doi: 10.1046/j.1529-8817.2003.00126.x

DNA Sequence Variation and Haplotype Structure of the ICAM1 and TNF Genes in 12 Ethnic Groups of India Reveal Patterns of Importance in Designing Association Studies

S. Sengupta[1], S. Farheen[1], N. Mukherjee[1], B. Dey[1], B. Mukhopadhyay[1], S. K. Sil[2], N. Prabhakaran[3], A. Ramesh[3], D. Edwin[4], M. V. Usha Rani[4], M. Mitra[5], C. T. Mahadik[6], S. Singh[7], S. C. Sehgal[7] and P. P. Majumder[1]

[1]*Anthropology & Human Genetics Unit, Indian Statistical Institute, Kolkata, India*
[2]*Tripura University, Agartala, India*
[3]*University of Madras, Chennai, India*
[4]*Bharathiar University, Coimbatore, India*
[5]*Pandit Ravishankar Shukla University, Raipur, India*
[6]*Research Society, B.J. Hospital for Children, Mumbai, India*
[7]*Regional Medical Research Centre, Indian Council of Medical Research, Port Blair, India*

Summary

We have examined the patterns of DNA sequence variation in and around the genes coding for ICAM1 and TNF, which play functional and correlated roles in inflammatory processes and immune cell responses, in 12 diverse ethnic groups of India. We aimed to (a) quantify the nature and extent of the variation, and (b) analyse the observed patterns of variation in relation to population history and ethnic background. At the *ICAM1* and *TNF* loci, respectively, the total numbers of SNPs that were detected were 28 and 12. Many of these SNPs are not shared across ethnic groups and are unreported in the dbSNP or TSC databases, including two fairly common non-synonymous SNPs at positions 13487 and 13542 in the *ICAM1* gene. Conversely, the TNF-376A SNP that is reported to be associated with susceptibility to malaria was not found in our study populations, even though some of the populations inhabit malaria endemic areas. Wide between-population variation in the frequencies of shared SNPs and coefficients of linkage disequilibrium have been observed. These findings have profound implications in case-control association studies.

Keywords: Single nucleotide polymorphism, Linkage disequilibrium, Genetic structure, Genetic affinity

Introduction

Analysis of DNA sequence variation within and between populations is useful for understanding the evolution and organization of the human genome, as well as the complex links between genotypic and phenotypic variation, including disease susceptibility and resistance. The most common form of DNA sequence variation is the single nucleotide polymorphism (SNP). Two recent studies (Carlson *et al.* 2003; Reich *et al.* 2003) have indicated various limitations of the data archived in the major SNP databases, dbSNP (http://www.ncbi.nlm.nih.gov/SNP) and TSC (http://snp.cshl.org/). These limitations include (a) bias towards SNPs present in European populations, (b) high rate of non-validation (12-35%), and (c) limited

*Address for Correspondence and Reprints: Partha P. Majumder Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India. Telephone: +91-33-25753209 Fax: +91-33-25773049. E-mail: ppm@isical.ac.in

availability of allele frequencies from diverse populations. Moreover, since a substantial number of common variations are population specific, the need for additional SNP discovery and validation studies in other large, diverse ethnic populations has been emphasized. The data from such studies can also be profitably utilized to understand the nature, extent and causes of genetic variation across ethnic groups.

In this study, we report a systematic survey of polymorphisms in and around two genes – the intercellular adhesion molecule 1 (*ICAM1*) and tumor necrosis factor α (*TNF*) genes – among 208 individuals drawn from 12 different ethnic groups of India. There is interaction between the *ICAM1* and *TNF* gene products in inflammatory processes and immune cell responses in a wide range of diseases. The cytokine TNF is known to upregulate the endothelial adhesion molecule ICAM1 (Meager, 1999). A large number of studies have reported associations of various diseases with polymorphisms in these genes, some of which are possibly of intrinsic functional relevance (Fernendez-Reyes *et al.* 1997; Giminez *et al.* 2003). The aims of this study were (a) to discover and validate SNPs in the *ICAM1* and *TNF* genes in multiple ethnic groups of India, (b) to identify the proportion of SNPs present in Indian populations that remain unreported in dbSNP, (c) to analyze the variation in SNP and SNP-haplotype frequencies across populations, with a view to quantifying genetic structure and understanding population relationships, and (d) to assess the extent of variation in linkage disequilibrium across populations.

Materials and Methods

Population Samples

Blood samples were collected from individuals unrelated to the first cousin level. These individuals belonged to 12 distinct ethnic groups inhabiting 5 different geographical regions of mainland India, and the Andaman and Nicobar Islands. Collection of blood samples was initiated from the populations of mainland India after approval of the Institutional Ethics Committees, and was carried out with informed consent of the participants. Blood samples from the Jarawas, who inhabit the Andaman and Nicobar Islands, were collected for medical purposes by the Regional Medical Research Centre, Indian Council of Medical Research, Port Blair, in collaboration with the Health Services Department of the Andaman & Nicobar Administration, when there was an outbreak of fever of unknown etiology some years ago. Before undertaking research using these collected blood samples, which were already stripped of all identifiers, approval of the Ethics Committee of the Regional Medical Research Centre, Port Blair, was gained. A list of the populations, with sample sizes and brief notes on their linguistic and socio-cultural backgrounds, are provided in the Table 1. The geographical locations of sampling are indicated in Figure 1.

Loci and Protocols

The *ICAM1* gene maps to 19p13.3-p13.2 and contains 7 exons. The *TNF* gene maps to 6p21.3 and contains 4 exons. Genomic sequences of these two genes were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu). The genomic region encompassing the *ICAM1* was repeat-masked using the program RepeatMasker2 (http://ftp.genome.washington.edu.cgi-bin/RepeatMasker). Appropriate primers to amplify the exons, introns, the 5' and a portion of the 3' untranslated regions (UTRs) of these genes (excluding the repeat-masked region of *ICAM1*) were designed. The total number of bases resequenced for each individual were 6000 and 3046, respectively, for the *ICAM1* and *TNF* genes.

