Genetic landscape of the people of India: a canvas for disease gene exploration

INDIAN GENOME VARIATION CONSORTIUM*

Abstract

Analyses of frequency profiles of markers on disease or drug-response related genes in diverse populations are important for the dissection of common diseases. We report the results of analyses of data on 405 SNPs from 75 such genes and a 5.2 Mb chromosome, 22 genomic region in 1871 individuals from diverse 55 endogamous Indian populations. These include 32 large (>10 million individuals) and 23 isolated populations, representing a large fraction of the people of India. We observe high levels of genetic divergence between groups of populations that cluster largely on the basis of ethnicity and language. Indian populations not only overlap with the diversity of HapMap populations, but also contain population groups that are genetically distinct. These data and results are useful for addressing stratification and study design issues in complex traits especially for heterogeneous populations.

[Indian Genome Variation Consortium 2008 Genetic landscape of the people of India: a canvas for disease gene exploration. *J. Genet.* **87**, 3–20]

Introduction

Genetically isolated populations are considered to be important in dissecting complex diseases and mapping underlying genes (Wright *et al*. 1999; Peltonen 2000; Heutink and Oostra 2002; Abecasis *et al.* 2005). However, the validation of results across populations has met with limited success. Population stratification, a consequence of differences in allele frequencies across populations arising mainly due to natural selection and genetic drift, is a major problem in association studies. It is, therefore, important to assess the nature and extent of population stratification in contemporary endogamous populations especially in the context of established or candidate disease genes. Indians, comprising about one-sixth of the world population, with large family sizes and high levels of endogamy, provide a unique resource for dissecting complex disease etiology and pathogenesis. Further, India provides a large patient pool with the majority being drug-naive. Historically, the Indian population is a conglomeration of multiple culture and evolutionary histories. Anatomically modern man is estimated to have reached the north-western periphery of the Indian subcontinent around 70,000 ybp and moved southward into Sri Lanka in the next 20,000 years (Habib 2001, 2002; Singh 2002). Modern human communities may also have migrated into eastern India from Myanmar around 4500 to 11,000 ybp (Habib 2001, 2002; Singh 2002). The evolutionary antiquity of Indian ethnic groups and subsequent migration from central Asia, west Asia and southern China has resulted in a rich tapestry of socio-cultural, linguistic and biological diversity. Broadly, Indians belong to Austro–Asiatic (AA), Tibeto–Burman (TB), Indo–European (IE) and Dravidian (DR) language families. Distinct religious communities, hierarchical castes and subcastes, and isolated tribal groups that comprise the people of India remain largely endogamous. Most of these groups have strict social rules governing mating patterns. Earlier studies using mitochondrial, Y-chromosomal and limited autosomal markers, that primarily addressed issues of origin and migrations, have demonstrated extensive genetic diversity in India (Bamshad *et al*. 2001; Roychoudhury *et al*. 2001; Basu *et al*. 2003; Kivisild *et al*. 2003; Cordaux *et al*. 2004; Kashyap *et al*. 2006; Sahoo *et al*. 2006; Sengupta *et al*. 2006; Thanseem *et al*. 2006). In contrast, a recent study based on autosomal microsatellite markers has inferred that Indian populations show low levels of genetic differentiation (Rosenberg *et al*. 2006). This inference was possibly due to biased recruitment

*For complete authors' list of the Indian Genome Variation Database Consortium please see Appendix. For correspondence. E-mail: skb@igib.res.in; Functional Genomics Unit, Institute of Genomics and Integrative Biology (CSIR), Mall Road, New Delhi 110 007, India.

Keywords. Indian genome variation; polymorphism; SNP; Asia; HapMap; complex disease; pharmacogenomics.

of study participants and insufficient classification based on language and ethnicity.

The Indian Genome Variation (IGV) Consortium (Indian Genome Variation Consortium 2005) was set up to build a resource that would enable us to address the following questions related to distribution of genetic variation and its relationship to complex disease in the Indian population: (i) Are the frequencies of SNPs putatively associated with complex diseases similar across different populations and can we identify clusters of populations which share similar SNP frequencies? (ii) Do these clusters correlate with ethnic, linguistic or geographical labels? (iii) What is the nature and extent of genetic differentiation within and among clusters? (iv) How are Indian populations related to HapMap populations? (v) Can we identify a subset of SNPs that help distinguish between ethnic groups? (vi) What directions can these data provide for the design of future studies on genomewide association *vis-à-vis* population stratification? and (vii) Can we identify 'at-risk' populations for complex disorders, poor drug response and predisposition to infectious diseases? In this study, we have primarily addressed the first five questions. We report the nature and extent of variation in 55 Indian populations based on 405 SNPs, selected from a set of 75 genes spread across all chromosomes and a 5.2 Mb segment of chromosome 22 spanning 49 genes. These populations are representative of the ethnic, linguistic and geographic diversity of India. The genes were selected based on pathway analysis and involvement in disparate molecular functions and biological processes, and are implicated in complex diseases as well as susceptibility to infection. A large number of noncoding SNPs were selected from the 5.2 Mb region of chromosome 22 continuous stretch to assess tag transferability in the study populations. This region harbours a susceptibility locus for schizophrenia and bipolar disorder that has been implicated in multiple studies (Papolos *et al.* 1996; Schwab and Wildenauer 1999; Kelsoe *et al.* 2001; Verma *et al.* 2004, 2005). To our knowledge, this is the largest single study conducted on Indian populations in terms of numbers of populations, candidate disease genes and biparental SNPs assayed.

Materials and methods

Selection of populations and sample collection

The initial study involved the identification of SNPs in a panel of 43 samples drawn from geographically and ethnically diverse populations. This was done to maximize the scope of discovery of novel SNPs (Indian Genome Variation Consortium 2005). Validation of SNPs was carried out on a panel of populations aimed at representing endogamous populations from AA, TB, IE and DR linguistic lineages from north, south, east, west, central and north-east India (see table 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). Instead of naming the populations, we followed a convention wherein each population was given a label of language, followed by geographical zone and ethnic category as LP, IP or SP. The LPs are caste groups, mostly large populations, the IPs are tribal isolated populations and the SPs are religious groups. Tribal populations can be large (a few of the IPs are also large groups). In each grid of language and geography, at least two IPs and LPs were identified for collection, wherever applicable. A minimum of 50 samples from LPs and 25 samples from IPs were collected (see table 2a in electronic supplementary material). From an initial number of 2014 unrelated samples, the final validation data are on 1871 samples comprising of 1240 males and 631 females (see table 2b of electronic supplementary material). The final set contains samples that qualified all quality control (QC) criteria including gender assignment as well as genotype success (explained below). Population identification and collection of samples were done with the help of trained anthropologists, and social and community health workers. Details of ethical clearance and establishment of ethnicity for the present study have been described earlier (Indian Genome Variation Consortium 2005). The endogamy for each population was established by taking extensive information about marriage patterns, gathered through pedigrees and interviews of family members of the donor as well as from published literature.

Selection of genes and genomic region

This study was aimed at understanding variability in SNPs across diverse individual populations with respect to functional and positional candidates, as well as understanding the underlying relatedness and ancestry of populations for linkage disequilibrium (LD) studies. A representative set of 75 genes and a large 5.2 Mb genomic region on chromosome 22 spanning 49 genes representing a susceptibility locus for schizophrenia and bipolar disorder was selected. Genes were selected based on their involvement in monogenic disorders, and their being positional as well as functional candidates for complex diseases. The representative set of genes included drug-response genes, genes involved in cancer and aging, eye diseases, allergy and asthma, neuro-psychiatric, metabolic and cardiovascular disorders as well as genes involved in susceptibility to infections etc. Details of the genes have been provided in table 3 of electronic supplementary material. These genes represent various biological processes and molecular functions (see figure 1 in electronic supplementary material). Nearly all the chromosomes are represented except the Y-chromosome.

Selection of SNPs

This study primarily focused on identification of functional polymorphisms and their associated haplotypes in the Indian population. By sequencing 730 amplicons of candidate gene loci in a multiethnic discovery panel of 43 samples, 170 novel and 560 reported SNPs were identified (Indian Genome Variation Consortium 2005). To prioritize SNPs for

validation in larger population samples, a set of filtering criteria were evolved. SNPs were first selected on the basis of frequency; novel variants detected in only one sample of the discovery (DSNP) panel or reported SNPs with a frequency of < 0.01 in the DSNP data were not taken forward. Following this, SNPs with reported frequencies of $> 10\%$ in at least three world populations or $> 20\%$ in at least two world populations along with all functional and novel SNPs were retained. This was followed by selection based on spacing between SNPs and minor allele frequency; all reported SNPs at least 1 kb apart and all novel and functional SNPs were retained. In closely spaced SNPs, the SNP with a higher frequency was selected. If required, additional SNPs were chosen from reported SNPs with uniform spacing, spanning the length of the gene.

Finally, 601 SNPs (including 17% novel SNPs) were taken for validation on 2014 samples. 10% of these SNPs failed during assay design and optimization on Sequenom and 126 were found to be monomorphic during validation. Twelve SNPs were dropped from the final validated set as they did not fulfill the threshold criteria of $> 80\%$ genotype success. The final validated dataset comprised of 420 SNPs, of which 405 were autosomal (including 90 SNPs from 5.2 Mb region of chromosome 22) and 15 were from the X-chromosome. From these, 276 SNPs have been typed in any one of the HapMap populations. Genotype frequencies for 231 SNPs for which complete genotype data were available for all the HapMap and 55 Indian populations are provided in table 4 of electronic supplementary material. Details of the 405 SNPs and their annotations (dbSNP build 125) are provided in table 5 of electronic supplementary material.

Genotyping and sequencing

The discovery of novel SNPs was carried out by bidirectional sequencing on a multi-ethnic Indian discovery panel of 43 samples (Indian Genome Variation Consortium 2005). For sequencing analysis, PCR primers were designed using DNASTAR Lasergene software (PrimerSelect 5.07, Madison, USA). Genomic DNA sequencing was carried out on ABI 3100 and ABI 3730 capillary based sequencers (Applied Biosystems, Foster City, CA). Genotyping of SNPs was performed using MALDI-TOF based chemistry on the Sequenom platform. Prior to validation on the entire sample set, the polymorphic status of both novel and database SNPs were revalidated on pools of samples from the discovery panel, as well as from individual populations. A number of QC filtering steps were performed prior to considering each SNP for analysis. These QC filters were applied with respect to genotype success ($> 80\%$), consistency in 10% duplicate controls (≤ 1 discrepancy in 5) and Hardy–Weinberg checks using Fisher's exact test at 5% significance level. The genotype error rate was estimated based on comparison with 10% duplicate controls. Blind QC was also performed for 24 samples. Gender QC was carried out based on heterozygosity checks of F-VIII and F-IX genes in male samples, Y-chromosome specific STRs as well as sex-specific genotyping of the *amelogenin* gene. The final data set thus comprised of genotypes of 1871 samples for 420 SNPs (both more than 80%).

The allele frequencies are reported with respect to the positive strand of the chromosome for the IGVdb reference allele (minor allele in more than 50% of the Indian populations). The HapMap alleles have also been converted with respect to the positive strand (see table 5 in electronic supplementary material) after confirmation of the strand information by BLAST analysis. This was specifically ensured when the variation resulted is transversion from A to T, or G to C.

