GENOMIC 'UNITY IN DIVERSITY' OF INDIA AND ITS IMPLICATIONS FOR GENOMIC MEDICINE

ARIJIT MUKHOPADHYAY
(ON BEHALF OF THE INDIAN GENOME VARIATION CONSORTIUM)

The Indian Genome Variation Consortium is one of the most ambitious endeavors taken up by the Council of Scientific and Industrial Research with funding from the Government of India. In this project we have acquired genetic information from hundreds of loci in more than 2000 individuals representing 55 populations of India. The data has revealed surprising similarities between populations that are contrasting to each other, geographically and culturally. It has shed new light on our population history and most importantly it has given us the handle to uncover hitherto unknown, private genetic changes that either protects or puts at-risk for a large number of common complex diseases. Based on this data future policies can be made which will lead to better, more efficient and population specific disease management for India.

Institute of Genomics and Integrative Biology, Council of Scientific and Industrial Research (CSIR), Mall Road, Delhi-110007, India. arijit@igib.res.in ; Phone 91-11-27666156, Fax 91-11-27667471.

Preamble

India is globally known for its hospitality and for spreading the message of VASUDHAIVA KUTUMBKAM (the world is a family). A lot of verses are written in many different Indian languages to showcase this since times immemorial. This central theme also unites the entire country across all social and cultural borders. At the same time, the diversity of India at all levels is simply mind-boggling. We are trying to live in peace and harmony with more than one billion people belonging to thousands of endogamous population groups which speak more than 300 different languages having more than 25 different scripts. This is what makes “Unity in Diversity” such an apt slogan to represent India. India is also a unique mixture of people and their cultures which resembles different parts of the world. On one hand we have populations who morphologically resemble more with people from other east-Asian countries, than with Indian mainland population; on the other hand we have people who are difficult to distinguish from the Europeans or Africans. Thus, India, owing to its diversity, provides a glimpse of the world, and serves as a melting pot of modern human civilization. The clue to the rich population diversity of India is in our societal rules that dictate our marriage patterns, thereby restricts the gene flow across the entire country keeping the diversity maintained. This also implicates that each community, with albeit homogeneous genetic make-up can have their own private genetic variation sets that shape the way they evolve and protect vis-à-vis predispose to various diseases. To decipher these genetic signals as well as to know our genomic roots, the Council of Scientific and Industrial Research (CSIR) had initiated the Indian Genome Variation Consortium (IGVC) back in 2002. In this article, the main highlights of the study will be discussed along with its implications, followed by the road ahead.

The partners in IGVC

This initiative by CSIR brought together six constituent laboratories, namely, Institute of Genomics and Integrative Biology (IGIB), Delhi (nodal laboratory), Centre for Cellular and Molecular Biology (CCMB), Hyderabad, Indian Institute of Chemical Biology (IICB), Kolkata, Central Drug Research Institute (CDRI), Lucknow, Industrial Toxicological Research Centre (ITRC, now called Indian Institute of Toxicology Research, IIITR), Lucknow and Institute of Microbial Technology (IMTECH), Chandigarh.
Apart from the CSIR laboratories, a key participant in the project was the Indian Statistical Institute (ISI), Kolkata. This project had participation of 152 scientific personnel directly involved with the study. The project also involved active participation of the Anthropological Survey of India that has helped in the identification of the various Indian subpopulations.

**The Strategy**

The human genome, consisting 6 billion nucleotides (diploid set), as it stands today, is more variable than conserved between any two individuals. Even within the two (haploid) sets in one individual there are differences in the genome structure and composition. Although majority of these changes are either innocuous or cause subtle perturbation that can be tolerated by our system, some of these variations cause change in the way the gene products function which can manifest into a disease. Out of all kinds of variations the genome harbor, single nucleotide polymorphisms (SNPs) are the most stable type. SNPs are allelic variations that arise by replication error of DNA polymerase refractory to the proof-reading activity. These variations then follow Mendelian segregation pattern through multiple generations and eventually stabilize in a population over a large time period. The effect of natural selection upon these variations also influence there frequency in a population. Currently, more than 15 million SNPs are known in the human genome (www.hapmap.org). SNPs are usually biallelic as the probability of a replication error occurring at the same nucleotide twice is extremely small. The biallelic nature makes the assays and downstream analyses simpler than other types of polymorphisms. In the IGVC project, for these reasons, we decided to study the genomic relatedness of various Indian populations through SNPs distributed across the human genome. The SNPs were mostly chosen from functionally important regions (coding region, promoter etc.) to study their disease-risk potential in different populations.

