jodhpur, were genotyped using Illumina GSA, out of which more than 2500 samples were deemed suitable for sequenced. Samples from these communities were sequenced using the Illumina NovaSeq 6000 platform. Raw data was processed to ensure they met the expected quality criteria, and sample wise variants were obtained following the DRAGEN pipeline. Sample level quality checks were performed to remove potential outliers, and those that passed the checks were taken further for population-scale joint genotyping.

Sample Collection

We have completed sample collection from our target communities from Tamil Nadu, Andhra Pradesh, Telangana, and Uttarakhand. The details are as follows:

The blood samples from the above were analyzed for various biochemical parameters including CRP, Liver function, Lipds, Kidney function among others. The blood samples were also sent to CBR, Bengaluru for biobanking.

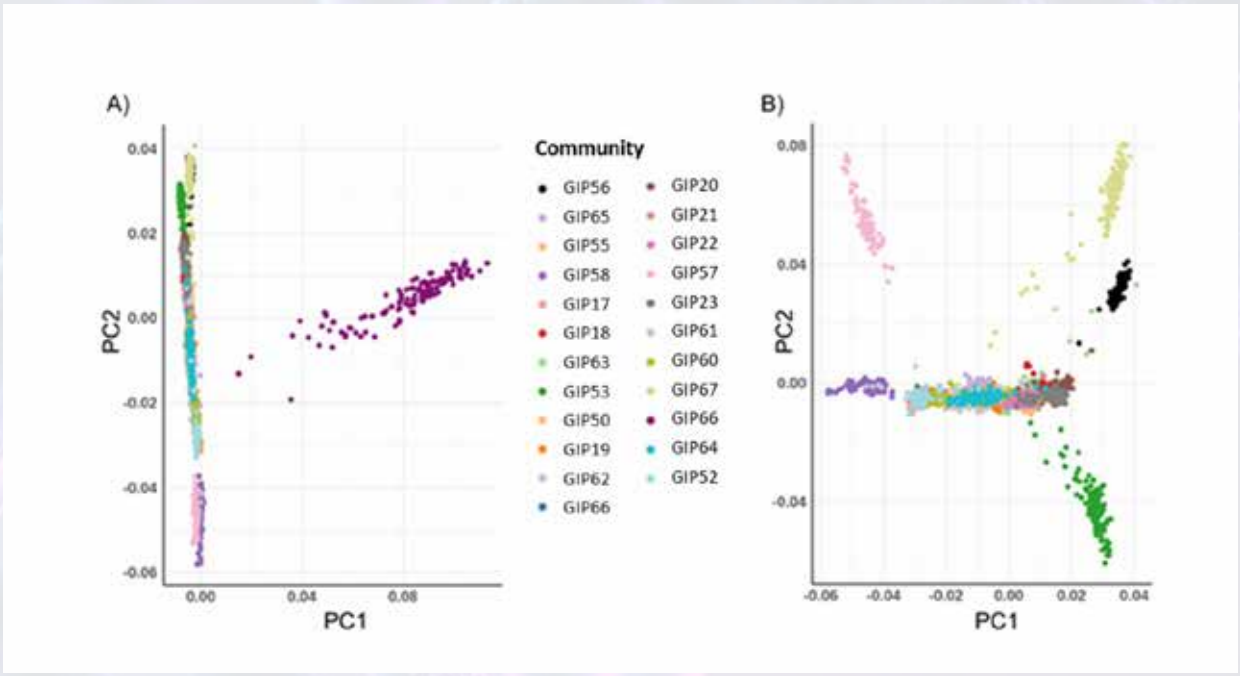


Figure 1. Population structure of various communities sequenced by CCMB depicting their distribution. A) Including GIP66 (purple dots). B) Excluding GIP66. As GIP66 are known outliers, including them in the PCA plot limits the resolution of other communities.

Sl. No.	Community	No. of samples collected
1	GIP17	206
2	GIP18	118
3	GIP19	319
4	GIP20	307
5	GIP21A	157
6	GIP22	235
7	GIP23A	206
8	GIP24	127
Total		1675

Sample collection by the CSIR-CCMB team

CSIR-Institute of Genomics and Integrative Biology

New Delhi

Principal Investigator: Dr. Mohammed Faruq

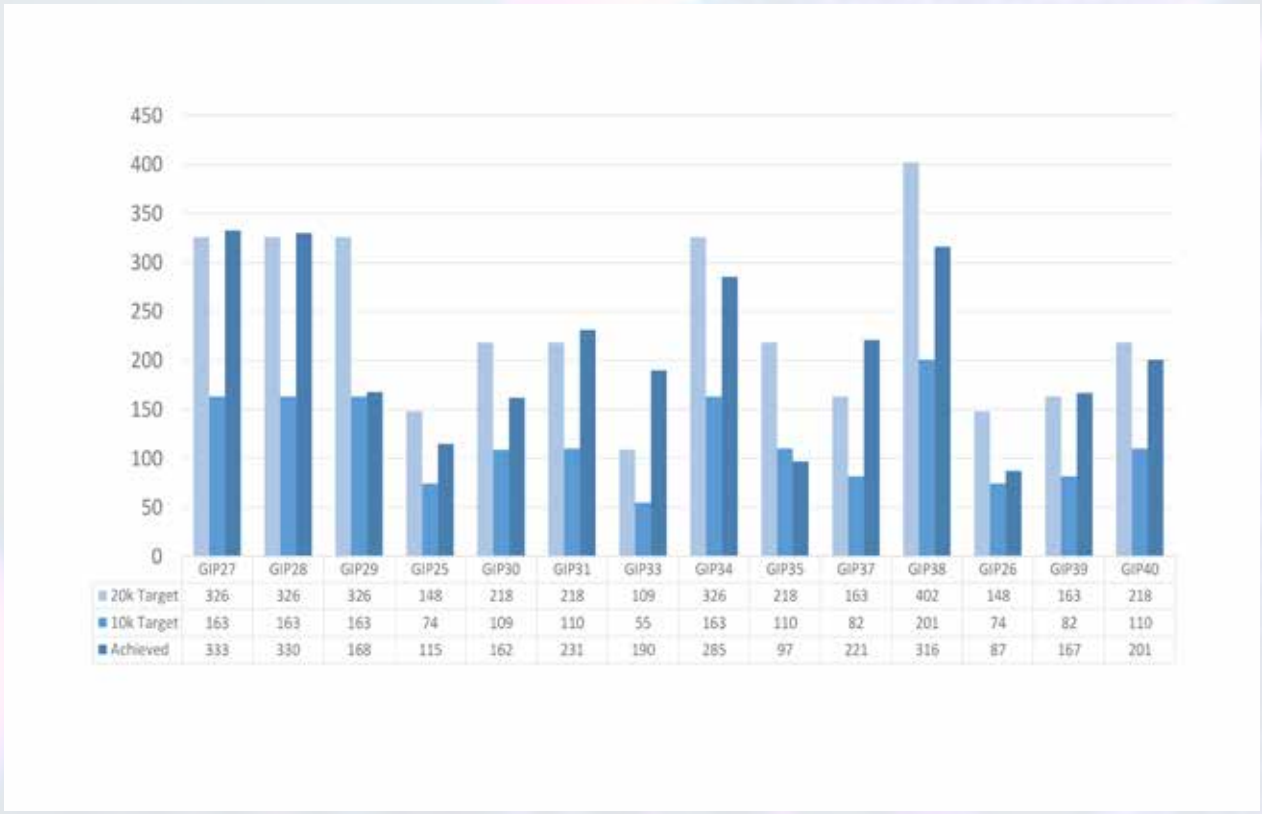
Contributors: Dr. Sridhar Sivasubbu, Dr. Vinod Scaria, Prof. Mitali Mukerji, Pooja Sharma, Shreya Bari, Tiyasha De, Bharathram Upilli, Sandeep Kumar Pal, Rahul Kumar Bhoyar, Bani Jolly, Arushi Batra, Shahrumi Reza, Dr. Simmy Kaur, Dr. Mahino Fatima, Mohammed Akbar, Aishwarya Shankar, Gayatri Singh, Suman Mudila, Divya Goel, Saima Iram, Chavvi Sharma, Aroosa Mir

Role of the Institution in the GenomeIndia Project: CSIR-IGIB played an important role in preparatory phase of GenomeIndia as well as towards the core deliverable of the program. CSIR-IGIB is one of the four whole genome sequencing centres and is responsible for the collection of samples (n=3390) and the sequencing of 2296 human genome samples for 10K targets of GenomeIndia.