Statistical methods

Analysis of genetic differentiation (Nei and Chesser 1983) was carried out using the large version 2.9.3.2 of FSTAT (courtesy Dr Jerome Goudet). Tests of significance of F_{ST} values were performed by bootstrapping, as implemented in FSTAT and Arlequin (http://lgb.unige.ch/arlequin/).

Estimation of D_A distance (Nei 1977) and phylogenetic analysis using the neighbour-joining (NJ) method (Saitou and Nei 1987), was done using DISPAN (available from http://iubio.bio.indiana.edu/soft/). Principal components, discriminant and classification analyses were carried out using SPSS for Windows (version 10). To identify population-cluster specific 'keystone' SNPs, we carried out a stepwise linear discriminant analysis (Rao 1952). In this analysis, uncorrelated linear combinations of allele frequencies (linear discriminant functions) of the loci that provide the best separation of the multidimensional scatter of the allele frequencies are estimated. Loci are entered into the discriminant function in a stepwise manner, starting with the locus that provides the best separation. This procedure is terminated when the next best locus to be entered into the discriminant function does not provide any further significant separation. These linear discriminant functions are then used to classify individual populations into groups on the basis of allele frequencies. To avoid 'over-fitting' (i.e., overestimating the proportion of populations correctly classified to its cluster), we adapted a half-sample approach. Initially, we found the subset of SNPs that can serve to identify populations belonging to specific clusters (ethnic, linguistic or geographical) from a randomly-chosen half-sample ('discovery half-sample'; 50% of populations from each cluster being randomly chosen). We then tested the performance of this subset of SNPs on the other disjoint half-sample ('validation half-sample').

Tag SNPs were identified in the HapMap data (http://www.hapmap.org) using Tagger (http://www.broad.mit.edu/mpg/tagger/). Haplotypes and their frequencies were statistically inferred from phase-unknown genotype data by using the software PHASE version 2.1 (Stephens *et al.* 2001). Mantel tests were carried out using zt, version 1.0 (http://www.psb.ugent.be/~erbon/mantel/).

System structure

A novel network analysis approach (http://physiol.eecs.cwru.edu:8802/~amit/) called system structure (SStr) was used for clustering populations. The SStr method uses a set-theoretic, distribution-free computational model for complex systems, from first principles. Based on the information contained in its SStr, a complex system can be partitioned into 'natural groups', without requiring any a priori ancestry information, including the number of groups. This characteristic distinguishes this approach from methods widely used for the analysis of population structure e.g. STRUCTURE (Pritchard *et al*. 2000). In contrast to Bayesian approaches, it is less model dependent. A detailed description of the method and definitions of memberships are provided in the supplementary note of electronic supplementary material.

Briefly, in SStr any system can be described by appropriately weighted interactions obtained, based on experimental measurements between system objects. Further, an appropriate measure of interaction needs to account for asymmetry in the relationship between any interacting pair taken into account, the system within which the interaction operates and further incorporate a propagation step that allows second and higher order interactions to diffuse and influence the measures between the system objects. Essentially, SStr is a weighted network where the weights, called the system measures, quantify the potentials associated with the nodes and arcs (Sinha 2001). A number of systems have been analysed within this framework (Sinha *et al*. 2004; Fogarty *et al*. 2006).

In the analysis of population genetic structure, the nodes of the network can be population and loci with their allele or genotype frequencies. System measures between populations (samples) and their cluster/partition memberships provide a useful description of population substructure and admixtures between populations without any a priori information/population labels. For each population one obtains fuzzy, possibility and typicality measures across all the clusters. These memberships are defined as follows: fuzzy, differential membership of a population across clusters; possibility, differential membership of a cluster across all populations relative to all population-cluster assignments; typicality, differential membership of a cluster only across its core members. Typicality needs to be interpreted carefully. This measure gives the significance of a core member in a specific cluster relative only to other core members of the same cluster (see supplementary note in electronic supplementary material).

For validation of the robustness of SStr, a comparison with Pritchard's STRUCTURE was carried out using the data on analysis of human populations by Rosenberg *et al*. (2006). The results obtained from both methods were highly concordant (see supplementary note in electronic supplementary material). Since STRUCTURE did not converge on the Indian data, SStr analysis in this study was particularly useful.

Results

Genetic differentiation

To determine the extent of genetic differentiation, we identified 55 representative populations drawn from four major linguistic groups (AA, TB, IE and DR), six geographical regions of habitat (N, north; NE, north-east; W, west; E, east, S, south; C, central) and different socio-cultural strata (LP, large population, caste; IP, isolated population, tribes; SP, special population, religious groups) (table 1). Pairwise F_{ST} values were calculated to determine the extent of differentiation among the populations, possibly derived from diverse ancestral lineages. F_{ST} values, calculated from allele frequencies at all autosomal loci between pairs of populations varied from 0.000 to 0.111 (figure 1).

Table 1. Details of the populations analysed in the current study and the average heterozygosity.

Sl. no	Population code	No. of samples	Caste/religious group/tribe	Mean heterozygosity	SE across loci
1	AA-C-IP1	46	Tribe	0.35	0.008
2	AA-C-IP4	23	Tribe	0.36	0.008
3	AA-C-IP5	23	Tribe	0.35	0.008
4	AA-E-IP1	49	Tribe	0.36	0.008
5	AA-E-IP2	23	Tribe	0.35	0.008
6	AA-E-IP3	23	Tribe	0.33	0.008
7	AA-NE-IP1	44	Tribe	0.35	0.008
8	AA-W-IP1	22	Tribe	0.36	0.007
9	DR-C-IP1	23	Tribe	0.35	0.007
10	DR-C-IP2	46	Tribe	0.32	0.009
11	DR-E-IP1	46	Tribe	0.35	0.007
12	DR-S-IP1	21	Tribe	0.32	0.009
13	DR-S-IP2	23	Tribe	0.35	0.008
14	DR-S-IP3	23	Tribe	0.36	0.007
15	DR-S-IP4	23	Tribe	0.34	0.008
16	DR-S-LP1	46	Caste	0.36	0.007

Table 1 (*contd*)

Sl. no	Population code	No. of samples	Caste/religious group/tribe	Mean heterozygosity	SE across loci
17	DR-S-LP2	46	Caste	0.35	0.007
18	DR-S-LP3	46	Caste	0.35	0.008
19	DR-S-LP4	23	Caste	0.36	0.007
20	DR-S-LP5	23	Caste	0.36	0.008
21	IE-E-LP1	46	Caste	0.36	0.007
22	IE-E-LP2	46	Caste	0.35	0.008
23	IE-E-LP3	23	Caste	0.36	0.007
24	IE-E-LP4	42	Caste	0.36	0.007
25	IE-NE-IP1	48	Tribe	0.34	0.008
26	IE-NE-LP1	23	Caste	0.36	0.008
27	IE-N-IP1	46	Tribe	0.36	0.007
28	IE-N-IP2	46	Tribe	0.35	0.007
29	IE-N-LP1	46	Caste	0.37	0.007
30	IE-N-LP10	46	Caste	0.37	0.007
31	IE-N-LP11	46	Caste	0.37	0.007
32	IE-N-LP18	46	Caste	0.36	0.007
33	IE-N-LP2	46	Caste	0.37	0.007
34	IE-N-LP3	46	Caste	0.37	0.007
35	IE-N-LP5	23	Caste	0.36	0.007
36	IE-N-LP6	46	Caste	0.35	0.008
37	IE-N-LP7	46	Caste	0.36	0.007
38	IE-N-LP8	46	Caste	0.37	0.008
39	IE-N-LP9	46	Caste	0.36	0.007
40	IE-N-SP2	18	Religious group	0.37	0.007
41	IE-N-SP3	46	Religious group	0.36	0.007
42	IE-N-SP4	23	Religious group	0.37	0.007
43	IE-N-SP5	46	Religious group	0.36	0.007
44	IE-S-IP1	46	Tribe	0.34	0.008
45	IE-W-IP1	23	Tribe	0.36	0.007
46	IE-W-IP2	23	Tribe	0.35	0.007
47	IE-W-LP1	23	Caste	0.36	0.008
48	IE-W-LP2	23	Caste	0.36	0.008
49	IE-W-LP3	46	Caste	0.37	0.007
50	IE-W-LP4	46	Caste	0.35	0.007
51	OG-W-IP	23	Out-group	0.34	0.008
52	TB-NE-LP1	46	Caste	0.34	0.008
53	TB-N-IP1	46	Tribe	0.33	0.008
54	TB-N-SP1	46	Religious group	0.35	0.008
55	TB-N-SP2	46	Religious group	0.32	0.009

The mean heterozygosity (H_o) was high for all populations and ranged from 0.32 to 0.37. Lower H_o was observed predominantly in isolated tribal populations (IPs) while, higher H_o was observed for large population (LPs). SE, Standard error.

The majority of the F_{ST} values between populations were significantly greater than zero ($P < 0.05$) indicating population differentiation. Mean F_{ST} (0.03 ± 0.0005) suggests that the extent of differentiation overall was low. However, it is possible that in some cases, due to small sample sizes, the F_{ST} values might not be significant even if there is differentiation. Mean F_{ST} values computed separately on the basis of frequencies of SNPs (see table 6 in electronic supplementary material) that were located in specific regions of the genome (e.g., promoter region, exon, intron and UTR) were not significantly different ($P = 0.063$). With respect to a few individual loci, the extent of genetic differentiation in India is high (see table 6 in electronic supplementary material) and of comparable magnitude to that observed among continental populations (0.14) (The International HapMap Consortium 2003; Tishkoff and Verrelli 2003; Watkins *et al*. 2003, 2005).

Maximum F_{ST} values were observed among the tribal populations of different linguistic lineages. On a pan-India level, when populations were grouped by language or by

