Single nucleotide polymorphisms in two genes among the Jarawa, a primitive tribe of the Andaman and Nicobar Islands

Sunita Singh*, Sanghamitra Sengupta[†], Kanchan M. Murhekar*, Subhash C. Sehgal* and Partha P. Majumder[†,#]

*Regional Medical Research Centre, Indian Council of Medical Research, Port Blair 744 101, India
[†]Anthropology and Human Genetics Unit, Indian Statistical Institute, Kolkata 700 108, India

Discovery and validation of single nucleotide polymorphisms (SNPs) in the human genome is an active area of interest, because of their usefulness in evolutionary and disease-association studies. We have resequenced 70 chromosomes from the Jarawa, a Negrito tribal group inhabiting the Andaman and Nicobar archipelago, for a ≈ 10 kb genomic region spanning two genes, *ICAM1* and *TNF*-α. We have discovered three new SNPs in the *TNF*-α gene. There is low sequence variability among the Jarawa, possibly due to founder effect and genetic drift.

GENETIC polymorphisms are useful tools for the reconstruction of evolutionary histories of populations and in the study of susceptibilities to common diseases[1]. The most common type of polymorphism in human genome is single nucleotide polymorphisms (SNPs). The great majority of these changes is stable and tends to remain biallelic. A database documenting these SNPs in global populations is dbSNP (www.ncbi.nlm.nih.gov/SNP/). Analysis of the data in dbSNP (www.ncbi.nlm.nih.gov/SNP/snp_summary.cgi) showed a density distribution of ≈ 18 SNPs per 10 kb of human genome, though the sequence-variation profiles of many populations, particularly those who are inhabitants of the Indian subcontinent, still remain under-represented. Since SNPs are being widely used in case-control association studies of common diseases[2], it is crucial to discover and validate these in multiple populations, especially because estimates of variation in SNP frequencies across populations can provide an assessment of the nature and extent of population substructuring, which need to be taken into account in disease-association studies[1].

In this communication we present a systematic survey of DNA sequence variation in two functionally coding regions in a nearly extinct ethnic group, Jarawa, who live in the Andaman Islands. The Jarawas are a numerically small (their current population size is estimated to be about 200 individuals), isolated autochthonous group representing remnants of one of the very few Negrito stocks in Asia. They primarily inhabit almost inaccessible regions of the middle and south Andamans. This ancient aboriginal, hunter–gatherer tribe may serve as a missing link in population genetics studies that aim to address the issues of settlement of eastern and southern Asia after anatomically modern human came out of Africa.

The target DNA regions in this study were the genes coding for tumour necrosis factor-α (TNF-α) and intercellular adhesion molecule 1 (*ICAM1*). Both these genes play important roles in a variety of human diseases. *TNF*-α is a pro-inflammatory cytokine known to be essential in the control of many intracellular infectious agents in humans[3]. Elevated levels of *TNF*-α have been implicated in the pathogenesis of infectious diseases such as malaria[4], meningococcal diseases[5,6], schistosomiasis[7] and some autoimmune diseases[8]. The endothelial receptor *ICAM1* is upregulated by TNF and other cytokines. The pathogenecity of *Plasmodium falciparum* has also been ascribed to the ability of the infected red blood cells to adhere to the endothelium[9]. A number of endothelial molecules, including *ICAM1* (CD54) have been identified as receptors of parasitized cells[10]. As mentioned earlier, since no data on polymorphisms in autosomal genes at the molecular level are available from this tribe and since these genes (TNF-α and *ICAM1*) play important roles in susceptibilities to infectious diseases, especially malaria which has already been found in this tribe, we chose these two genes. However, the primary objective of our study was to generate base-line polymorphism data that can be used for future evolutionary and disease-association studies.