The next challenge was to choose the right population sets to study. As mentioned above, India is home to thousands of communities and sampling all of them was impractical. So the first major hurdle was to choose the minimum number of population that will represent the extent of genomic unity in diversity of India. Without too many repetitions of nearly identical populations. We formed a matrix of language (Austro-Asiatic, Dravidian, Indo-European and Tibeto-Burman) and ethnicity (caste and tribe) versus geographical regions (North, South, East, West, North-East and Central) and from each cell of the matrix chose the two most contrasting population. For example, a caste population from north India speaking Indo-European language was paired with a tribal population from north speaking Indo-European language. In case of absence of a contrasting pair, a matrix cell would be represented by only one population. Using this strategy we had finalized on 55 populations for this study that comprised of 32 large (caste) populations and 23 isolated (tribal) populations. We had also adopted a strategy of encoding the population names where each population was represented by its linguistic background, followed by its geographical location and then its ethnicity. For example, IE-N-LP would mean an Indo-European (IE) language speaking population from the northern part of India (N) which is a caste population (LP, Large Population, as caste populations are usually >10 million). Similarly, IE-N-IP would mean an isolated/tribal population (IP) with IE language from north. Details about all population codes and their positions can be obtained from the two major publications from this study1,2.

Initially, we hypothesized that the unique population pool of India will also harbor many novel SNPs. To test that we selected a panel of 43 samples (‘discovery panel’) from 43 different populations chosen from the matrix described above. We had sequenced randomly selected regions from the entire genome to detect novel SNPs as well as to validate the known SNPs in these samples. We were surprised to find that only 4% of the detected SNPs were novel and hence we proceeded mainly with estimating the SNPs already known from other world populations.

**Choice of SNP Markers to Study**

As mentioned in the beginning, the focus of this study was to identify potential SNPs where an allelic variation might alter the risk of manifesting a disease phenotype. In addition another aim was to analyze the distribution of reported at-risk alleles across the country. Keeping that focus in mind, we selected putative and reported functional SNPs from various genes responsible for a variety of complex diseases. Thus in the first phase of the study we had analysed 405 SNPs from 75 candidate genes and a 5.2 Mbp region from chromosome 22 in a total of 1871 individuals. This data allowed us to evaluate their possible role in disease and drug-treatment outcomes in studies in the future. Further the data has also been used to construct a genetic map of India that seeks to understand the relationships and differences between these 55 populations2.
Genomic Unity in Diversity

We had estimated the genetic differentiation to analyze diversity across these 55 populations. The majority of the populations were genetically different from each other - that is they exhibit "population differentiation". However, the quantitative measure of this differentiation, which is expressed as genetic distance ($F_{ST}$) suggested that the extent of differentiation overall was low. With respect to a few individual SNP loci, the extent of genetic differentiation in India was found to be high and of comparable magnitude to that observed among continental populations.

Maximum differentiation was observed between tribal populations which show different linguistic lineages. On a Pan-India level, when populations were grouped by language or by geographical region of habitat, the extent of genetic differentiation among linguistic or geographical groups was not statistically significant. However, grouping by ethnicity (caste and tribe) indicated significant differentiation. This might be due to the fact that the historically tribal populations are usually highly inbred compared to the cast populations (antiquity and isolation). The spectrum of genetic differentiation within geographical regions or ethnic-linguistic groups revealed that the extent of genetic difference between Dravidian (DR) castes and tribes was low while the Indo-European (IE) caste and tribe groups were significantly differentiated. Within tribes, but not castes, the IE- and DR-speakers showed statistically significant differentiation; however, the DR tribal groups were not significantly differentiated from the AA-speaking groups. The above results implicate that if one uses a sample cohort from multiple populations without considering all the three criteria (namely; ethnicity, linguistic affiliation and geographical location) for any population based genetic study in India, then it is likely that the inferences will be wrong. This is because two genetically different populations (but similar by either ethnicity, or language or geography) can have very different frequency of SNPs thereby a mixture will not represent a true state of the variation in the cohort.

What has been described above pertains to the "diversity" in India - this was followed up by measurement of genetic affinities to assess "unity". We used multiple statistical techniques to analyze the extent of genetic relatedness among the populations. Our analyses demonstrated there are only few major and distinctly different clusters present in the Indian population. The first cluster primarily comprised of Austro-Asiatic (referred to as AA) and Dravidian (referred to as DR) populations consistent with the earlier observation of a statistically non-significant genetic differentiation between AA and DR tribals. The second cluster included Tibeto-Burman (referred to as TB) speaking populations, irrespective of their geographical region of habitat which is made up of three Indo-European isolated populations (IE-IPs) and two IE-Large Populations (IE-LPs) who reside in the Himalayan belt. Another separate cluster is formed with the DR-speaking large populations and isolated populations predominantly from southern India. Thus, ethnicity (tribal/non-tribal) and language seem to be the major determinants of genetic affinities between the populations of India rather than the geography except for a population called OG-W-IP which was derived from an African population. The other noteworthy feature is that many clusters contain one of more populations that are socially or geographically distinct from the other populations belonging to that cluster. These exceptions are not unexpected in a country like India with history of genetic admixtures between diverse lineages. For example, it has been thought that Dravidian speakers, now geographically confined to southern India, were more widespread throughout India prior to the arrival of the Indo-European speakers and possibly underwent genetic admixture with the Indo-European population before retreating to southern India. This hypothesis has been supported by mitochondrial DNA analyses. This feature is present in other populations in India as well where high degree of genetic heterogeneity has resulted from admixture among populations. It is surprising that in spite of such high levels of admixtures; enough genetic differences still remain to distinguish one population from the other in terms of their genetic heterogeneity.