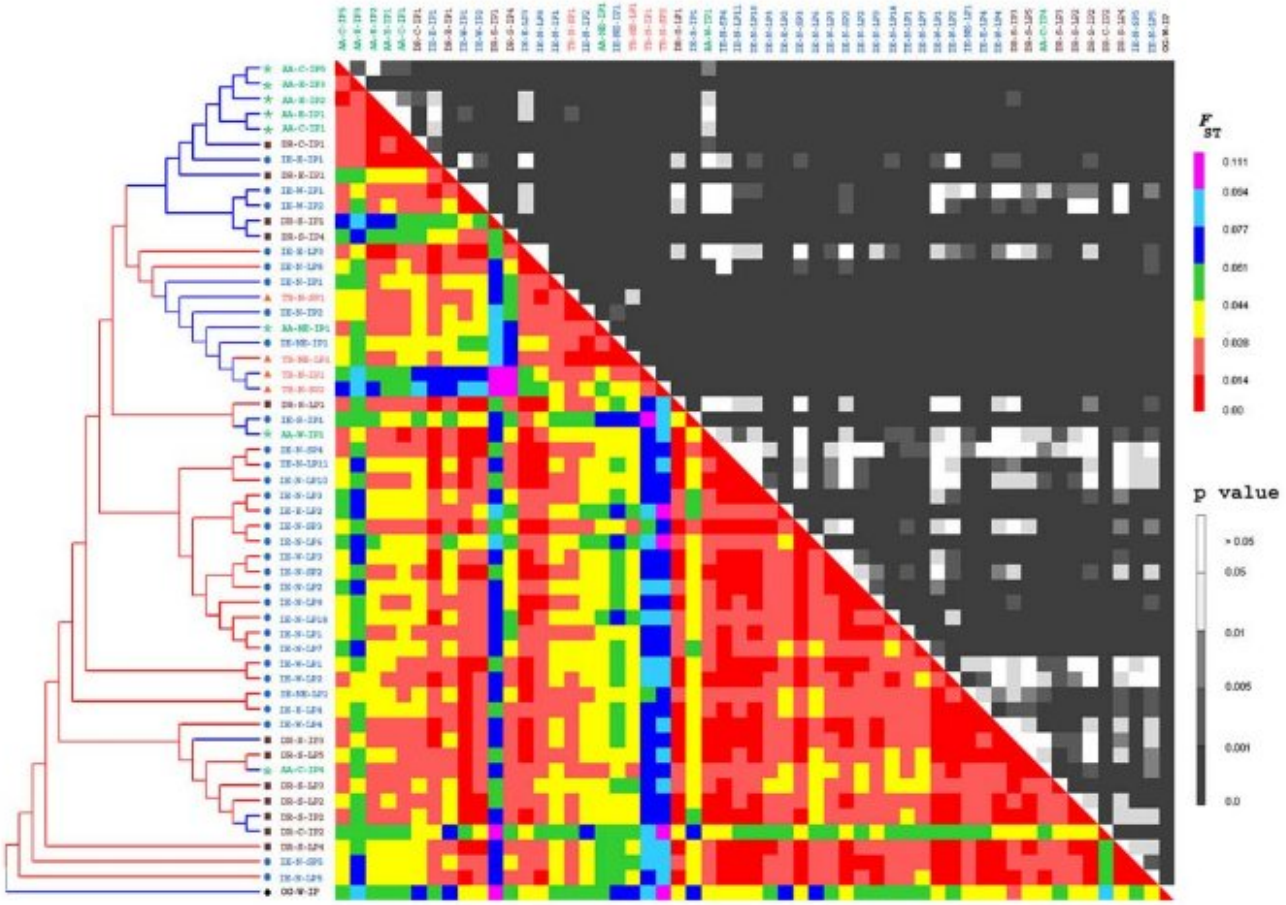

Figure 1. Heterogeneity and inter-relatedness among Indian populations. Heatmap of pairwise F_{ST} (colour scale) between populations and corresponding P values (gray scale) depicted with neighbour-joining tree illustrating population affinities based on Nei's D_A distance. The red and blue branches in the tree represent large and isolated populations, respectively and the symbols represent the linguistic lineage of a population. The colour legends for F_{ST} and P values are indicated. The populations are coded by linguistic lineage (AA, Austro–Asiatic; IE, Indo–European; DR, Dravidian and TB, Tibeto–Burman) followed by geographical location (N, north; NE, north-east; W, west; E, east; S, south and C, central) and ethnic category (LP, castes; SP, religious groups and IP, tribes). Population codes are also coloured on the basis of language family.

geographical region of habitat, the extent of genetic differentiation among linguistic or geographical groups was not statistically significant ($P > 0.1$; see table 7 in electronic supplementary material). However, grouping by ethnicity (caste and tribe) indicated significant differentiation ($P < 0.05$; see table 7 in electronic supplementary material) possibly due to antiquity and isolation of the tribal compared to the caste populations.

The picture of genetic differentiation within geographical regions or ethno-linguistic groups presented some interesting features. While DR-speaking LPs and IPs did not show significant genetic differentiation ($P = 0.9$), the IE-speaking LP and IP groups were significantly differentiated ($P = 0.01$). Within tribes, but not the castes, the IE-speakers and DR-speakers showed statistically significant differentiation ($P = 0.01$). However, the DR-speaking tribal groups were not significantly differentiated ($P = 0.93$) from the AA-speaking groups. The patterns of genetic differentiation estimated by AMOVA were similar to the above inferences (table 2). From the above results, it is clear that pooling populations without considering ethnicity and linguistic affiliations that contribute to population stratification can result in false inferences in genetic association studies.

Genetic affinities

We used cluster-analytic, principal-components and SStr-based approaches to analyse the extent of genetic relatedness among the populations. Few major clusters of the study populations were identified from the tree of genetic relationships computed on the basis of Nei's genetic distance D_A and F_{ST} (figure 1). The first cluster primarily comprised of AA–IPs and DR–IPs consistent with the earlier observation of a statistically nonsignificant F_{ST} value between AA and DR tribals. The second cluster included TB-speaking populations, irrespective of their geographical region of habitat. This cluster also comprised of three IE-speaking isolated populations (IE–IPs) and two IE–LPs. Majority of

Table 2. Extent of genetic differentiation estimated by AMOVA.

Based on language			
	Among populations within groups	Among groups	Among individuals
IE – TB	1.91	2.80	95.29
IE – AA	1.98	1.25	96.77
IE – DR	2.11	0.30	97.59
TB – AA	1.95	2.20	95.85
TB – DR	2.31	3.04	94.65
AA – DR	2.41	0.71	96.88
Based on geography			
	Among populations within groups	Among groups	Among individuals
North – north east	2.41	1.23	96.36
North – east	2.52	0.38	97.10
North – central	2.46	1.22	96.32
North – west	2.23	0.24	97.53
North – south	2.45	0.48	97.07
North east – east	2.44	1.19	96.37
North east – central	2.23	1.18	96.59
North east – west	1.33	2.28	96.39
North east – south	2.25	2.08	95.67
East – central	2.59	0.29	97.12
East – west	1.92	0.22	97.86
East – south	2.51	0.30	97.19
Central – west	1.57	0.93	97.50
Central – south	2.39	0.80	96.81
West – south	1.87	0.08	98.05
Based on ethnicity			
	Among populations within groups	Among groups	Among individuals
IE Large – IE isolated	1.70	0.75	97.55
DR Large – DR isolated	2.57	0.01	97.42

these populations reside in the Himalayan belt. There were a larger number of smaller clusters that predominantly comprised of IE–LPs and IE–SPs. The DR-speaking LPs and IPs, predominantly from southern India, formed a separate cluster (figure 1). There seems to be a considerable diversity among IE-speaking populations in different geographical regions, as reflected by the large number of smaller clusters to which they belong and also by our finding of F_{ST} values significantly greater than zero between several IE populations ($P < 0.05$). Thus, although there are no clear geographical grouping of populations, ethnicity (tribal/nontribal) and language seem to be the major determinants of genetic affinities between the populations of India. This is concordant with an earlier finding based on allele frequencies at blood group, serum protein and enzyme loci (Piazza *et al*. 1980). Besides, within the IE group, LPs and SPs (religious groups) exhibited high genetic affinities. Similar affinity was also observed between TB–IPs and TB–SPs. The population OG-W-IP, known to have been derived from an African population (Singh 2002) was an outlier on the phylogenetic tree (figure 1). We also carried out principal component analysis (PCA) to examine the patterns of variations among populations. The first two principal components (PCs) explained about 25% of the variation in allele frequencies. The pattern of genetic affinities was largely in accord with that observed in the cluster analysis, but highlighted the heterogeneity among the DR populations (figure 2). It should be noted however, 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 is contented that the Dravidian speakers, now geographically confined to southern India, were more widespread throughout India prior to the arrival of the Indo–European speakers (Thapar 1966). They, possibly after a period of social

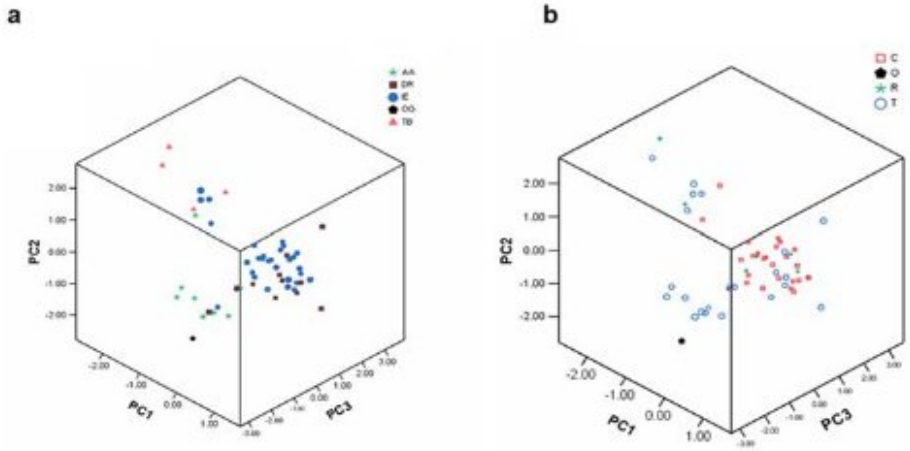

Figure 2. Principal component analysis (PCA) plots depicting separation of each population in different components. The populations have been coloured on the basis of their (a) linguistic group: AA, Austro–Asiatic; DR, Dravidian; TB, Tibeto Burman; IE, Indo–European OG is an out-group population of more recent African ancestry and (b) ethnicity: C, caste; T, tribe; R, religious group; O, out-group.

and genetic admixture with the Indo–Europeans, retreated to southern India, a hypothesis that has been supported by mitochondrial DNA analyses (Basu *et al*. 2003). Our results showing genetic heterogeneity among the Dravidian speakers further supports the above hypothesis. The Indo–European speakers also exhibit a similar or higher degree of genetic heterogeneity possibly because of different extents of admixture with the indigenous populations over different time periods after their entry into India. It is surprising that in spite of such a high levels of admixtures, the contemporary ethnic groups of India still exhibit high levels of genetic differentiation and substructuring.

Ideally, inferences regarding genetic affinities are drawn based on a random set of loci from the genome. On the contrary, our study included SNPs from genes that are possibly associated with disease outcomes and therefore could bias our inferences. To address this concern we recomputed distances (D_A) between populations after removing 73 (18%) loci that belonged to the upper and lower 20% tails of the joint distribution of heterozygosity (H_o) and F_{ST} (see table 6 in electronic supplementary material). We then compared the distance matrix generated above with the matrix of all 405 loci using a nonparametric Mantel test. The correlation between the matrices was 0.99 ($P < 0.0001$; based on 10,000 permutations), indicating that the inferences on genetic affinities among populations based on all loci are not significantly altered by inclusion of highly differentiated SNPs from disease candidate genes. We used a novel system-theoretic network analysis approach (see supplementary note in electronic supplementary material) in addition to dimension-reducing tree-based and principal-component approaches, to understand relationships between populations belonging to different clusters. Based on genotype frequencies in each population, SStr analysis identified five optimal groups (figure 3a; see figure 2 in electronic supplementary material). Fuzzy measures revealed two near homogeneous groups (1 and 2) where at least 80% of the populations shared >75% membership. These were derived mostly from the IE-speaking LPs and SPs (group 1), and IPs and SPs of the Himalayan belt (group 2). Groups 3, 4 and 5 were more heterogeneous. Group 5 predominantly consisted of AA members that shared membership with isolated populations of group 2, indicative of admixture with the latter. As with F_{ST} and D_A analyses, the DR populations were distributed across all clusters, indicating high heterogeneity and diverse ancestry of the DR group. The heterogeneity in DR populations is also evident when the fuzzy memberships of each population are overlaid on the linguistic map of India (figure 3b). The map also depicts genetic correlates to linguistic histories in a large number of populations. According to the distribution on the map (figure 3b), populations in group 2 cover the Himalayan belt extending from the north to the north-east; group 1 covers most of northern India; group 5 is the IP belt of central and eastern India while group 4 represents populations from the southern part of India as well as some IE populations of the northern, eastern and western regions. Group 3, which is a mixture of some IE–LPs, IE–IPs and DR–IPs, and spread mostly across central India, seems to be a 'bridge' between groups 1 and 4. Admixture of AA-speaking and DR-speaking IPs of the tribal belt with TB populations from the Himalayan belt is also evident from the map.

