doi: 10.1111/j.1469-1809.2006.00383.x

Portability of Tag SNPs Across Isolated Population Groups: An Example from India

N. Sarkar Roy[1], S. Farheen[2], N. Roy[2], S. Sengupta[2] and P. P. Majumder[1,2,*]

[1]*TCG-ISI Centre for Population Genomics, Bengal Intelligent Park Ltd., Building B, 3rd Floor, Block EP & GP, Sector V, Salt Lake Electronics Complex, Kolkata 700091, India*

[2]*Human Genetics Unit, Indian Statistical Institute, 203 B. T. Road, Kolkata 700108, India*

Summary

Isolated population groups are useful in conducting association studies of complex diseases to avoid various pitfalls, including those arising from population stratification. Since DNA resequencing is expensive, it is recommended that genotyping be carried out at tagSNP (tSNP) loci. For this, tSNPs identified in one isolated population need to be used in another. Unless tSNPs are highly portable across populations this strategy may result in loss of information in association studies. We examined the issue of tSNP portability by sampling individuals from 10 isolated ethnic groups from India. We generated DNA resequencing data pertaining to 3 genomic regions and identified tSNPs in each population. We defined an index of tSNP portability and showed that portability is low across isolated Indian ethnic groups. The extent of portability did not significantly correlate with genetic similarity among the populations studied here. We also analyzed our data with sequence data from individuals of African and European descent. Our results indicated that it may be necessary to carry out resequencing in a small number of individuals to discover SNPs and identify tSNPs in the specific isolated population in which a disease association study is to be conducted.

Keywords: linkage disequilibrium, haplotype, genetic affinity, complex disease, association study

Introduction

In genetic association studies for complex diseases, although resequencing of cases and controls is ideal, it is not a feasible option for most researchers. Therefore, the current approach is to rely on linkage disequilibrium (LD) among polymorphic markers – commonly, single nucleotide polymorphisms (SNPs) – and to choose markers that are not in complete or strong LD. Various algorithms have been developed (Zhang et al. 2005; Stram, 2005) to select a maximally informative set of markers, called tagSNPs (tSNPs), in a population. All these algorithms rely on estimates of LD among SNPs. However, the strength of LD among marker pairs in a genomic region is strongly influenced by the demographic history of the population (Stumpf & Goldstein, 2003). Thus, populations with differing demographic histories exhibit different LD patterns for the same genomic region. This implies that a set of tSNPs selected in one population may not represent a maximally informative set of SNPs in another population. Thus, in terms of study design, unless there is portability of tSNPs across populations it may not be wise to use tSNPs chosen in one population for conducting an association study in another population. The HapMap database is a rich resource for selecting tSNPs, but their portability to other populations is uncertain. One recent study (Huang et al. 2006) has shown that tSNPs selected from continental populations represented in the HapMap database are portable to other populations in East Asia with reasonable efficiency.

Not many studies have yet been carried out to investigate tSNP portability across isolated ethnic populations, and to compare these results with continental populations. In a study (Conrad et al. 2006) carried out using 52 geographically dispersed populations and a 12Mb genomic region, it was found that 83% of common 20kb haplotypes in a population are also common in the HapMap population with the highest genetic affinity. However, even this study used some ethnically ill-defined populations (e.g. Russian,

* Corresponding author: Professor Partha P. Majumder, Human Genetics Unit, Indian Statistical Institute, 204 B. T. Road, Kolkata 700108, India. E-mail: ppm@isical.ac.in

Cambodian). Investigation of tSNP portability across isolated populations is important, because the usefulness of isolated populations for mapping complex traits has been recognized and emphasized (Peltonen et al. 2000). Further, population stratification can have a major impact on the results of an association study (Cardon & Palmer, 2003; Freedman et al. 2004); therefore, it has been recommended that cases and controls for an association study be chosen from a defined population group and not from a conglomerate population. However, it is hardly possible to carry out resequencing and identify tSNPs in every isolated population from which samples will be drawn for an association study. Thus, the nature and extent of tSNP portability among isolated populations living in geographic proximity is an issue that deserves attention.

In this study we resequenced three genes, comprising a total of 12Kb, in 160 individuals chosen from 10 ethnically and geographically disparate populations from India. We introduced a measure of tSNP portability across populations and carried out an empirical study among the 10 Indian ethnic groups. We examined whether the pattern of portability among the populations correlates with the pattern of genetic affinities estimated using a different set of markers from these populations. We also performed comparative analyses with two continental populations for which sequence data were available from the SeattleSNPs Project (http://pga.gs.washington.edu/).