Accomplishments and Outcomes

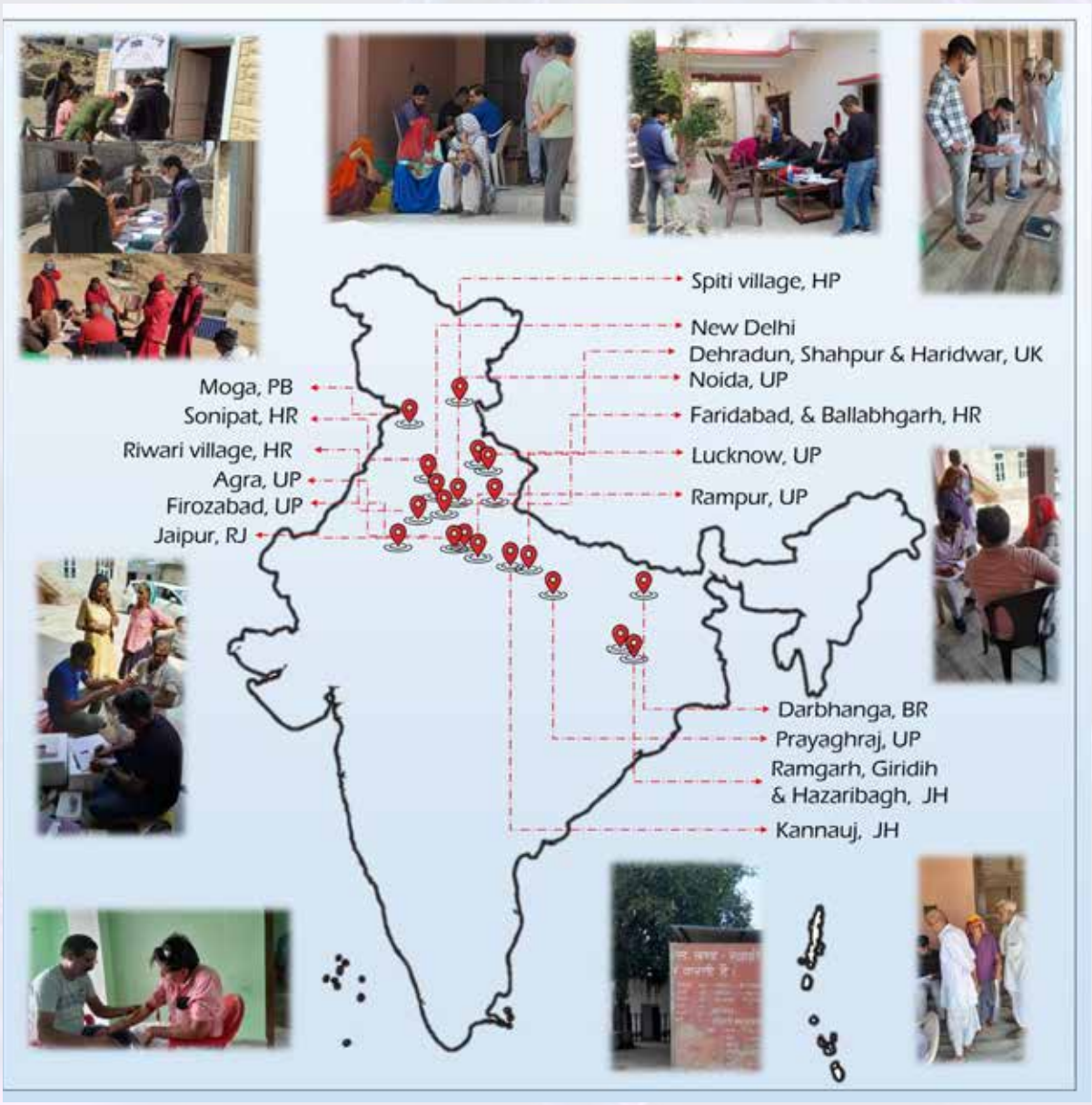
Subject Enrollment and Sample Collection

Central to the success of this project is the meticulous process of sample collection. IGIB has successfully collected 3006 samples across 17 ethnicities. The collection camps were set in various districts and remote villages in around 7 Indian states. The team ensured the collection of maximum unrelated samples and Trios. All the collected samples have been processed for biobanking. Biochemistry and phenotype data has been submitted to CBR for all the collected samples.



	GIP27	GIP28	GIP29	GIP25	GIP30	GIP31	GIP33	GIP34	GIP35	GIP37	GIP38	GIP26	GIP39	GIP40
20k Target	326	326	326	148	218	218	109	326	218	163	402	148	163	218
10k Target	163	163	163	74	109	110	55	163	110	82	201	74	82	110
Achieved	333	330	168	115	162	231	190	285	97	221	316	87	167	201

Sample Collection Sites



Sample collection by the CSIR-IGIB team

Whole Genome Sequencing

IGIB has completed the whole genome sequencing of 2515 samples along with 1691 samples for genotyping array. Out of the 2515 sequenced samples, GVCF generation has been completed for 1761 samples. Data of 1532 samples in the form of FASTQ and GVCF has been successfully transferred to CBR and IBDC.

Medical Genomics Data Analysis

CSIR-IGIB is leader and pioneer in the area of medical genomics. CSIR-IGIB has carried out the analysis of Clinically relevant variations (ACMG actionable genes), Pharmacogenics marker analysis in GenomeIndia, frequency data and their implications as well as finding frequency of pathogenic variations in various rare genetic disorders like Sickle cell anemia, thalessemia etc.

iBRIC - National Institute of Biomedical Genomics

Kalyani

Principal Investigator: Prof. Analabha Basu

Co-Investigators: Dr. Nidhan K. Biswas, Prof. Arindam Maitra

Contributors: Dr. Suman K Paine, Dr. Chandrika Bhattacharya, Poulomi Ghosh, Mahabub Alam, Sourav Gangopadhyay, Azad Ali, Debashree Tagore, Parveena Choudhury, Sukanya Dhar, Saurav Roy, Shouvanik Sengupta, Nasrin Parvin, Rahul Modak, Sayan Bhowmick, Devashish Tripathi, Vinay More, Haya Afreen

Role of the Institution in the GenomeIndia Project: NIBMG is in the forefront in design and execution of the project and implemented the responsibility of collection, storage, sequencing of the samples and analysis of the data.

Accomplishments and Outcomes

NIBMG is the leader in defining the uniform protocols for data generation, variant calling, quality control parameters and downstream population-scale analysis. In Phase 1, the joint call of the variants has been carried out independently by NIBMG and CBR on whole genome sequence data of 5750 individuals. Salient outcomes on preliminary analysis are:

- Unprecedented diversity detected unrevealing a lot of new understanding of population history.
- More than 4.7million common variants identified, which are candidates for association study design (Construction of genotyping chip)
- Discovery of 27 million rare variants, will completely transform the definition of rare disease burden in India.
- There is moderate correspondence between genetic and geographical distances (Figure 1)