The pattern of clustering of Indian population groups in our analysis suggests that the effects of population stratification in disease association studies may be small, if cases and controls are both drawn from the same cluster even if they do not belong to the same ethnic group. Thus, correction for population stratification would be needed if cases and controls were drawn from populations that belong to different

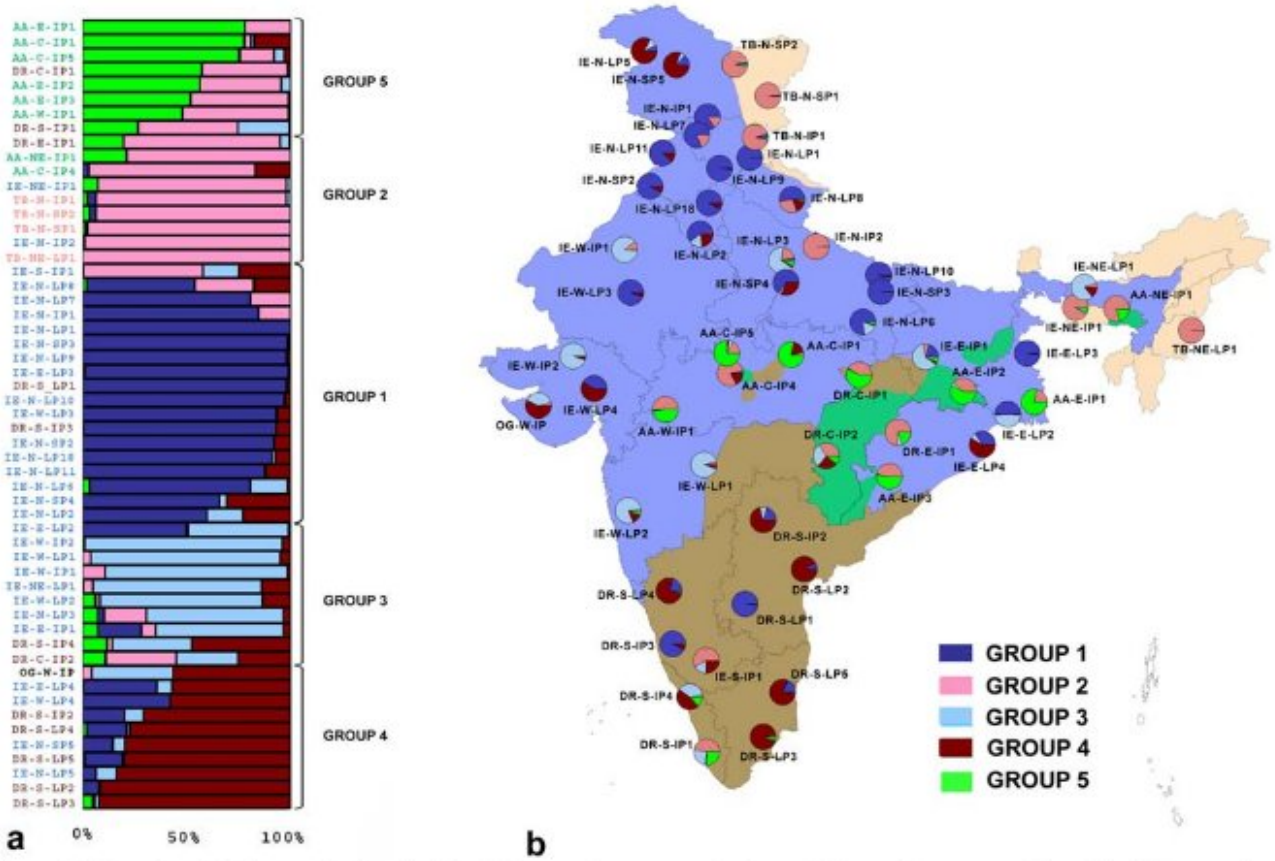

Figure 3. Grouping of Indian populations by SStr (a) System Structure analysis reveals five major groups depicted by different colour schemes. The relative fuzzy memberships of populations in each group are depicted by horizontal bars. (b) Representation of system structure-derived membership information on the linguistic map of India. Blue, brown, green and pink backgrounds indicate regions where languages of predominantly IE, DR, AA and TB lineages are spoken, respectively. A key to the group colour based on typicality membership (see figure 2c in electronic supplementary material) is given. The pie-charts represent fuzzy membership information for each population inferred from SStr analysis.

clusters. This analysis further highlights the requirement for sampling populations belonging to each of these clusters in order to capture the entire genetic spectrum of India.

SNP signatures of population clusters

We observed that on a pan-India level, the tribal and caste populations are significantly differentiated. Besides, within some geographical regions, tribes and castes subclassified by language are also well differentiated. In tune with the active search for ethnicity-specific or ancestry-informative markers (Akey *et al.* 2002; Shriver and Kittles 2004), we sought to identify a set of SNPs that could classify populations in terms of ethnicity, language or geographical region of habitat. Though we recognize that except for ethnicity (tribe/caste), the other determinants of genetic affinity did not turn out to be significant in the present data set, inferences on affinity might have been influenced by summarization of data (genetic distances, principal components, etc) pertaining to >400 genetic markers. In principle, it is possible to identify a small number of SNPs that can serve as signatures of population ancestry. To explore this, we used stepwise linear discriminant and classification analysis (see materials and methods section) using allele frequencies of the top 100 loci that exhibited high interpopulation variance in allele frequency. Our analysis revealed that a very small set of SNPs sufficed to identify populations with a high degree of accuracy to the broad clusters of ethnicity and language (table 3). Allele frequencies at 12 SNP loci (termed 'keystone SNPs') were sufficient to identify a population with unknown ethnicity as IP (predominantly tribal) or LP (predominantly caste) with 100% accuracy. Spatial maps of allele frequencies at representative keystone SNP loci are shown in figure 4. This perhaps is a reflection of the anthropological notion that IPs of India unlike for example the LPs are relatively unadmixed. The success in identifying linguistic lineage was 85.2% based on eight keystone SNPs. However, the success in classifying a population to a geographical region based on such keystone SNPs was low (~56%). These results underscore that it is possible to classify a population into a larger socio-geographical cluster with a reasonable degree of accuracy using a small number of SNPs. However,

Table 3. List of keystone* SNPs useful for classification of Indian populations based on ethnic, linguistic and geographical assignments.

Grouping of populations	Discriminating SNP ID	Gene name	% of samples correctly classified using keystone SNPs
Ethnicity (2 groups: IPs and LPs)	rs4147536	*ADH1B*	100
	rs712700	*PAX4*	
	rs747672	*OPTC*	
	rs713689	Chr22 region	
	rs1056827	*CYP1B1*	
	rs1799971	*OPRM1*	
	rs327516	*PAX4*	
	rs1801368	*MAD1L1*	
	rs2274976	*MTHFR*	
	I000050	*IL4 R*	
	rs2239704	*TNF2*	
	rs1801274	*FCGR2A*	
Language (4 groups: AA, DR, IE, TB)	rs4934028	*MAT1A*	85.2
	rs3753868	*APCS*	
	rs1169289	*TCF1*	
	rs133335	Chr22 region	
	rs738534	Chr22 region	
	rs2267432	*ACO2*	
	rs445122	*PPP2R2B*	
	rs1317944	*COPA*	
Geography (6 regions: C, E, N, NE, S, W)	rs17107315	*SPINK1*	55.6
	rs133335	Chr22 region	
	rs2234926	*MYOC*	
	rs1317944	*COPA*	
	rs5021654	*TYR*	
	rs137116	Chr22 region	

*Keystone SNPs were discovered on the basis of a first half-sample (discovery sample) and their performance was assessed on the basis of a disjoint half-sample (validation sample).

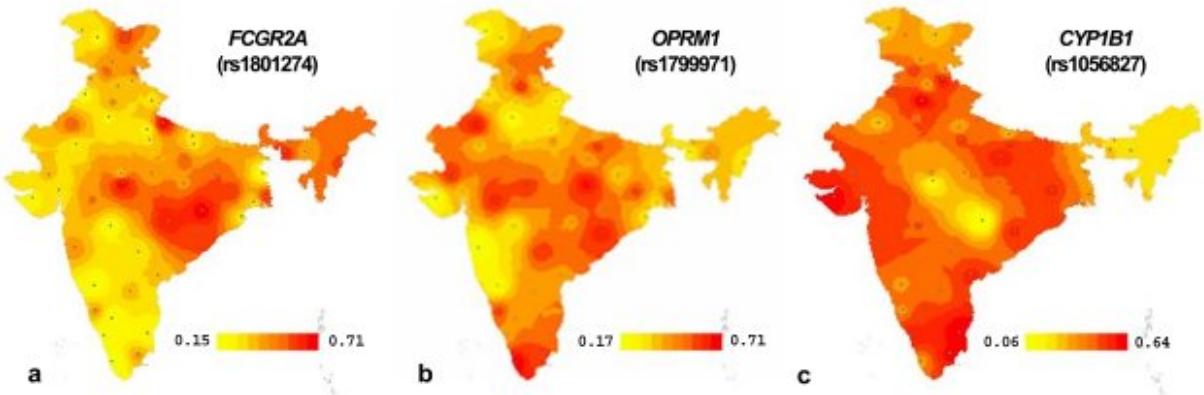

Figure 4. Spatial frequency maps depicting distribution of minor allele frequencies of selected keystone SNPs. Minor allele frequency distribution is plotted for SNPs of: (a) *FCGR2A* (rs1801274; p.R131H), (b) *OPRM1* (rs1799971; p.D102N) and (c) *CYP1B1* (rs1056827; p.A119S). The colour gradient below each map depicts the range of observed frequency from minimum to maximum.

we would like to emphasize that the keystone SNPs identified by us are not unique; a different set of SNPs investigated may yield a different set of keystone SNPs. The well-validated *OPRM1* SNP (rs1799971) that influences binding of β-endorphin to μ-opioid receptor (van den *et al.* 2007) and the SNP (rs1056827) in the drug metabolizing enzyme *CYP1B1* (Hanna *et al.* 2000) distinguished most of the LPs from IPs (figure 4).