DNA amplification conditions for PCR were optimized using control samples. PCR products were cleaned using Exonuclease I and Shrimp Alkaline Phosphatase and subjected to sequencing on an ABI-3100 automated sequencer using dye-terminator chemistry. (Primer sequences and PCR conditions are available on request.) ABI trace files thus generated were analyzed using the PHRED software (http://www.mbt.washington.edu/phrap.docs/phred.html) which assigns quality scores to each base. The PHRED outputs for all the individuals for any given PCR amplicon were aligned using PHRAP software. The resulting assemblies were viewed using CONSED that allows identification of putative

Table 1 Names of study populations, sample sizes, geographical locations of habitat and socio-linguistic information

Population Name [Code]	No. of Individuals sampled	Linguistic Affiliation	Geographical Region (State)	Social Category
Bhutia [BHU]	13	Tibeto-Burman	North-East (Sikkim)	Tribe
Mizo [MZO]	21	Tibeto-Burman	North-East (Mizoram)	Tribe
Manipuri (Meitei) [MNP]	11	Tibeto-Burman	North-East (Manipur)	Caste
Santal [SAN]	16	Austro-Asiatic	East (Bihar and West Bengal)	Tribe
West Bengal Brahmins [WBR]	16	Indo-European	East (West Bengal)	Caste
Kadar [KAD]	16	Dravidian	South (Tamilnadu)	Tribe
Iyer [IYR]	17	Dravidian	South (Tamilnadu)	Caste
Muria [MUR]	16	Dravidian	Central (Chattisgarh)	Tribe
Saryupari Brahmins [SBR]	16	Indo-European	Central (Chattisgarh)	Caste
Maratha [MRT]	15	Indo-European	West (Maharashtra)	Caste
Konkan Brahmins [KBR]	16	Indo-European	West (Maharashtra)	Caste
Jarawa[a] [JAR]	35	Jarawa Language[b]	Middle Andaman (Andaman & Nicobar Islands)	Tribe

[a]Data on the Jarawa have been published in Singh et al.[28]
[b]The Jarawas speak a dialect that remains unclassified

sequence variants. All samples with putative variant alleles were resequenced using the reverse primers for confirmation.

Statistical analysis

Allele frequencies at each variant site were computed by the gene-counting method. Maximum likelihood estimates of haplotype frequencies from the *ICAM1* and *TNF* polymorphic sites were obtained via the EM algorithm using the program HAPLOPOP (Majumdar & Majumder, 1999). Standard diversity indices and coefficients of pairwise linkage disequlibrium (D') were estimated using the Arlequin package (http://anthropologie.unige.ch/arlequin). Population structure analysis was also performed using Arlequin. Genetic affinities were estimated by the standard principal components analysis and neighbour-joining phylogenetic analysis (Saitou & Nei, 1998).

Results

Sequence Variation

At the *ICAM1* locus, 29 variant sites were identified by resequencing the *ICAM1* gene in 208 individuals drawn from the 12 different ethnic groups (Table 1) inhabiting different geographical regions of India. Allele frequency and relevant characteristics of each variant site are given in Table 2. Transition substitutions are more prevalent (64%) than transversions (35%); one insertion/deletion (indel) polymorphism was observed. All variant sites are biallelic, except for one site (Table 2, rs5030352) where a third T-allele appeared in two Konkan Brahmins of Maharashtra that were GT heterozygotes. (We removed these two individuals from allele frequency estimation for that site and also for haplotype reconstruction.) Interestingly, we observed two fairly common non-synonymous SNPs in our samples, at nucleotide positions 13487 and 13542, that have not been reported previously. The 29 polymorphic sites detected by resequencing represent an overall occurrence of 1 site per 213 bp; 1 per 207 bp in introns and 1 per 177 bp in exons. The minor allele frequencies of 6 of the 7 non-synonymous SNPs are above 5% in one or more ethnic group in our sample. Only 5 out of 29 sites are shared among the 11 ethnic groups inhabiting mainland India. Wide differences in allele frequencies across groups were observed (Table 2). The Jarawas are monomorphic for 25 out of 29 sites (Table 2). The locations of the SNPs on the map of the *ICAM1* gene are provided in Figure 2(a).

At the *TNF* locus, 12 SNPs (9 transitions and 3 transversions) and 2 indels were identified (Table 3). Four new SNPs were discovered, of which 3 are present only in the Jarawa. One of these private sites among the Jarawa (C500T) is highly polymorphic; the frequency of