Materials and Methods

Study Populations and Samples

We studied 10 ethnic population groups from India. These populations were chosen so as to represent the social, cultural, linguistic and geographical diversity of India. The tribal groups are the autochthonous populations of India. The vast majority of Indian populations belong to the Hindu caste fold. Individuals belonging to these populations, both tribal and caste, predominantly marry within their own group and are therefore relatively unadmixed. Dialects belonging to the Austro-Asiatic linguistic family are spoken exclusively by the tribal people of India. The other languages used in India by caste and tribal groups are Dravidian, Tibeto-Burman and Indo-European. Descriptions of the populations are provided in Table 1; from each population we sampled phenotypically normal individuals (comprising approximately equal numbers of males and females) who were unrelated at least to the first-cousin level; sample sizes are provided in Table 1. Ten mL blood was drawn by venipuncture, with institutional ethical approval and informed consent, from each individual. High molecular weight genomic DNA was isolated from each blood sample, either by the salting-out method (Miller et al. 1988) or using a column (QIAGEN).

For purposes of comparison of results obtained from the Indian population groups studied we downloaded relevant genotype data – derived on the basis of resequencing – from the SeattleSNPs database (http://pga.gs.washington.edu/). The SeattleSNPs data pertain to individuals of European (EUR) and African (AFR) descent.

Genes

We considered three genes in this study: *ADRB2* (chromosomal location: 5q31-q32, nucleotide positions: 148186369..148188379), *TNF* (6p21.3, 31651329..31654091) and *ICAM1* (19p13.3-p13.2, 10242779..10258291). In connection with a study on cardiovascular diseases, undertaken by us, we chose these genes because of their relevance. Parenthetically, we may add that the questions investigated in this study arose in connection with designing this study in ethnic groups of India.

In addition, to test the hypothesis that the extent of tSNP portability among populations positively correlates with genetic affinities, we used data for 69 binary Y-chromosomal markers to estimate frequencies of Y-chromosomal haplogroups in the study populations (Sengupta et al. 2006).

DNA Analysis

After repeat-masking we designed overlapping primers for the exons of these genes and carried out double-pass resequencing for every individual. (Primer sequences and amplification

Table 1 Names, sample sizes, ethnic characteristics and geographical locations of Indian population groups included in this study

Population Name (Code)	Geographical region of habitat and sampling	Social group	Linguistic group	Sample size
Iyer (IYR)	South (neighbourhood of Chennai city)	Caste	Dravidian	17
Kadar (KAD)	South (Nilgiri Hills)	Tribe	Dravidian	16
Koknasth Brahmin (KBR)	West (neighbourhood of Mumbai city)	Caste	Indo-European	16
Manipuri Brahmin (MNP)	North-east (rural areas of Manipur State)	Caste	Tibeto-Burman	11
Maratha (MRT)	West (neighbourhood of Mumbai city)	Caste	Indo-European	15
Muria (MUR)	Central (neighbourhood of Raipur city)	Tribe	Dravidian	16
Mizo (MZO)	North-east (neighbourhood of Aizawl city)	Tribe	Tibeto-Burman	21
Santal (SAN)	East (rural areas on the border of West Bengal and Bihar States)	Tribe	Austro-Asiatic	16
Saryupari Brahmin (SBR)	Central (neighbourhood of Raipur city)	Caste	Indo-European	16
Bengali Brahmin (WBR)	East (neighbourhood of Kolkata city)	Caste	Indo-European	16

protocols are available from PPM on request.) The number of bases resequenced for each individual were ~3kb, ~3kb and ~6kb, respectively, for the *ADRB2*, *TNF* and *ICAM1* genes (including about 1 kb upstream of each gene). If any genotype differed between the forward and reverse sequencing passes, the sample was sequenced again and all differences were eventually resolved.

Males drawn from these populations were screened for 69 binary Y-chromosomal markers, following the protocol given in Sengupta et al. (2006). Based on these data we classified individuals into haplogroups following the Y Chromosome Consortium (2002) nomenclature, and estimated haplogroup frequencies for each population. It may be noted that only a subset of males used for estimating Y-haplogroup frequencies was used for resequencing autosomal genes.

