J Vector Borne Dis 46, September 2009, pp. 230–236

Evolutionary insights into duffy gene in mammalian taxa with comparative genetic analysis

Gauri Awasthi[a], Aditya P. Dash[b] & Aparup Das[a]

[a]*National Institute of Malaria Research, New Delhi;* [b]*World Health Organization-South East Asia Regional Office, New Delhi, India*

Abstract

Background & objectives: Evolutionary analyses of genes conserved across taxa are keys to understand the complexity of gene and genome variation. Since malaria is a highly infectious human disease and its susceptibility in human is genetically controlled, characterization and evolutionary analyses of such genes are of prime importance to understand genetic mechanisms of disease susceptibility. In the present study we have characterized and performed comparative genomic analyses of the human Duffy gene responsible for malaria pathogenesis in nine different mammalian taxa.

Methods: DNA sequences of human duffy gene were downloaded from public domain and have been characterized in detail and compared with eight other different mammalian taxa (*Pan troglodytes, Macaca mulatta, Pongo pygmaeus, Rattus norvegicus, Mus musculus, Monodelphis domestica, Bos taurus* and *Canis familiaris)*. Comparative and evolutionary analyses were performed using statistical software and tools.

Results: We observed that the genetic architecture of this gene was entirely different across all the nine taxa and a close similarity between *Homo sapiens* and *Pan troglodytes* (chimpanzee) was evident for several aspects of this gene. Comparisons on several aspects, such as ratio of coding and non-coding regions, total gene length number and size of introns and difference of number of nucleotides in human and chimpanzees have revealed interesting features. Phylogenetic inferences based on the Duffy gene among nine different taxa were found to be different than other genes previously studied.

Interpretation & conclusion: Most remarkably, human and chimpanzee were only 0.75% different in this gene. The results were discussed on the similarities between human and chimpanzee and gain of introns in human-chimpanzee clade with an inference on the role of evolutionary forces (mainly natural selection) in maintaining such variations across closely-related mammalian taxa.

Key words Duffy gene – malaria – natural selection – phylogenetic trees

Introduction

Genes of essential functions often are conserved across taxa but might take different functions according to need and thus evolve to perform similar or slightly dissimilar function specific to the taxa. Evolutionary forces play crucial roles in the architectural differences of such genes and comparative genetic analyses could unravel the precise roles of such forces[1]. Evolutionary forces may act in the whole gene or in a part of the gene (coding or non-coding) and a detail characterization of such genes across taxa could pinpoint where exactly evolution might have acted[2]. Determination of precise functional changes

that might have brought about by the organisms and the type of evolutionary forces that might have acted to preserve such essential functions could thus be possible with comparative gene analyses[3,4]. Thus, the gimmicks of evolution at the molecular level could only be understood by studying the characterization of genes in detail and compare them with closely related taxa.

The Duffy blood group system in human has been well-characterized. The Duffy antigen/receptor for chemokine (DARC) is a promiscuous chemokine receptor that also binds *Plasmodium vivax.* DARC belongs to a family of heptahelical chemokine receptors that includes specific (IL-8RA) and shared (IL-8RB) IL-8 receptors[5]. Duffy (FY) gene is a receptor on RBCs for the human malaria parasite *P. vivax* and is located in chromosome-1[6]. Three major alleles of this gene are known, FY*A, FY*B and FY*O and molecular detection of these alleles is possible by PCR[7]. The FY*A and FY*B alleles are distinguished by a missense mutation, which results in a single amino acid difference[8–11.] The FY*B is the ancestral allele while the non-FY*B alleles are derived ones[8]. The FY*O allele, with the Fy (a-b-) phenotype is due to a T-46C point mutation[11,12] on the FY*B gene promoter, which abolishes the erythroid gene expression by disrupting a binding site for the GATA-1 erythroid transcription factor[12,13]. Moreover, the FY (a-b-) phenotype has become resistant to *P. vivax* infection in sub-Saharan African population and the FY gene is conserved across mammalian species[14]. The sequence of this gene is available in majority of the mammalian taxa present in public domain. However, detailed genetic characterization and evolutionary analyses of this gene has not been done so far.