Blood samples were collected by the Regional Medical Research Centre, Indian Council of Medical Research, Port Blair, in collaboration with the Health Services Department of the Andaman and Nicobar Administration, when there was an outbreak of unknown fever among the Jarawas some years ago. About 60 blood samples were collected in the Kadamtala Public Health Centre, located within the Jarawa habitation in the Middle Andaman. These samples were analysed and it was established that many of these individuals were suffering from malaria. Further studies showed[11] that there was also on alarming prevalence of hepatitis-B infection among the Jarawa. When these results were brought to the attention of the Government of India, an Expert Group constituted by the Ministry of Home Affairs recommended that further genetic studies be conducted among all primitive tribes of the Andaman and Nicobar Islands, including the Jarawa. Before undertaking research using these collected blood samples, approval of the Ethics Committee of the Regional Medical Research Centre, Port Blair, was taken. The blood samples of 35 Jarawa individuals analysed in the present study are a subset of the original collection. Since it is almost impossible to communicate with the Jarawas, except in body language, no explicit informed consent could be obtained from any of the individuals. Except for obvious parent–offspring relationships, other forms of relationship among the individuals from whom blood sam-

[#]For correspondence. (e-mail: ppm@isical.ac.in)

ples were drawn were impossible to ascertain. To the best of our knowledge, therefore, parents and their offspring are not included in this subset of 35 individuals. Genomic DNA was isolated using a standard procedure (phenol–chloroform extraction) at Regional Medical Research Centre of the Indian Council of Medical Research, Port Blair. DNA sequencing and data analysis were performed at Indian Statistical Institute, Kolkata.

The genomic sequences which also include the 1000 bp upstream region of *ICAM1* (15 kb) and *TNF-α* (3085 bp) were obtained from UCSC Genome Browser (http://genome.ucsc.edu). The genes map to the chromosomal region 19p13.2–p13.3 and 6p21.3, respectively. The *ICAM1* gene contains seven exons and the *TNF-α* gene contains four exons. The genomic region encompassing *ICAM1* was repeat-masked using the program RepeatMasker2 (http://ftp.genome.washington.edu.cgi-bin/RepeatMasker). Appropriate primers for DNA amplification of the genomic regions were designed after repeat-masking. The 5′ and 3′ UTRs, and the exons and introns were completely sequenced, except the repeat-masked regions. The total number of bases resequenced was 6088 and 3085, respectively for the *ICAM1* and *TNF-α* genes.

DNA amplification (PCR) conditions were optimized using control samples. PCR products were treated with exonuclease I and shrimp alkaline phosphatase to remove the excess dNTPs and unused primers, 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 analysed using the PHRED software (http://www.mbt.washington.edu/phrap.docs/phred.html) which assigns quality scores to each base. The PHRED outputs for any given PCR amplicon were then aligned using the PHRAP software. The resulting assemblies were viewed using CONSED which allows identification of sequence differences as well as access to the individual chromatograms to scrutinize each putative variant. A subset of samples in which variant alleles were observed was sequenced using reverse primers to validate the variant.

After scoring the variants by inspecting CONSED output files, genotype frequencies were tabulated. Allele frequencies at each variant site were computed by the gene-counting method. Haplotypes were constructed and maximum-likelihood estimates of their frequencies were obtained via the EM-algorithm using HAPLOPOP[12].

A total of six polymorphic sites were detected in the *TNF-α* gene, among which two single nucleotide and one insertion/deletion (Indel) polymorphisms present in the gene have already been reported in NCBI's dbSNP database. We have discovered three new polymorphisms in Jarawas (Table 1). The only exonic (exon 4 and 3′ UTR) SNP found in *TNF-α* harbours an A-to-C transversion, while the remaining five changes lie either in the upstream region of the gene or in the intronic region. The minor allele frequencies at all six polymorphic loci are all ≥ 10%. Nine haplotypes based on these six polymorphic sites were observed; among which four were observed only once (Table 2). HT-1, HT-2, HT-3 and HT-5 together account for 85% of the chromosomes in the Jarawas (Table 2). They are related to each other by a single mutational difference in the haplotype string. HT-4 which accounts for 10% of the total chromosomes in the Jarawas, occupies a phylogenetically distant position from the cluster formed by HTs-1, 2, 3 and 5. The remaining haplotypes are possible products of recombination.