Statistical Analysis

The DNA sequences were analysed using the Polyphred package (http://droog.mbt.washington.edu/PolyPhred.html), and genotypes of individuals were determined. Loci with minor allele frequency (MAF) <0.05 in all population groups were not considered for further analysis. Based on genotype data at the polymorphic loci, we identified tSNPs separately for every population using the Tagger package (http://www.broad.mit.edu/mpg/tagger/) with cut-off for r^2 set at 0.8. Downloaded genotype data pertaining to the three genes under consideration from the SeattleSNPs database for individuals of European and African descent were also analyzed using these packages and cut-off values.

We used the following measure of portability, $\pi(P_1 > P_2)$, for tSNPs chosen in population P_1 to population P_2. Our measure of tagSNP portability stems from a definition of a tagSNP, originally proposed by Johnson et al. (2001) [see also, Zhang et al. (2005)]. Suppose there are L polymorphic sites in population P_1. Based on these L sites, suppose the estimated number of haplotypes in the population P_2 is M, and the estimated frequency of the i-th haplotype in this population is f_i ($i = 1, 2, \ldots, M$). Then, the haplotype diversity in population P_2 is: $H_{P_2} = 1 - \sum_{i=1}^{M} f_i^2$. Suppose in population P_1, of the L polymorphic SNPs, T are tagSNPs. If we wish to use these T tSNPs identified in population P_1 to carry out a study in population P_2, then these tagSNPs are appropriate for use in P_2 they explain a "large" proportion of the total haplotype diversity of P_2. Suppose, based on the T tSNPs, the set of M haplotypes in population P_2 reduces to M^*. Let the frequency of the j-th haplotype (each of length $\leq$ T SNPs) be g_j; $j = 1, 2, \ldots, M^*$. (Obviously, each of the M^* haplotypes is obtained by collapsing a certain number of the original M haplotypes, and the frequency g_j of the j-th "collapsed" haplotype is obtained by adding the estimated frequencies of the corresponding original haplotypes.) Then, the haplotype diversity in P_2 based on the frequencies of the M^* "collapsed" haplotypes is: $H_{P_2}* = 1 - \sum_{j=1}^{M*} g_j^2$. Hence, the proportion of the total haplotype diversity in P_2, H_{P_2}, explained by the haplotypes based on the tagSNPs identified in P_1 is: $\frac{H_{P_2}*}{H_{P_2}}$. We call this ratio the *portability index* $\pi(P_1 > P_2)$. Obviously, the higher this ratio is the better the usefulness in population P_2 of the tagSNPs identified in P_1. We note: (a) $0 \leq \pi(P_1 > P_2) \leq 1$; (b) $\pi(P_1 > P_1) \leq 1$ and is <1 if some of the non-tagged SNPs are not in complete linkage disequilibrium with the selected tSNPs; (c) $\pi(P_1 > P_2) = 0$ if all of the tSNPs selected in P_1 are non-polymorphic in P_2; and, (d) $\pi(P_1 > P_2)$ may not be equal to $\pi(P_2 > P_1)$. If multiple unlinked genes or genomic regions are considered, then for each gene or genomic region the portability index can be calculated separately and averaged to yield an overall index of tagSNP portability. A determination of the quality of portability may be based on the empirical distribution of $\pi(P_2 > P_1)$. We suggest the quality of portability be classified as (a) "High", if an observed value of $\pi > \overline{\pi} + s_\pi$; (b) "Low", if $\pi < \overline{\pi} - s_\pi$; and, (c) "Moderate", if $\overline{\pi} - s_\pi < \pi < \overline{\pi} + s_\pi$, where $\overline{\pi}$ is the average of π and s_π is the standard deviation of π. We recognize that the cut-off points on the distribution of π are somewhat arbitrary; other possible cut-off points may also be used.

Haplotypes were constructed using Haploview (http://www.broad.mit.edu/mpg/haploview/). In this software haplotypes are estimated using an accelerated EM algorithm similar to the partition/ligation method (Qin et al. 2002).

To estimate genetic affinities among the populations we used Y-haplogroup frequencies and calculated the ($1-D_A$) index (Nei, 1987) between pairs of populations. A neighbor-joining tree (Saitou & Nei, 1987) depicting genetic affinities among the populations was also computed based on the matrix of pairwise distances. To examine whether there the extent of tSNP-portability among populations correlates with genetic affinities, we carried out a non-parametric Mantel (1967) test using the two matrices corresponding to π and ($1-D_A$) between pairs of populations. The test was carried out using the zt software (Bonnet & van de Peer, 2002).