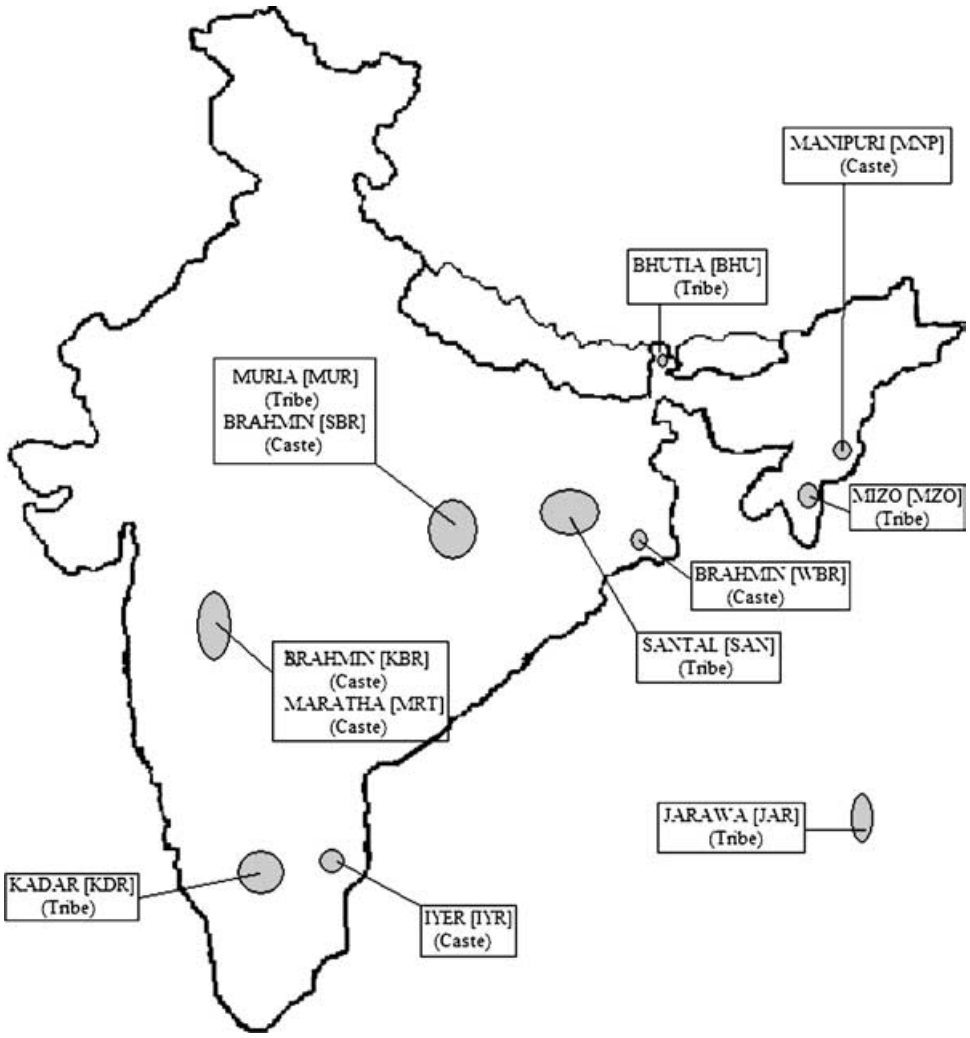

Figure 1 Geographical locations of sampling for the 12 ethnic groups of India studied.

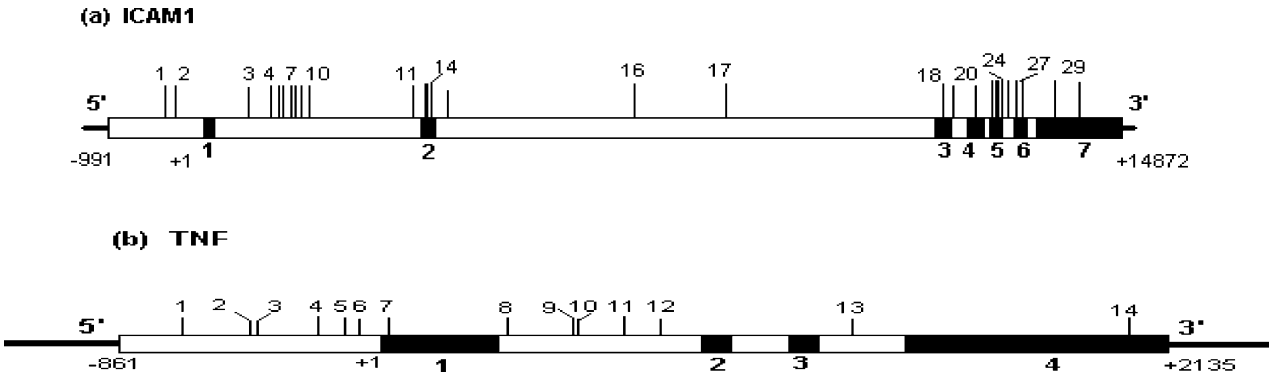

Figure 2 Structures of (a) *ICAM1* and (b) *TNF* genes showing the locations of the SNPs identified in this study.

the rarer allele at this site is 0.343. Wide variation in allele frequencies across populations were observed (Table 3). The Manipuris and the Santals are monomorphic at all except one (C56T) and two (G489A and A2053C) sites, respectively. The locations of the SNPs on the map of the *TNF* gene are provided in Figure 2(b).

The gene diversity across populations varies between 5–10% at the *ICAM1* locus (Table 2), while there is a

Table 2 Characteristics of observed single nucleotide polymorphisms in and around the ICAM1 gene and their estimated frequencies in 12 ethnic groups from India

Sl No.	Position & Nucleotide Change[a]	Characteristics Region	Amino Acid Change	Whether new or reported dbSNP ID	Nucleotides flanking the SNP site 5′	3′	Frequency of the Minor Allele[b] (±se) Population Code BHU	MZO	MNP	SAN	WBR	KAD	IYR	MUR	SBR	MRT	KBR	JAR
1	**A-785del**	Promoter		Reported rs5030389	GCC	GCG	0.038 ±.038	0.024 ±.024			0.031 ±.031	0.063 ±.043	0.088 ±.049	0.031 ±.031	0.031 ±.031	0.033 ±.033	0.094 ±.052	
2	C-667T	Promoter		New	GCC	TCG			0.045 ±.044									
3	A493C	Intron-1		New	TAC	GTT							0.029 ±.029					
4	**C503T**	Intron-1		Reported rs5030340	CAG	TGT	0.038 ±.038	0.024 ±.024			0.031 ±.031	0.063 ±.043	0.088 ±.049	0.031 ±.031	0.031 ±.031	0.033 ±.033	0.094 ±.052	
5	T840C	Intron-1		New	GTT	GGG								0.031 ±.031				
6	**C958G**	Intron-1		New	GGG	GAA								0.063 ±.043				
7	**C1066G**	Intron-1		New	ATC	CAG		0.024 ±.024	0.045 ±.044	0.031 ±.031	0.031 ±.031	0.031 ±.031		0.063 ±.043				
8	G1076A	Intron-1		New	CTC	GGA							0.029 ±.029					
9	G1110C	Intron-1		New	CAC	AGG		0.024 ±.024	0.045 ±.024	0.031 ±.031		0.045 ±.044		0.033 ±.033				
10	G1195C	Intron-1		Reported rs3093035	AGC	TTC							0.029 ±.029					
11	C3642T	Intron-1		New	CGC	TCT											0.033 ±.033	
12	**G3757A**	Exon-2	Q54Q	New	CCA	CCC						0.063						

Table 2 Continued.