Our Global Genomic Relations

The next obvious question was, how do we relate to the rest of the world from a genomic perspective? Considering that in the global SNP analysis (www.hapmap.org) the Indian data was absent, we assessed the proximity of populations included in the HapMap study with Indian populations using SNPs for which frequencies were available in both the Indian and HapMap populations. The isolated populations of the Himalayan belt (irrespective of their linguistic background) were closest to the Chinese (CHB) and Japanese (JPT) populations and separated out from the rest of the populations. As expected, YRI, a population of African descent was an outlier and closest to OG-W-IP from India. CEU (Europeans) was most proximal to the IE-LPs, the majority of which were from north India. The Austro-Asiatic
Southeast Asia was the major geographic source of East and Dravidian language speaking populations, predominantly from the tribal belt and inhabiting the central and southern regions of India were distinct from HapMap populations implicating these as our indigenous population groups. This indicates that populations included as Asian (CHB and JPT) and CEU in HapMap do not capture the entire diversity of the Indian subcontinent. Thus, it may not be wise to directly use the HapMap data to design genetic epidemiological studies for entire population of India.

**Implications**

This project has tremendous implications in future study designs and even in future policy decisions involving health and disease management of the Indian population. The impact of this work is already evident from more than 30 research publications citing it in two years. Also, the availability of the data amongst consortium participants has boosted the quality of individual research projects immensely where an investigator can check the distribution of the SNP marker (allele frequency) of his/her interest across the country without doing a single experiment or collecting a single additional sample. Numerous publications have come out where the consortium data has been used to reinforce a finding using a patient-control cohort. Below some of the key findings using this resource has been discussed.

*India in Asian perspective:* Members of the IGVC consortium have been involved in a similar study as a part of a PAN-ASIAN consortium involving the genetic mapping of Asia. This Pan-Asian SNP Consortium of the Human Genome Organization (HUGO) recently analyzed human genetic diversity in 73 Southeast Asian and East Asian populations. Besides IGIB scientists, the consortium of 90 scientists has participants from Singapore, India, China, Japan, Thailand, Philippines, South Korea, Taiwan, Japan and Indonesia. Here the goal was to develop a better understanding of the Asian population structure and the way it correlated to geography, language and demographic history. This genetic mapping of Asia was also aimed at providing help in understanding migratory patterns in human history and the study of human genetics and diseases. The study showed that genetic ancestry was highly correlated with ethnic and linguistic groups. The study also suggested that there was one major inflow of human migration into Asia arising from Southeast Asia, rather than multiple inflows from both southern and northern routes as proposed before. This indicates that Southeast Asia was the major geographic source of East Asian and North Asian populations. The most recent common ancestors of Asians out of Africa arrived first in India. From here the population moved to South East Asia and East Asia which show common genetic origin. This work together with the IGVC study shows that India represents a microcosm of Asia’s genetic diversity. The genetic make-up of the population of India encompasses the diversity of Asia and so is the perfect setting for the study of disease-genetics, pharmaco-genomics and clinical trials for the world.

*Genomic variations, Ayurveda and Personalized medicine:* The 3500 years old texts of Ayurveda describing the principles of analyzing individual constitution types for health and disease are perhaps Man’s first stint with personalized medicine. IGIB has been trying to find correlations between principles of Ayurveda and genomic variations with the goal of linking traditional medicine with modern genomics for a better understanding of what personalized medicine can offer. Their research have shown that classifying people according to their constitution types can bring more genomic homogeneity thereby reducing the false positive discoveries using modern genomic tools. They have also shown that Ayurveda based method of phenotypic classification of extreme constitutional types can uncover genes that may contribute to macro level differences in normal individuals which could lead to differential disease predisposition. They have shown that ‘homogenous’ groups of populations identified by the IGVC study can still have large genomic heterogeneity owing to their different constitution types. This implies that even if individuals from the same population background are used to detect genetic risk alleles, the results might be inconclusive unless one looks for further sub-phenotyping of the individuals under study. Thus the integration of the IGVC data with the concepts of Ayurveda highlights the importance of deep phenotyping and also shows evaluation of individual constitution types as an effective tool towards that goal (more details can be found in a another article in the same issue).