Results

Details pertaining to the number of SNPs in the various study populations, and the number of SNPs shared among individuals drawn from Indian ethnic groups and of AFR and EUR descent are provided in Table 2. Except for *ICAM1*, in which a large number of SNPs present among individuals of AFR and EUR descent were not found among Indian ethnic groups, the extent of shared SNPs for the other two genes (*ADRB2* and *TNF*) was high. Indian ethnic groups seem to harbour a small number of "private" SNPs. For *ICAM1*, the extent of SNP-sharing was higher between individuals of AFR and EUR descent compared to the extent of sharing with Indian ethnic groups; this feature was not observed for *ADRB2* or *TNF*. Details of SNP loci and allele frequencies at these loci are provided in online Supplementary Table 1. We noted that there is considerable variability in allele frequencies across populations; at some loci an allele that is "minor" in several populations is the "major" allele in other populations.

SNPs that were found to be tagging in each of the populations are presented in Figure 1. The matrix of portability

	ADRB2	*TNF*	*ICAM1*
Total no. of SNPs in Indian, AFR and EUR populations	14	11	39
No. of SNPs present in at least one Indian ethnic group	13	11	17
No. of SNPs in Indian ethnic groups that are unreported in dbSNP*	2	1	5
No. of SNPs shared between:			
Indian and AFR populations	10	3	4
Indian and EUR populations	10	5	3
AFR and EUR populations	10	2	12

Table 2 Statistics pertaining to the SNPs in the three genes studied

*These unreported SNPs are also absent among individuals of AFR and EUR descent resequenced in the SeattleSNPs project.

index, averaged over the three genes, between pairs of populations is presented in Table 3. These matrices separately for each of the three genes are presented in online Supplementary Table 2. The distribution of observed values of the portability index is presented in online Supplementary Figure 1. The mean and standard deviation of the observed values of portability index were, respectively, 0.816 and 0.131. Using the cut-off points suggested by us, the values of which were 0.947 and 0.686, it is clear from Table 3 that portability of tagSNPs from MNP to most other populations was low (since the values of portability index from MNP to the other populations were <0.686). Similarly, the tSNP portability from most other populations to MNP was also low. The vast majority of populations showed moderate levels of tSNP portability ($0.947 > \pi > 0.686$). The nature or extent of portability does not seem to depend on ethnic or geographical characteristics of the populations. That is, we did not find that tSNPs selected in a caste (tribal) population were portable across other caste (tribal) populations. Nor did we find that tSNPs selected in a population inhabiting north-east India were portable to other populations living in that geographical area, but not to those living in other areas. We also noted that portability of tSNPs across (both from and to) the European (EUR) population and Indian ethnic populations was higher compared to the African (AFR) population. However, the portability of tSNPs from the EUR population to most Indian ethnic populations was only moderate. This obviously has implications for the use of tSNPs chosen from the HapMap database for use in designing studies in Indian population groups.

These features regarding portability of tagSNPs are obviously a reflection of the variability of LD values across isolated populations, and with continental groups such as AFR and EUR. Variability of LD values across populations is strongly dependent on the variability of demographic histories of the populations. We, therefore, sought to examine whether the pattern of tSNP portability correlated with the pattern of genetic affinities among a set of isolated populations, assuming that genetically similar extant populations had similar demographic histories. The Mantel test of correlation between tSNP portability and genetic similarity, based on frequencies of Y-chromosomal haplogroups between pairs of populations (Table 4), was not significant ($\bar{r} = 0.178$; one-tailed $\bar{p}$-value $= 0.155$).

Discussion

Since the usefulness of isolated population groups in the study of complex traits has been emphasized (Peltonen et al. 2000), it is of importance to assess whether tagSNPs are portable across such populations. With this aim in mind we resequenced 3 genes in 160 individuals drawn from 10 ethnic groups from India, and downloaded available sequence data pertaining to individuals of African and European descent. In each group we identified tagSNPs and calculated a measure of tSNP portability between populations. We found that there was considerable variation in the extent of tSNP portability between Indian ethnic groups, and to AFR and EUR populations (Table 3). This indicates that there is considerable variation among Indian ethnic groups in the pattern of linkage disequilibrium among SNPs within genomic regions. There is no clear concordance between the magnitudes of portability (Table 3) and geographical region of habitat, linguistic affinity or socio-cultural affinity (Table 1). The pattern of tSNP portability did not significantly correlate (Mantel test p-value $= 0.155$) with genetic affinities among the Indian populations. This finding is somewhat surprising, and is indicative of differential evolutionary pressures across populations shaping genetic affinities and linkage disequilibria among SNPs.

