RESEARCH ARTICLE

Utilizing linkage disequilibrium information from Indian Genome Variation Database for mapping mutations: SCA12 case study

SAMIRA BAHL[1], IKHLAK AHMED[2], THE INDIAN GENOME VARIATION CONSORTIUM[3] and MITALI MUKERJI[1]

[1]*Functional Genomics Unit, Institute of Genomics and Integrative Biology (CSIR), Mall Road, New Delhi, 110 007, India*
[2]*G. N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology (CSIR), Mall Road, New Delhi, 110 007, India*
[3]*Composition first described in Hum. Genet. 2005,* ***118****, 1–11*

Abstract

Stratification in heterogeneous populations poses an enormous challenge in linkage disequilibrium (LD) based identification of causal loci using surrogate markers. In this study, we demonstrate the enormous potential of endogamous Indian populations for mapping mutations in candidate genes using minimal SNPs, mainly due to larger regions of LD. We show this by a case study of the *PPP2R2B* gene (~400 kb) that harbours a CAG repeat, expansion of which has been implicated in spinocerebellar ataxia type 12 (SCA12). Using LD information derived from Indian Genome Variation database (IGVdb) on populations which share similar ethnic and linguistic backgrounds as the SCA12 study population, we could map the causal loci using a minimal set of three SNPs, without the generation of additional basal data from the ethnically matched population. We could also demonstrate transferability of tagSNPs from a related HapMap population for mapping the mutation.

[Bahl S., Ahmed I., Indian Genome Variation Consortium and Mukerji M. 2009 Utilizing linkage disequilibrium information from Indian Genome Variation Database for mapping mutations: SCA12 case study. *J. Genet.* **88**, 55–60]

Introduction

The Indian population is extremely diverse due to the presence of different linguistic lineages (Grierson 1927), ethnically distinct tribal and large caste groups (Singh 1998), and varying extents of admixtures with different world populations (Habib 2001, 2002). However, stringent mating patterns have led to the existence of thousands of endogamous populations within this diverse pool. This provides an enormous resource for carrying out association studies and mapping mutations. Once a positional candidate is identified either through linkage or association, the next step is to identify the causal mutation responsible for the disease. A simpler approach is to carry out re-sequencing of the candidate gene in a number of cases and controls. However, given the average size of a gene is ~ 25 kb, this step is extremely time consuming and cost intensive. With the availability of a large amount of SNP information, one could use linkage disequilibrium (LD) for demarcating minimal regions that could harbour causal mutations. This approach is expected to be more successful in endogamous populations where fewer markers are needed to map the causal region owing to higher extents of LD. The Indian Genome Variation database (IGVdb) provides polymorphism data on a large number of disease candidate genes from a large number of ethnically and linguistically diverse Indian populations belonging to different geographical regions of habitat (Indian Genome Variation Consortium 2005, 2008). If the relatedness of the study populations to the IGV populations is established, it is possible that tagSNPs from these reference datasets can be used for mapping mutations using LD-based methods without additional generation of data.

In this study, we demonstrate how LD information from basal population variation data from both IGV (http://www.igvdb.res.in) and HapMap (http://www.hapmap.org) databases can be effectively utilized to demarcate causal mutations. We used information on the genetic relatedness between the Indian populations and that with the HapMap populations from a recently reported analysis by the IGV

*For correspondence. E-mail: mitali@igib.res.in.

Keywords. Indian genome variation; SNP; linkage disequilibrium; mutation mapping; SCA12; HapMap.

consortium (Indian Genome Variation Consortium 2008). We tested the applicability of this approach taking the example of a monogenic, autosomal dominant disorder, spinocerebellar ataxia type 12 (SCA12) (Holmes *et al.* 1999), which has previously been shown to share a common founder in an Indo–European speaking endogamous population from north India by our group (Bahl *et al.* 2005). Such an approach is likely to be useful in reducing the amount of re-sequencing efforts required to identify the causal mutations.

