Hum Genet (2005)
DOI 10.1007/s00439-005-0009-9

REVIEW ARTICLE

The Indian Genome Variation Consortium

The Indian Genome Variation database (IGVdb): a project overview

Received: 28 February 2005 / Accepted: 26 May 2005

Abstract Indian population, comprising of more than a billion people, consists of 4693 communities with several thousands of endogamous groups, 325 functioning languages and 25 scripts. To address the questions related to ethnic diversity, migrations, founder populations, predisposition to complex disorders or pharmacogenomics, one needs to understand the diversity and relatedness at the genetic level in such a diverse population. In this backdrop, six constituent laboratories of the Council of Scientific and Industrial Research (CSIR), with funding from the Government of India, initiated a network program on predictive medicine using repeats and single nucleotide polymorphisms. The Indian Genome Variation (IGV) consortium aims to provide data on validated SNPs and repeats, both novel and reported, along with gene duplications, in over a thousand genes, in 15,000 individuals drawn from Indian subpopulations. These genes have been selected on the basis of their relevance as functional and positional candidates in many common diseases including genes relevant to pharmacogenomics. This is the first large-scale comprehensive study of the structure of the Indian population with wide-reaching implications. A comprehensive platform for Indian Genome Variation (IGV) data management, analysis and creation of IGVdb portal has also been developed. The samples are being collected following ethical guidelines of Indian Council of Medical Research (ICMR) and Department of Biotechnology (DBT), India. This paper reveals the structure of the IGV project highlighting its various aspects like genesis, objectives, strategies for selection of genes, identification of the Indian subpopulations, collection of samples and discovery and validation of genetic markers, data analysis and monitoring as well as the project's data release policy.

Keywords Indian population · Ethnicity · Genetic structure · Single nucleotide polymorphism · Repeat polymorphism · Indian genome variation database

Introduction

India has served as a major corridor for the dispersal of human beings that started from Africa about 100,000 years ago (Cann 2001). Though the date of entry of modern humans into India remains uncertain, evidence from archaeological studies suggest that by the middle paleolithic period [50,000–20,000 years before present (ybp)], humans appear to have spread to many parts of India (Misra 1992, 2001). Molecular genetic evidence also supports the fact that a major population expansion of modern humans took place within India (Majumder et al. 1999). The time of this demographic expansion has been speculated to be about 60,000–85,000 ybp (Mountain et al. 1995). Contemporarily, there is an extensive social, cultural, linguistic and biological diversity in the Indian population, nurtured to a large extent by the varied topography of the country (Gadgil and Guha 1992). The vast majority of the people of India (~80%) belong to the Hindu religious fold. Hindus are hierarchically arranged into four socio-cultural clusters of groups (castes) and there are set rules governing marriage within the Hindu religious fold. About 8% of the population is constituted by tribals, who are ancestor worshippers and are largely endogamous. The remaining belongs to other religious groups, including Muslims, Christians, Buddhists, Jews, etc. Primarily, marriages occur within the religious groups. In addition, language and geographical location of habitat serve as barriers to free gene flow. These factors

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00439-005-0009-9

The Indian Genome Variation Consortium
Functional Genomics Unit,
Institute of Genomics and Integrative Biology (CSIR),
Mall Road, Delhi, 110 007, India
E-mail: skb@igib.res.in
Tel.: +91-11-27667806
Fax: +91-11-27667471

have resulted in the formation of a several thousand endogamous groups in India. Indian population, comprising of more than a billion people, consists of 4,693 communities with several thousands of endogamous groups, 325 functioning languages and 25 scripts (Singh 2002).

In some geographical regions of India, inbreeding is practised. The population-inbreeding coefficient in India varies from 0.00 to 0.20 (Rao 1984; Malhotra and Vasulu 1993; Bittles and Neel 1994). Besides, different waves of migration have led to admixture of different ethnic groups, cultures and languages, with the native population, thereby contributing significantly to the present day gene pool of the subcontinent (Sankalia 1974; Allchin and Allchin 1982; Mishra 1992; Bhasin et al. 1994; Gadgil et al. 1998). With the exception of Africa, such an extent of genetic diversity is not observed in comparable global regions (Majumder 1998). Indian population can be, to a large extent, substructured on the basis of their ethnic origin as well as linguistic lineages. All the four major morphological types—Caucasoid, Mongoloid, Australoid and Negrito are present in the Indian population (Malhotra 1978). The ‘‘Caucasoid’’ and ‘‘Mongoloid’’ populations are mainly concentrated in the north and northeastern parts of the country. The ‘‘Australoids’’ are mostly confined to the central, western and southern India, while the ‘‘Negritos’’ are restricted only to the Andaman Islands (Cavalli-Sforza et al. 1994) (Fig. 1). Linguistically, Indian populations belong to four major language families: Indo-European, Dravidian, Tibeto-Burman and Austro-Asiatic. The Indo-European and Dravidian languages are spoken in the northern and southern parts of the subcontinent, respectively (Gadgil et al. 1998). The Tibeto-Burman speakers are supposedly immigrants to India from Tibet and Burma (now, Myanmar) and are concentrated in the northeastern parts of the country (Guha 1935). The Austro-Asiatic speakers are exclusively tribals and are dispersed mostly in the central and eastern parts of the country (Fig. 2). Molecular diversity studies have revealed that the Austro-Asiatic speakers are likely to have been the most ancient inhabitants of India (Majumder 2001; Roychoudhury et al. 2001).

In common complex diseases, gene mapping is often difficult due to sampling from genetically heterogeneous populations. This is circumvented in isolated populations wherein due to genetic as well as environmental homogeneity, there are fewer variants of the disease as well as the extent of linkage disequilibrium is generally larger than outbred populations. Studies in the Indian