Our finding on portability of tSNPs raises caution, because it indicates that although isolated populations are useful in the study of complex disorders (Peltonen et al. 2000), tSNPs identified in one isolated population

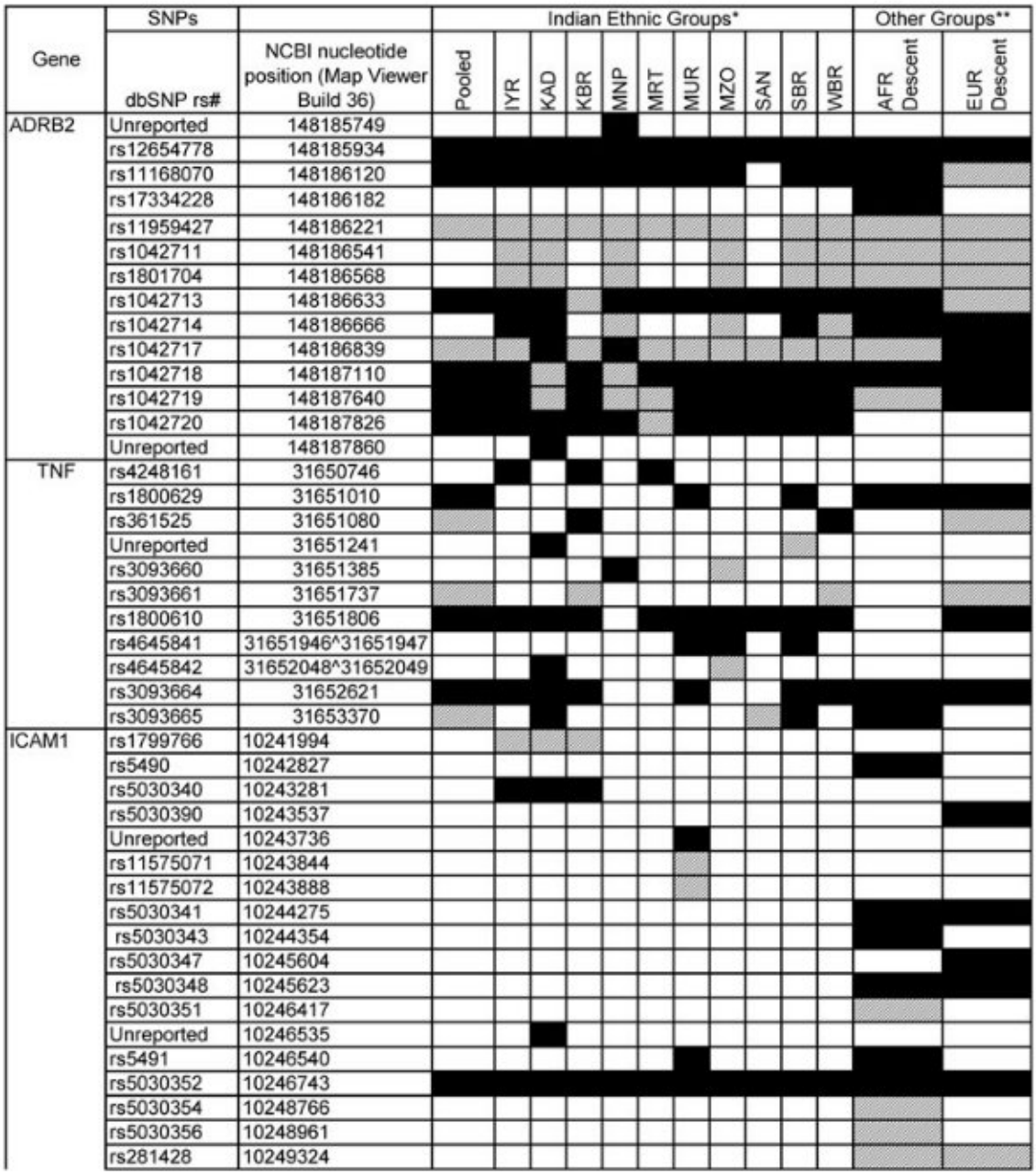
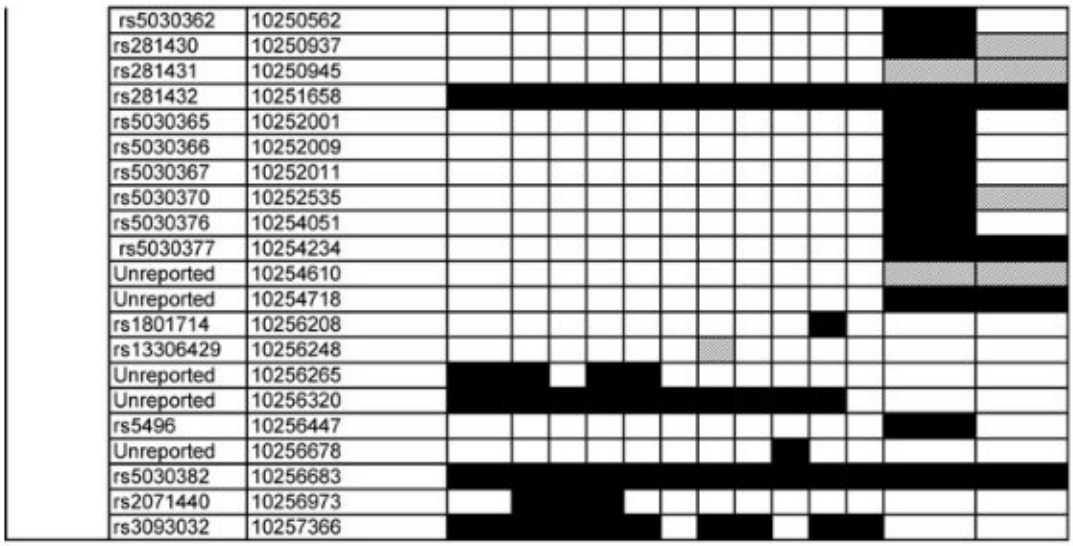