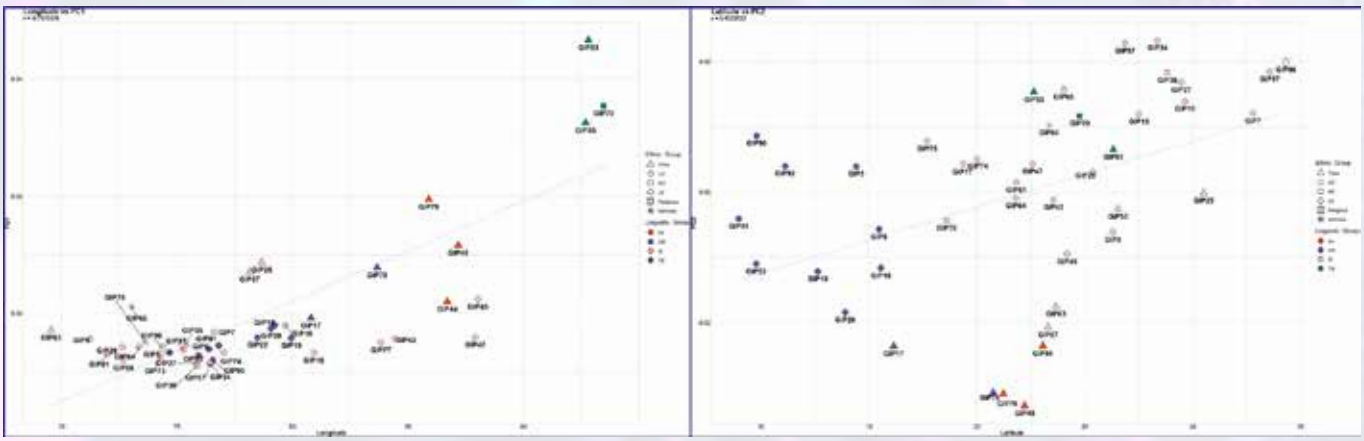


Figure 1: The two dimensional summary of genetic variation in Indian populations when correlated with latitude and longitude produces modest correlation [0.49 between Latitude and Principal component 2 and 0.73 between Longitude and Principal component 1].

Details of the samples collected by NIBMG

Sl. No.	Community	No. of samples collected
1	GIP42	37
2	GIP43	104
3	GIP44	271
4	GIP45	256
5	GIP46	170
6	GIP47	210
7	GIP48	189
8	GIP49	158
Total		1395

Sample collection by the iBRIC-NIBMG team

Sample Collection Centres

All India Institute of Medical Sciences

Jodhpur

Principal Investigator: Dr. Kuldeep Singh

Co-Investigators: Dr. Praveen Sharma

Contributors: Dr. Dolat Singh

Role of the Institution in the GenomeIndia Project: AIIMS Jodhpur is involved in conceptualization, participants recruitment including community engagement & involvement, physical & biochemical phenotyping as per the study protocol.

Accomplishments and Outcomes

A systematic approach including dialogues with regulatory authorities, community leaders and stakeholders, a medical professional in team helped in enrollment. Approaching people outside the state (UP, MP, Haryana) and tribals was challenging but helped the team to devise strategies. Multiple lessons learnt by our team defying challenges of pandemic and funding issues. Each participant provided with routine reports.

Sl. No.	Community	No. of samples collected
1	GIP50	187
2	GIP51	167
3	GIP52	122
4	GIP53	115
5	GIP54	331
6	GIP55	162
7	GIP56	119
8	GIP57	145
9	GIP58	151
10	GIP59	156
Total		1655

Sample collection by the AIIMSJ team

Gujarat Biotechnology Research Centre

Gandhinagar

Principal Investigator: Dr. Madhvi Joshi

Contributors: Prof. Chaitanya G. Joshi

Role of the Institution in the Genome India Project: To collect 1688 samples of 10 different communities from Gujarat and Madhya Pradesh states.

Accomplishments and Outcomes

- A total of 1720 samples were collected (101.89%) from 10 different communities from Gujarat and Madhya Pradesh.
- Sequencing Genome in a Bottle samples showed >99.5% and 98.4% precision for calling SNPs and INDELs, respectively.
- Sequencing of total of 172 samples done and 150 in progress.

Sl. No	Community	No. of samples collected
1	GIP60	166
2	GIP61	326
3	GIP62	163
4	GIP63	148
5	GIP64	148
6	GIP65	154
7	GIP66	160
8	GIP67	155
9	GIP68	150
10	GIP69	150
Total		1720

Sample collection by the GBRC team

iBRIC - Institute of Bioresources and Sustainable Development

Imphal

Principal Investigator: Dr. Nanaocha Sharma

Role of the Institution in the GenomeIndia Project: Sample collection of four assigned communities and data analysis using genetic variants obtained from the project.

Accomplishments and Outcomes

The team has collected samples from various districts of Manipur, Ranchi and Jarkhand successfully.

Sl.No.	Community	Number of samples collected
1	GIP70	370
2	GIP71	148
3	GIP72	148
4	GIP73	109
TOTAL		775

Sample collection by the iBRIC-IBSD team

Indian Institute of Science Education and Research

Pune

Principal Investigator: Dr. Mayurika Lahiri

Co-Investigators: Santosh Dixit and L.S. Shashidhara

Role of the Institution in the GenomeIndia Project: IISER Pune was entrusted with collecting 1142 samples from four different ommunities.

Accomplishments and Outcomes

The required number of samples from each of the communities were collected and sent to CBR, Bengaluru for sequencing. DNA extraction from the samples have been done and kept at IISER Pune as back-up.

Sl.No.	Community	No. of samples collected
1	GIP74	326
2	GIP75	326
3	GIP76	327
4	GIP89	163
TOTAL		1142

Sample collection by the IISER Pune team

iBRIC - Institute of Life Sciences

Bhubaneswar

Principal Investigator: Dr. Sunil K. Raghav

Co-Investigators: Dr. Punit Prasad

Role of the Institution in the GenomeIndia Project: ILS is involved in sample collection with phenotyping and secondary data analysis using the variants generated in the project.

Accomplishments and Outcomes

ILS has collected five populations with a total of 940 samples. All computational pipelines are ready for secondary data analysis for functional association of variants.

Sl. No.	Community	No. of samples collected
1	GIP77	353
2	GIP78	150
3	GIP79	125
4	GIP80	151
5	GIP81	161
6	GIP82	In progress
Total		940

Sample collection by the iBRIC-ILS team

Mizoram University

Aizawl

Principal Investigator: Prof. N. Senthil Kumar

Co-Investigators: Dr. H. Lalhruaitluanga, Dr. J. Zohmingthanga, Dr. C Lalchhandama, Prof. Lalnundanga

Contributors: Ranjan Jyoti Sarma, Andrew Vanlallawma, Baby Lalrintluangi, T. Lahriatpuii

Role of the Institution in the GenomeIndia Project: Mizoram University (MZU) was assigned to collect 592 samples from four different communities including 130 unrelated and 6 trio samples from each population and data analysis. MZU associated with different working groups to strategize the downstream analysis for pharmacogenomics, population and evolutionary genetics and disease associated variants, to be performed using the GenomeIndia data.

Accomplishments and Outcomes

- MZU completed collecting 592 samples from four designated populations from Mizoram, Assam and Meghalaya and conducting blood biochemistry and anthropometry tests. Sample sets with relevant data were sent to NIBMG for WGS, to CBR for biobanking, and also stored locally.
- As a part of GenomeIndia project, the data analyst in MZU has successfully developed pipelines for generating VCF files from FASTQ files automatically and also pipeline for pharmacogenomic study. Moreover, the data analyst also developed python scripts for statistical analysis of SNPs for cancer association comparing healthy and cancer patients.

Sl. No	Community	No. of samples collected
1.	GIP83	148
2.	GIP84	148
3.	GIP85	148
4.	GIP86	148
Total		592

Sample collection by the MZU team

National Institute of Mental Health and Neurosciences

Bengaluru

Principal Investigator: Dr. Shivakumar V.

Co-Investigator: Dr. G. Venkatasubramanian

Contributors: Dr. Naren P. Rao

Role of the Institution in the GenomeIndia Project: The National Institute of Mental Health and Neurosciences (NIMHANS) had the role of sample collection from various communities within Karnataka.

Accomplishments and Outcomes

NIMHANS was tasked with the sample collection from six communities. NIMHANS completed the due process of obtaining permission from the state Government to collect samples from these populations and achieved the required sample collection target. The same has been sent to the sequencing center.