Materials and methods

Subjects

As a part of the ongoing IGV consortium effort, 2014 control samples drawn from 55 different Indian populations have been genotyped for SNPs in the *PPP2R2B* gene. These populations belong to different linguistic lineages, namely Indo–European (IE), Austro–Asiatic (AA), Tibeto–Burman (TB) and Dravidian (DR), and cover six major geographical zones of the Indian subcontinent, namely north, east, west, central, south and northeast (Indian Genome Variation Consortium 2008). There are 25 caste populations (large populations, LPs), 24 tribal populations (isolated populations, IPs) and six populations belonging to different religious groups (special populations, SPs). Data from these samples were used for LD studies.

In order to test the applicability of the IGV data for mutation mapping, we genotyped 19 SCA12 affected individuals from 18 unrelated affected families (Bahl *et al.* 2005). All these families were derived from a single endogamous IE population of the north (IE-N-LP18), which has also been included in IGV. Therefore, for the case–control study, we utilized data ($n = 111$) from 92 chromosomes genotyped in IGV along with chromosomes not carrying the expanded repeat from SCA12 families. DNA was isolated from peripheral blood leukocytes using the salting-out procedure (Miller *et al.* 1988).

SNP selection, genotyping and quality control

We genotyped 17 SNPs spanning the entire length of the *PPP2R2B* gene at an average spacing of 20 kb on 2014 controls described above as a part of the IGV consortium work. These 17 SNPs were also genotyped on all SCA12 probands using the Sequenom MassARRAY platform (Sequenom, San Diego, USA). All the SNPs were genotyped with ≥ 80% success. Hardy–Weinberg equilibrium (HWE) was tested for each locus individually in all the 55 populations. All SNPs were found to be in HWE at 1% level of significance in more than 90% of the populations.

Statistical analysis

Allele and genotype frequencies were computed by the gene-counting method. HWE was calculated using Fisher's exact test. LD measure (D′) was computed by performing pairwise comparisons for all SNP loci using the Haploview software, version 3.2 (Barrett *et al.* 2005). Haplotypes and their frequencies were statistically inferred from phase-unknown genotype data by use of the software PHASE version 2.1 (Stephens *et al.* 2001). We used parameter values of 100 iterations, a thinning interval of 10, and a burn-in value of 100 in the Markov chain Monte Carlo simulations. The distributions of allele and haplotype frequencies were compared in affected and normal individuals using Pearson's chi-squared (χ^2) test. TagSNPs were identified using Tagger (http://www.broad.mit.edu/mpg/tagger/) with pair-wise tagging at an r^2 cut-off of 0.8.

Results and discussion

The diverse populations represented in IGV form five broad, genetically homogeneous clusters, which group primarily on the basis of language and ethnicity (Indian Genome Variation Consortium 2008). Since SCA12 has been reported in a large caste population of Indo–European background, we considered samples from IE–LPs as control for this study. We used genotype information of *PPP2R2B* SNPs from IGVdb for carrying out association studies in SCA12. We selected six sets of three SNPs each (figure 1,a) based on the LD (D′) pattern observed in control samples pooled from the cluster of IE-LPs (figure 1,b). Of these, three SNP sets were in LD and the other three not in LD with each other (figure 1,a). The founder haplotype could be traced using all the six SNP sets (table 1), probably due to the monogenic nature of the disease and the highly endogamous nature of the population in which the disease prevails. We also tested the applicability of LD patterns derived from two contrasting clusters to look for association with the disease (data not shown). The common founder could be traced using information from both these clusters as well, again probably due to the highly endogamous nature of the SCA12 population. Also, some of the SNP sets showed near conserved LD patterns across different clusters.