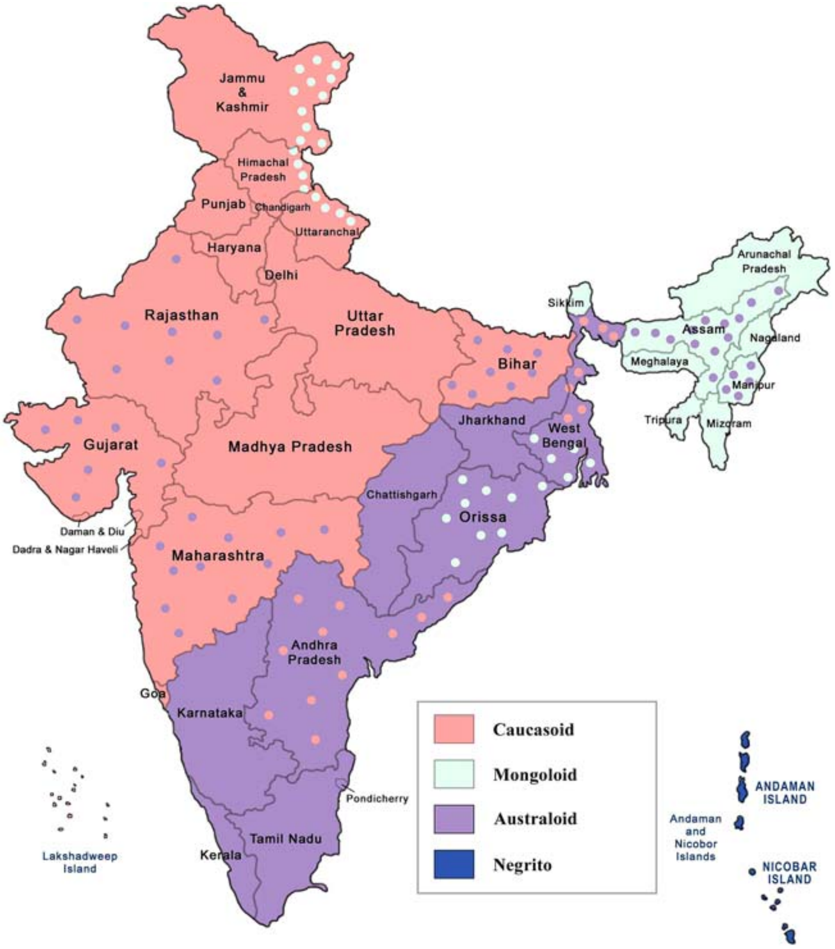


Fig. 1 *Map of India* showing the distribution of the four major morphological types, namely, Caucasoid, Mongoloid, Australoid and Negrito. Each morphological type is represented by a different *colour*, ‘‘Caucasoid’’ by *pink*, ‘‘Mongoloid’’ by *green*, ‘‘Australoid’’ by *mauve* and ‘‘Negrito’’ by *blue*. Admixture between the ethnic groups is shown by *dots*

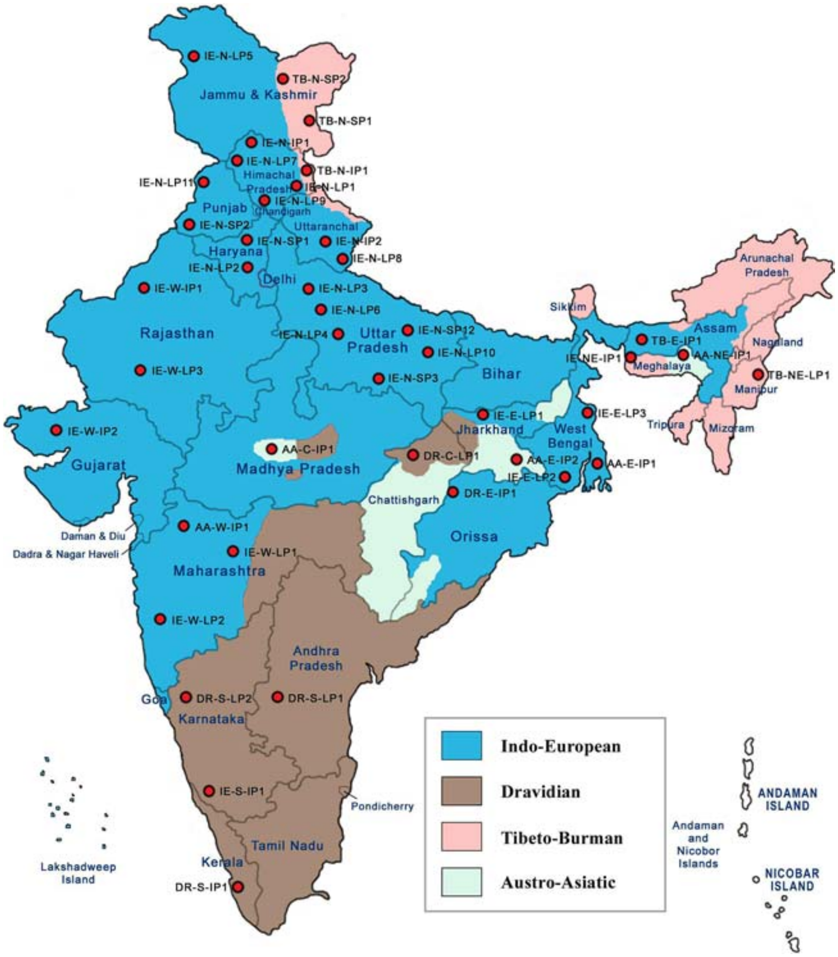


Fig. 2 *Map of India* showing the broad distribution of the four major language families, namely, Indo-European, Dravidian, Tibeto-Burman and Austro-Asiatic and populations selected for the first phase of validation in the IGV project. Each of the four linguistic lineages is represented by a different *colour*, Indo-European by *blue*, Dravidian by *brown*, Tibeto-Burman by *pink* and Austro-Asiatic by *green*. Admixture between the different linguistic families is not shown. Each selected population is depicted as a *red circle* along with a particular code revealing the linguistic affinity of the population, the geographic zone to which the population belongs as well as the type of population, viz, large endogamous population or isolated population. *IE*,*DR*, *TB* and *AA* represent the Indo-European, Dravidian, Tibeto-Burman and Austro-Asiatic language families, respectively; *N*, *E*, *W*, *C*, *S* and *NE* represent the *north*, *east*, *west*, *central*, *south* and *north-east* zones of the country, respectively; *LP*, *IP* and *SP* represent large endogamous population, isolated population and special population, respectively

subpopulations provide an enormous opportunity for complex disease gene mapping. For instance, it has recently been found (Majumder PP, unpublished results) that heterozygosity values at many genetic loci in several Indian ethnic groups are lower than those of the Icelandic population that is considered genetically isolated and homogeneous. However, between subpopulations, the extent of haplotype diversity is highly variable. Therefore, genomic diversity of the Indian ethnic groups coupled with an underlying genomic unity also provides an opportunity to replicate findings in groups of genetically similar populations (Sengupta et al. 2004).

The Indian subpopulations' wide diversity shows extensive variations in frequencies of alleles and haplotypes of common polymorphisms (Bhattacharyya et al. 1999; Roy et al. 2001, 2003; Chandak et al. 2002; Nagarkatti and Ghosh 2002; Nagarkatti et al. 2004a, 2004b; Pasha et al. 2002; Batra et al. 2005) as well as occurrence of some private polymorphisms and mutations (Pramanik et al. 2000; Mahajan et al. 2004, Mukherjee et al. 2004, Sengupta et al. 2004). Several studies on Indian subpopulations have revealed founders for disease mutations and helped in understanding the origins of monogenic diseases in India (Saleem et al. 2000, 2002, 2003; Mittal et al. 2005; Bahl et al. 2005). These findings suggest the significance of population stratification in designing case-control association studies. Case-control studies carried out in the Indian population sometimes reveal observations distinct from other world populations (Ramana et al. 2000; Kukreti et al. 2002; Nagarkatti et al. 2002; Thangaraj et al. 2002; Sengupta et al. 2004; Batra et al. 2005; Rajput et al. 2005). Using SNPs and repeat polymorphisms, both novel and reported, studies have been successful in identifying at-risk markers or haplotypes for particular diseases (Choudhry et al. 2001; Ahsan et al. 2004; Nagarkatti et al. 2004a, 2004b; Sharma and Ghosh 2004; Sharma et al. 2004; Verma et al. 2004; Verma et al. (in press)).

Though comparative molecular studies using polymorphisms in mitochondria as well as Y chromosome have revealed a fundamental genomic unity in the Indian population, there is a considerable extent of diversity at the genetic level between subpopulations (Bamshad et al. 1998, 2001; Ramana et al. 2001; Roychoudhury et al. 2001; Basu et al. 2003; Kivisild et al. 2003). To allow accurate estimation of ancestral sizes and bottleneck times, it therefore, becomes imperative to determine the extent of genetic variability and affinity within and between the Indian subpopulations as well as their affinity to other worldwide populations. Thus, the analysis of sequence variation through single nucleotide as well as repeat polymorphisms in the context of genomic diversity would help in identifying genetic substructures, which in turn would be invaluable for identification of markers for linkage or association studies, detecting candidate susceptibility regions for many common and complex diseases as well as drug-response studies.