Sl No.	Position & Nucleotide Change[a]	Characteristics Region	Amino Acid Change	Whether new or reported dbSNP ID	Nucleotides flanking the SNP site 5′	3′	Frequency of the Minor Allele[b] (±se) Population Code BHU	MZO	MNP	SAN	WBR	KAD	IYR	MUR	SBR	MRT	KBR	JAR
13	**A3762T**	Exon-2	K56M	Reported rs5491	CCA	GTT		0.024 ±.024	0.045 ±.044	0.031 ±.031	0.031 ±.031	±.043 0.031 ±.031		0.063 ±.043				
14	G3784A	Exon-2	P63P	New	CCC	TTG	0.038 ±.038	0.024 ±.024										
15	**C3965G**[c]	Intron-2		Reported rs5030352	ACC	GGT	0.385 ±.095	0.286 ±.070	0.227 ±.089	0.438 ±.088	0.375 ±.086	0.281 ±.079	0.412 ±.084	0.375 ±.086	0.313 ±.082	0.214 ±.078	0.462 ±.098	0.206 ±.049
16	T7175C	Intron-2		New	ACA	GAC			0.045 ±.044	0.031 ±.031								
17	**G8880C**	Intron-2		Reported rs281432	TTT	TGA	0.462 ±.098	0.619 ±.075	0.545 ±.106	0.250 ±.077	0.438 ±.088	0.500 ±.088	0.441 ±.085	0.406 ±.087	0.531 ±.088	0.367 ±.088	0.500 ±.088	0.514 ±.060
18	G12625A	Exon-3	R193Q	New	TGA	GCC								0.031 ±.031				
19	C12739T	Intron-3		New	ATC	GGT		0.024 ±.024										
20	**G13014A**	Exon-4	G241R	Reported rs1799969	GAC	GGC					0.063 ±.043						0.094 ±.052	
21	**C13430T**	Exon-5	P352L	Reported rs1801714	GCC	GAG									0.063 ±.043			

Table 2 Continued.

Sl No.	Position & Nucleotide Change[a]	Characteristics Region	Amino Acid Change	Whether new or reported dbSNP ID	Nucleotides flanking the SNP site 5'	3'	Frequency of the Minor Allele[b] (±se) Population Code BHU	MZO	MNP	SAN	WBR	KAD	IYR	MUR	SBR	MRT	KBR	JAR
22	**C13470T**	Exon-5	N365N	New	CAA	GGG			0.045 ±.044	0.031 ±.031		0.031 ±.031		0.063 ±.043				
23	**G13487T**	Exon-5	C371F	New	CCT	CTC			0.136 ±.073			0.031 ±.031	0.206 ±.069			0.067 ±.046	0.063 ±.043	
24	**G13542T**	Exon-5	E389D	New	GGA	CTT	0.308 ±.091	0.368* ±.078	0.364 ±.103	0.063 ±.043	0.031 ±.031	0.156 ±.064	0.147 ±.061	0.375* ±.086	0.094 ±.052	0.467* ±.091	0.219 ±.073	
25	C13668T	Intron-5		New	CAT	GTG						0.031 ±.031						
26	**C13900T**	Exon-6	T467T	New	TCA	CCG				0.063 ±.043								
27	**A13905G**	Exon-6	E469K	Reported rs5030382	CGC	AGG	0.192 ±.024	0.286 ±.070	0.409 ±.105	0.531 ±.088	0.594 ±.087	0.406 ±.087	0.471 ±.086	0.531 ±.088	0.469 ±.088	0.533 ±.091	0.313 ±.082	0.486 ±.060
28	**G14195A**	Exon-7 (3'UTR)		Reported rs2071440	CCC	GGA					0.031 ±.031	0.094 ±.052	0.059 ±.040	0.031 ±.031	0.033 ±.033		0.125 ±.058	
29	**C14588T**	Exon-7 (3'UTR)		Reported rs3093032	AGG	CCC	0.346 ±.093	0.200 ±.063	0.136 ±.073	0.031 ±.031	0.219 ±.073	0.094 ±.052	0.147 ±.061	0.094 ±.052	0.094 ±.052	0.033 ±.033	0.156 ±.064	0.071 ±.031
Gene Diversity(±se)							0.085 ±.032	0.084 ±.029	0.100 ±.030	0.069 ±.026	0.080 ±.029	0.097 ±.027	0.103 ±.031	0.104 ±.031	0.074 ±.028	0.075 ±.030	0.104 ±.030	0.051 ±.026

[a]Nucleotide positions have been counted from the transcriptional start site. SNPs indicated in boldface have been considered for haplotype determination.
[b]The allele with a lower frequency in the pooled sample is designated as the minor allele. Blank cells frequencies indicate zero frequencies.
[c]A third allele T was detected as GT heterozygotes in two KBR individuals. These two individuals have been excluded from allele frequency estimation.
*Significantly ($p<0.05$) deviated from Hardy-Weinberg equilibrium.