Figure 1 Characteristics of SNPs in three genes in 10 ethnic groups of India and in groups of individuals of African (AFR) and European (EUR) descent

*Details of Indian ethnic groups are provided in Table 1.

**Based on resequencing data downloaded from SeattleSNPs (http://pga.gs.washington.edu/)

may not be profitably used to carry out a similar study in another isolated population. This finding is at variance with findings from more conglomerate populations (Huang et al. 2006). It is also inconsistent with the finding of Rosenberg et al. (2006) that there is a low level of genetic divergence among the population groups of India. With low genetic divergence among populations one would expect a high level of tSNP portability across these populations, which is not what we found. While the reason for this inconsistency is unclear to us, we note that in Rosenberg et al.'s (2006) study (a) no Indian tribal populations were included, and (b) Indian linguistic groups that were included (e.g, Bengali, Hindi, etc.) are actually "mixed" ethnic groups from a



* Details of Indian ethnic groups are provided in Table 1.
** Based on resequencing data downloaded from SeattleSNPs (http://pga.gs.washington.edu/)

Figure 1 Continued

Table 3 Values of portability index, π ($P_1 > P_2$), from population P_1 along the columns to population P_2 along the rows*

	IYR	KAD	KBR	MNP	MRT	MUR	MZO	SAN	SBR	WBR	AFR	EUR
IYR	**1**	0.9	**0.999**	*0.653*	0.89	0.902	0.851	0.823	0.91	0.889	0.695	0.815
KAD	0.885	**0.998**	0.864	*0.618*	*0.681*	0.855	0.716	0.702	0.887	0.842	0.769	0.765
KBR	**0.995**	0.898	**0.997**	*0.628*	0.756	0.904	0.742	0.716	0.904	0.88	0.724	0.787
MNP	*0.667*	*0.641*	*0.62*	**0.962**	*0.597*	*0.642*	*0.642*	*0.636*	*0.642*	*0.579*	*0.562*	*0.537*
MRT	**1**	0.894	**1**	*0.615*	**0.988**	0.946	0.946	0.886	0.946	0.892	*0.6*	0.88
MUR	0.87	0.841	0.87	*0.653*	0.782	**1**	0.869	0.795	**0.999**	0.865	0.827	0.88
MZO	0.872	0.898	0.869	0.725	0.849	**0.963**	**0.963**	0.866	**0.964**	0.838	*0.608*	0.79
SAN	0.946	**0.98**	0.945	*0.64*	0.881	0.946	0.946	**0.952**	**0.994**	0.941	0.689	0.856
SBR	0.9	0.927	0.897	*0.646*	0.738	**0.998**	0.892	0.791	**1**	0.887	0.77	0.84
WBR	0.926	0.886	**0.997**	*0.646*	0.699	0.926	0.785	0.752	0.926	**0.997**	0.763	0.804
AFR	0.722	0.701	0.707	*0.539*	*0.539*	0.814	*0.545*	*0.557*	0.842	0.707	0.877	0.801
EUR	0.787	0.755	0.78	*0.585*	*0.681*	0.947	0.703	0.687	**0.951**	0.78	0.905	**0.962**