In this purview, SNPs, being the markers of choice, owing to their high density in the human genome, are being used extensively for linkage disequilibrium studies towards the identification of candidate genes and disease predisposition markers. SNP databases exist containing polymorphism information on worldwide populations, e.g., dbSNP (Sherry et al. 2001), HGVBase (Fredman et al. 2004), HapMap (The International HapMap Consortium 2003), Celera (Kerlavage et al. 2002) etc. and specific populations, e.g., JSNP (Hirakawa et al. 2002). However, in the major target populations for which SNP databases exist in public domain, the Indian subcontinent is not represented. Since studies relevant from the point of view of disease, evolution and demography have already been initiated in the Indian population by different research groups, as is evident from the few representative examples discussed above, it gives us enormous scope to track the uniqueness of the Indian population by constructing a database, which would provide information on the variability in the Indian population, which is one-sixth of the world population. Thus, the IGV project aims at developing informative markers for predictive medicine using both repeats and single nucleotide polymorphisms within genes in Indian subpopulations.

Genesis of the Indian consortium for repeats and single nucleotide polymorphisms

The IGV initiative is a network program on predictive medicine that focuses on repeats and single nucleotide polymorphisms, initiated in 2003 and tenured for 5 years, by six constituent laboratories of the Council of Scientific and Industrial Research (CSIR), with funding from the Government of India. The laboratories include Institute of Genomics and Integrative Biology (IGIB), Delhi, 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), Lucknow and Institute of Microbial technology (IMTECH), Chandigarh (Box 1). These laboratories are involved in studies related to asthma, diabetes, neuropsychiatric disorders, cancer, coronary artery disease, clot disorders, high altitude disorders, retinitis pigmentosa, predisposition to malaria as well as other infectious diseases and drug metabolism. The consortium deemed it pertinent to understand the inherent genetic variability of the subpopulations as a first step towards identifying susceptible biomarkers for any disease or understanding drug response in different subpopulations. This consortium, therefore, aims at providing information on variations in the subpopulations representing the entire country. These variations would be useful for investigators for specific candidate gene studies conducted in any part of the country.

Apart from the CSIR laboratories, a key participant in the project is the Indian Statistical Institute (ISI), Kolkata, which will help us in the analysis of our data. The institute has an established expertise in human genetic variation data analysis (Tapadar et al. 2000; Basu et al. 2005). The project also involves active participation of the Anthropological Survey of India (Singh 2002) that has helped in the identification of the various Indian subpopulations. In addition to the institutional facility, the project also has collaborations with The Centre for Genomic Application (TCGA), established through the support of Department of Science and Technology (DST), CSIR with The Chatterjee Group (TCG) for high throughput sequencing and genotyping and SilicoGene Informatics Private Limited along with LabVantage, India for development of a comprehensive platform for IGV database management, analysis and portal development.

Objectives of the project

The project was initiated with well-defined objectives wherein it was planned to collect samples from 15,000 individuals drawn from different subpopulations, select ~1000 genes relevant to common diseases and drug response and identify at least 5–10 informative markers per gene in the Indian subpopulations, which would be useful for candidate gene studies. In addition, several SNP/repeat markers in the regions where there was no adequate gene representation, are also being mapped along with gene duplications to cover the entire genome. The aim is to estimate allele frequencies of the novel as well as reported SNPs and microsatellites, construct haplotypes and determine the extent of linkage disequilibrium within and across genes and across various Indian subpopulations. The ultimate goal is to create a DNA variation database of the people of India and make it available to researchers for understanding human biology with respect to disease predisposition, adverse drug reaction, population migration etc. To achieve these objectives, the project activities were divided into two phases. Details of these activities are described in the subsequent sections.

Selection of genes

The distinctive feature of common complex diseases is the absence of a single specific predisposition factor. Individual genetic background, which can be defined as a unique combination of thousands of common/rare variants of genes governing metabolic pathways and regulatory systems at different levels may be influenced by numerous environmental factors. Besides, different complex diseases may also share common pathways or genes. For this reason, an effort has been made by the members of the consortium to select ~1000 genes, which may include functional candidate genes, positional candidates and members of candidate pathways. Certain amount of flexibility has been maintained in the list of the genes, keeping in mind the dynamism in the field of research itself.

Identification of populations

The project aims to validate SNPs in 15,000 individuals drawn from well-defined ethnic groups that have been chosen to represent the entire spectrum of diversity within the Indian population. Considering the population diversity, two issues had to be addressed. One, defining the composition of the population substructure, which captures the entire genetic diversity and the other, the composition of a small panel of samples for SNP discovery that would ensure representation of SNPs from the entire Indian population. Taking into account the cost involved in the above endeavour, we decided to carry out the initial study in two steps; discovery of SNPs on a small panel (discovery panel) of 43 samples followed by estimation of their frequency in a larger set of samples, which constitutes the validation panel (details described below). It was felt that this would give an estimate of the genetic heterogeneity that would help us in further substructuring of the population.

Composition of the discovery panel

With a view to discover novel SNPs as well as to determine the presence/frequency of the reported SNPs in the Indian population, an initial panel was made comprising of representatives drawn from 43 different subpopulations (Supplementary figure). This discovery panel included samples, both tribal and non-tribal, belonging to diverse geographical zones and linguistic backgrounds to maximize novel SNP discovery. Though 43 individuals per se do not represent the entire Indian population, such a diverse set does increase the heterogeneity in terms of SNP discovery as compared to a set of samples from a single subpopulation.

Composition of the validation panel

The populations for validation panel have been identified based on the following criteria; geographical zones, linguistic groups, practice of endogamy, presence of minority communities from different religious groups and existence of populations of different sizes. Four major linguistic lineages, namely, Indo-European, Dravidian, Tibeto-Burman and Austro-Asiatic have been considered. We also categorized the populations as small if their size was <1 million and large if >10 million. The ultimate target of sample collection for our project is 15,000 unrelated individuals drawn from distinct Indian subpopulations so as to capture the entire genetic diversity of the population. We plan to collect 100 such subpopulations, 192 samples from large populations (LP or SP) and 96 samples from populations of smaller sizes (IP or SP). This strategy is being followed to ensure maximum coverage of the Indian population as well as to capture the minor alleles in large outbred populations. However, in the first phase of validation, we wanted to determine the extent of diversity and heterogeneity prevalent in the Indian population. For this purpose, populations were categorized based on the different geographical zones and linguistic categories and two contrasting populations in terms of their sizes were minimally selected and prioritized from each category wherever available (Table 1). We identified 42 subpopulations from which it was decided, to collect, on average 40 samples from each population in the first phase. Individual samples from 31 of these are also represented in our discovery panel. In addition, eight large, one isolated and two special populations are also included in this validation panel to maximize representation of people of India. The diversity (allele frequency) information generated from the first phase of validation will be utilized as an additional criterion for further collection of samples to attain the target of 15,000.