We have characterized the Duffy gene in humans (*Homo sapiens*) and eight other mammalian taxa (*Pan troglodytes*–chimpanzee, *Macaca mulatta*–rhesus monkey, *Pongo pygmaeus*– orangutan, *Rattus norvegicus*–brown rat, *Mus musculus*–mouse, *Monodelphis domestica*–opposum, *Bos taurus* – cow and *Canis familiaris*–dog) in detail and herewith present different analyses and infer phylogenetic positions of taxa in this gene. While species-specific differences were apparent, human and chimpanzee were clear outliers at the Duffy gene locus. Phylogenetic relationships among nine mammalian taxa revealed interesting patterns. The data were analyzed keeping the Duffy gene into account and known evolutionary relationships among the mammalian taxa into consideration.

Methods

Nucleotide sequences of the Duffy gene of nine mammalian taxa were downloaded from Ensembl database *(http://www.ensembl.org)*. These DNA sequences were subjected to different computational and statistical analyses. Total gene lengths, coding and non-coding sequences of this gene in different taxa were calculated and compared. Statistical analyses were conducted and correlation coefficients between different variables (total gene length, exon length and intron length) were compared across taxa. In order to infer evolutionary inter-relationships among various taxa at the Duffy gene, phylogenetic tree was constructed using the DNASTAR software[15]. Multiple sequence alignments and construction of neighbor-joining (NJ) phylogenetic tree were performed using MegAlign, a part of Lasergene (DNASTAR) software. Individual branch lengths were calculated using phylogeny option in the statistical programme VEGAZZ *(http://www.vegazz.net)*. Pearson's correlation coefficients were calculated using 'Analyze-it' software, an add-on to the MS Excel Software. For all statistical analyses $p<0.05$ was considered as level of significance. Bootstrap values and branch length values were calculated by using phylogeny option in DNASTAR software.

Results & Discussion

Characterization of Duffy gene across nine taxa revealed several interesting features. A schematic representation of the detailed coding versus non-coding contents of this gene is shown in Fig. 1. It is apparent that the lengths of coding exons vary across taxa as also the untranslated regions (UTRs) at both the

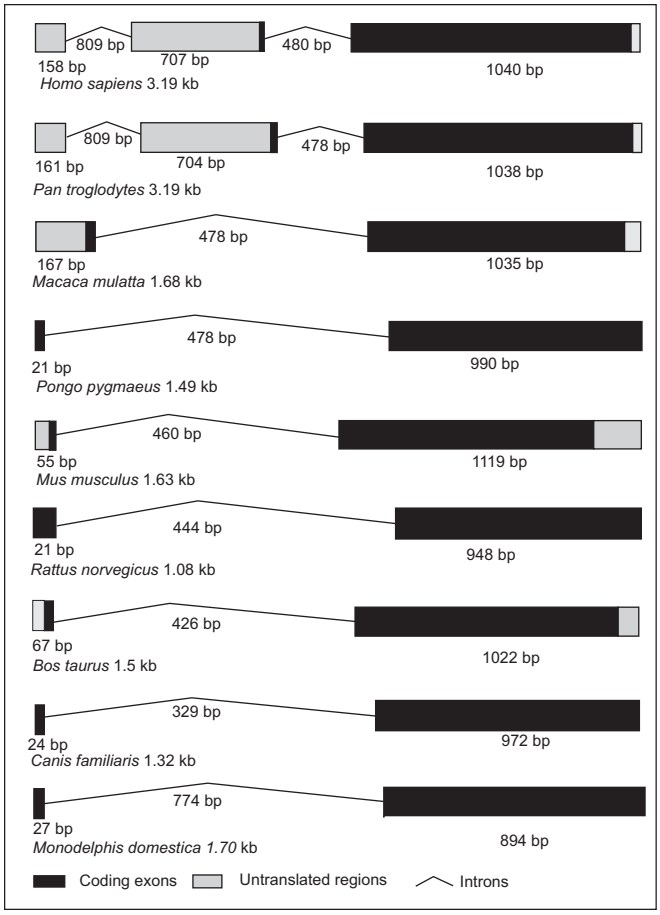

Fig. 1: Characterization of all nine taxa (Not in scale)