Table 3 Characteristics of observed single nucleotide polymorphisms in and around the *TNF* gene and their estimated frequencies in 12 ethnic groups from India

Sl	Position & Nucleotide	Characteristics		Nucleotides flanking the SNP site		Frequency of the Minor Allele[b] (±se) Population Code											
No.	Change[a]	Region	Whether new or reported dbSNP ID	5'	3'	BHU	MZO	MNP	SAN	WBR	KAD	IYR	MUR	SBR	MRT	KBR	JAR
1	A-572C	Promoter	Reported (rs4248161)	GCT	CCC							0.059 ± .04		0.031 ± .031	0.067* ± .046	0.067* ± .046	
2	G-308A	Promoter	Reported (rs1800629)	ATG	GGA	0.038 ± .038				0.031 ± .031	0.033 ± .033	0.029 ± .029	0.200* ± .073	0.061 ± .043			
3	G-303A	Promoter	New	GAC	GGG											0.100	± .036
4	G-238A	Promoter	Reported (rs361525)	ATC	GAG	0.154 ± .071	0.024 ± 024			0.125 ± .058	0.031 ± .031	0.029 ± .029	0.031 ± .031	0.033 ± .033	0.033 ± .033	0.094 ± .052	
5	T-1717A	Promoter	New	AGT	TTC					0.031 ± .031	0.125 ± .058	0.029 ± .029	0.031 ± .031	0.067 ± .046			
6	C-14T	Promoter	New	ACA	CCA												0.157 ± .043
7	C56T	Exon1 (5'UTR)	Reported (rs3093660)	AGA	CCC		0.048 ± .033	0.182 ± .082									
8	G420A	Intron1	Reported (rs3093661)	GAC	CAA	0.154 ± .071	0.024 ± .024			0.125 ± .058	0.031 ± .031	0.029 ± .029	0.031 ± .027	0.033 ± .033	0.033 ± .033	0.094 ± .052	
9	G489A	Intron1	Reported (rs1800610)	AAC	TGG	0.077 ± .052	0.167 ± .058		0.219 ± .073	0.094 ± .052	0.094 ± .052	0.118 ± .054	0.187 ± .069	0.133 ± .062	0.2 ± .073	0.062 ± .043	

Table 3 Continued.

Sl	Position & Nucleotide	Characteristics		Nucleotides flanking the SNP site		Frequency of the Minor Allele[b] (±se) Population Code											
No.	Change[a]	Region	Whether new or reported dbSNP ID	5'	3'	BHU	MZO	MNP	SAN	WBR	KAD	IYR	MUR	SBR	MRT	KBR	JAR
10	C500T	Intron1	New	AGA	GGG												0.343 ±.057
11	AATG Indel at 625	Intron1	Reported (rs4645841)	GAA	CAA		0.095 ±.045			0.031 ±.031	0.031 ±.031		0.062 ±.043	0.067 ±.046			
12	AG Indel at 731	Intron1	Reported (rs4645842)	GAG	CGG		0.048 ±.033			0.031 ±.031	0.094 ±.052		0.031 ±.031	0.033 ±.033			0.129 ±.04
13	A1304G	Intron3	Reported (rs3093664)	GGG	TTG	0.154 ±.071	0.024 ±.024			0.156 ±.064	0.187 ±.069	0.059* ±.040	0.062 ±.043	0.100 ±.046	0.036 ±.035	0.133 ±.062	0.147 ±.043
14	A2053C	Exon4 (3'UTR)	Reported (rs3093665)	CTC	ACC				0.062* ±.043	0.031 ±.031	0.187 ±.069		0.031 ±.031	0.067 ±.046			0.143 ±.042
Gene Diversity(±se)						0.074 ±.030	0.056 ±.022	0.022 ±.022	0.033 ±.026	0.086 ±.026	0.104 ±.031	0.048 ±.017	0.085 ±.029	0.084 ±.020	0.047 ±.024	0.060 ±.023	0.117 ±.041

[a]Nucleotide positions have been counted from the transcriptional start site.

[b]The allele with a lower frequency in the pooled sample is designated as the minor allele. Blank cells frequencies indicate zero frequencies.

*Significantly ($p<0.05$) deviated from Hardy-Weinberg equilibrium.

larger variation (9–12%) across populations at the *TNF* locus.

Haplotype Frequencies

Frequencies of haplotypes at the *ICAM1* locus were estimated (Table 4) using allele frequency data from only those 17 polymorphic sites at which the frequency of the rarer allele exceeded 0.05 in at least one population. A total of 61 haplotypes were present, about 34% (19 of 61) of which are shared by at least two groups. Three haplotypes – H1 (21%), H5 (14%) and H9 (12%) – are the most frequent. Notable are two haplotypes (ACCCGAGCGCCGGCAGC and ACCCGAGCGCCGGCAGT) with frequencies 5% and 10%, respectively, that are present among the Jarawa. The southern-Indian Brahmin group, the Iyer, harbours the largest number of haplotypes (16), while the Jarawas harbour the lowest number (8).

At the *TNF* locus, 36 haplotypes were observed (Table 5), of which 11 are shared among groups. Haplotype H1 constitutes 62.5% of the *TNF* gene pool in India, outnumbering all other haplotypes. Similar to the *ICAM1* locus, at the *TNF* locus also the Jarawas revealed a deviant haplotype frequency distribution compared to the other populations from mainland India.

Haplotype diversities at both loci showed similar patterns as those of gene diversities.

Linkage Disequibrium

At each locus, we estimated the coefficient of linkage disequilibrium, D', for every pair of polymorphic sites, separately for each population. Both loci show considerable variation in the estimates of D' across populations (detailed results not shown). At the *TNF* locus, there are only two SNPs that are present in all 12 populations. Therefore, results pertaining to variation in LD values across populations are not shown for this locus. Further, for those SNPs that are present in multiple populations, the extent of variation in LD across populations at this locus is not as pronounced as for the *ICAM1* locus. In Figure 3, therefore, we have presented the values of D' for all pairs of sites that are polymorphic in the vast majority of the ethnic groups at the *ICAM1* locus.