Further, to evolve a strategy for tracing the causal mutation (CAG repeat in this case) in the ~400-kb *PPP2R2B* gene using LD information, we screened for a risk haplotype for the disease by determining the frequencies of the disease-associated haplotypes using the three SNP sets observed to be in LD in the IE–LPs. From these, the C-A-T haplotype defined by set B spanning 119 kb was extremely rare across all the 55 populations and hence defined a near-risk haplotype. Incidentally, this region spans the CAG repeat, expansion of which is implicated in disease. In contrast, the founder haplotype defined by set A, which is distant from the CAG repeat, was present in many control ethnic groups with considerable frequencies (figure 2). Thus, using SNPs in set B, we could localize the causal region from an entire gene of ~400 kb to ~119 kb. Genotyping with a subset of SNPs within this region may further reduce the length to be sequenced for identifying the causal mutation. These results thus demonstrate that common SNPs and LD patterns

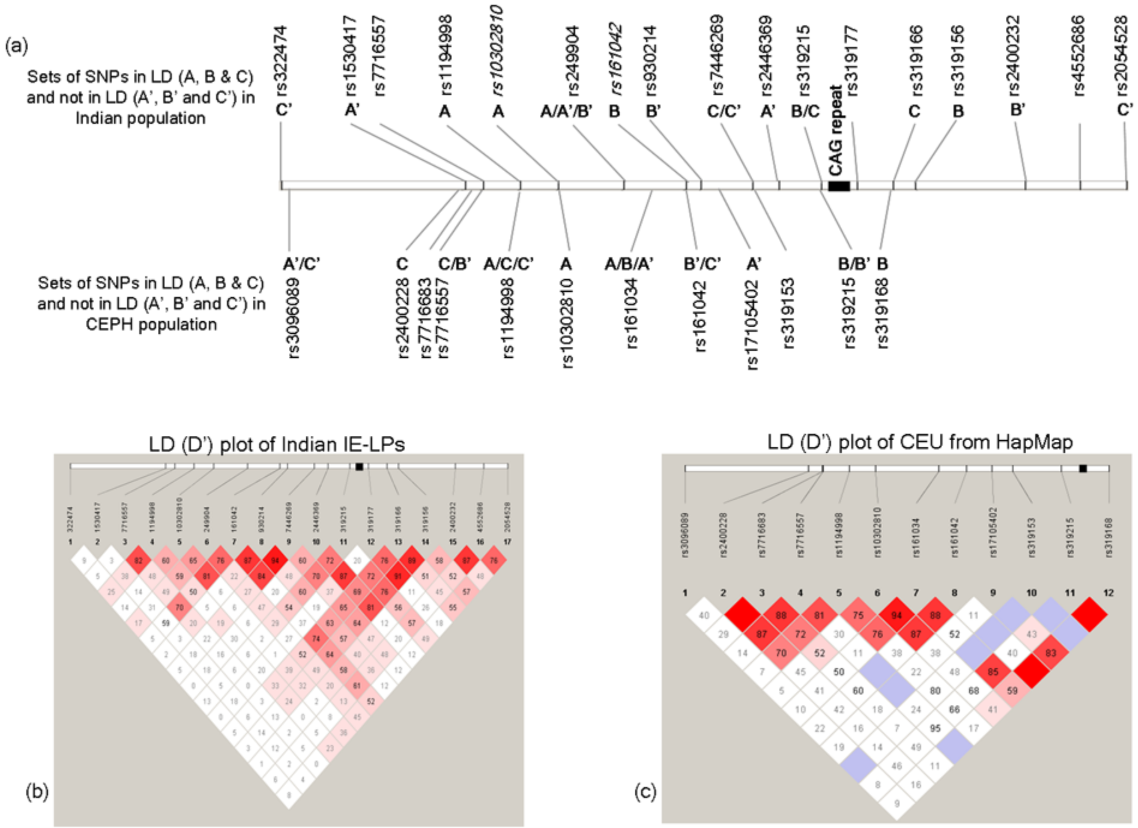

Figure 1. LD based marker selection for identification of common founder and causative mutation in SCA12. (a) Line diagram represents the SNPs selected in Indian (upper panel) and CEPH (lower panel) populations arranged in an increasing order of their chromosomal positions from left-to-right. The causative mutation (CAG repeat) is depicted by the filled box. Sets of SNPs in LD and not in LD are indicated by A, B, C and A′, B′, C′, respectively for both the Indian and CEU populations. (b) LD (D′) pattern of IE–LPs from India using 17 SNPs genotyped in Indian population. (c) LD (D′) pattern of CEU using 12 SNPs that serve as tagSNPs for those genotyped in (b). In case a SNP genotyped in Indian population was not tagged by any of the CEU SNPs, the nearest genotyped SNP was considered from the CEPH population. CAG