3' and 5' ends. In seven out of nine taxa, only two exons were present where as in other two taxa, i.e. human and chimpanzee, three exons were found. Although there are no much variations in the total gene size across seven taxa and the size of human and chimpanzee gene were much higher (almost twice) than the rest seven (Table 1). Even the intron sizes were quite variable among taxa and the taxa where only one intron is present, the size varies from 329 bp (*C. familiaris*) to 774 bp (*M. domestica*). In *H. sapiens* and *P. troglodytes*, two introns were seen due to the presence of a non-coding exon (97% redundant). Even the first exon is completely non-coding (158 bp). The size of the third exon of these two taxa (1040 bp in humans and 1038 bp in chimps) is however, comparable to the size of the second (and the only coding) exon of other seven taxa. It has been demonstrated that introns are biased towards the 52 ends of genes in intron-poor genomes but are evenly distributed in intron-rich genomes[16]. Overall, not much difference was noted in the total gene size of *H. sapiens* and *P. troglodytes*, whereas other seven taxa were clearly different in all aspects of data. In order to note the nucleotide substitutions between the two taxa, we aligned the two sequences and found only 24 nucleotide substitutions (Table 2). Out of these 24 substitutions, 18 were located in the non-coding regions and rest six were located in the coding regions. From these six substitutions, three were synonymous mutations (silent) while other three were non-synonymous (Table 3). It is widely known that the genetic distance between humans and the chimpanzee is probably too small to account for their substantial organismal differences[17], our study focuses on the differences which are mainly due to change in non-

Table 1. Characterization of nine taxa with details on exons length, introns length, total gene length, exon-intron ratios, exon and intron numbers, and length of coding and non-coding regions

Taxa/Length (bp)	E-1	I-1	E-2	I-2	E-3	Total genes	Ratio (E/I)	CR	NCR+UTRs
H. sapiens	158	809	707	480	1040	**3194**	1.47	**1011**	**2183**
P. troglodytes	161	809	704	478	1038	3190	1.47	**1011**	2179
M. mulatta	167	478	1035	–	–	1680	2.51	1008	672
P. pygmaeus	21	478	990	–	–	1489	2.11	**1011**	478
R. norvegicus	21	444	948	–	–	1413	2.18	969	444
M. musculus	55	460	1119	–	–	1634	2.55	1005	629
C. familiaris	24	329	972	–	–	**1325**	**3.02**	996	**329**
B. taurus	67	426	1022	–	–	1515	2.55	993	522
M. domestica	27	774	894	–	–	1695	**1.18**	**921**	774

E–Exon; I–Intron; CR–Coding region; NCR–Non-coding region; UTR–Untranslated region; Highest and lowest values are highlighted in bold.

Table 2. Twenty four SNPs in *H. sapiens* and *P. troglodytes*

Nucleotide substitution(bp)	275	285	372	392	417	876	878	919	1061	1084	1440	1519	1625	1728	1789	1801	1912	2136	2157	2256	2258	2476	2706	2940
H. sapiens	C	G	G	C	T	C	C	A	G	T	C	G	C	G	T	T	T	C	G	T	G	G	T	G
P. troglodytes	T	C	C	A	C	T	T	G	A	G	T	T	G	C	C	–	–	T	A	C	A	A	A	C

coding region rather than protein-coding mutations[17]. Recent studies have demonstrated that changes in both amino acid[18] and regulatory sequences[19] have both been involved in the evolution of uniquely human phenotypes but the present data on Duffy gene do not corroborate to the fact that substitutions in the protein coding regions of the gene, have contributed in human chimpanzee differences as the difference contribution ratio of non-coding to coding is 3:1 (75% and 25%) in Duffy. However, the role of coding regions in creating differences between human and chimps cannot be completely ruled out as three non-synonymous mutations are present in both the taxa (Table 3). Divergence for the taxa was also calculated and not much difference was noted for coding and non-coding regions. Length of coding regions contributes only 0.59% (Length=1011 bp; 6 nucleotide substitutions) whereas non-coding contribute 0.82% (Length=2183 bp; 18 nucleotide substitutions) towards divergence. Genome-level similarities between these taxa are highly variable, as in reported estimates vary in between 1.5 to 2.5%. However, according to a recent estimate with gene families, the difference was as high as 6%[20]. Present results from Duffy gene indicate a mere 0.70% differences between *H. sapiens* and *P. troglodytes*. Furthermore, increased intron gain and decreased intron loss in evolutionarily conserved genes indicate that intron insertion often might be adaptive[21]. This fact seems to hold true in case of Duffy gene, particularly in human-chimp lineage.