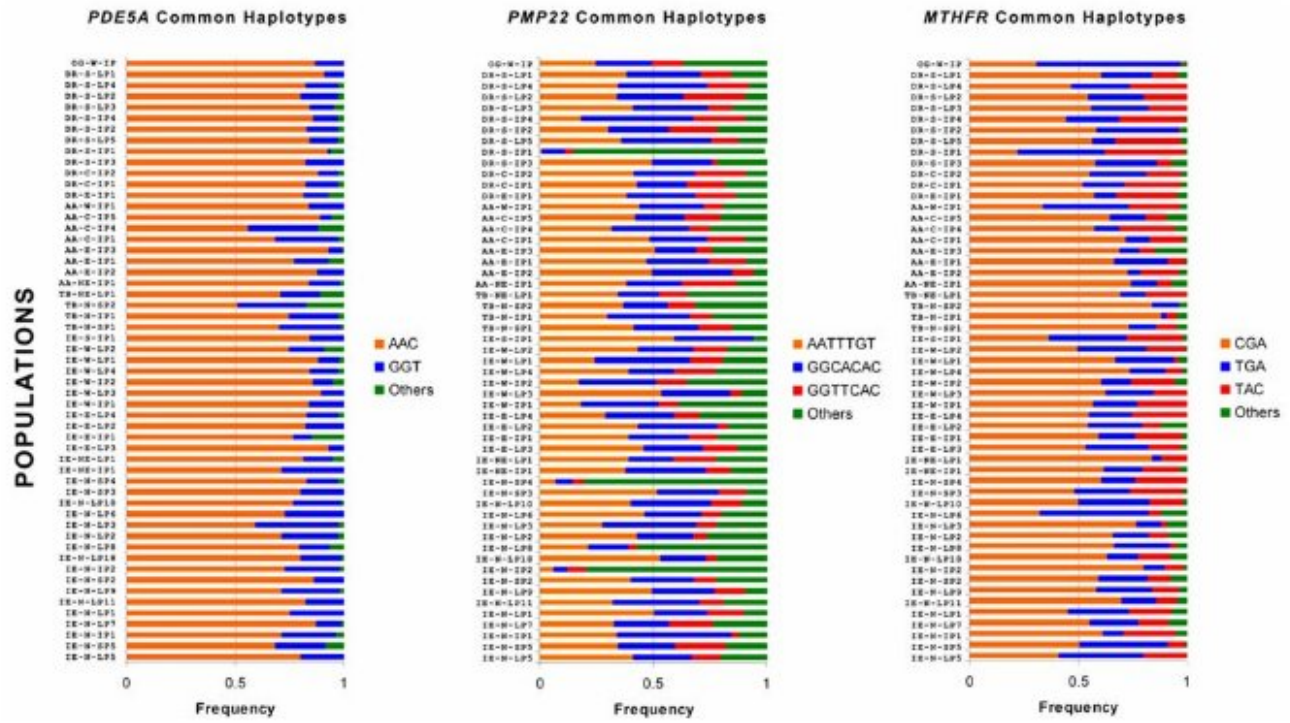

Figure 5. Haplotype sharing among Indian populations. Three examples of distribution of major haplotypes frequency > 0.05 across all populations. All the haplotypes with frequency < 0.05 have been pooled as others.

Haplotype diversity

We also estimated the extent of haplotype diversity across populations since this has relevance to complex disease gene identification. Haplotypes were reconstructed for a diverse set of 21 genes in each population and mean haplotype diversities were calculated. Mean haplotype diversities across populations were observed to be high for the majority of the genes (see table 8 in electronic supplementary material), except for phosphodiesterase gene *PDE5A*, for which the estimate of haplotype diversity was < 0.5. The extent of variation in haplotype diversity across populations was low for most genes. Despite the presence of highly differentiating SNPs, most of the genes had two to five major haplotypes (with frequencies > 0.05) shared across all populations (figure 5).

Indian genome variation in a global perspective

Considering the heterogeneity described above as well as the absence of any Indian population in the HapMap dataset, we assessed the proximity of populations included in the HapMap study with Indian populations using SNPs for which allele frequencies were available in both the Indian and HapMap populations (see table 4 in electronic supplementary material). The relatedness of the Indian to HapMap was estimated using D_A distance and PCA (figure 6, a&b). The first two principal components (PCs) explained about 31% of the variation in allele frequencies. The isolated populations of the Himalayan belt (figure 3) were closest to the Chinese (CHB) and Japanese (JPT) populations, and separated out from the rest of the populations in PC1. As expected, YRI, a population of African descent was an outlier and closest to OG-W-IP and CEU was most proximal to the IE–LPs, the majority of which were from north India. The AA and DR speaking populations, predominantly from the tribal belt and inhabiting the central and southern regions of India were distinct from HapMap populations (figure 6). 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 be difficult to directly use the HapMap data to design genetic epidemiological studies for entire population of India.

The differential affinity of the Indian populations to the various HapMap populations is also pertinent to the choice of tagSNPs (tSNPs) identified from the HapMap database for genetic epidemiological studies and design of genomewide association studies in India. To estimate differences in LD between Indian and HapMap populations, we compared the mean r^2 values between tSNPs chosen from each of the HapMap populations in the Indian populations. We chose the 5.2 Mb contiguous stretch in chromosome 22 spanning 49 genes which also harbours the schizophrenia and bipolar disorder susceptibility locus. The number of SNPs relevant to this analysis was 66 (MAF ≥ 0.05 in all the 55 Indian populations). Of these, the number of SNPs common with the HapMap populations was between 44 and 51, from which tSNPs were identified in each of the HapMap populations.

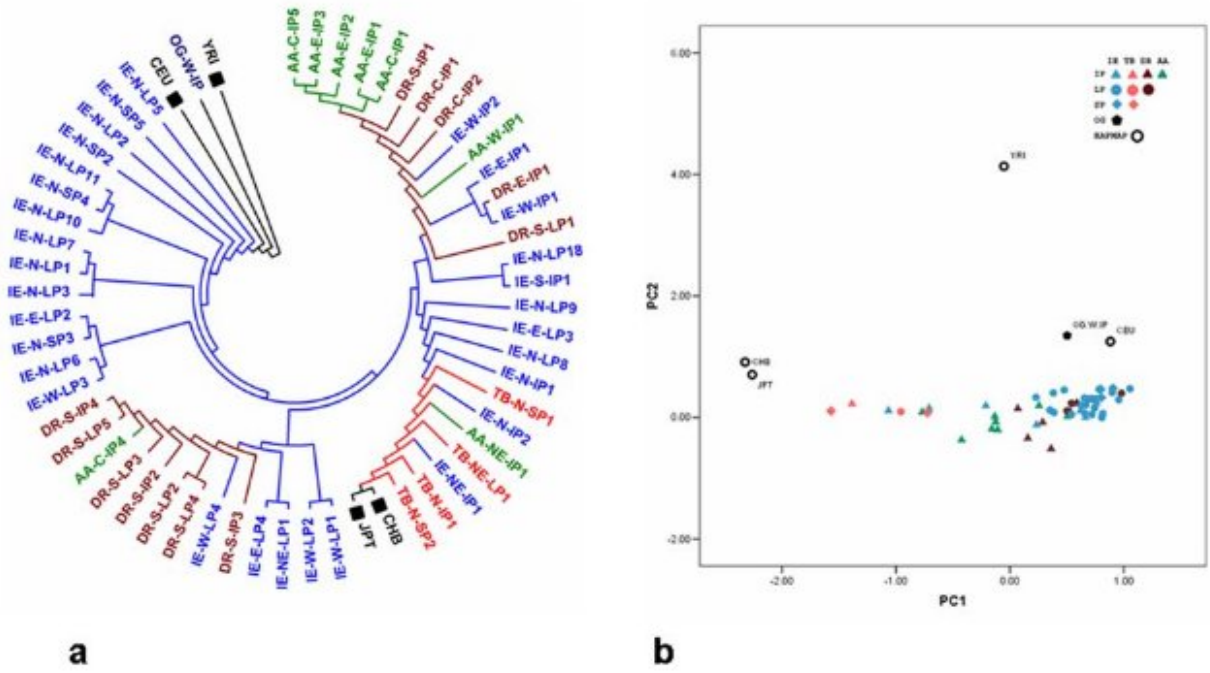

Figure 6. Relatedness between Indian and HapMap populations. (a) Neighbour joining tree based on Nei's D_A distance and (b) Principal component analysis computed using 230 shared SNPs depict affinities between Indian and HapMap populations. OG-W-IP1 is the out-group population of known African ancestry. Colour key for the populations and their ethno-linguistic affinities are provided in the figure.

We calculated the mean r^2 between adjacent tSNPs (Carlson *et al*. 2004) chosen in a specific HapMap population and also computed it for each of the 55 populations included in this study (figure 7). The mean r^2 values between tSNPs selected from all the four HapMap populations were found to be higher in most of the Indian population, implying that Indian populations in general have larger LD regions than HapMap populations. As expected, mean r^2 values for the tSNPs selected from YRI were observed to be the highest in all the Indian populations including OG-W-IP, a population of known African ancestry. Even with respect to CHB and JPT the mean r^2 values were higher in the Indian population albeit less strikingly than YRI. However, with respect to CEU no such consistent pattern was observed. A few populations like IE-N-LP8, IE-N-SP4, DR-C-IP1 and IE-E-IP1 where the mean r^2 values of adjacent tSNPs were comparatively lower than in most of the HapMap populations, had indications of admixture (figure 3a). Higher LD between HapMap tSNPs suggests the potential for LD-based disease gene mapping in some of the Indian populations. A more rigorous analysis of tag transferability to substantiate these observations on a larger dataset is underway.

Indian genome variation data: distribution of functional polymorphisms

Analysis of distribution of the functional alleles, which consistently show association in diseases across studies in different populations, also provides information for future validation studies in India. This would also be useful for identification of appropriate cohorts for pooling samples. In this context, we describe the distribution of some validated functional polymorphisms across Indian populations, taking specific examples.

Trends of selection: identifying populations for genotype–phenotype correlations

A SNP in the *MTHFR* gene (rs1801133, Ala222Val) along with folate and vitamin B12 deficiencies is a key factor that elevates levels of homocysteine. This SNP lies in the catalytic domain of the enzyme and in heterozygous (CT) and homozygous (TT) individuals the enzyme activity is reduced by about 35% and 70%, respectively (Weisberg *et al*. 1998). The *MTHFR* C677T homozygous genotype has been associated with premature coronary artery disease and also with neural tube defects, pre-eclampsia and other complications of pregnancy especially in conjunction with folate deficiency. In the Indian population, the overall MAF of this SNP was found to be 0.14, considerably lower than that reported for CEU (0.24), CHB (0.51) and JPT (0.36) and close to YRI (0.11) population (see table 6 in electronic supplementary material). Only 3% of the subjects genotyped had the homozygous variant (TT) and this variant was not observed in 29 out of the 55 populations studied. The homozygous mutant genotype was most prevalent in the TB group followed by IE of north, DR and AA populations (figure 8).