repeat is denoted by filled boxes in (b) and (c). In the Haploview generated LD plots, red squares indicate statistically significant LD between the pairs of SNPs; dark red squares indicate higher values of D′, up to a maximum of 1. White squares indicate pair-wise D′ values < 1 with no statistically significance evidence of LD.

Table 1. Founder for SCA12 identified using different sets of markers selected on the basis of LD pattern of IE–LPs and CEPH population from HapMap.

Set	SNPs included	Haplotype	Patients (frequency)	Controls (frequency)	χ^2	*P*
Based on LD pattern of IE–LPs						
Sets of SNPs in LD with each other						
Set A	rs1194998, rs10302810, rs249904	T-G-T	19/19 (1.000)	34/111 (0.306)	32.329	0.0000
Set B	rs161042, rs319215, rs319156	C-A-T	17/19 (0.895)	0/111 (0.000)	114.257	0.0000
Set C	rs7446269, rs319215, rs319166	T-A-G	18/19 (0.947)	18/111 (0.162)	49.952	0.0000
Sets of SNPs not in LD with each other						
Set A′	rs1530417, rs249904, rs2446369	C-T-T	16/19 (0.842)	49/111 (0.441)	10.417	0.0012
Set B′	rs249904, rs930214, rs2400232	T-C-C	19/19 (1.000)	20/111 (0.180)	51.922	0.0000
Set C′	rs322474, rs7446269, rs2054528	C-T-A	18/19 (0.947)	54/111 (0.486)	13.946	0.0001
Based on LD pattern of CEU						
Sets of SNPs in LD with each other						
Set A	rs1194998, rs10302810, rs249904	T-G-T	19/19 (1.000)	34/111 (0.306)	32.329	0.0000
Set B	rs249904, rs319215, rs319156	T-A-T	17/19 (0.895)	1/111 (0.009)	106.692	0.0000

Table 1 *(contd.)*

Set C	rs1530417, rs7716557, rs1194998	C-A-T	16/19 (0.842)	40/111 (0.360)	15.354	0.0000
Sets of SNPs not in LD with each other						
Set A′	rs322474, rs10302810, rs930214	C-G-C	19/19 (1.000)	24/111 (0.054)	45.022	0.0000
Set B′	rs7716557, rs161042, rs319215	A-C-A	14/19 (0.737)	6/111 (0.216)	58.099	0.0000
		T-C-A	4/19 (0.211)	0/111 (0.000)	24.110	0.0000
Set C′	rs322474, rs1194998, rs161042	C-T-C	19/19 (1.000)	15/111 (0.135)	62.83	0.0000