*Common diseases and IGVC study:* As discussed multiple times in this article, the main goal of the IGVC study was to generate a basal SNP map of India in relation with the prevalence/incidence and risk of various diseases vis-à-vis its unique population resources. As expected, several investigators who had initially contributed to the creation of the IGVC resource have also utilized the data for a better understanding of the diseases they study from a population perspective. The prevalence of malaria has
been mapped along with the SNP frequency in various candidate genes which has provided novel insights of population specific signatures of natural selection predisposing and/or protecting from malaria\textsuperscript{13-15}. The IGVC resource has been used for better success of carrier detection for Hemophilia A\textsuperscript{16} and for genotype to phenotype correlation in Hemophilia B\textsuperscript{17} across India. This data has also been used in metabolic diseases where genetic variations have been linked to homocysteine levels and with vegetarian or non-vegetarian diets\textsuperscript{18}. In neurodegenerative diseases, IICB, Kolkata has been using the IGVC resource and made significant contributions for Glaucoma\textsuperscript{19}, Parkinson’s disease\textsuperscript{20}, Wilson disease\textsuperscript{21} and Albinism\textsuperscript{22}. They have documented population specific natural selection of cSNPs in the \textit{CYP1B1} gene in relation to risk of glaucoma for the entire country\textsuperscript{19} and reported usefulness of SNP markers in family based studies of Parkinsonism\textsuperscript{20}.

As evident from the title of this article, the study was heavily focused on taking a step ahead in the field of personalized genomic medicine or pharmacogenomics. In one of the first reports of using the IGVC data in pharmacogenomics, IGIB has reported the analysis of 19 functional variants in 12 genes with known roles in drug metabolizing across India as well as in epilepsy patients and have shown significant interethnic differences of allele frequencies indicating population specific responses for drugs\textsuperscript{23}. More studies of this nature are underway and will be continued in future. For personalized medicine, one of the major concerns is population heterogeneity in a clinical setting. This is especially relevant in populations of countries like Africa, India where there are extremes of diversity in cohabiting populations due mainly to language and ethnicity coupled with endogamous nature of majority of the populations. In IGVC, a functional polymorphism in the β-2 adrenergic receptor which confers differential response to the bronchodilator, salbutamol has been linked to population level heterogeneity with respect to genotypes for poor or good responses in relation to asthma\textsuperscript{24,25}. This highlights the requirement of genetic map of functional polymorphisms of drug response genes in diverse populations. This would also enable identification of prospective cohorts for validation studies. The IGVC study shows that Indian subcontinent is a suitable place to develop pharmacogenomic clinical trials of drugs not only for major world populations but also other genetic isolates. Such genomic information based analysis will allow several cheap drugs to be found suitable for a cross section of the population without any side effects.

The entire data generated under this consortia effort is also publicly available for further analysis on this data by the global research community. The interested reader can visit the websites, www.igvdb.res.in and http://igvbrowser.igib.res.in to access the data. The latter site houses the entire dataset including the global data for comparison\textsuperscript{26}.

\textbf{The Road Ahead}

The field of scientific research (or any research) has always generated more questions than answers and that is the inherent property of research in any field. The field of human disease genomics is no exception. In the last decade with the launch of the Human Genome Project, we thought we will solve many problems, but it only made the problems more complex. This followed by worldwide polymorphism mapping initiatives (e.g. HAPMAP) with the hope that mapping all the variations in the genome will lead us to understand the molecular basis of common complex diseases, but this again led us to further questions. With the recent addition of personal genome sequences in the list, we are already realizing that also will not address the issues. But at the same time, all these efforts have opened up completely new possibilities and dimensions of the genome based on which we are able to take the next step. Similarly, the IGVC data also had prepared us to peep into the enormous wealth of genomic signatures that India can offer and we have to now take it in our stride with the confidence that we are prepared today to ask our own genomic questions, to make our own personalized medicines that will solve our very own disease burdens. Thus as India is on its way from infancy to a self-dependent adult in almost every aspect, the IGVC project has made it possible in the field of Genomics and personalized medicine.

\textbf{Acknowledgements}

As this is a consortia activity, first of all I would like to acknowledge all the people that were associated with the project. This involves more than 150 active researchers around the country including scientists and students. The support from the Anthropological Survey of India is duly acknowledged. The service provided by the non-scientific support staff at every centre to run the administration and logistics of this large project is also acknowledged. Finally, the participation of all the individuals who happily agreed to donate their samples for this cause across the country is highly appreciated and acknowledged. Without that this project would not succeed.
References