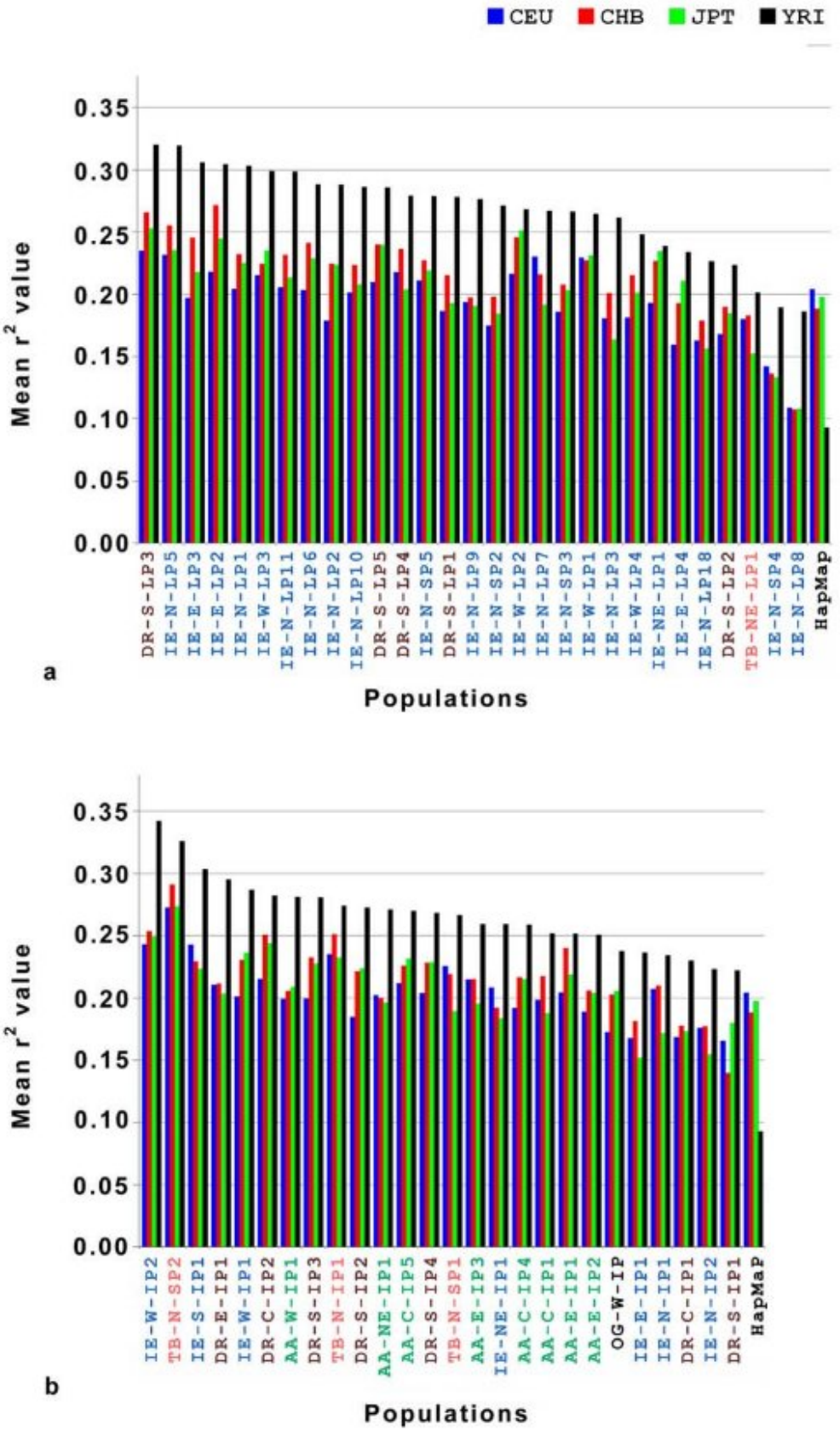

Figure 7. Extent of linkage disequilibrium between adjacent HapMap tagSNPs in Indian populations. Mean r^2 values between adjacent tSNPs chosen from each of the four HapMap populations are plotted in: (a) LP and (b) IP populations of India.

HIV susceptibility and the CCR5Δ32 mutation

A 32 bp deletion (*CCR5*Δ32, rs333) in the *CCR5* gene has been found to protect individuals against HIV infection. The frequency of *CCR5*Δ32 was extremely low in the Indian population [pooled allele frequency 0.01 (see table 6 in electronic supplementary material) with the maximum of 5.8% in a north Indian IE population] compared to the Caucasian population (16%). Only a cluster of populations from IE-N, IE-W and TB were found to have a moderate frequency while it was completely absent in IE-NE, IE-E, AA and DR populations (except from DR-S-LP4) (figure 9a). Thus, there is a high-to-low gradient from north to south. These results are consistent with (i) the observations made earlier by Majumder and Dey (2001), and (ii) the 2005 antenatal clinical HIV prevalence survey that reports a high frequency of HIV in south Indian populations (Steinbrook 2007). The allele frequencies of the Δ32 mutation presented in diverse populations of India may, therefore, provide guidance to future studies seeking to examine the nature and extent of correlation between *CCR5*Δ32 genotype and HIV infection.

Mapping populations for adverse drug response

The β2-adrenergic receptor (*ADRB2*) is the target for β-2-agonist drugs used for bronchodilation in asthma and other respiratory diseases. Detailed functional analysis of SNPs has clearly suggested that some variants of *ADRB2* may act as disease modifiers in asthma or may be the basis for known interindividual variation in the bronchodilating response to β-agonists (Drysdale *et al*. 2000; Israel *et al*. 2004). In an earlier study, a strong allelic/genotypic association of a nsSNP (rs1042713; p.R16G) with response to salbutamol in the Indian population (Kukreti *et al*. 2005) was observed. Though, locus-wise F_{ST} analysis did not reveal high differentiation, a difference in frequency of the risk genotype was observed in a few Indian populations (figure 9b). The extremes were observed in DR-S-LP3 and AA-C-IP4, that had the highest (69%) and lowest (4.8%) frequencies of the AA genotype. These data provide a framework for designing future epidemiological studies to identify populations with differential response to a given drug or a class of drugs, that is potentially useful in pharmacogenomics and personalized medicine.

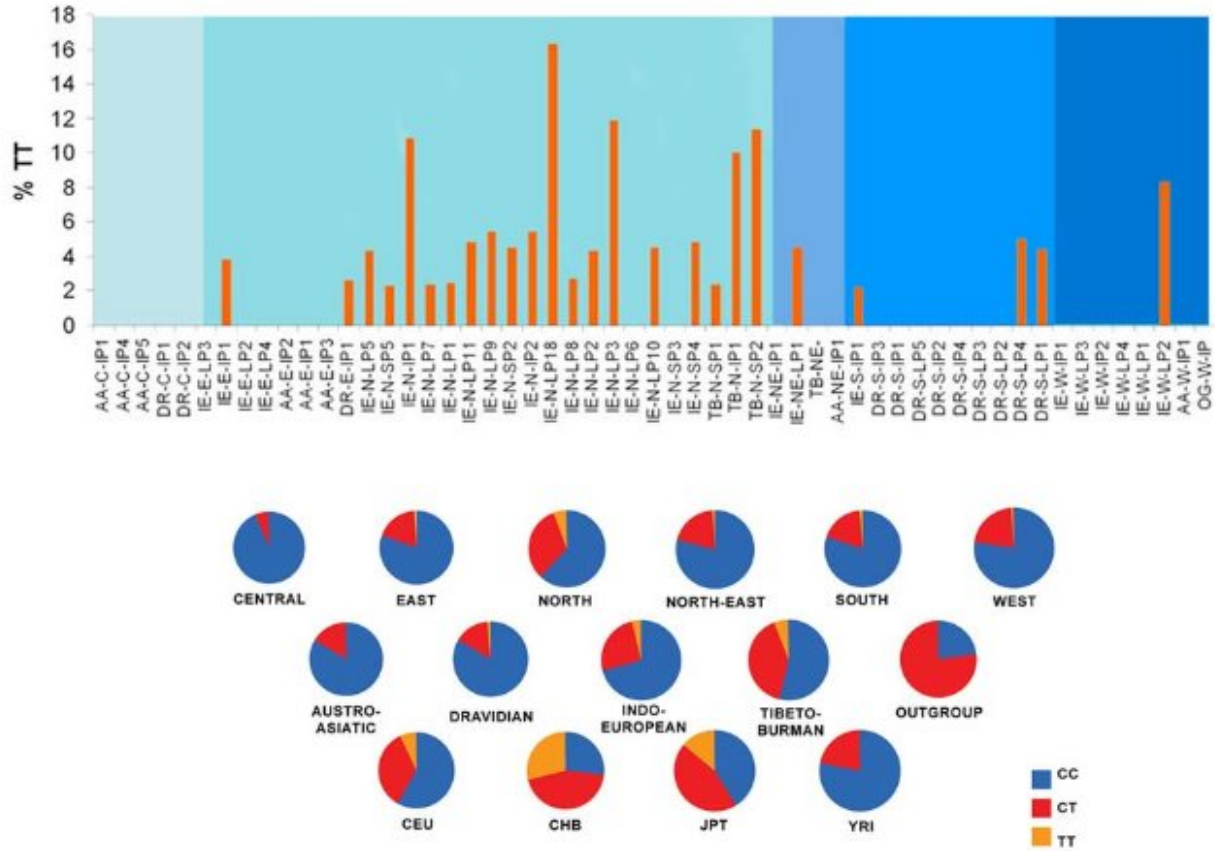

Figure 8. Distribution of *MTHFR* polymorphism. Frequencies of the TT genotype of *MTHFR* nsSNP (rs1801133, C/T, p.A222V) across populations from different geographical regions of India i.e., C, central; E, east; N, north; NE, north-east; S, south and W, west are shown and distinguished by a graded colour scheme. Genotype distributions of SNPs in different population groups are illustrated in the form of pie-chart by the colour scheme shown.

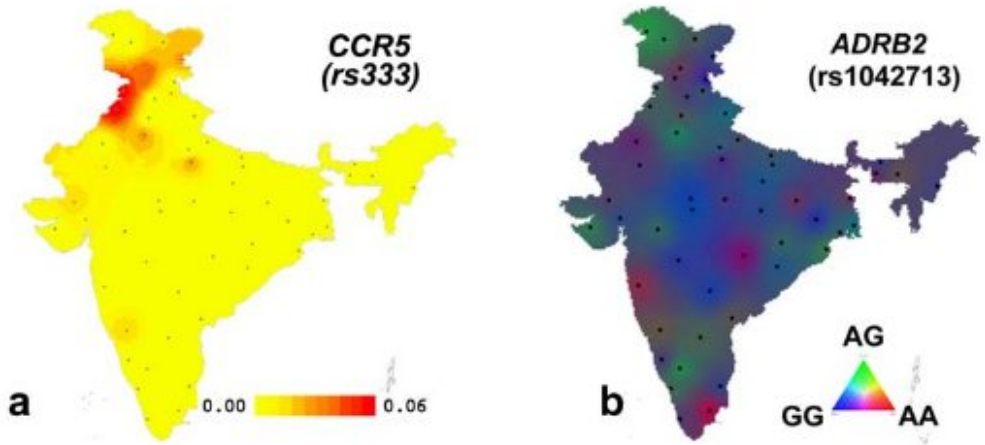

Figure 9. Distribution of SNP associated with (a) susceptibility to HIV-spatial frequency map of *CCR5* (rs333). The colour gradient depicted below the map from minimum to maximum frequency observed. (b) Response to salbutamol-colour (RGB) composite map of genotype frequency of rs1042713, a non-synonymous A/G variant (p.R16G) in the *ADRB2* gene. Variants of this SNP have been shown to confer different responsiveness to salbutamol in world populations. Red, genotype frequency of AA (poor responder) and blue, genotype frequency of GG (good responder); green, genotype frequency of AG.

Discussion

This is the largest study conducted on genomic variation in India in terms of its population and genomic coverage. The study included 32 large populations (of sizes >10 million) and 23 isolated tribal populations, representing a vast ethnic, linguistic and geographical diversity of India and provides data on the nature and extent of variation pertaining to a large number of genes and a genomic region related to disease susceptibility and response to drugs. Our study reveals a high degree of genetic differentiation among Indian ethnic groups and suggests that pooling of endogamous populations without regard to ethno-linguistic factors will result in false inferences in association studies. We note that the people of India are referred as 'Indian' in many population genetic studies. The implication of such usage is that the Indian population is genetically homogeneous, which, as the results of our study indicate, is evidently not true. However, we have also shown that it is possible to identify large clusters of ethnic groups that have substantial genetic homogeneity. Additionally, the SStr approach has indicated levels of admixture as well as assigned group memberships to populations, enabling us to identify a reduced number of reference populations for future disease-association studies. Our results also enable identification of population groups from which cases and controls may be sampled, and their data analysed in genomewide association studies without additional corrections or confounding effects of population stratification, thereby increasing the power of association studies. This is of paramount importance, because it is difficult to gather sufficient number of cases from individual isolated populations to obtain the required statistical power. We also observed a number of SNPs, in some cases from within the same gene, which belonged to extremes of H_o and F_{ST} distribution. If tested against a background of neutral variations to determine selection, such analyses may help prioritize disease candidates. As revealed from our study, Indian populations thus form a continuum of genetic spectrum bridging CEU and JPT/CHB, the two distinct HapMap populations. The observed affinities with the HapMap populations coupled with the highly endogamous nature of some Indian populations provide a potential resource of cohorts for coarse as well as fine mapping of disease genes. It is anticipated that the Indian Genome Variation data along with epidemiological and associated phenotype data will help in the construction of specific drug response/disease predisposition maps to aid policy level decision making for drug dosage interventions and disease risk management especially for complex as well as infectious diseases.