Sl. No.	Community	Number of samples collected
1	GIP1B	114
2	GIP2B	121
3	GIP3B	14
4	GIP87	150
5	GIP88	150
6	GIP4B	3
Total		552

Sample collection by the NIMHANS team

iBRIC - Rajiv Gandhi Centre for Biotechnology

Thiruvananthapuram

Principal Investigator: Dr. E.V. Soniya

Co-Investigators: Dr. Abdul Jaleel K.A.

Role of the Institution in the GenomeIndia Project: Sample collection of seven distinct population groups from Kerala, followed by secondary analysis of variants. Aims to unveil genetic insights and enhance understanding of regional diversity.

Accomplishments and Outcomes

A total of 1,636 blood samples have been successfully collected, from seven different communities. A notable achievement includes the submission of blood samples to the CBR, procured the essential equipment for sample collection and a computing cluster for the secondary analysis of the data, complemented by an operational registration web portal for GenomeIndia.

Sl. No	Community	No. of samples collected
1	GIP21	47
2	GIP23B	48
3	GIP90	448
4	GIP91	372
5	GIP92	354
6	GIP96	185
7	GIP94	182
Total		1,636

Sample collection by the iBRIC-RGCB team

Sher-i-Kashmir Institute of Medical Sciences

Srinagar

Principal Investigator: Prof. Mohd Ashraf Ganie

Co-Investigators: Prof. Bashir Ahmad Charoo, Dr. Mahrukh Hameed Zargar.

Contributors: Dr. Imtiyaz Ahmad Wani

Role of the Institution in the GenomeIndia Project: Sher-i-Kashmir Institute of Medical Sciences, (SKIMS), Srinagar has a role of sampling and phenotyping of five ethnic groups from Jammu, Kashmir and Ladakh divisions of UT of Jammu and Kashmir.

Accomplishments and Outcomes

The SKIMS site has completed its target of sampling and phenotyping is as: GIP 95 (N=326); GIP 96 (326); GIP 97 (326); GIP 98 (148) and GIP 99 (148). Personal information, socio-demographic details and other relevant information has been captured as per the approved questionnaire. All the subjects are done with laboratory tests and samples were shipped to IGIB, Delhi and CBR, Bengaluru for sequencing and biobanking, respectively.

Sample collection by the SKIMS team

Method Development Centres

iBRIC - Centre for DNA Fingerprinting and Diagnostics

Hyderabad

Principal Investigator: Dr. Ashwin Dalal

Co-Principal Investigator: Dr. Murali Dharan Bashyam

Role of the Institution in the GenomeIndia Project: Data analysis using genetic variants obtained from the project.

Accomplishments and Outcomes

We successfully showed the utility of GenomeIndia sequence data in improving mutation calling in sporadic (cancer) and familial diseases, by whole genome sequence data of a small set of 28 'normal' samples. More importantly, this utility was proven to be over and above other 'non-Indian' genome data such as gnomAD and GenomeAsia.

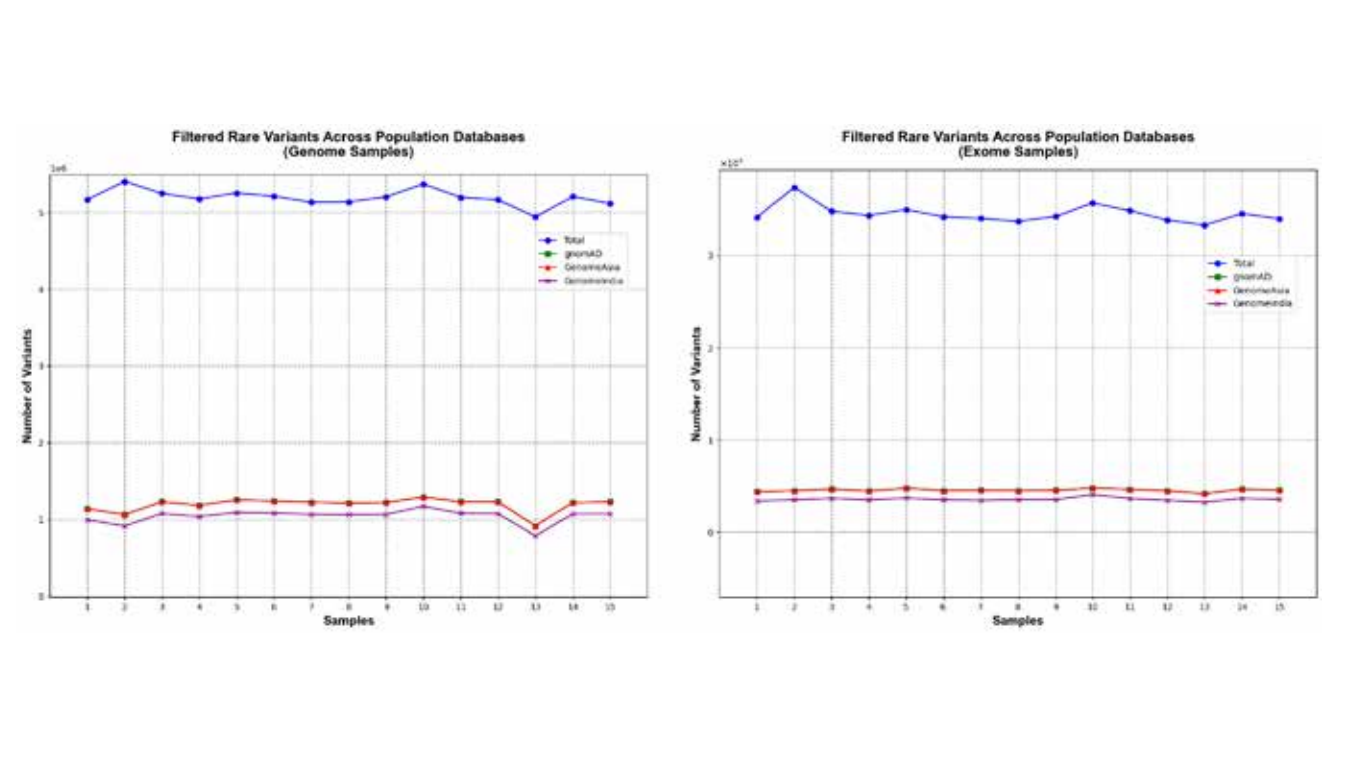


Figure 1: Variant filtering using gnomAD at 1% frequency removed average 77% of variants in genome and 86% of variants in exome, GenomeAsia based filtering had minimal impact, and GenomeIndia based filtering further reduced variants by about 2.86% (Approx. 149,536 variants in Genomes and 988 in Exomes)

Indian Institute of Information Technology
Allahabad

Principal Investigator: Prof. Pritish Kumar Varadwaj

Co-Investigators: Dr. B.S. Sanjeev

Contributors: Dr. Imlimaong Aier

Role of the Institution in the GenomeIndia Project: To create tools for large-scale data analysis and to create a knowledgebase for association mining on gene-disease-drug interplay

Accomplishments and Outcomes

The IIITA team has developed tools for fast and accurate clustering of large-scale data using DASK and RAPIDSAI frameworks. These tools are optimized for chromosome-based and complete genome-based clustering. Further, the team have optimized the pipelines for IBD analysis to identify common subpopulation SNPs. In addition, the team has developed a gene-disease-drug association database with six network mining and visualization tools (GDDExplorer, DGisExplorer, Disease2Gene, Disease2Drug, GDisexplorer, Gene2Disease). The detailed notes and algorithm are available on GitHub.