Sample collection

The identification of populations as well as collection of samples have been carried out with the help of trained anthropologists, social workers and community health workers, as their participation is essential for establishing rapport with the general public. Also, individuals fluent in the local language of the concerned populations are consulted and actively involved in the study in order to get maximum and authentic information from the donors and also to help them to better understand the purpose of carrying out such an investigation. Endogamy of the populations is established by taking extensive information about the marriage pattern, gathered through pedigrees and interview of family members of the donor as well as published literature (Bittles 2002). A general template to obtain informed consent from the donors of the samples is used and in cases where the donor is illiterate, thumb impression is used. In addition, verbal tape-recorded consent of the donors is also taken. It is ensured that the individuals are unrelated at least to the first cousin level and both males and females are being collected in equal numbers. Although large sample sizes would add robustness to our inferences, it is expected that even small sample sizes in highly endogamous small cohorts would be equally informative with respect to allele frequency. All the institutes are participating in the collection of samples, with three nodal

Table 1 Categorization of the Indian subpopulations on the basis of geographic zones and the four major language families for the validation panel

Zone/Linguistic category	Indo-European	Dravidian	Tibeto-Burman	Austro-Asiatic
North	IE-N-IP1, IE-N-IP2, IE-N-LP1, IE-N-LP2, IE-N-LP3, IE-N-LP4, IE-N-LP5, IE-N-LP6, IE-N-LP7, IE-N-LP8, IE-N-LP9, IE-N-LP10, IE-N-LP11, IE-N-SP1, IE-N-SP2, IE-N-SP3, IE-N-SP4		TB-N-SP1, TB-N-SP2, TB-N-IP1	
East	IE-E-LP1, IE-E-LP2, IE-E-LP3	DR-E-IP1	TB-E-IP1	AA-E-IP1, AA-E-IP2
West	IE-W-LP1, IE-W-LP2, IE-W-LP3, IE-W-IP1,IE-W-IP2			AA-W-IP1
Central		DR-C-LP1		AA-C-IP1
South	IE-S-IP1	DR-S-LP1, DR-S-LP2, DR-S-IP1		
North-east	IE-NE-IP1		TB-NE-LP1	AA-NE-IP1

IE,*DR*, *TB* and *AA* represent the Indo-European, Dravidian, Tibeto-Burman and Austro-Asiatic language families, respectively; *N*,*E*, *W*,*C*, *S* and *NE* represent the *north*, *east*, *west*, *central*, *south* and *north-east* zones of the country respectively; *LP*,*IP* and *SP* represent large endogamous population, isolated population and special population, respectively

centres, IGIB, CCMB and IICB, which are connected to the other centres. IGIB, CDRI, IMTECH and ITRC have ensured collection of samples from the northern and central parts of India, IGIB and CCMB from western part, IICB from eastern part and CCMB from the southern part of the country (Fig. 2). Each institute has obtained prior ethical clearance from the Institutional Bioethics Committee (IBC) for the collection of samples following the guidelines of Indian Council of Medical Research (ICMR) (http://icmr.nic.in/ethical.pdf) for the complete period of 5 years. A uniform bar-coded, detailed questionnaire was developed, containing information pertaining to ethnicity, family history of diseases and other phenotypic traits of the sample donor (Supplementary text). The filled forms are scanned; the scanned data is extracted using Omni Extract 4.0 software (Newgen Software Technologies Limited, India), which is uploaded on to the database after verification. Prior to sample collection, it is explained to the participants that the personal identifiers in the questionnaire are confidential and are not available to the researchers. Also, the samples are irretrievably coded. It is also explained to the volunteers that the project aims at understanding the extent of variability and diversity in different subpopulations and the basal data generated in this study would be used for disease-specific association studies. We also ensure that the participation is entirely voluntary and no materialistic promises were made to the donors. Also, no promise for a genetic test is provided.

Ten 10 ml of blood sample is collected from each individual and isolation of DNA is carried out using a standard procedure. Each laboratory has access to all the samples collected. Extreme care is taken to ensure that only high quality samples that meet the purity criteria are used for the study. All the collected samples are quantified for their DNA content by fluorometric analysis using Picogreen dye (Molecular probes, United States). The collection of 43 samples from different Indian subpopulations for the discovery phase has been completed. So far, 1,700 samples from 42 different subpopulations have been collected for the first phase of validation.

Managing ethical issues

Although the project will include no personal identifiers, each sample is identifiable through a sample code as well as a population code. There is a provision for the volunteers to withdraw from the study at their will. Though naming the population with a particular set of tag SNPs allows a better interpretation of the biological significance to be used in future studies of association, population history and population relatedness, it does, however, have important ethical and social ramifications. To avoid any social backlash that could destabilize the very fabric of Indian society, i.e. unity in diversity, a decision was taken against disclosing the identity of the populations. This is because, the way a population is labelled in this project and described in publications will have implications for all members of the population, as all of them (and all members of closely related populations) might be affected by the interpretation and use of findings of future studies. However, this study clearly spells out that individual variation within an at-risk population for a given predisposition cannot be ruled out. The samples collected from different populations are bar-coded with each population being given a specific code revealing the linguistic affinity of the population, the geographic zone to which the population belongs as well as the type of population, viz, large endogamous population, isolated population or special population (Table 1).

Strategy and methods for marker discovery and validation

Though a large number of studies have been carried out using mitochondrial or Y chromosome polymorphisms and microsatellite markers in the Indian population (Thangaraj et al. 1999; Mukherjee et al. 2001; Ramana et al. 2001; Roychoudhury et al. 2001; Roy et al. 2003), there is very little data that gives an estimate of the extent of SNPs shared both within and between Indian subpopulations and other world populations. Therefore, it was decided that in the initial phase of the project, screening for novel SNPs would be carried out in 71 genes on the discovery panel of 43 samples. For this purpose, amplicons were generated in exonic regions spanning nearly the entire gene. This would not only provide the data on the SNPs shared between the different Indian subpopulations and Indian and other world populations but also reveal population specific indigenous SNPs. In addition, it provides data on 86 chromosomes, thus enabling the identification of SNPs with an overall minimal allele frequency (MAF) >0.05 to be used for further validation.

For the discovery of novel SNPs, bi-directional sequencing of the 43 samples of the discovery panel was carried out. A few selection criteria were evolved for prioritizing the SNPs for validation based on the data on novel and putative functional SNPs as well as minor allele frequencies of the SNPs in the discovery panel. Information on the frequencies of SNPs in different databases like dbSNP, Celera, RealSNP and HapMap along with the information on haplotype block structures and tag SNPs were taken into consideration during selection of SNPs for validation. Also, though flexible, spacing between the different selected SNPs within a gene was taken care of depending upon the size of the gene so as to uniformly cover the entire gene. After going through these series of filters, additional gaps were filled, if required, by SNPs reported in the database based on different validation criteria such as multiple submissions.

Following the above criteria, in the first phase of validation, 447 SNPs are observed in 71 genes, from which a total of 276 SNPs have been selected. Based on the length, 5–8 SNPs have been selected from each gene. In the first phase of the project, these SNPs are being validated on 1,700 samples collected from different populations using the Sequenom massarray system. The validation process is being carried out in two steps—the initial confirmation of SNPs is being done in population pools followed by estimation of frequency in the individual samples. These data would give us insights for further identification of informative SNPs and substructuring of the population, which would enable a judicious collection of the 15,000 samples.