Genetic Affinities and Differentiation

Based on the haplotype frequencies of the *ICAM1* and *TNF* loci, we carried out a principal components analysis. The bidimensional plot depicting affinities among the populations based on the values of the first two principal components (that explain about 30% of the total variance in haplotype frequencies) is presented in Figure 4. No strong clustering of populations belonging to the same social, geographical, or linguistic group is observed. This finding was also corroborated by a cluster analysis performed using the neighbor-joining method (results not presented). The F_{st} values among the 11 populations of mainland India (the Jarawas were excluded from the analysis because they possess many private polymorphisms), grouped by geographical region of habitat, socio-cultural category and linguistic affiliation, indicated similar levels of genetic differentiation for the various groupings. Genetic differentiation at the *ICAM1* locus is higher than the *TNF* locus (Table 6). The analysis of molecular variance (AMOVA) results (Table 6) indicated that the extent of genetic variation attributable to between-group differences is quite low; and that among populations within groups is only slightly higher. At both loci, most of the genetic variation is attributable to differences between individuals within populations.

Discussion

In this study, we have examined the patterns of DNA sequence variation in and around the genes coding for *ICAM1* and *TNF*, in 12 diverse ethnic groups of India, with a view to quantifying the nature and extent of the variation, and to analyze the patterns of variation with respect to population history and ethnic background. The primary motivations for undertaking this study were (a) the recent emphasis for the need of SNP discovery and validation studies in disparate global populations (Carlson *et al.* 2003; Reich *et al.* 2003), (b) the need to explore variation in linkage disequilibrium across populations, to provide a clearer understanding of the statistical intricacies of disequilibrium mapping of human diseases (Chattopadhyay *et al.* 2003), and (c) to examine the causes of maintenance of variation at functionally important genomic regions. The *ICAM1*

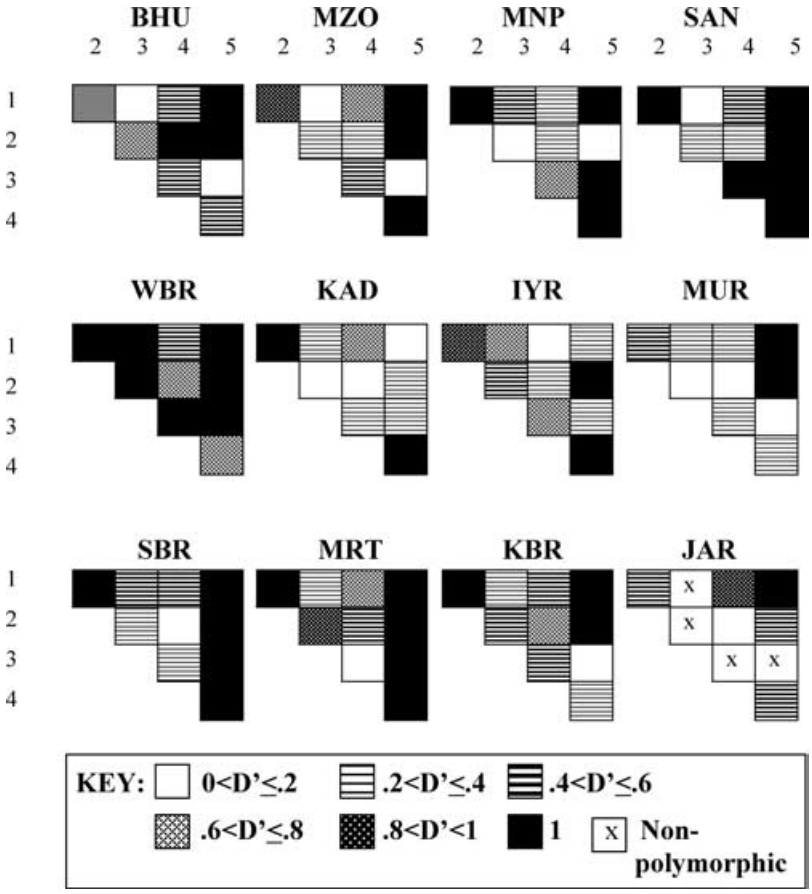

Figure 3 Population-wise variation in estimated coefficients of linkage disequilibrium (D') between pairs of the 5 *ICAM1* SNP-loci which are polymorphic across most ethnic groups. (In this figure, loci 1, 2, 3, 4 and 5 correspond, respectively, to C3965G, G8880C, G13542T, A13905G and C14588T.)

and *TNF* genes were selected in view of their functional and correlated roles in inflammatory processes and immune cell responses in a wide range of diseases (Dobbie *et al.* 1999; Striz *et al.* 1999; Bjornsdottir & Cypcar, 1999).

Our study has shown that Indian ethnic groups harbour SNPs that remain unreported in the major SNP databases. Some reported SNPs have not been found in our study populations. The SNP frequencies also show wide variation across populations, including some private polymorphisms among the Jarawa. To summarize, a comparison of the variant sites observed in the two genomic regions among Indian populations with those catalogued in dbSNP (Build No. 120) revealed that: (i) 21 polymorphic sites found in individuals of either African and/or European descent are also common to Indian samples; (ii) 22 new SNPs were discovered in Indian samples, of which 11 are rare and private to one group or region; (iii) 45 variable sites reported in dbSNP, of which 6 have frequencies greater than 10% in either European or African Americans from the same target region, could not be validated in our samples. These findings have obvious implications for case-control studies and, in part, may explain why disease-marker associations reported in one population cannot be replicated in another population. To exemplify, an association between the E469K gene polymorphism at the *ICAM1* locus and Alzheimer's disease (AD) was reported among Italian patients, indicating the role of the *ICAM1* gene in the pathophysiology of neuro-degenerative diseases (Pola *et al.* 2003). However, this association was not found among Finish patients (Mattila *et al.* 2003). We, in this study, have detected a wide variation in allele frequency for the E469K polymorphism among the groups

Table 4 Estimated frequencies of major[a] haplotypes at the *ICAM1* locus in 12 ethnic groups from India