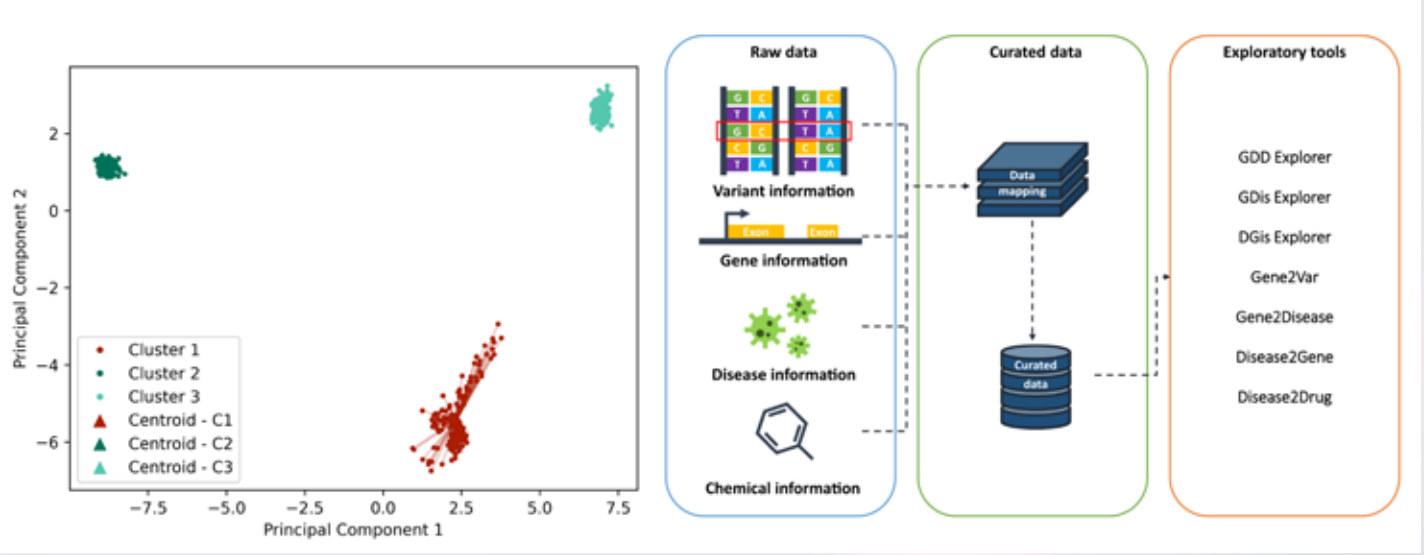


Figure 1: Clustering based on whole genome sequence

Figure 2: In-house designed Gene-Disease-Drug database

Indian Institute of Science

Bengaluru

Principal Investigator: Prof. Y. Narahari

Co-Investigators: Prof. Yogesh Simmhan and Prof. Arun Kumar

Contributors: Prof. Chirag Jain

Role of the Institution in the GenomeIndia Project: Developing novel algorithms based on big data analytics for compression and decompression of Whole Genome Sequence (WGS) datasets for efficient data storage and transfer.

Accomplishments and Outcomes

We have developed pipelines based on advanced bioinformatics algorithms for seamless and guaranteed lossless compression and decompression of GenomeIndia uBAM datasets. Such compression tools can help dramatically reduce storage costs for such large genomics repositories and also reduce the data transfer time over the Internet, which can hasten the time to effective science. These leverage parallel optimizations to achieve a 5x reduction in size (from ~50GB to ~5GB per sequence) that saves on storage and transfer costs, and a parallelized time of 120mins. The pipeline is modular and allows newer algorithms to be rapidly incorporated over time. The success of this tool can further be enhanced through concurrency and I/O reduction techniques, and expanded to methods that which trade-off metadata loss and read ordering against performance.

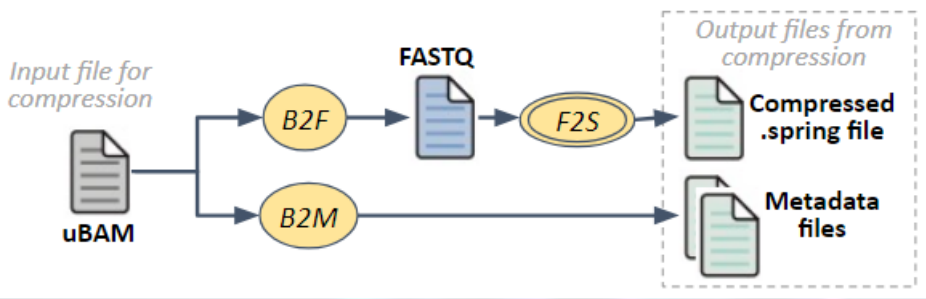


Figure 1: Compression Pipeline. F2S stage is parallelized

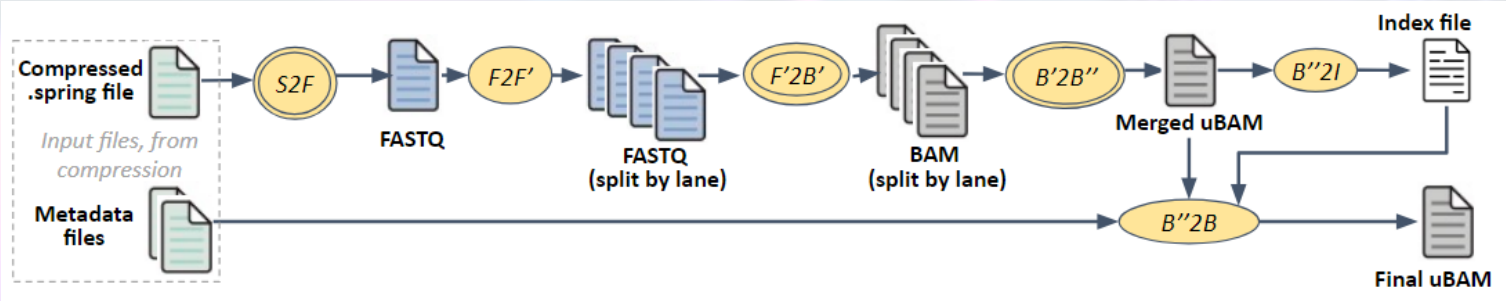


Figure 2: Decompression Pipeline. Parallel tasks are shown in double ovals.