For identification of informative microsatellite markers, initially the region of interest is scanned using RepeatMasker program to identify simple repeats. Subsequently, repeats with a particular threshold (di >15, tri >10, tetra >8 repeat units) are selected based on the nature of the repeat. The polymorphism status of these repeats is determined in pools of 100 samples taken from the validation panel using fluorescent PCR-based methods followed by fragment sizing on ABI 3100. The spread of the peaks gives a coarse estimation on the polymorphic status of the microsatellite marker and subsequently the frequency of different alleles is determined through individual genotyping. Markers having heterozygosity value >0.5 would be considered for construction of LD maps.

Data analysis and monitoring

We have developed a comprehensive platform for SNP data management and analysis for High Throughput-SNP Sequencing and Screening (HT-SSS). This was developed with the aim of creating a web-based information management infrastructure, which would be capable of handling comprehensive data management requirements for SNP analysis. This includes activities ranging from sample collection/tracking, primer design/synthesis/tracking, and SNP discovery/validation to data analysis and publication of the SNPs on the portal within and across the participating institutes (Fig. 3).

The platform architecture (Fig. 4) supports distributed information management over secured networks including Secured Socket Layer (SSL) enabled internet. HT-SSS helps to minimize manual data entry by enabling quick and easy import of data to information sources such as instruments (e.g. primer synthesizers, liquid handling robotic stations, sequencers, dHPLC, MALDI-TOF, etc.) and third-party analysis applications (e.g. SAS) by generating compatible input data formats. HT-SSS also helps to curate output data generated by these instruments and analysis softwares in Oracle 9i relational database system. This capability greatly streamlines data entry and minimizes errors that may occur with manual input file generation. The platform allows elaborate plate management functionality for implementation of high throughput pooling and multiplexing of samples across diverse plate formats thereby reducing errors in sample handling and maintains uniformity in plate management as well as labelling across all the participating institutes. It implements integrated support for PCR, sequencing, dHPLC analysis, SNP discovery and validation. Further, the platform enables analysis of discovered SNP data for allele/genotype/haplotype frequency calculations, determination of heterozygosity and determination of haplotype blocks for single and multiple populations. The analysis of the data can also be visualized through different graphical user interfaces. A portal has been developed, which would house all the information generated in the course of SNP discovery and validation during this project.

Fig. 3 High Throughput—SNP Sequencing and Screening (*HT-SSS*) system overview depicting the activities performed by the HT-SSS portal

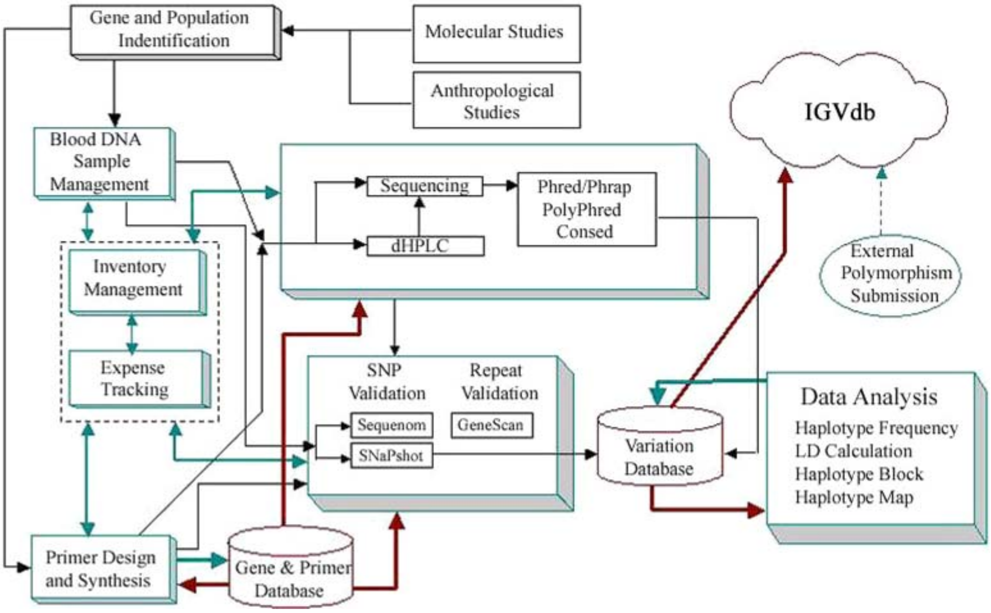
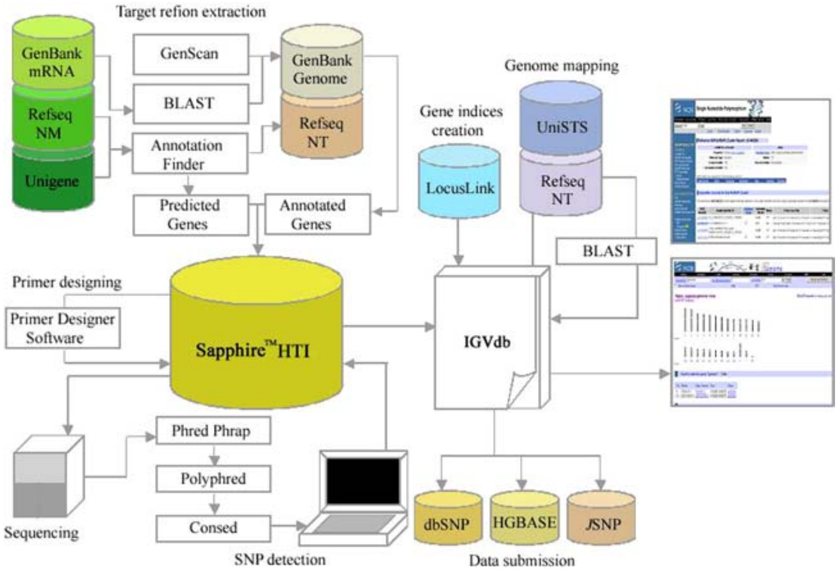


Fig. 4 HT-SSS software architecture, showing the connectivity between different databases (represented by *cylinders*) and the activities (represented by *rectangles*). It depicts the functioning of HT-SSS solution for management and publication of the data generated and submitted to the portal



Data release policy of the Indian Genome Variation project

It is envisaged that the IGV project would eventually be useful for identifying predisposed haplotypes for common and complex disorders or the common functional polymorphisms, which might be useful for pharmacogenomics studies. It would be a resource that catalogues the common patterns of genetic variation in important complex disease candidate genes. There is provision for incorporating or widening the scope of the project as more and more information on the human genome variations is being made available with additional information on patterns of linkage disequilibrium, as well asdevelopment of cost-effective high throughput technologies. Though nearly 11 million SNPs have been released in the public database and the HapMap data is going to be available soon, the selection of the appropriate set of markers for identifying susceptibility haplotypes for different complex genetic diseases is still debatable. Moreover, these databases do not include Indian samples. In the IGV consortium, there is also a provision for parallel research to determine factors, which could lead to generation of informative repeats and SNPs suitable for designing case-control association studies. These inputs can be incorporated during the development of the SNP database of the Indian population.

Usage of the portal will be freely available for all academic users around the world. However, the discoveries arising out of the IGV project will be IPR protected and will be licensed for commercial exploitation.

Acknowledgements This work was supported by grants from the Council of Scientific and Industrial Research (CSIR), Government of India for the project "Predictive medicine using repeat and single nucleotide polymorphisms" (CMM0016).