Additional data are available in The Indian Genome Variation database (http://igvdb.res.in).

Appendix

The Indian Genome Variation Database Consortium authors' list with their contributions.

Core manuscript writing group

Samir K. Brahmachari[1], Partha P. Majumder[7], Mitali Mukerji[1], Saman Habib[4], Debasis Dash[1], Kunal Ray[3,1] and Samira Bahl[1]

Project conceptualization

Samir K. Brahmachari[1,*] and Lalji Singh[2,*]

Project planning

Samir K. Brahmachari[1], Abhay Sharma[1,*], Mitali Mukerji[1,*], Kunal Ray[3,*], Susanta Roychoudhury[3,*], Lalji Singh[2], G. R. Chandak[2,*], K. Thangaraj[2,*], Saman Habib[4,*], D. Parmar[5,*] and Partha P. Majumder[7,*]

Implementation
Samir K. Brahmachari[1], Mitali Mukerji[1], Shantanu Sengupta[1,*], Dwaipayan Bharadwaj[1,*], Abhay Sharma[1], Debasis Dash[1,*], Kunal Ray[3], Susanta Roychoudhury[3], G. R. Chandak[2], Saman Habib[4], Srikanta K. Rath[4,*], D. Parmar[5] and Jagmohan Singh[6,*]

Population identification
Partha P. Majumder[7], Ganga Nath Jha[1], Komal Virdi[1], Samira Bahl[1], Mitali Mukerji[1], Samir K. Brahmachari[1], V. R. Rao[2], K. Thangaraj[2], Saman Habib[4], Srikanta K. Rath[4], Swapnil Sinha[4], Ashok Singh[4], Amit K. Mitra[4], Shrawan K. Mishra[4] and D. Parmar[5]

Sample collection and community engagement
Ganga Nath Jha[1], Shantanu Sengupta[1], Dwaipayan Bharadwaj[1], Mitali Mukerji[1], Qadar Pasha[1,*], Abhay Sharma[1], Sridhar Sivasubbu[1], Samira Bahl[1], Komal Virdi[1], Rajesh Pandey[1], Aradhita Baral[1], Prashant K. Singh[1], Amitabh Sharma[1], Jitender Kumar[1], Tsering Stobdan[1], Yasha Bhasin[1], Chitra Chauhan[1], Ashiq Hussain[1], Elyanambi Sundaramoorthy[1], S. P. Singh[1]; Arun Bandyopadhyay[3*], Susanta Roychoudhury[3], Krishanu Dasgupta[3], Lalji Singh[2], K. Thangaraj[2], G. R. Chandak[2], A. K. Reddy[2], Charles J Spurgeon[2], M. Mohd Idris[2], Saman Habib[4], Srikanta K. Rath[4], Swapnil Sinha[4], Ashok Singh[4], Shrawan K. Mishra[4], Amit K. Mitra[4], Vinay Khanna[5], Alok Dhawan[5], Mohini Anand[5], R. Shankar[5,*], R. S. Bharti[5], Madhu Singh[5], Arvind P. Singh[5], Anwar J. Khan[5], Parag P. Shah[5], A. B. Pant[5], D. Parmar[5], R. S. Bharti[5], Jagmohan Singh[6], Rupinder Kaur[6], Kamlesh K. Bisht[6], Ashok Kumar[6], Victor Rajamanickam[8], Eugene Wilson[8] and Antony Thangadurai[8]

Sample management
Mitali Mukerji[1], Rajesh Pandey[1], Aradhita Baral[1], Samira Bahl[1], Komal Virdi[1], Pankaj K. Jha[1], Mahua Maulik[3], Kunal Ray[3], Susanta Roychoudhury[3], G. R. Chandak[2] and K. Thangaraj[2]

Sequencing, genotyping and primer synthesis
Mitali Mukerji[1], Prashant K. Singh[1], Shantanu Sengupta[1], Neelam Makhija[1], Abdur Rahim[1], Sangeeta Sharma[9], Rupali Chopra[9], Pooja Rana[9], M. Chidambaram[9], Arindam Maitra[9], Ruchi Chawla[9], Suruchika Soni[9], Preeti Khurana[9], Mohamed Nadeem Khan[9], Sushanta Das Sutar[9], Amit Tuteja[9], K. Narayansamy[9,*]; Rachna Shukla[2], Swami Prakash[2], Swapna Mahurkar[2], K Radha Mani[2], J. Hemavathi[2], Seema Bhaskar[2], Pankaj Khanna[2], G. S. Ramalakshmi[2], Shalini Mani Tripathi[2], Nikita Thakur[2]; Swapnil Sinha[4], Ashok Singh[4], Shrawan K. Mishra[4] and Amit K. Mitra[4]

SNP discovery, validation and analysis
Mitali Mukerji[1], Dwaipayan Bharadwaj[1], Balaram Ghosh[1,*], Shantanu Sengupta[1], Qadar Pasha[1], Abhay Sharma[1], Ritushree Kukreti[1,*], Taruna Madan[1,*], Samir K. Brahmachari[1], Samira Bahl[1], Chitra Chauhan[1], Ranjana Verma[1], Prashant K. Singh[1], G. Sudheer[1], Jitender Kumar[1], Anubha Mahajan[1], Sreenivas Chavali[1], Rubina Tabassum[1], Sandeep Grover[1], Meenal Gupta[1], Jyotsna Batra[1], Amrendra Kumar[1], Pankaj K. Jha[1], Tsering Stobdan[1], Abdoulazim Nejatizadeh[1], Mudit Vaid[1], Swapan K. Das[1], Shilpy Sharma[1], Mamta Sharma[1], Rajshekhar Chatterjee[1], Jinny A. Paul[1], Pragya Srivastava[1], Rupali Chopra[1], Aradhita Baral[1], Charu Rajput[1], Uma Mittal[1], Mridula Singh[1], Manoj Hariharan[1]; Kunal Ray[3], Susanta Roychoudhury[3], Sumantra Das[3,*], Keya Chaudhuri[3,*], Mahua Maulik[3], Mainak Sengupta[3], Moulinath Acharya[3], Ashima Bhattacharyya[3], Atreyee Saha[3], Arindam Biswas[3], Moumita Chaki[3], Arnab Gupta[3], Saibal Mukherjee[3], Suddhasil Mookherjee[3], Ishita Chattopadhyay[3], Taraswi Banerjee[3], Meenakshi Chakravorty[3], Chaitali Misra[3], Gourish Monadal[3], Shiladitya Sengupta[3], Dipanjana Dutta De[3], Swati Bajaj[3], Ishani Deb[3], Arunava Banerjee[3], Rajdeep Chowdhury[3], Debalina Banerjee[3], Krishanu Dasgupta[3], Deepak Kumar[3], Sumit Ranjan Das[3]; G. R. Chandak[2], Shrish Tiwari[2], Anshu Bharadwaj[2], Rachna Shukla[2], Swami Prakash[2]; Saman Habib[4], Srikanta K. Rath[4], Swapnil Sinha[4], Ashok Singh[4], Shrawan K. Mishra[4], Amit K. Mitra[4]; Sangeeta Sharma[9], M. Chidambaram[9] and Rupali Chopra[9]

Data QC
Debasis Dash[1], Mitali Mukerji[1], Prashant K. Singh[1], Samira Bahl[1], Sangeeta Khanna[1], Rajesh Pandey[1], Ikhlak Ahmed[1], Pankaj K. Jha[1], Sumera Parveen[1], Nivedita Singh[1], Samir K. Brahmachari[1], G. R. Chandak[2], Rachna Shukla[2], Swami Prakash[2]; Mahua Maulik[3] and Kunal Ray[3].

IGV data handling, database and portal development
Debasis Dash[1], Mitali Mukerji[1], Dipayan Dasgupta[1], Siddharth Singh Bisht[1], Ikhlak Ahmed[1], Sangeeta Khanna[1], Rashmi Rajput[1], Biswaroop Ghosh[1], Naveen Kumar[1], Amit Chaurasia[1], Sumera Parveen[1], Nivedita Singh[1] and James K. Abraham[1]

Project management
Samir K. Brahmachari[1], Mitali Mukerji[1], Neelam Makhija[1] and S. P. Singh[1]

Global data analysis
Partha P. Majumder[7], Samir K. Brahmachari[1], Mitali Mukerji[1], Amit Sinha[1,*], Debasis Dash[1], Amit Chaurasia[1], Samira Bahl[1], Ikhlak Ahmed[1], Abhay Sharma[1], Prashant K. Singh[1], G. Sudheer[1], Rajesh Pandey[1], Yasha Bhasin[1], Vinod Scaria[1], Tav Pritesh Sethi[1], Amit K. Mandal[1], Arijit Mukhopadhyay[1], Saman Habib[4] and Swapnil Sinha[4]

[1]Institute of Genomics and Integrative Biology (CSIR), Mall Road, Delhi, 110 007, India
[2]Centre for Cellular and Molecular Biology (CSIR), Uppal Road, Hyderabad 500 007, India
[3]Indian Institute of Chemical Biology (CSIR), Kolkata 4, Raja S. C. Mullick Road, Kolkata 700 032, India
[4]Central Drug Research Institute (CSIR), Chattar Manzil Palace, P.B. No. 173, Lucknow 226 001, India
[5]Indian Toxicology Research Centre (CSIR), P.B. No. 80, Mahatma Gandhi Marg, Lucknow 226 001, India
[6]Institute of Microbial Technology (CSIR), Sector 39-A, Chandigarh 160 036 India
[7]Anthropology and Human Genetics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700 035, India
[8]Shanmugha Arts Science Technology and Research Academy, Thanjavur 613 402 India
[9]The Centre for Genomic Application (An IGIB-IMM Collaboration), 254 Ground Floor, Phase-III Okhla Industrial Estate, New Delhi 110020, India
*Principal Investigators

Acknowledgements

Financial support from Council of Scientific and Industrial Research (CSIR) Government of India, Task Force Project on 'Predictive Medicine using repeat and single nucleotide polymorphisms' (CMM0016); Department of Science and Technology (DST) for the TCGA facility and Department of Biotechnology (DBT) for 'Programme support on Functional Genomics to IGIB' is duly acknowledged. Consortium members would like to acknowledge the help of a large number of anthropologists and other members for help in community engagement and sample collection - Profs. P. K. Das and Jagannath Das, Dept. of Anthropology, Utkal University, Bhubaneshwar, Orissa; Thakur Prasad Murmu, Sunil Baraik, Ranchi; Prof. Jebon Singh, Dept. of Anthropology, Manipur University, Imphal; Dr A. V. Arakeri, Anthropological Survey of India (ASI), Udaipur; Dr F. A. Kulirani (ASI), Shillong; Dr B. N. Sarkar (ASI), Kolkata; Dr B. R. K. Shukla, Dept. of Anthropology, Lucknow University, Lucknow; Prof. Pradeep K. Singh, Dept. of Anthropology, Ranchi College, Jharkhand; Biren Hajong, Tura, Meghalaya; V. S. Upadhye and his team in LABINDIA, Shekhar Ranjan Ghoshal, Western Railways, Mumbai; Mr Rajesh Bhandari, Sundernagar; Devi Singh and Karan Thakur, Chambi village, Himachal; Nitin Maurya, Leena Kalla, Delhi University;