*Bold entries indicate "high portability" and italicized entries indicate "low portability"; other values indicate "moderate portability".

Table 4 Genetic similarity index ($1-D_A$) between pairs of populations based on Y-haplogroup frequencies

	IYR	KAD	KBR	MNP	MRT	MUR	MZO	SAN	SBR	WBR
IYR*	***	0.159	0.777	0.447	0.764	0.166	0.036	0.136	0.823	0.642
KAD		***	0.248	0.000	0.439	0.747	0.572	0.712	0.146	0.111
KBR			***	0.490	0.731	0.179	0.039	0.054	0.874	0.847
MNP				***	0.224	0.000	0.401	0.000	0.572	0.601
MRT					***	0.518	0.096	0.134	0.704	0.597
MUR						***	0.392	0.506	0.175	0.211
MZO							***	0.638	0.038	0.045
SAN								***	0.143	0.063
SBR									***	0.864
WBR										***

genetic viewpoint, because there are multiple endogamous castes within a linguistic group. Service et al. (2007) also reported that HapMap CEU tSNPs performed more poorly in samples from Costa Rica (CR), the Azores (AZO) and Antioquia (ANT) than in samples from, for example, Finland or Newfoundland. They surmised that the reason for this may be that the CR, AZO and ANT populations are characterized by more extensive derivation from non-European populations compared to the other populations analyzed by them. We note that the primary contemporary issue is whether tSNPs from the populations included in HapMap are portable to the Indian ethnic groups. Since the HapMap data were not generated by resequencing, in the genes considered by us there were only a small number of SNPs that were interrogated in the HapMap project. Therefore, we are unable to comment on the portability of tSNPs identified from the HapMap data to Indian ethnic groups. However, our finding that portability of tSNPs from individuals of European descent (based on resequencing) to Indian ethnic groups is generally "moderate", suggests that portability of HapMap tSNPs will possibly be moderate, at least from the CEU population. Similarly, our results from individuals of African descent indicate that the portability of tSNPs chosen from the YRI population to Indian ethnic groups will be low. We also note that there are some large-scale resequencing projects that are ongoing in multiple Indian populations (Indian Genome Variation Consortium, 2005). One objective of these projects is to identify tSNPs in some population groups, so that these can be used in case-control association studies in other populations. However, as our results indicate, the feasibility of this may only be moderate.

We recognize two important limitations of our study: (a) we have studied short genomic regions, totalling ~12 kb, and (b) the sample size of each population included in this study was limited. Our conclusions, therefore, need to be validated with larger sample sizes and longer genomic regions. Our finding that the portability of tSNPs in the EUR population to Indian ethnic groups is higher compared to the AFR population is supported by earlier genetic studies, which indicated that Indian populations, particularly those belonging to the Hindu caste fold, are closer to European than to African populations (Bamshad et al. 2001). Of course it is possible that a pair of genetically close extant populations may have considerably different demographic histories. Our study, therefore, indicates that it may be necessary to exercise some caution in using tSNPs identified in one isolated population for disease association studies in another isolated population. It may be necessary to carry out resequencing in a small number of individuals to discover SNPs and identify tSNPs in the specific isolated population in which the disease association study is to be conducted.

Acknowledgments

We are grateful to A. Ramesh, M.V. Usha Rani, Chitra Mahadik Thakur, S. K. Sil and M. Mitra for help in obtaining informed consent and collection of samples, and to B. Dey for help in DNA sequencing. We are also grateful to the Department of Biotechnology, Government of India, and the Indian Statistical Institute for financial and infrastructural support.