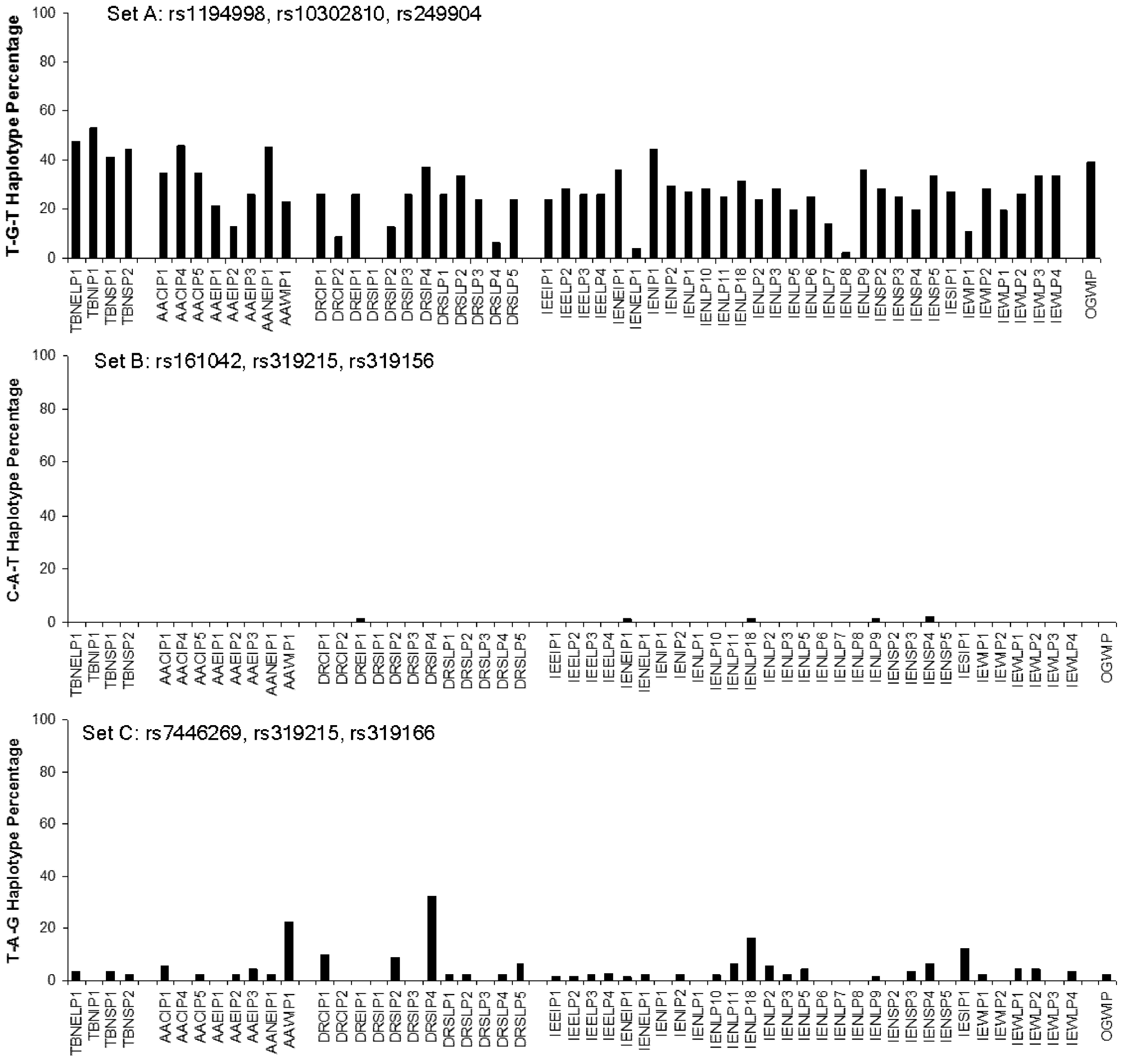

Figure 2. Applicability of LD pattern of IE–LPs in identifying SCA12 risk haplotype. The plots represent the frequency distributions of associated haplotypes for SCA12 across 55 Indian populations using marker sets selected based on the LD pattern of IE–LPs. The three sets comprised of SNPs in considerably high LD with each other. The SNP sets are as follows: set A (rs1194998, rs10302810, rs249904), set B (rs161042, rs319215, rs319156) and set C (rs7446269, rs319215, rs319166).

from genetically related populations can help to map mutations or associated haplotypes in endogamous Indian populations.