Indian Institute of Technology Delhi

New Delhi

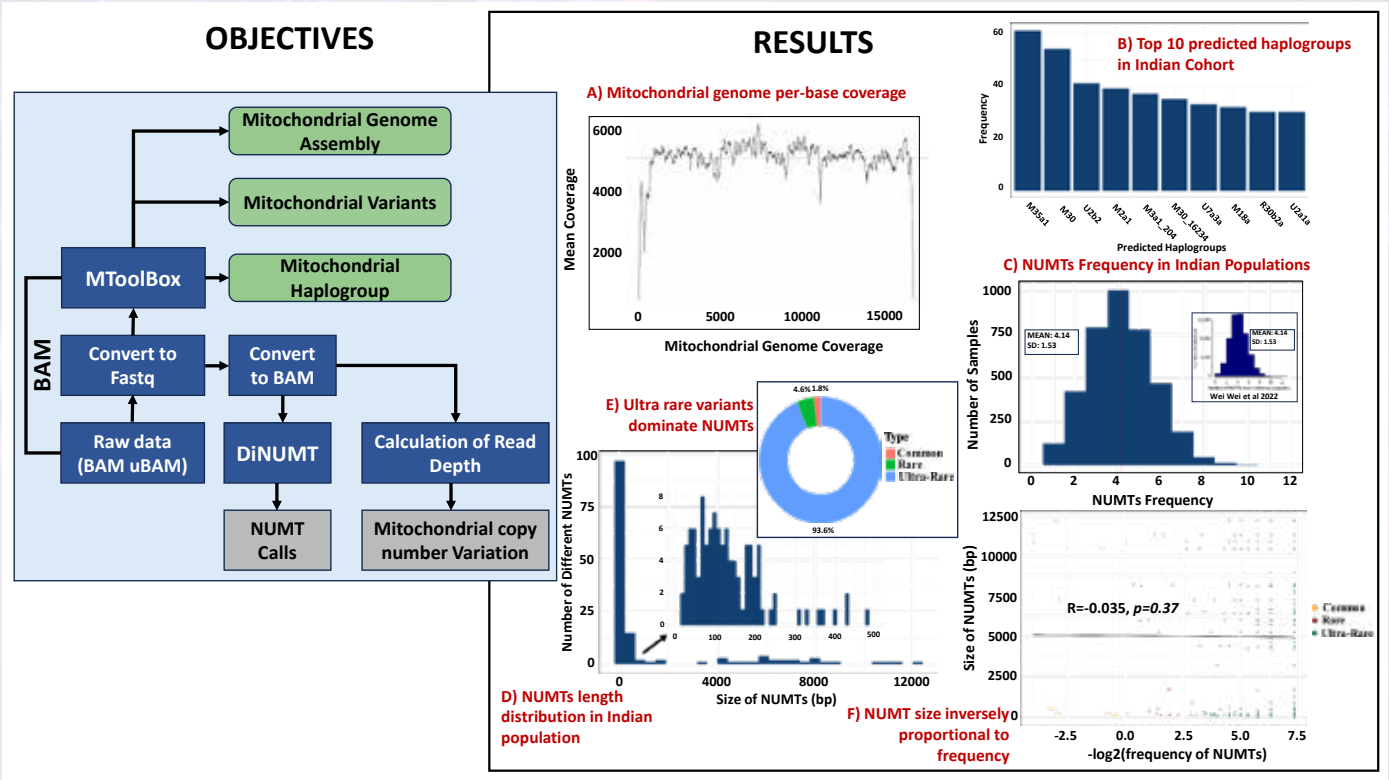
Principal Investigator: Prof. D. Sundar

Co-Investigators: Prof. Ishaan Gupta

Role of the Institution in the GenomeIndia Project: We are working on algorithms and pipelines for de novo assembly of the mitochondrial genome. We aim to generate the Indian mitochondrial genome reference to identify mitochondrial disorders, enumerate variation, and perform population genetics and genomics.

Accomplishments and Outcomes

After benchmarking all the available tools (Singh et al. BMC Bioinformatics, 2023), we have optimised our pipeline (available on GitHub linked to GenomeIndia). This is being used to process 50-75 genomes per day at IITD and CBR, having completed the analysis of 5500 genomes. The analysis includes de novo assembly, enumerating mitochondrial heteroplasmy and copy number variation, and unique Nuclear encoded mitochondria patterns. This data is used for population genetics and functional genomics of measured traits.



A) Per-base coverage plot of mitochondrial genome for 5500 individuals. The line represents the mean coverage per base across the population. B) The top three mitochondrial haplogroups for the Indian population are M35a1, M30 and R5a2b. C) Individuals had an average of 4.18 NUMTs (S.D. = 1. 1.53) absent in the reference sequence. D) NUMT length distribution in the Indian population. Most of the NUMTs are less than 500 bp. E) Out of 615, 574 (93.3%) were ultra-rare(F<0.1%), 29 (4.7%) were rare(0.1%<=F<1%) and 12 were common(F>1%). F) NUMT size is inversely proportional to the population frequency (R2= -0.017).

Indian Institute of Technology Jodhpur

Jodhpur

Principal Investigator: Dr. Pankaj Yadav

Co-Investigators: Dr. Sushmita Jha

Contributors: Vaishnavi Jangale, Rajveer Singh Shekhawat, Soham Biswas, Samarpita Saha

Role of the Institution in the GenomeIndia Project: Develop a variant prioritization pipeline based on machine learning algorithms for genome-wide association studies (GWAS)

Accomplishments and Outcomes

We developed a comprehensive and robust variant prioritization pipeline involving data quality control and feature selection, followed by association analysis using machine learning methods such as Support Vector Regression (SVR). We tested our pipeline on both simulated and real datasets. Our pipeline could determine the top SNPs using permutation scores. Further, our pipeline uses a variety of tools to evaluate the biological importance of identified SNPs. These includes GRASP for literature p-values, GTEx for SNP expression analysis, Disgenet for gene-disease associations, and the Panther classification system for biological, molecular, and cellular studies. This extensive investigation of SNP association ensures a complete examination, illuminating their possible importance and contributes to a more detailed understanding of their biological implications for given trait.

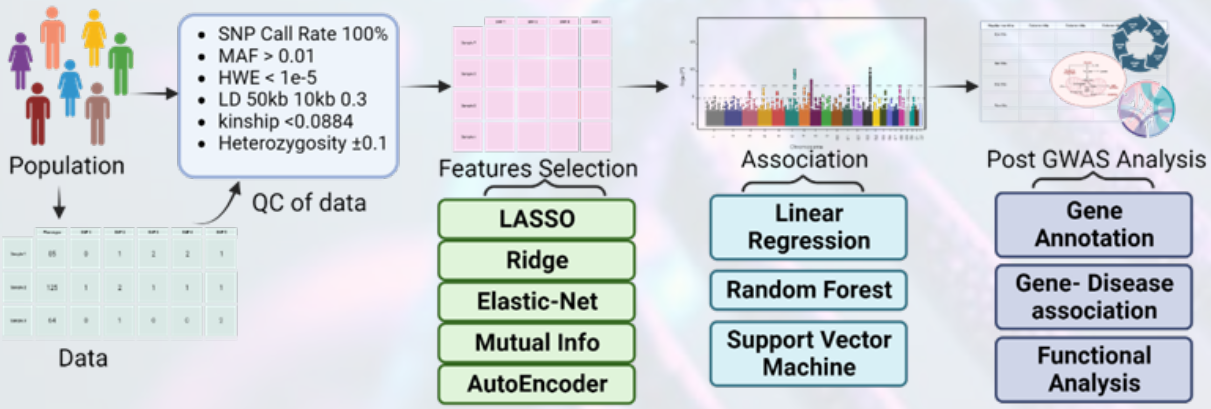


Figure 1: Proposed pipeline for SNP-phenotype association

Indian Institute of Technology Madras

Chennai

Principal Investigator: Dr. Himanshu Sinha

Co-Investigators: Dr. Karthik Raman, Dr. Manikandan Narayanan

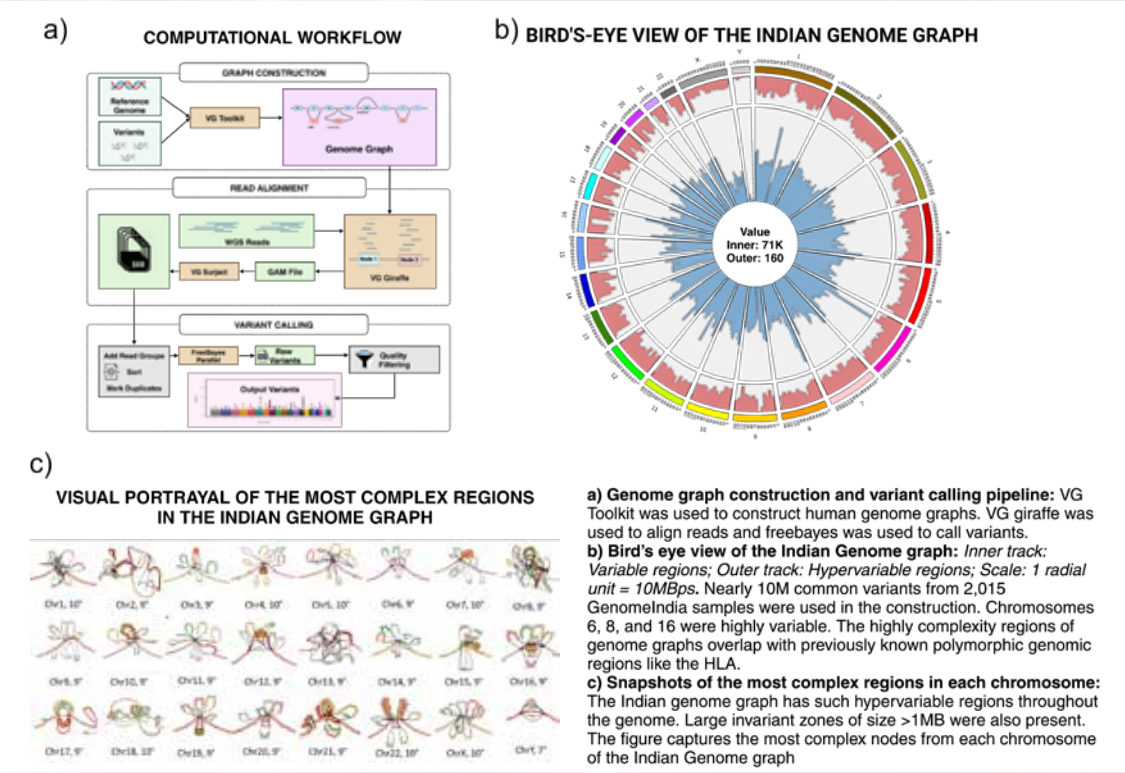
Contributors: Venkatesh Kamaraj, Ayam Gupta, Harshita Agarwal

Role of the Institution in the GenomeIndia Project: Design and implement graph-based algorithms for constructing Indian-population-specific reference genome structures and comprehensively analyse the variants and heterogeneity of Indian subpopulations.