Rana Nagarkatti, Kamya Mehla, Tej P. Singh, S. Siva, Aarif Ahsan, Karamjit Singh Dolt, Chitra Chauhan, Mridula Singh, Dr Souvik Maiti, Inder Singh, Ravishankar Roy (IGIB); V. Prasad Kolla, V. N. S. Prathyusha, Inder Deo Mali (CCMB), Bipin C. Mishra, J. P. Srivastava, R. K Gupta (CDRI); Somnath Dutta, and Siddiq Sarkar (IICB), clinicians Drs. Saurabh Malhotra and Ajay Vidhani. For HTSSS implementation we would like to acknowledge 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) and Raghunath Chatterjee (IICB) for help in informatics support. Administrative support from Dr Rekha Chaturvedi, T. V. Joshua, V. P. Bharadwaj, Ajit Singh, Hemant Kulkarni, P. Bansal, (IGIB) and Arpita Sengupta and Pabitra Patnaik (TCGA) is duly acknowledged. Critical prereviewing and comments on the drafts of the manuscript from Dr Satyajit Rath, National Institute of Immunology, Delhi; Prof. Irfan Habib, Aligarh Muslim University and Prof. Vani Brahmachari, Ambedkar Centre for Biomedical Research, Delhi University are acknowledged. Authors also thank Dr Ashis K. Saha and Department of Geography, Delhi University, for help in map preparation and Drs R. A. Mashelkar, V. S. Ramamurthy, M. K. Bhan, G. Padmanabhan, V. C. Vora for support and encouragement. Authors are indebted to Myles Axton, Alan Packer and the anonymous referees of Nature Genetics for their constructive suggestions and several rounds of reviewing of the manuscript. But for their efforts, the manuscript would not be in the present form.

We would like to thank all the members who donated their blood samples and key respondents in villages who helped in community engagement, without their voluntary participation this work would not have been possible.

All genotype information are available online at The Indian Genome Variation database (http://igvdb.res.in).

References

Abecasis G. R., Ghosh D. and Nichols T. E. 2005 Linkage disequilibrium: ancient history drives the new genetics. *Hum. Hered.* **59**, 118–124.

Akey J. M., Zhang G., Zhang K., Jin L. and Shriver M. D. 2002 Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814.

Bamshad M., Kivisild T., Watkins W. S., Dixon M. E., Ricker C. E., Rao B. B. *et al.* 2001 Genetic evidence on the origins of Indian caste populations. *Genome Res.* **11**, 994–1004.

Basu A., Mukherjee N., Roy S., Sengupta S., Banerjee S., Chakraborty M. *et al.* 2003 Ethnic India: a genomic view, with special reference to peopling and structure. *Genome Res.* **13**, 2277–2290.

Carlson C. S., Eberle M. A., Rieder M. J., Yi Q., Kruglyak L. and Nickerson D. A. 2004 Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120.

Cordaux R., Aunger R., Bentley G., Nasidze I., Sirajuddin S. M. and Stoneking M. 2004 Independent origins of Indian caste and tribal paternal lineages. *Curr. Biol.* **14**, 231–235.

Drysdale C. M., McGraw D. W., Stack C. B., Stephens J. C., Judson R. S., Nandabalan K. *et al.* 2000 Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc. Natl. Acad. Sci. USA* **97**, 10483–10488.

Fogarty M. S., Sinha A. K., and Jaffe A. B. 2006, ATP and the U.S. Innovation System: A Methodology for Identifying Enabling R&D Spillover Networks Gaithersburg, NIST GCR 06–895.

Habib I. 2001 *People's History of India - Part 1: Prehistory* Aligarh Historians Society and Tulika Books, Aligarh.

Habib I. 2002 *People's History of India Part 2: The Indus Civilization.* Aligarh Historians Society and Tulika Books, Aligarh.

Hanna I. H., Dawling S., Roodi N., Guengerich F. P. and Parl F. F. 2000 Cytochrome P450 1B1 (CYP1B1) pharmacogenetics: association of polymorphisms with functional differences in estrogen hydroxylation activity. *Cancer Res.* **60**, 3440–3444.

Heutink P. and Oostra B. A. 2002 Gene finding in genetically isolated populations. *Hum. Mol. Genet.* **11**, 2507–2515.

Indian Genome Variation Consortium 2005 The Indian Genome Variation database (IGVdb): a project overview. *Hum. Genet.* **118**, 1–11.

Israel E., Chinchilli V. M., Ford J. G., Boushey H. A., Cherniack R., Craig T. J. *et al.* 2004 Use of regularly scheduled albuterol treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. *Lancet* **364**, 1505–1512.

Kashyap V. K., Guha S., Sitalaximi T., Bindu G. H., Hasnain S. E. and Trivedi R. 2006 Genetic structure of Indian populations based on fifteen autosomal microsatellite loci. *BMC Genet.* **7**, 28.

Kelsoe J. R., Spence M. A., Loetscher E., Foguet M., Sadovnick A. D., Remick R. A. *et al.* 2001 A genome survey indicates a possible susceptibility locus for bipolar disorder on chromosome 22. *Proc. Natl. Acad. Sci. USA* **98**, 585–590.

Kivisild T., Rootsi S., Metspalu M., Mastana S., Kaldma K., Parik J. *et al.* 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., Bhatnagar P., Rao C., Gupta S., Madan B., Das C. *et al.* 2005 Beta(2)-adrenergic receptor polymorphisms and response to salbutamol among Indian asthmatics. *Pharmacogenomics* **6**, 399–410.

Majumder P. P. and Dey B. 2001 Absence of the HIV-1 protective Delta ccr5 allele in most ethnic populations of India. *Eur. J. Hum. Genet.* **9**, 794–796.

Nei M. 1977 F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet.* **41**, 225–233.

Nei M. and Chesser R. K. 1983 Estimation of fixation indices and gene diversities. *Ann. Hum. Genet.* **41**, 253–259.

Papolos D. F., Faedda G. L., Veit S., Goldberg R., Morrow B., Kucherlapati R. *et al.* 1996 Bipolar spectrum disorders in patients diagnosed with velo-cardio-facial syndrome: does a hemizygous deletion of chromosome 22q11 result in bipolar affective disorder? *Am. J. Psychiatry* **153**, 1541–1547.

Peltonen L. 2000 Positional cloning of disease genes: advantages of genetic isolates. *Hum. Hered.* **50**, 66–75.

Piazza A., Menozzi P. and Cavalli-Sforza L. L. 1980 The HLA-A,B gene frequencies in the world: migration or selection? *Hum. Immunol.* **1**, 297–304.

Pritchard J. K., Stephens M. and Donnelly P. 2000 Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959.

Rao C. R. 1952 p. **390** Advanced Statistical Methods in Biometric Research, John Wiley, New York.

Rosenberg N. A., Mahajan S., Gonzalez-Quevedo C., Blum M. G., Nino-Rosales L., Ninis V. *et al.* 2006 Low Levels of Genetic Divergence across Geographically and Linguistically Diverse Populations from India. *PLoS Genet.* **2**, 215.

Roychoudhury S., Roy S., Basu A., Banerjee R., Vishwanathan H., Usha Rani M. V. *et al.* 2001 Genomic structures and population histories of linguistically distinct tribal groups of India. *Hum. Genet.* **109**, 339–350.

Sahoo S., Singh A., Himabindu G., Banerjee J., Sitalaximi T., Gaikwad S. *et al.* 2006 A prehistory of Indian Y chromosomes: evalu-

ating demic diffusion scenarios. *Proc. Natl. Acad. Sci. USA* **103**, 843–848.

Saitou N. and Nei M. 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Schwab S. G. and Wildenauer D. B. 1999 Chromosome 22 workshop report. *Am. J. Med. Genet.* **88**, 276–278.

Sengupta S., Zhivotovsky L. A., King R., Mehdi S. Q., Edmonds C. A., Chow C. E. *et al*. 2006 Polarity and temporality of high-resolution Y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of central asian pastoralists. *Am. J. Hum. Genet.* **78**, 202–221.

Shriver M. D. and Kittles R. A. 2004 Genetic ancestry and the search for personalized genetic histories. *Nat. Rev. Genet.* **5**, 611–618.

Singh K. S. 2002 *People of India*: *introduction national series*. Oxford University Press, Delhi.

Sinha A. K. 2001 A Fuzzy Measure Theoretic Quantum Approximation of an Abstract System, Case Western Reserve University Cleveland, Ohio.

Sinha A. K., Richoux W. J. and Loparo K. A. 2004 A system-theoretic state description for temporal transitions in electroencephalogram data of severe epileptic patients, 43rd IEEE Conference on Decision and Control, Atlantis, Paradise Island, Bahamas.

Steinbrook R. 2007 HIV in India–a complex epidemic. *N. Engl. J Med.* **356**, 1089–1093.

Stephens M., Smith N. J. and Donnelly P. 2001 A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* **68**, 978–989.

Thanseem I., Thangaraj K., Chaubey G., Singh V. K., Bhaskar L. V., Reddy B. M. *et al*. 2006 Genetic affinities among the lower castes and tribal groups of India: inference from Y chromosome and mitochondrial DNA. *BMC Genet.* **7**, 42.

Thapar R. 1966 *A history of India*. Penguin Books, London.

The International HapMap Consortium 2003 The International HapMap Project. *Nature* **426**, 789–796.

Tishkoff S. A. and Verrelli B. C. 2003 Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* **4**, 293–340.

van den W. E., Wiers R. W., Dessers J., Janssen R. G., Lambrichs E. H., Smeets H. J. *et al*. 2007 A functional polymorphism of the mu-opioid receptor gene (OPRM1) influences cue-induced craving for alcohol in male heavy drinkers. *Alcohol Clin. Exp. Res.* **31**, 1–10.

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

Verma R., Kubendran S., Das S. K., Jain S. and Brahmachari S. K. 2005 SYNGR1 is associated with schizophrenia and bipolar disorder in southern India. *J. Hum. Genet.* **50**, 635–640.

Watkins W. S., Rogers A. R., Ostler C. T., Wooding S., Bamshad M. J., Brassington A. M. *et al*. 2003 Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. *Genome Res.* **13**, 1607–1618.

Watkins W. S., Prasad B. V., Naidu J. M., Rao B. B., Bhanu B. A., Ramachandran B. *et al*. 2005 Diversity and divergence among the tribal populations of India. *Ann. Hum. Genet.* **69**, 680–692.

Weisberg I., Tran P., Christensen B., Sibani S. and Rozen R. 1998 A second genetic polymorphism in methylenetetrahydrofolate reductase (MTHFR) associated with decreased enzyme activity. *Mol. Genet. Metab.* **64**, 169–172.

Wright A. F., Carothers A. D. and Pirastu M. 1999 Population choice in mapping genes for complex diseases. *Nat. Genet.* **23**, 397–404.

Received 21 February 2008, in revised form 28 February 2008; accepted 28 February 2008
Published on the Web: 9 April 2008