It is possible that variation data on all loci are not available in IGVdb. However, the Indian genome variation data provides information on its relatedness with the HapMap populations. We used this information to test the applicability of tagSNPs from HapMap for mutation mapping in Indian populations. We carried out an exploratory analysis of mapping the SCA12 causative CAG mutation, using information from the CEPH (CEU) population and genotype data of the 17 *PPP2R2B* SNPs available in IGVdb. We selected the CEU population since it has been shown to have affinity with the IE–LPs from northern India (Indian Genome Variation Consortium 2008), the cluster in which the disease population is represented. As a first step, we identified tagSNPs in the CEU population from an initial set of 298 SNPs (MAF ≥ 0.05) typed from *PPP2R2B* gene, at an average median spacing of ~0.6 kb. From these, SNPs, which served as tags for the 17 SNPs genotyped in the Indian populations, were selected. In case, if a SNP genotyped in the Indian population was not tagged by the CEU tagSNPs, the data from next proximal SNP in IGVdb was taken. As a result, 12 tagSNPs were selected from CEU at an average median spacing of ~20 kb (figure 2,a). Based on the LD pattern of CEU using these 12 tags (figure 2,c), six sets of markers, three in significant LD and three not in LD with each other were selected. The common founder could be identified using all SNP sets (table 1). The SNP set, which was in near complete LD and most proximal to the CAG repeat was the most informative (figure 3). This suggests that, based on LD pattern from a related HapMap population, it is possible to map mutations using minimal number of tag SNPs. This is especially applicable to populations with high endogamy, as is the case for Indian populations.

Thus, the highly endogamous nature of even large Indian populations provides an enormous resource for identifying risk haplotype/causative mutation for disease as exemplified by the SCA12 case study. Additionally, we could also demonstrate that sets of limited number of tagSNPs from related HapMap populations can be used for disease gene mapping. We would like to emphasize that the endogamous