Accomplishments and Outcomes

To explore a genome, the first step is to align the sequenced reads with a reference genome. The human reference genome currently in use (GRCh38) is a DNA sequence representing each chromosome as a linear, continuous string of nucleotide bases. This reference genome serves as a blueprint for comparison of newly sequenced genomes. GRCh38 is limited in capturing the genetic diversity of various populations, leading to reference allele bias in the analysis. To address this, genome graphs, augmented by known variants, offer a more comprehensive representation.

Our research focuses on creating genome graphs tailored to the Indian population. We've developed computational pipelines integrating DRAGEN-called variants to construct these graphs. The initial Indian genome graph includes common variants from 2,015 GenomeIndia samples. Additionally, we've devised novel frameworks for complex structural analysis and extensive functional and sample-level annotation of genome graphs. We will see advancements in mapping reads, calling of genetic variants, and analysis of human genomes through genome graphs.



a) Genome graph construction and variant calling pipeline: VG Toolkit was used to construct human genome graphs. VG giraffe was used to align reads and freebayes was used to call variants.
b) Bird's eye view of the Indian Genome graph: *Inner track: Variable regions; Outer track: Hypervariable regions; Scale: 1 radial unit = 10MBps.* Nearly 10M common variants from 2,015 GenomeIndia samples were used in the construction. Chromosomes 6, 8, and 16 were highly variable. The highly complexity regions of genome graphs overlap with previously known polymorphic genomic regions like the HLA.
c) Snapshots of the most complex regions in each chromosome: The Indian genome graph has such hypervariable regions throughout the genome. Large invariant zones of size >1MB were also present. The figure captures the most complex nodes from each chromosome of the Indian Genome graph

National Centre for Biological Sciences

Bengaluru

Principal Investigator: Prof. Raghu Padinjat

Co-Investigators: Dr. Sabarinathan Radhakrishnan

Role of the Institution in the GenomeIndia Project: Constructing a genome-wide map of constrained genomic regions in Indian population. Supporting genome sequencing through the NCBS NGS facility.

Accomplishments and Outcomes

We have developed a bioinformatic pipeline to identify constrained genomic regions (indicating the signals of purifying selection). The pipeline was validated using publicly available dataset such as gnomAD and GenomeAsia. Further, we have applied it to the GenomeIndia dataset (joint variant calls from 5750 samples) and currently exploring the results.

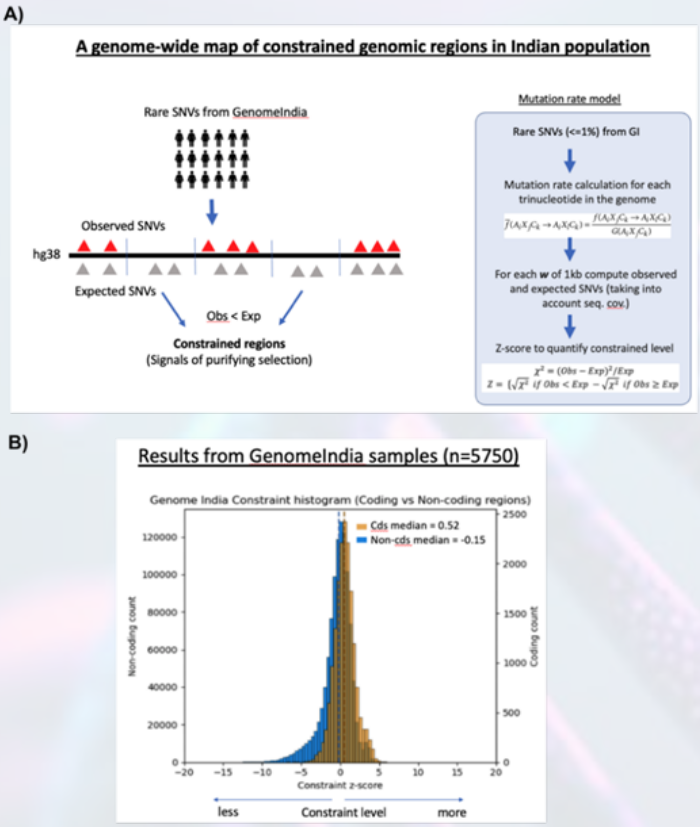


Figure 1: A) Overview of the method to detect constrained genomic regions, and B) Distribution of constrained z-score for genomic windows of 1kb generating using GenomeIndia dataset.

Biobanking and Data Archival Centres

GenomeIndia Biobank at the Centre for Brain Research

Bengaluru

CBR biobank hosts biofluids from the cohort studies that are being undertaken at the centre. It is the primary repository to store and safeguard samples from the Upon the receipt of the samples, they are stored in minus 30-degree freezers and are taken up for DNA isolation. The isolated DNA reconstituted in the elution buffer is further made into small aliquots and the aliquots are stored at minus 80-degree freezers at CBR biobank. A systematic protocol is followed for sample collection, storage, and processing helps to maintain the integrity and traceability of the samples, while preserving the confidentiality of the study participants.

Sl. No	Institutes Name	No. of samples received for biobanking
1	CBR	3003
2	CCMB	1506
3	IGIB	3131
4	NIBMG	1260
5	AIIMS-J	1505
6	IISER-P	1153
7	RGCB	1851
8	NIMHANS	519
9	ILS-B	940
10	GBRC	1720
11	MZU	592
12	SKIMS	1274
13	IBSD	777
Total		19231

Figure 1: Biobanking facility at CBR, Bengaluru

Indian Biological Data Centre, Regional Centre for Biotechnology

Faridabad

Lead Coordinator: Dr. Arvind Sahu

Co-Investigators: Dr. Debasisa Mohanty (NII), Dr. Dinesh Gupta (ICGEB) and Dr. Deepak T Nair (RCB)

Contributors: Mr. Sanjay Deshpande (RCB-IBDC) and Mr. Kalpanath Paswan (RCB-IBDC)

Role of the Institution in the GenomeIndia Project:
The Indian Biological Data Centre, hosted at the Regional Centre for Biotechnology (RCB), Faridabad, serves as a digital repository for data generated in India in the area of life sciences. The role of the IBDC is to archive and provide access to the datasets generated by the GenomeIndia project.

Accomplishments and Outcomes

RCB-IBDC has provided efficient access to all the Sequencing and Analysis centres by creating FTP accounts for the Upload/Downloading of the (UBAM, FASTQ and GVCF) datasets. RCB-IBDC is connected with all the stake-holding institutes by a dedicated NKN-GenomeIndia VRF line for secure data exchange. RCB-IBDC has created a dedicated portal of GenomeIndia for the metadata of data generated. RCB-IBDC will also aid development of policy to enable access to the deposited data to appropriate stakeholders.

Role of CBR:
The overall data transfer activities were coordinated by a team led by Mr. Jothibasu from the Centre for Brain Research.