Four SNPs were detected by sequencing 6088 bp unique sequence segments embedded in 15,781 bp genomic region coding *ICAM1*. All four SNPs present in *ICAM1* have already been reported in dbSNP. One of these polymorphisms (rs5030382) harbours a non-conservative

Table 1. Description and frequency of SNPs found in *TNF-α* and *ICAM1* genes in Jawaras

Gene	Description of variation: nucleotide change (amino acid change, if any)	Whether new or reported (dbSNP ID)	Frequency (± SE) of major allele
TNF-α	G-483A	New	0.900 ± 0.036
	C-257T	New	0.842 ± 0.043
	C320T	New	0.657 ± 0.057
	AG Indel at 551	Reported (rs464582)	0.871 ± 0.040
	A1124G	Reported (rs3093664)	0.853 ± 0.043
	A1873C	Reported (rs3093665)	0.851 ± 0.042
ICAM1	C3908 G	Reported (rs5030352)	0.794 ± 0.049
	C8823G	Reported (rs281432)	0.515 ± 0.060
	A13848G (E469K)	Reported (rs5030382)	0.514 ± 0.060
	G14501A	Reported (rs2071440)	0.928 ± 0.031

Table 2. Haplotype and its frequency in *TNF-α* and *ICAM1* genes in Jarawas

Gene	Haplotype sl. no.	Haplotype description	Frequency (± SE)
TNF-α	HT1	G–C–T–D–A–A	0.342 ± 0.003
	HT2	G–C–C–D–A–A	0.315 ± 0.003
	HT3	G–T–C–D–A–A	0.142 ± 0.001
	HT4	G–C–C–I–G–C	0.098 ± 0.001
	HT5	A–C–C–D–A–A	0.041 ± 0.0005
	HT6	A–C–C–I–G–C	0.015 ± 0.0002
	HT7	G–C–C–D–G–C	0.014 ± 0.0002
	HT8	A–C–C–I–G–A	0.014 ± 0.0002
	HT9	A–T–C–I–G–C	0.014 ± 0.0002
ICAM1	HT1	C–G–G–T	0.236 ± 0.0025
	HT2	C–C–G–T	0.220 ± 0.0024
	HT3	C–G–A–C	0.192 ± 0.0022
	HT4	C–C–C–C	0.151 ± 0.0018
	HT5	G–C–C–C	0.088 ± 0.0011
	HT6	G–C–C–T	0.054 ± 0.0007
	HT7	G–G–C–C	0.040 ± 0.0005
	HT8	G–G–C–T	0.017 ± 0.0002

D, AG deletion; I, AG insertion.

amino acid change from glutamic acid (E) to lysine (K) at position 469 of the CD54 protein, while another (rs30933032) lies in the exon 7 of the gene (3′UTR). The remaining two polymorphisms are intronic. Like the SNPs present in *TNF*-α, the frequencies of *ICAM1* SNPs are also moderate to high (Table 1). A total number of eight haplotypes were observed, among which HTs-1 and 2 are equally frequent. Except for one haplotype (HT-8), all others occur more than once in the Jarawa gene pool. All *ICAM1* haplotypes can be joined sequentially by single-site mutational difference, though the possibility of recombination in this gene cannot be ruled out.