Table 3. Three non-synonymous mutations present in humans and chimps

Codon position	48	115	269
H. sapiens	Gly (G)	Val (V)	Lys (K)
P. troglodytes	Aspartic acid (D)	IL (I)	Asp (N)

Gly–Glycine; Val– Valine; Lys–Lysine; IL–Isoleucine; Asp–Asparagine.

It was evident from Table 1 that the coding regions contribute more to the total gene length in comparison to the length of the non-coding region, a feature opposite to studies done in TNF-α and CD36 genes, important for malaria pathogenecity[22,23]. The ratio of exon to intron varies across taxa, minimum in *M. domestica* and maximum in *C. familiaris* (Table 1). Further, with a view to know whether the size of the Duffy gene in mammalian taxa had increased due to accumulation of introns, Pearson's correlation coefficient values were calculated for the total gene lengths and the intron sizes as well as the exon sizes. Positive correlation was obtained in both the cases with a value of $r = 0.97$; $p < 0.0001$. Proportionately, low exon/intron ratio in *H. sapiens* and *P. troglodytes* among the studied mammalian taxa might be due to gain of a second intron. Intron gain and loss in organisms are highly discussed[24] and appears that recombination causes the widely observed but poorly understood phenomenon of internal intron loss and that DNA repeat expansion can create new introns[25]. In humans, this loss and gain is of much importance[24] especially for a gene responsible for diseased pathogenecity like Duffy.

From Table 1 it was also evident that exon 3 of *H. sapiens* and *P. troglodytes* and exon 2 of other seven taxa were almost similar in length, therefore, we were interested to see if these sequences are similar in conservation as well. Thus, we aligned these sequences and constructed a neighbour-joining phylogenetic tree (Fig. 2) and as predicted, *H. sapiens* and *P. troglodytes* were placed in one clade. It appears from Fig. 1 that possibly insertion of first intron (809 bp) and second exon (707 bp) has taken place in humans and

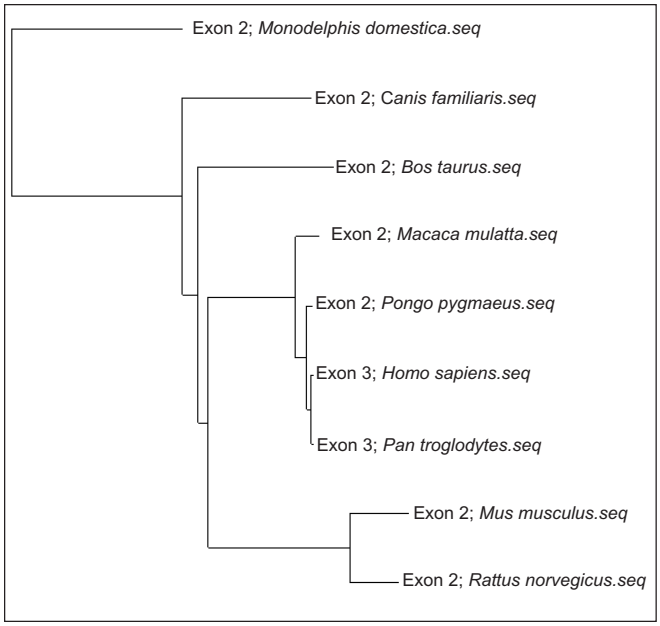

Fig. 2: Phylogenetic tree constructed for the third exon of *H. sapiens* and *P. troglodytes* and compared with the second exon of all the seven taxa. Bootstrapping was also done and absolute values were found