Appendix

Box 1 The Indian Genome Variation database consortium

Project conceptualization

Samir K. Brahmachari (IGIB) and Lalji Singh (CCMB)

Project planning

Samir K. Brahmachari, Abhay Sharma, Mitali Mukerji (IGIB); Kunal Ray, Susanta Roychoudhury (IICB); Lalji Singh, G. R. Chandak, K. Thangaraj (CCMB); Saman Habib (CDRI) D. Parmar (ITRC); Partha P. Majumder (ISI)

Implementation

Samir K. Brahmachari, Mitali Mukerji, Shantanu Sengupta, Dwaipayan Bharadwaj, Debasis Dash (IGIB); Kunal Ray, Susanta Roychoudhury (IICB); G. R. Chandak (CCMB); Saman Habib, Srikanta K. Rath (CDRI); D. Parmar, R. Shankar (ITRC); Jagmohan Singh (IMTECH)

Population identification

Partha P. Majumder (ISI); Mitali Mukerji, Komal Virdi, Samira Bahl (IGIB); V. R. Rao, K. Thangaraj (CCMB); Saman Habib, Srikanta Rath, Swapnil Sinha, Ashok Singh, Amit Mitra, Shrawan K. Mishra (CDRI); D. Parmar (ITRC); B. R. K. Shukla (Lucknow University)

Sample collection

Shantanu Sengupta, Dwaipayan Bharadwaj, Mitali Mukerji, Qadar Pasha, Souvik Maiti, Abhay Sharma, Samira Bahl, Komal Virdi, Amitabh Sharma, Jitender Kumar, Aarif Ahsan, Tsering Stobdan, Chitra Chauhan, Saurabh Malhotra, Ajay Vidhani, S. Siva, Aradhita Baral, Rajesh Pandey, Ravishankar Roy, Mridula Singh, S. P. Singh (IGIB); Nitin Maurya (University of Delhi); Arun Bandyopadhyay, Susanta Roychoudhury, Ganga Nath Jha, Somnath Dutta, Gautam Ghosh, Tufan Naiya (IICB); K. Thangaraj, G. R. Chandak, Manoj Jain (CCMB); Saman Habib, Srikanta Rath, Swapnil Sinha, Ashok Singh, Amit Mitra, Shrawan K. Mishra, J. P. Srivatava, J. R. Gupta (CDRI); Vinay Khanna, Alok Dhawan, Mohini Anand, R. Shankar, R. S. Bharti, Madhu Singh, Arvind P. Singh, Anwar J. Khan, D. Parmar (ITRC); Kamlesh Kumar Bisht, Ashok Kumar (IMTECH)

SNP discovery

Shantanu Sengupta, Dwaipayan Bharadwaj, Mitali Mukerji, Qadar Pasha, Balaram Ghosh, Abhay Sharma, Swapan Kumar Das, Taruna Madan, Chitra Chauhan, Ranjana Verma, Uma Mittal, Samira Bahl, Amitabh Sharma, Jitender Kumar, Anubha Mahajan, Sreenivas Chavali, Rubina Tabassum, Vijaya Banerjee, Jyotsna Batra, Rana Nagarkatti, Shilpy Sharma, Mamta Sharma, Rajshekhar Chatterjee, Jinny A. Paul, Pragya Srivastava, Rupali Chopra, Aradhita Baral, Ankur Saxena, Charu Rajput, Prashant Kumar Singh, Aarif Ahsan, Tsering Stobdan, Mudit Vaid (IGIB); Kunal Ray, Susanta Roychoudhury, Sumantra Das, Keya Chaudhuri, Rukhsana Chowdhury, Arun Bandyopadhyay, Arijit Mukhopadhyay, Moulinath Acharya, Ashima Bhattacharyya, Atreyee Saha, Arindam Biswas, Moumita Chaki, Arnab Gupta, Saibal Mukherjee, Suddhasil Mookherjee, Ishita Chattopadhyay, Taraswi Banerjee, Meenakshi Chakravorty, Chaitali Misra, Gourish Monadal, Shiladitya Sengupta, Ishani Deb, Arunava Banerjee, Rajdeep Chowdhury, Amalendu Ghosh, Kalidas Paul, Priyanka De, Sumita Mishra (IICB); G. R. Chandak, K. Thangaraj, Rachna Shukla, G. S. Ramalakshmi, Pankaj Khanna, M. Mohd. Idris, K. Radha Mani, Seema Bhaskar, Swapna Mahurkar, Shalini Mani Tripathi, V. N. S. Prathyusha, V. Prasad Kolla, J. Hemavathi, Nikita Thakur (CCMB); Swapnil Sinha, Ashok Singh, Amit Mitra, Shrawan K. Mishra (CDRI); Madhu Singh, Arvind P. Singh, Anwar J. Khan, R. Shankar, Deepa Agarwal, D. Parmar (ITRC); Jagmohan Singh, Rupinder Kaur, Kamlesh Kumar Bisht, Ashok Kumar (IMTECH)

Repeat discovery and analysis

Mitali Mukerji, Komal Virdi, Uma Mittal, Aradhita Baral, Rajesh Pandey (IGIB); Kunal Ray (IICB)

HTSSS implementation

Samir K. Brahmachari, Mitali Mukerji, Debasis Dash, Manoj Hariharan, Shantanu Sengupta, Siddharth Singh Bisht, Dipayan Dasgupta, Mridula Singh, Sangeeta Khanna (IGIB); Keya Chaudhuri, Susanta Roychoudhury, Kunal Ray (IICB); Saman Habib (CDRI); Ranjan Basu, Biswajit Das, Shuvankar Mukherjee, Jhuma Mukherjee, Debasish Saha (Silicogene); Pallab Banerjee, Bijoyesh Saha, Anirban Chatterjee, S. R. Moquim, Navneet Kwarta, Manish Kumar, Debkumar Sinha (Labvantage Asia)

Data analysis, monitoring and database development

Partha P. Majumder (ISI); Mitali Mukerji, Swapan Kumar Das, Chitra Chauhan, Samira Bahl, Komal Virdi, Uma Mittal, Ranjana Verma, Debasis Dash, Manoj Hariharan, Mridula Singh, Rajesh Pandey, Shantanu Sengupta, Dwaipayan Bharadwaj, Balaram Ghosh, Qadar Pasha, Taruna Madan, Samir K. Brahmachari (IGIB); Kunal Ray, Susanta Roychoudhury, Sumantra Das, Keya Chaudhuri, Rukhsana Chowdhury, Arun Bandyopadhyay (IICB); Shrish Tewari, G. R. Chandak, Lalji Singh (CCMB); Swapnil Sinha, Ashok Singh, Amit Mitra, Shrawan K. Mishra (CDRI); Madhu Singh, Arvind P. Singh, Anwar J. Khan, R. Shankar, A. Dhawan, V. K. Khanna, D. Parmar (ITRC); Jagmohan Singh, Balvinder Singh, G. P. S. Raghava (IMTECH)