ID #	Haplotype[c]	Frequency[b]											
		BHU	MZO	MNP	SAN	WBR	IYR	KAD	MUR	SBR	MRT	KBR	JAR
H1	ACCCGACCGCCGGCAGC	.417	.371	.227	.104	.046	.284	.244	.187	.266	.040	.423	.069
H2	. . . . . GC . . . . . T	.189	.125		.031	.147				.100			
H3	. . . . . GG . . . . T . . . T	.118	.085	.090		.031			.046		.035		
H4	. . . . . G . . . . T . G . .	.107	.132					.025			.064		
H5	. . . . . G . . . . . G . .	.046		.227	.312	.067	.058	.093	.249	.133	.324		.183
H6	. . . . . T	.044	.076	.136							.205		
H9	. . . . . G . .		.083	.136	.051	.352	.068	.062		.193			.272
H10	. . . . . GG		.026	.022	.270	.021	.176	.187		.073	.131	.038	.032
H14	. . . . . T . G . .		.022		.031			.036	.125	.040	.075		
H18	. . . . . GG . . . . T			.022				.031	.140	.060	.011		
H25	. . . . . G					.065		.036			.040		.268
Other 50 Haplotypes		.079	.080	.140	.201	.271	.414	.286	.253	.135	.075	.539	.176
No. of Haplotypes		8	11	11	10	14	16	15	13	10	11	14	8
Haplotype		.781	.829	.902	.824	.860	.899	.885	.881	.873	.843	.824	.812
Diversity		±	±	±	±	±	±	±	±	±	±	±	±
(± se)		.064	.046	.034	.044	.049	.034	.04	.032	.034	.045	.074	.022

[a]A haplotype with an estimated frequency > 5 in the pooled sample is designated as a major haplotype.
[b]Blank cells represent zero frequencies.
[c]Based on 17 polymorphic sites corresponding to serial numbers 1, 4, 6, 7, 12, 13, 15, 17, 20, 21, 22, 23, 24, 26, 27, 28 and 29 of Table 2.

Table 5 Estimated frequencies of major[a] haplotypes in *TNF* gene in 12 ethnic groups from India

ID #	Haplotype	Frequency[b]											
		BHU	MZO	MNP	SAN	WBR	IYR	KAD	MUR	SBR	MRT	KBR	JAR
H1[c]	AGGGTCCGGC6 I A A	.726	.706	.818	.750	.647	.781	.533	.567	.665	.714	.750	.279
H2	. . . A . . . A . . . . G .	.082				.027			.033	.033	.035	.071	
H3	. . . . . A	.043	.103		.188	.065	.031	.100	.133	.100	.178	.071	
H6	. A	.038				.031			.167	.035			
H16	. . . . . .DGC					.031		.100	.033				.118
H33	. . . . . T												.338
H34	. . . . . T												.147
Other 29 Haplotypes		.111	.191	.182	.062	.199	.188	.267	.067	.167	.073	.108	.118
No. of Haplotypes		7	9	2	4	9	8	8	7	9	5	5	7
Haplotype		.470	.486	.311	.413	.570	.701	.395	.650	.556	.470	.436	.775
Diversity		±	±	±	±	±	±	±	±	±	±	±	±
(±se)		.119	.093	.106	.094	.102	.084	.11	.084	.106	.102	.112	.025

[a]A haplotype with an estimated frequency > 5 in the pooled sample is designated as a major haplotype.
[b]Blank cells represent zero frequencies.
[c]6 indicates (AATG) copy number at position 625; I and D represent AG insertion and deletion, respectively, at position 731.

studied, clearly showing that unless the issue of population stratification is adequately addressed in designing case-control association studies, false positive and false negative error rates may be very high. Another example is that a well-studied polymorphism at the *TNF* locus, that results in a G to A transition at position -308, was found to be strongly associated with cerebral malaria (Wilson *et al.* 1997). Two other alleles at this locus, TNF-376A and TNF-238A, are also reported to be associated with susceptibility to severe malarial anemia among children in Gambia and Kenya (Knight *et al.* 1999; McGuire *et al.* 1999). We did not find the TNF-376A polymorphism in our populations, and detected large variations in population frequencies of

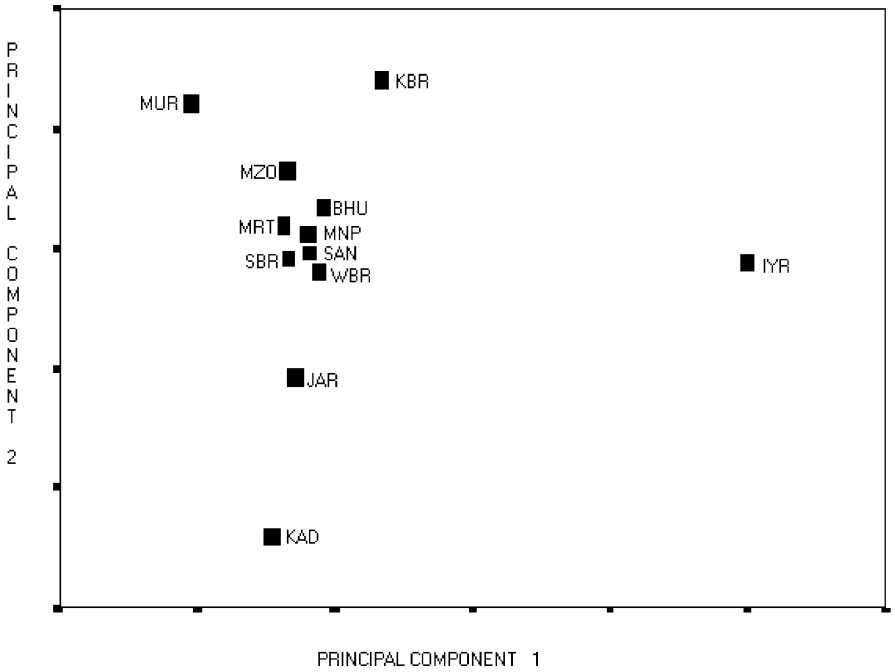

Figure 4 Bidimensional plot of the first two principal components extracted from the haplotype frequencies at the *ICAM1* and *TNF* loci, depicting the affinities among the 12 ethnic groups.