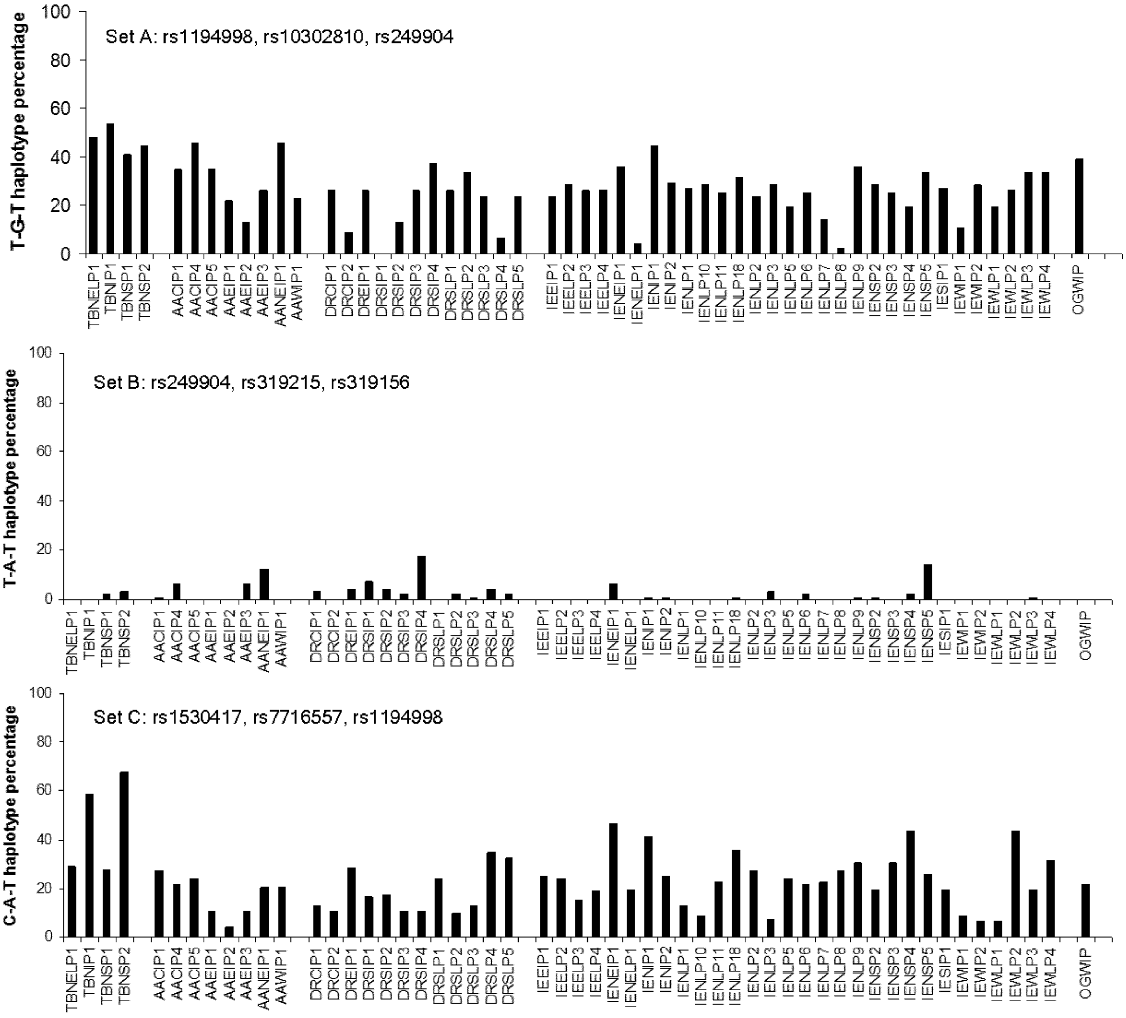

Figure 3. Applicability of HapMap tagSNPs in identifying SCA12 risk haplotype in Indian population. The plots represent the frequency distributions of associated haplotypes for SCA12 across 55 Indian populations using marker sets selected based on the LD pattern of HapMap CEPH population, which shows affinities with IE–LPs to which the diseased endogamous population belongs. The three sets comprised of SNPs (tagSNPs in CEU) in considerably high LD with each other in the CEPH population. The SNP sets are as follows: set A (rs1194998, rs10302810, rs249904), set B (rs249904, rs319215, rs319156) and set C (rs1530417, rs7716557, rs1194998).

nature of Indian populations is an additional advantage to this mapping strategy. The approach adapted in this study is especially relevant to other world populations, which are as diverse as the Indian population.

Acknowledgements

The authors thank S. K. Brahmachari, P. P. Majumder and D. Dash for valuable scientific inputs; D. Dasgupta, M. Singh, D. Dash, A. Chaurasia and B. Ghosh for computational assistance; S. Sharma, M. Chidambaram, R. Chopra and P. Rana from TCGA facility for SNP genotyping. This work was supported by grant from the Council of Scientific and Industrial Research (CSIR), Government of India, for the project 'Predictive medicine using repeat and single nucleotide polymorphisms' (CMM0016).

References

Barrett J. C., Fry B., Maller J. and Daly M. J. 2005 Haploview, analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265.

Bahl S., Virdi K., Mittal U., Sachdeva M. P., Kalla A. K., Holmes S. E. *et al*. 2005 Evidence of a common founder for SCA12 in the Indian population. *Ann. Hum. Genet*. **69**, 528–534.

Grierson G. 1927 *Linguistic survey of India*. Motilal Banarsidas, Calcutta.

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.

Holmes S. E., O'Hearn E. E., McInnis M. G., Gorelick-Feldman D. A., Kleiderlein J. J., Callahan C. *et al*. 1999 Expansion of a novel CAG trinucleotide repeat in the 5′ region of PPP2R2B is associated with SCA12. *Nat. Genet*. **23**, 391–392.

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

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

Miller S. A., Dykes D. D. and Polesky H. F. 1988 A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. **16**, 1215.

Singh K. S. 1998 *India's communities*. People of India National Series, Anthropological Survey of India. Oxford University Press, New Delhi.

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.

Received 2 June 2008, in revised form 8 October 2008; accepted on 7 November 2008

Published on the Web: 30 March 2009