Genotyping, sequencing, primer synthesis management

Mitali Mukerji, Shantanu Sengupta, Neelam Makhija, Abdur Rahim (IGIB); K. Narayanasami, Arindam Maitra, Sangeeta Sharma, Ruchi Chawla, Suruchika Soni, Preeti Khurana, Sushanta Das Sutar, Amit Tuteja, Mohd. Nadeem Khan, Abhishek Chandragupta, Pooja Rana, M. Chidambaram (TCGA); Kunal Ray, Susanta Roychoudhury, Sumantra Das, Keya Chaudhuri, Rukhsana Chowdhury, Arun Bandyopadhyay, Arijit Mukhopadhyay, Moulinath Acharya, Ashima Bhattacharyya, Atreyee Saha, Arindam Biswas, Moumita Chaki, Arnab Gupta, Saibal Mukherjee, Ishita Chattopadhyay, Taraswi Banerjee, Meenakshi Chakravorty, Chaitali Misra, Gourish Monadal, Shiladitya Sengupta, Ishani Deb, Arunava Banerjee, Rajdeep Chowdhury, Amalendu Ghosh, Kalidas Paul, Priyanka De, Sumita Mishra (IICB); K. Thangaraj, G. R. Chandak (CCMB); Swapnil Sinha, Ashok Singh, Amit Mitra, Shrawan K. Mishra, Bipin C. Mishra (CDRI); Madhu Singh, Arvind P. Singh, Anwar J. Khan, Deepa Agarwal, R. Shankar, D. Parmar (ITRC); Jagmohan Singh (IMTECH)

Project management

Hemant Kulkarni (IGIB), O. P. Aggarwal (CSIR), Samir K. Brahmachari (IGIB)

References

Ahsan A, Charu R, Pasha MAQ, Norboo T, Afrin F, Baig MA (2004) eNOS allelic variants at the same locus associate with HAPE and adaptation. Thorax 59:1000–1002

Allchin B, Allchin R (1982) The rise of civilization in India and Pakistan. Cambridge University Press, Cambridge

Bahl S, Virdi K, Mittal U, Sachdeva MP, Kalla AK, Holmes SE, O'Hearn E, Margolis RL, Jain S, Srivastava AK, Mukerji M (2005) Evidence of a common founder for SCA12 in the Indian population. Ann Hum Genet (in press)

Bamshad MJ, Watkins WS, Dixon ME, Jorde LB, Rao BB, Naidu JM, Prasad BV, Rasanayagam A, Hammer MF (1998) Female gene flow stratifies Hindu castes. Nature 395:651–652

Bamshad M, Kivisild T, Watkins WS, Dixon ME, Ricker CE, Rao BB, Naidu JM, Prasad BV, Reddy PG, Rasanayagam A, Papiha SS, Villems R, Redd AJ, Hammer MF, Nguyen SV, Carroll ML, Batzer MA, Jorde LB (2001) Genetic evidence on the origins of caste populations. Genome Res 11:994–10004

Basu A, Mukherjee N, Roy S, Sengupta S, Banerjee S, Chakraborty M, Dey B, Roy M, Roy B, Bhattacharyya NP, Roychoudhury S, Majumder PP (2003) Ethnic India: a genomic view, with special reference to peopling and structure. Genome Res 13:2277–2290

Basu A, Chaudhuri P, Majumder PP (2005) Identification of polymorphic motifs using probabilistic search algorithms. Genome Res 15:67–77

Batra J, Niphadkar PV, Sharma SK, Ghosh B (2005) Uteroglobin-related protein 1(UGRP1) gene polymorphisms and atopic asthma in the Indian population. Int Arch Allergy Immunol 136:1–6. Epub 2004 December 08

Batra J, Sharma M, Chatterjee R, Sharma S, Mabalirajan U, Ghosh B (2005) Chemokine receptor 5 (CCR5) D32 deletion and atopic asthma in India. Thorax 60:85

Bhasin MK, Walter H, Danker-Hopfe (1994) People of India: an investigation of biological variability in ecological ethnoeconomic and linguistic groups. Kamla-Raj Enterprises, Delhi

Bhattacharyya NP, Basu P, Das M, Pramanik S, Banerjee R, Roy B, Roychoudhury S, Majumder PP (1999) Negligible male gene flow across ethnic boundaries in India, revealed by analysis of Y-chromosomal DNA polymorphisms. Genome Res 9:711–719

Bittles AH (2002) Endogamy, consanguinity and community genetics. J Genet 81:91–98

Bittles AH, Neel JV (1994) The costs of human inbreeding and their implications for variations at the DNA level. Nat Genet 8:117–121

Cann RL (2001) Genetic clues to dispersal in human populations: retracing the past from the present. Science 291:1742–1748

Cavalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton University Press, Princeton, NJ, pp 208–213

Chandak GR, Sridevi MU, Vas CJ, Panikker DM, Singh L (2002) Apolipoprotein E and presenilin-1 allelic variation and Alzheimer's disease in India. Hum Biol 74:683–693

Choudhry S, Mukerji M, Srivastava AK, Jain S, Brahmachari SK (2001) CAG repeat instability at SCA2 locus: anchoring CAA interruptions and linked single nucleotide polymorphisms. Hum Mol Genet 10:2437–2446

Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ (2004) HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32:D516–D519

Gadgil M, Guha R (1992) The fissure land: an ecological history of India. Oxford University Press, New Delhi

Gadgil M, Joshi NV, Prasad UV, Manoharan S, Patil S (eds) (1998) In the Indian human heritage. Universities Press, Hyderabad, pp 100–129

Guha BS (1935) The racial affinities of the people of India; in census of India, 1931, Part III—Ethnographical. Government of India Press, Simla

Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T, Nakamura Y (2002) JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30:158–162

Kerlavage A, Bonazzi V, di Tommaso M, Lawrence C, Li P, Mayberry F, Mural R, Nodell M, Yandell M, Zhang J, Thomas P (2002) The celera discovery system. Nucleic Acids Res 30:129–136

Kivisild T, Rootsi S, Metspalu M, Mastana S, Kaldma K, Parik J, Metspalu E, Adojaan M, Tolk HV, Stepanov V, Golge M, Usanga E, Papiha SS, Cinnioglu C, King R, Cavalli-Sforza L, Underhill PA, Villems R (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. Am J Hum Genet 72:313–332

Kukreti R, C BR, Das SK, De M, Talukder G, Vaz F, Verma IC, Brahmachari SK (2002) Study of the single nucleotide polymorphism (SNP) at the palindromic sequence of hypersensitive site (HS)4 of the human beta-globin locus control region (LCR) in Indian population. Am J Hematol 69:77–79

Mahajan A, Chavali S, Kabra M, Chowdhury MR, Bharadwaj D (2004) Molecular characterization of hemophilia B in North Indian families: identification of novel and recurrent molecular events in the factor IX gene. Haematologica 89:1498–1503

Majumder PP (1998) People of India: biological diversity and affinities. Evol Anthrop 6:100–110

Majumder PP (2001) Ethnic populations of India as seen from an evolutionary perspective. J Biosci 26:533–545

Majumder PP, Roy B, Banerjee S, Chakraborty M, Dey B, Mukherjee N, Roy M, Thakurta PG, Sil SK (1999) Human-specific insertion/deletion polymorphisms in Indian populations and their possible evolutionary implications. Eur J Hum Genet 7:435–446

Malhotra KC (1978) Morphological composition of the people of India. J Hum Evol 7:45–63

Malhotra KC, Vasulu TS (1993) Structure of human populations in India. In: Majumdar PP (ed) Human population genetics: a centennial tribute to JBS Haldane. Plenum Press, New York, pp 207–233

Mishra S (1992) The age of the Acheulian in India: new evidence. Curr Anthropol 33:325