References

Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., Naidu, J. M., Prasad B. V. R., Reddy, P. G., Rasanayagam, A., Papiha, S. S., Villems, R., Redd, A. J., Hammer, M. F., Nguyen, S. V., Carroll, M. L., Batzer, M. A. & Jorde, L. J. (2001) Genetic Evidence on the Origins of Indian Caste Populations. *Genome Res* **11**, 994–1004.

Bonnet, E. & van de Peer, Y. (2002) zt: a software tool for simple and partial Mantel tests. *J Stat Software* **7**, 1–12.

Cardon, L. R. & Palmer, L. J. (2003) Population stratification and spurious allelic association. *Lancet* **361**, 598–604.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A. & Pritchard, J. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* **38**, 1251–1260.

Freedman, M. L., Reich, D., Penney, K. L., McDonald, G. J., Mignault, A. A., Patterson, N., Gabriel, S. B., Topol, E. J., Smoller, J. W., Pato, C. N., Pato, M. T., Petryshen, T. L., Kolonel, L. N., Lander, E. S., Sklar, P., Henderson, B., Hirschhorn, J. N. & Altshuler, D. (2004) Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**, 388–93.

Huang, W., He, Y., Wang, H., Wang, Y., Liu, Y., Wang, Y., Chu, X., Wang, Y., Xu, L., Shen, Y., Xiong, X., Li, H., Wen, B., Qian, J., Yuan, W., Zhang, C., Wang, Y., Jiang, H., Zhao, G., Chen, Z. & Jin, L. (2006) Linkage disequilibrium sharing and haplotype-tagged SNP portability between populations. *Proc Natl Acad Sci USA* **103**, 1418–1421.

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

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., DiGenova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G. & Todd, J. A. (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet* **29**, 233–237.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Res* **27**, 209–220.

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

Nei, M. (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Peltonen, L., Palotie, A. & Lange, K. (2000) Use of population isolates for mapping complex traits. *Nat Rev Genet* **1**, 182–190.

Qin, Z. S., Niu, T. & Liu, J. S. (2002). Partition-ligation EM algorithm for haplotype inference with single nucleotide polymorphisms. *Am J Hum Genet* **71**, 1242–1247.

Rosenberg, N. A., Mahajan, S., Gonzalez-Quevdo, C., Blum M. G. B., Nino-Rosales, L., Ninis, V., Das, P., Hegde, M., Molinari, L., Zapata, G., Weber, J. L., Belmont, J. W. & Patel, P. I. (2006) Low levels of genetic divergence across geographically and linguistically diverse populations from India. *PLoS Genet* **2**, 2052–2061.

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

Sengupta, S., Zhivotovsky, L. A., King, R., Mehdi, S. Q., Edmonds, C. A., Chow, C.-E. T., Lin, A. A., Mitra, M., Sil, S. K., Ramesh, A., Usha Rani, M. V., Thakur, C. M., Cavalli-Sforza, L. L., Majumder, P. P. & Underhill, P. A. (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 of Hum Genet* **78**, 202–221.

Service, S., the International Collaborative Group on Isolated Populations, Sabatti C. & Freimer, N. (2007) Tag SNPs chosen from HapMap perform well in several population isolates. *Genet Epid* **31**, 189–194.

Stram, D. O. (2005) Software for tag single nucleotide polymorphism selection. *Hum Genomics* **2**, 1–8.

Stumpf M. P. H. & Goldstein, D. B. (2003) Demography, recombination hotspot intensity, and the block structure of linkage disequilibrium. *Curr Biol* **13**, 1–8.

The Y Chromosome Consortium (2002) A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. *Genome Res* **12**, 339–348.

Zhang, K., Qin, Z., Chen, T., Liu, J. S., Waterman, M. S. & Sun, F. (2005) HapBlock: haplotype block partitioning and tagSNP selection software using a set of dynamic programming algorithms. *Bioinformatics* **21**, 131–134.

Supplementary Material

The following supplementary material is available for this article:

Figure S1. Values of Portability Index (Y-axis), sorted in descending order of magnitude, for the 144 values presented in Table 3.

Table S1. Characteristics of variations in the three genes studied, minor allele designations and frequencies in 10 Indian ethnic groups and in individuals of African and European descent.

Table S2. Values of Portability Index for Individual Genes, TNF, ADRB2 and ICAM1.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1469-1809.2006.00383.x
(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Received: 12 December 2006
Accepted: 8 June 2007