This DNA-marker study carried out among the Jarawas, is based on autosomal markers. The primary objective of this study was to examine the nature and extent of polymorphisms in two known autosomal genes that are known to play important roles in determining susceptibilities to infectious diseases, especially malaria which has been detected in this tribe. Other studies[13,14] have dealt with mitochondrial and Y-chromosomal markers in Andamanese tribes. In this study we have resequenced approximately 10 kb region from two autosomal loci in each of the thirty-five Jarawa individuals. The major observations emerging from the pattern of sequence variability of *TNF*-α and *ICAM1* genes in Jarawas are: (i) there is low DNA sequence variability, (ii) rare SNPs are absent, (iii) the group harbours new SNPs, and (iv) there is restricted variation in haplotypes. The general paucity of low frequency variations among the Jarawas indicates that they have not undergone any demographic expansion. The lack of large number of haplotypes suggests that this population is genetically homogeneous. Both these findings corroborate the fact that the Jarawas constitute a small and highly isolated, closed population. Possibly because of strong founder effect and small demographic size of the population, that results in strong genetic drift, the level of sequence variability is low. Alternatively, the paucity of sequence variation in *TNF*-α and *ICAMI* may be due to purifying selection. Both *TNF*-α and *ICAM1* are involved in inflammation and have significant roles in controlling the pathogenesis of various infectious and parasite-mediated diseases, particularly cerebral malaria. Therefore, it is likely that selection pressure is operative on these two autosomal genes. Our discovery of three new polymorphisms with relatively high heterozygosities in *TNF*-α gene suggests that SNPs already posted in the major databases, such as dbSNP, do not comprise an exhaustive list, and indicates the need to carry out more systematic SNP search in many more populations.

1. Brookes, A. J., The essence of SNPs. *Gene*, 1999, **234**, 177–186.
2. Latsi, P., Pantelidis, P., Vassilakis, D., Sato, H., Welsh, K. I. and Du Bois, R. M., Analysis of IL-12 p40 subunit gene and IFN-γ G5644A polymorphism in idiopathic pulmonary fibrosis. *Respir. Res.*, 2003, **4**, 6–13.
3. Beutler, B. and Grau, G. E., Tumor necrosis factor in the pathogenesis of infectious disease. *Crit. Care Med. (Suppl.)*, 1993, **21**, S423–S435.
4. Kwiatkowski, D. *et al.*, TNF concentration in fatal cerebral, non-fatal cerebral, and uncomplicated *Plasmodium falciparum* malaria. *Lancet*, 1999, **336**, 1201–1204.
5. Wensdrop, R. G., Langermans, J. A., de Bel, C. E., Meinders, A. E. and Van Dissel, J. T., Release of tumor necrosis factor: An innate host characteristic that may contribute to the outcome of meningococcal disease. *J. Infect. Dis.*, 1995, **171**, 1057–1060.
6. Riordan, F. A., Marzouk, O., Thompson, A. P. J., Sills, J. A. and Hart, A. C., Proinflammatory and anti-inflammatory cytokines in meningococcal disease. *Arch. Dis. Child.*, 1996, **75**, 453–454.
7. Amiri, P., Lockley, R. M., Parslow, T. G., Sadick, M., Rector, E., Ritter, D. and McKerrow, J. H., Tumor necrosis factor alpha restores granulomas and induces parasite egg-laying in schistosome-infected SCID mice. *Nature*, 1992, **356**, 565–566.
8. Wilson, A. G., di Giovine, F. S. and Duff, G. W., Genetics of tumor necrosis factor in autoimmune, infectious and neoplastic diseases. *J. Inflamm.*, 1995, **45**, 1–12.
9. Pasloske, B. L. and Howard, R. J. D., Malaria, the red cell, and the endothelium. *Annu. Rev. Med.*, 1994, **45**, 283–295.
10. Berendt, A. R. *et al.*, Molecular mechanism of sequestration in Malaria. *Parasitology*, 1994, **108**, S19–S28.
11. Murhekar, M. V., Murhekar, K. M. and Sehgal, S. C., Alarming prevalence of hepatitis-B infection among the Jarawas – a primitive Negrito tribe of Andaman and Nicobar Islands, India. *J. Viral Hepat.*, 2003, **10**, 232–233.
12. Majumdar, P. and Majumder, P. P., HAPLOPOP: A computer program package to estimate haplotype frequencies from genotype frequencies via the EM algorithm, AHGU Tech. Rep. 1/99, Indian Statistical Institute, Kolkata, 1999.
13. Thangaraj, K. *et al.*, Genetics affinities of the Andaman islanders, a vanishing human population. *Curr. Biol.*, 2003, **13**, 86–89.
14. Endicott, P. *et al.*, The genetic origins of the Andaman islanders. *Am. J. Hum. Genet.*, 2003, **72**, 178–184.