Table 6 Estimates of F_{st} and AMOVA results based on *ICAM1* and *TNF* haplotypes for different groupings of the populations studied

			% variation attributable to[a]			
	F_{st}		Among groups within groups		Among populations	
Grouping[b]	*ICAM1*	*TNF*	*ICAM1*	*TNF*	*ICAM1*	*TNF*
5 groups: Geographical Region	0.058 0.011	0.011	0.70	0.00	5.10	1.21
2 groups: Caste and Tribe	0.049 0.014	0.014	0.00	0.31	5.31	1.12
4 groups: Linguistic Category	0.058 0.011	0.011	0.73	0.07	5.14	1.05

[a] The percentages of variation attributable to among individuals within groups are obtainable by subtracting from 100 the sum of the percentages of total variation attributable to the other two sources of variation shown here.
[b]The Jarawa was excluded from this analysis.

TNF-308A and TNF-238A. While it is possible that we have missed the TNF-376A polymorphism because of small sample sizes of individual ethnic groups, this possibility seems unlikely since our total sample size is reasonably large.

Indian populations show high, but variable, levels of genomic diversity (Tables 2-5). Large variation is also observed in the extent of linkage disequilibrium at the *ICAM1* locus (Figure 3). These features can be explained in part by the variable evolutionary histories of Indian ethnic groups (Basu *et al.* 2003), including strong founder and drift effects, but nevertheless underscore their importance in designing case-control association studies.

Acknowledgements

This study was supported in part by a grant from the Department of Biotechnology, Government of India, to PPM.

References

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

Bjornsdottir, U. S. & Cypcar, D. M. (1999) Asthma: An inflammatory mediator soup. *Allergy* **54**, 55–61.

Carlson, C. S., Eberle, M. A., Reider, M. J., Smith, J. D., Kruglyak, L. & Nickerson, D. A. (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* **33**, 518–521.

Chattopadhyay P., Patkis, A. J., Mukherjee, N., Iyengar, S., Odunsi, A., Okonofua, F., Bonne-Tamir, B., Speed, W., Kidd, J. R. & Kidd, K. K. (2003) Global survey of haplotype frequencies and linkage disequilibrium at the RET locus. *Eur J Hum Genet* **10**, 760–769.

Dobbie, M. S., Hurst, R. D., Klein, N. J. & Surtees, R. A. (1999) Upregulation of intercellular adhesion molecule-1 expression on human endothelial cells by tumour necrosis factor-alpha in an in vitro model of the blood-brain barrier. *Brain Res* **830**, 330–336

Fernandez-Reyes, D., Craig, A. G., Kyes, S. A., Peshu, N., Snow, R. W., Berendt, A. R., Marsh, K. & Newbold, C. I. (1997) A high frequency African coding polymorphism in the N-terminal domain of ICAM-1 predisposing to cerebral malaria in Kenya. *Hum Mol Genet* **6**, 1357–1360.

Gimenez, F., de Lagerie, S. B., Fernandez, F., Pino, P. & Mazier, D. (2003) Tumor necrosis factor α in the pathogenesis of cerebral malaria. *Cell Mol Life Sci* **60**, 1623–1635.

Knight, J. C., Udalova, I., Hill, A. V., Greenwood, B. M., Peshu, N., Marsh, K. & Kwiatkowski, D. (1999) A polymorphism that affects OCT-1 binding to the TNF promoter region is associated with severe malaria. *Nat Genet* **22**, 145–150.

Majumdar, P. & Majumder, P. P. (1999) HAPLOPOP: A computer program package to estimate haplotype frequencies from genotype frequencies via the EM algorithm. *AHGU Tech Rep*. Indian Statistical Institute, Kolkata.

Mattila, K. M., Hiltunen, M., Rinne, J. O., Mannermaa, A., Röyttä, M., Alafuzoff, I., Laippala, P., Soininen, H. & Lehtimäki, T. (2003) Absence of association between an intercellular adhesion molecule 1 gene E469K polymorphism and Alzheimer's disease in Finnish patients. *Neurosci Lett* **337**, 61–63.

McGuire, W., Knight, J. C., Hill, A. V., Allsopp, C. E., Greenwood, B. M. & Kwiatkowski, D. (1999) Severe malarial anemia and cerebral malaria are associated with different tumor necrosis factor promoter alleles. *J Infect Dis* **179**, 287–289

Meager, A. (1999) Cytokine regulation of cellular adhesion molecule expression in inflammation. *Cytokine Growth Factor Rev* **10**, 27–39.

Pola, R., Flex, A., Gaetani, E., Papaleo, P., De Martini, D., Gerardino, L., Serricchio, M., Pola, P. & Berbabei, R. (2003) Intercellular adhesion molecule-1 K469E gene polymorphism and Alzheimer's disease. *Neurobiol Aging* **24**, 385–387.

Reich, D. E., Gabriel, S. B. & Altshuler, D. (2003) Quality and completeness of SNP databases. *Nat Genet* **33**, 457–458.

Saitou, N. & Nei, M. (1998) The neighbour-joining method: a new method for reconstrucing phylogenetic trees. *Mol Biol Evol* **4**, 406–425

Striz, I., Mio, T. & Adachi, Y. (1999) IL-4 induces ICAM-1 expression in human bronchial epithelial cells and potentiates TNF-alpha. *Am J Physiol* **277**, L58–64.

Wilson, A. G., Symons, J. A., McDowell, T. L., McDevitt, H. O., Duff, G. W. (1997) Effects of a polymorphism in the human tumor necrosis factor alpha promoter on transcriptional activation. *Proc Natl Acad Sci USA* **94**, 3195–3199.

Received: 15 April 2004
Accepted: 01 June 2004