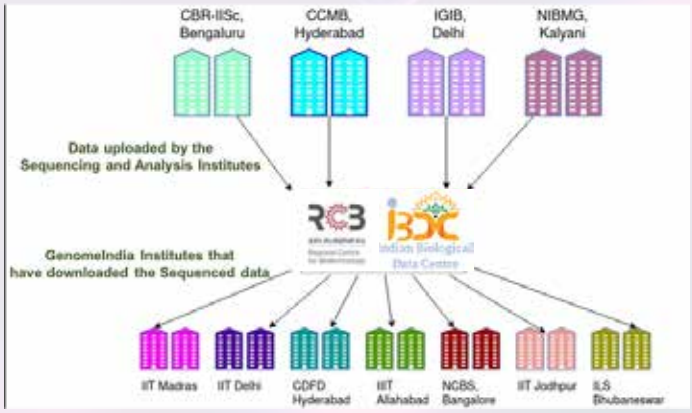


Figure 1: Data sharing enabled by IBDC

Monitoring Committee & Investigators

GenomeIndia Technical Monitoring and Assessment Committee (TMAC)

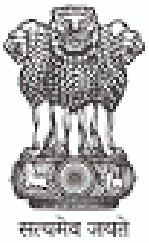
Sl. No.	Name of the Experts	Chair/Member
1.	**Prof. Partha P. Majumder** iBRIC-NIBMG, Kalyani	Chair
2.	**Prof. Shiv Kumar Sarin** ILBS, New Delhi	Co-Chair
3.	**Dr. Debasisa Mohanty** NII, New Delhi	Co-Chair
4.	**Shri. Vishvajit Sahay** AS & FA, DBT	Member
5.	**Dr. Suchita Ninawe** Adviser, DBT	Member
6.	**Prof. Shinjini Bhatnagar** THSTI, Faridabad	Member
7.	**Prof. B.K. Thelma** UDSC, New Delhi	Member
8.	**Prof. Madhulika Kabra** AIIMS, New Delhi	Member
9.	**Dr. Vivek Kumar Singh** BHU, Varanasi	Member
10.	**Dr. G.R. Chandak** CSIR-CCMB, Hyderabad	Member
11.	**Dr. S. Shivaji** LVPEI, Hyderabad	Member
12.	**Dr. T. Rajkumar** Cancer Institute (WIA), Chennai	Member
13.	**Dr. Binay Panda** JNU, New Delhi	Member
14.	**Dr. Janesh Kumar** NCCS, Pune	Member
15.	**Dr. Richi V. Mahajan** Scientist D, DBT	Member Secretary

Our sincere tributes to the erstwhile Chair of the TMAC, Late Professor M.R.S. Rao, Jawaharlal Nehru Centre for Advanced Scientific Research (JNCASR), Bengaluru, for his excellent and scholarly guidance

GenomeIndia Investigators

1.	Prof. Y. Narahari, CBR & IISc, Bengaluru	National Coordinator
2.	Prof. K. Thangaraj, CSIR-CCMB, Hyderabad	National Coordinator
3.	Prof. Bratati Kahali, CBR, IISc Campus, Bengaluru	Principal Investigator
4.	Dr. Vinay K. Nandicoori, CSIR-CCMB, Hyderabad	Principal Investigator
5.	Dr. Mohammed Faruq, CSIR-IGIB, Delhi	Principal Investigator
6.	Prof. Analabha Basu, iBRIC-NIBMG, Kalyani	Principal Investigator
7.	Prof. Kuldeep Singh, AIIMSJ, Jodhpur	Principal Investigator
8.	Prof. Madhvi Joshi, GBRC, Gandhinagar	Principal Investigator
9.	Dr. Nanaocha Sharma, iBRIC-IBSD, Imphal	Principal Investigator
10.	Dr. Mayurika Lahiri, IISER-Pune, Pune	Principal Investigator
11.	Dr. Sunil K. Raghav, iBRIC -ILSB, Bhubaneswar	Principal Investigator
12.	Prof. N. Senthil Kumar, MZU, Aizwal	Principal Investigator
13.	Dr. Shivakumar V., NIMHANS, Bengaluru	Principal Investigator
14.	Dr. E.V. Soniya, iBRIC-RGCB, Thiruvananthapuram	Principal Investigator
15.	Prof. Mohd Ashraf Ganie, SKIMS, Srinagar	Principal Investigator
16.	Dr. Ashwin Dalal, iBRIC-CDFD, Hyderabad	Principal Investigator
17.	Prof. Pritish Kumar Varadwaj, IIITA, Allahabad	Principal Investigator
18.	Prof. D. Sundar, IITD, Delhi	Principal Investigator
19.	Dr. Pankaj Yadav, IITJ, Jodhpur	Principal Investigator
20.	Dr. Himanshu Sinha, IITM, Madras	Principal Investigator
21.	Prof. Raghu Padinjat, NCBS, Bengaluru	Principal Investigator
22.	Dr. Divya Tej Sowpati, CSIR-CCMB, Hyderabad	Co-Investigator
23.	Dr. Karthik Bharadwaj Tallapaka, CSIR-CCMB, Hyderabad	Co-Investigator

24.	Dr. Nidhan K. Biswas, iBRIC-NIBMG, Kalyani	Co-Investigator
25.	Prof. Arindam Maitra, iBRIC-NIBMG, Kalyani	Co-Investigator
26.	Dr. Praveen Sharma, AIIMSJ, Jodhpur	Co-Investigator
27.	Dr. Santosh Dixit, IISER-Pune, Pune	Co-Investigator
28.	Dr. L.S. Shashidhara, IISER-Pune, Pune	Co-Investigator
29.	Dr. Punit Prasad, iBRIC-ILS, Bhubaneswar	Co-Investigator
30.	Dr. H Lalhruaitluanga, MZU, Aizawl	Co-Investigator
31.	Dr. J. Zohmingthanga, MZU, Aizawl	Co-Investigator
32.	Dr. C. Lalchhandama, MZU, Aizawl	Co-Investigator
33.	Prof. Lalnundanga.MZU, Aizawl	Co-Investigator
34.	Dr. G. Venkatasubramanian, NIMHANS, Bengaluru	Co-Investigator
35.	Dr. Abdul Jaleel K.A, iBRIC-RGCB, Thiruvananthapuram	Co-Investigator
36.	Dr. Murali Dharan Bashyam, iBRIC-CDFD, Hyderabad	Co-Investigator
37.	Dr. B.S. Sanjeev, IIITA, Allahabad	Co-Investigator
38.	Prof. Yogesh Simmhan, IISc, Bengaluru	Co-Investigator
39.	Prof. Arun Kumar, IISc, Bengaluru	Co-Investigator
40.	Prof. Ishaan Gupta, IITD, Delhi	Co-Investigator
41.	Dr. Sushmita Jha, IITJ, Jodhpur	Co-Investigator
42.	Dr. Karthik Raman, IITM, Madras	Co-Investigator
43.	Dr. Manikandan Narayanan, IITM, Madras	Co-Investigator
44.	Dr. Sabarinathan Radhakrishnan, NCBS, Bengaluru	Co-Investigator
45.	Dr. Prathima Arvind, CBR, IISc Campus, Bengaluru	Project Manager
46.	Mr. Jothibasu, CBR, IISc Campus, Bengaluru	IT team



DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology
Government of India

dbtindia.govt.in /dbtindia @dbtindia @dbtindia

Contact Information

Dr. Suchita Ninawe
Adviser
Department of Biotechnology
Ministry of Science and Technology
Government of India
E-mail: suchita.ninawe@dbt.nic.in

Dr. Richi V. Mahajan
Scientist 'D'
Department of Biotechnology
Ministry of Science and Technology
Government of India
E-mail: rv.mahajan@dbt.nic.in

DEPARTMENT OF BIOTECHNOLOGY
Ministry of Science & Technology
Government of India

GenomeIndia Collaborating Institutes

BRIC Institutions
a DBT Organization

Institute of Life Sciences Bhubaneswar

RGCB
DISCOVERIES FOR A BETTER TOMORROW

CBR

CCMB

IGIB
INSTITUTE OF GENOMICS & INTEGRATIVE BIOLOGY
Genomics Knowledge Partner

MIZORAM UNIVERSITY

GBRC

IISER PUNE

SHER-I-KASHMIR INSTITUTE OF MEDICAL SCIENCES, SGR

भारतीय विज्ञान संस्थान

ncbs
National Centre for Biological Sciences
Tata Institute of Fundamental Research

RCB
Regional Centre for Biotechnology

IBDC
Indian Biological Data Centre