Misra VN (1992) Stone age in India: an ecological perspective. Man Env 14:17–64

Misra VN (2001) Prehistoric human colonization of India. J Biosci 26:491–531

Mittal U, Srivastava AK, Jain S, Jain S, Mukerji M (2005) Founder haplotype for Machado-Joseph disease in the Indian population: novel insights from history and polymorphism studies. Arch Neurol 62:637–640

Mountain JL, Hebert JM, Bhattacharyya S, Underhill PA, Ottolenghi C, Gadgil M, Cavalli-Sforza LL (1995) Demographic history of India and mtDNA-sequence diversity. Am J Hum Genet 56:979–992

Mukherjee N, Nebel A, Oppenheim A, Majumder PP (2001) High-resolution analysis of Y-chromosomal polymorphisms reveals signatures of population movements from Central Asia and West Asia into India. J Genet 80:125–135

Mukherjee S, Mukhopadhyay A, Banerjee D, Chandak GR, Ray K (2004) Molecular pathology of haemophilia B: identification of five novel mutations including a LINE 1 insertion in Indian patients. Haemophilia 10:259–263

Nagarkatti R, Ghosh B (2002) Identification of single-nucleotide and repeat polymorphisms in two candidate genes, interleukin 4 receptor (IL4RA) and signal transducer and activator of transcription protein 6 (STAT6), for Th2 mediated diseases. J Hum Genet 47(12):684–687

Nagarkatti R, Rao C-B, Rishi JP, Chetiwal R, Shandilya V, Vijayan V, Kumar R, Pemde HK, Sharma SK, Sharma S, Singh AB, Gangal SV, Ghosh B (2002) Association of IFNG gene polymorphism with asthma in the Indian population. J Allergy Clin Immunol 110:410–412

Nagarkatti R, C BR, Vijayan V, Sharma SK, Ghosh B (2004a) Signal transducer and activator of transcription 6 haplotypes and asthma in the Indian population. Am J Respir Cell Mol Biol 31:317–321. Epub 2004 April 22

Nagarkatti R, Kumar R, Sharma SK, Ghosh B (2004b) Association of IL4 gene polymorphisms with asthma in North Indians. Int Arch Allergy Immunol 134:206–212. Epub 2004 June 01

Pasha MAQ, Khan AP, Kumar R, Ram RB, Grover SK, Srivastava KK, Selvamurthy W, Brahmachari SK (2002) Variations in angiotensin-converting enzyme gene insertion/deletion polymorphism in Indian populations of different ethnic origins. J Biosci 27:67–70

Pramanik S, Basu P, Gangopadhaya PK, Sinha KK, Jha DK, Sinha S, Das SK, Maity BK, Mukherjee SC, Roychoudhuri S, Majumder PP, Bhattacharyya NP (2000) Analysis of CAG and CCG repeats in Huntingtin gene among HD patients and normal populations of India. Eur J Hum Genet 8:678–682

Rajput C, Makhijani K, Norboo T, Afrin F, Sharma M, Pasha ST, Pasha MAQ (2005) *CYP11B2* gene polymorphisms and hypertension in highlanders accustomed to high salt intake. J Hypertens 23:79–86

Ramana GV, Chandak GR, Singh L (2000) Sickle cell gene haplotypes in Relli and Thurpu Kapu populations of Andhra Pradesh. Hum Biol 72:535–540

Ramana GV, Su B, Jin L, Singh L, Wang N, Underhill P, Chakraborty R (2001) Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. Eur J Hum Genet 9:695–700

Rao PSS (1984) Inbreeding in India: concepts and consequences. In: Lukacs JR (ed) The people of South Asia. Plenum Press, New York, pp 239–268

Roy B, Majumder PP, Dey B, Chakraborty M, Banerjee S, Roy M, Mukherjee N, Sil SK (2001) Ethnic differences in distributions of GSTM1 and GSTT1 homozygous “null” genotypes in India. Hum Biol 73:443–450

Roy S, Thakur Mahadik C, Majumder PP (2003) Mitochondrial DNA variation in ranked caste groups of Maharashtra (India) and its implication on genetic relationships and origins. Ann Hum Biol 30:443–454

Roychoudhury S, Roy S, Basu A, Banerjee R, Vishwanathan H, Usha Rani MV, Sil SK, Mitra M, Majumder PP (2001) Genomic structures and population histories of linguistically distinct tribal groups of India. Hum Genet 109:339–350

Saleem Q, Choudhry S, Mukerji M, Bashyam L, Padma MV, Chakravarthy A, Maheshwari MC, Jain S, Brahmachari SK (2000) Molecular analysis of autosomal dominant hereditary ataxias in the Indian population: high frequency of SCA2 and evidence for a common founder mutation. Hum Genet 106:179–187

Saleem Q, Muthane U, Verma IC, Brahmachari SK, Jain S (2002) Expanding colonies and expanding repeats. Lancet 359:895–896

Saleem Q, Roy S, Murgood U, Saxena R, Verma IC, Anand A, Muthane U, Jain S, Brahmachari SK (2003) Molecular analysis of Huntington’s disease and linked polymorphisms in the Indian population. Acta Neurol Scand 108:281–286

Sankalia HD (1974) Prehistory and protohistory of India and Pakistan. Deccan College, Poona

Sengupta S, Farheen S, Mukherjee N, Dey B, Mukhopadhyay B, Sil SK, Prabhakaran N, Ramesh A, Edwin D, Usha Rani MV, Mitra M, Mahadik CT, Singh S, Sehgal SC, Majumder PP (2004) DNA sequence variation and haplotype structure of the ICAM1 and TNF genes in 12 ethnic groups of India reveal patterns of importance in designing association studies. Ann Hum Genet 68:574–587

Sharma S, Ghosh B (2004) Association of an intragenic microsatellite marker in the CC16 gene with asthma in the Indian population. J Hum Genet 49:677–683. Epub 2004 November 12

Sharma S, Nagarkatti R, B-Rao C, Niphadkar PV, Vijayan V, Sharma SK, Ghosh B (2004) A three locus haplotype, A_16_C, in Fc epsilon RI beta gene confers higher risk for atopic asthma. Clin Genet 66:417–425

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311

Singh KS (2002) People of India: introduction national series. Anthropological Survey of India-Oxford University Press, Delhi

Tapadar P, Ghosh S, Majumder PP (2000) Haplotyping in pedigrees via a genetic algorithm. Hum Hered 50:43–56

Thangaraj K, Ramana GV, Singh L (1999) Y-chromosome and mitochondrial DNA polymorphisms in Indian populations. Electrophoresis 20:1743–1747

Thangaraj K, Joshi MB, Reddy AG, Gupta NJ, Chakravarty B, Singh L (2002) CAG repeat expansion in the androgen receptor gene is not associated with male infertility in Indian populations. J Androl 23:815–818

The International HapMap Consortium (2003) The international HapMap Project. Nature 426:789–796

Verma R, Chauhan C, Saleem Q, Gandhi C, Jain S, Brahmachari SK (2004) A nonsense mutation in the synaptogyrin 1 gene in a family with schizophrenia. Biol Psychiatry 55:196–199

Verma R, Mukerji M, Grover D, Rao CB, Das SK, Kubendran S, Jain S, Brahmachari SK (2005) MLC1 gene is associated with schizophrenia and bipolar disorder in Southern India. Biol Psychiatry (in press)