similarly insertion of first intron (809 bp) and second exon (704 bp) has taken place in chimps some 6–10 million years ago, i.e. almost before the divergence and separation of these two taxa. This 1516 bp long stretch in humans and 1513 bp in chimps must have been present as one single stretch before divergence. Even after the split of these taxa, this stretch is conserved in the lineage till date thus emphasing the fact that natural selection is responsible in its maintenance. Similarly, *M. musculus* and *R. norvegicus* were also present in one clade only and *M. domestica* seems to be an outlier sequence. Bootstrap values were also calculated (1000 simulations) and found to be absolute for all clades (data not shown). Similar NJ tree was also constructed taking the whole Duffy gene into consideration (Fig. 3). Almost similar tree topologies were obtained which clearly demonstrate that most of the variations are in fact located in the coding nucleotides (exon 2 of seven taxa and exon 3 of *H. sapiens* and *P. troglodytes*) for Duffy gene in the mammalian taxa studied here and are in agreement with earlier findings of Mikkelsen *et al*[18].

The above observations have remarkable implications on the research of the erythrocyte chemokine receptor (Duffy) gene which is of high importance to vivax malaria. Duffy blood group polymorphisms are important in areas where *P. vivax* predominates, because this gene acts as a receptor for this protozoan. Studies have shown that natural selection favours a particular allele of this gene (FYO) in sub-Saharan Africa[9,26] justifying the fact that this gene is under constant selection pressure in humans to protect from *P. vivax* malaria. However, in other taxa, similar but unknown and unexplained mechanisms might exist for conservation of the Duffy gene. The present study, thus, is a bold step to unravel the fact that functional constraint might be responsible in evolution of Duffy gene across mammalian taxa. If this is true, the role of natural selection in shaping variation in different taxa and uniqueness in human-chimpanzee lineage seems inevitable.

Conclusion

In conclusion, the present study provides a deeper understanding of the genetic architecture and evolu-

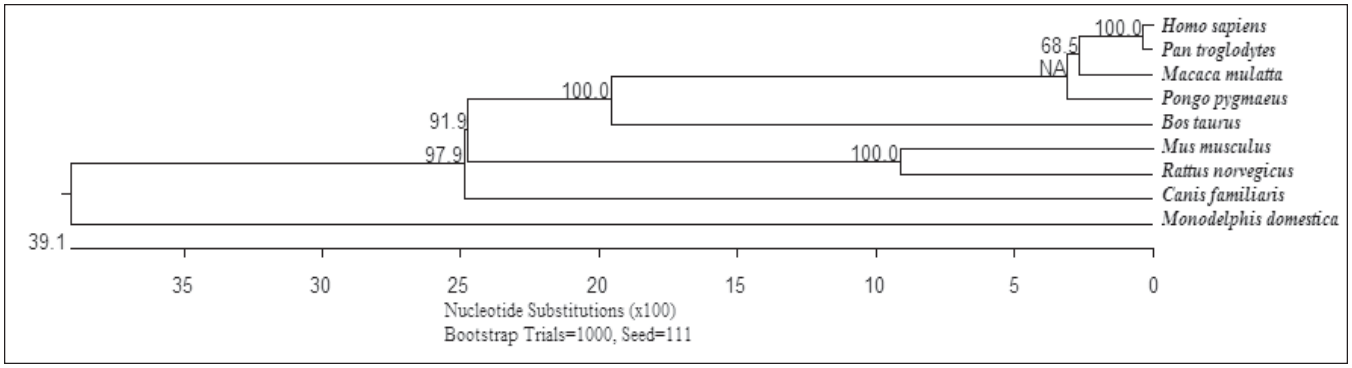

Fig. 3: Phylogenetic tree constructed for nine taxa. Bootstrap values obtained for 1000 simulation were shown in the figure

tionary aspects of one of the major human erythrocyte receptor gene that is of high relevance to human malaria pathogenecity using computational approaches. Similar studies in two genes (CD36 and TNF-α) responsible for malaria pathogenecity reveal many contrasting evolutionary patterns, although phylogenetic positions of taxa were found to be almost similar as found for Duffy[19,20]. Thus, evolutionary patterns of host genes responsible for malaria pathogenecity might be different and this might be due to different environmental and interactional events occurring between host and malaria parasites. Further work involving population genetic approaches would surely unravel several interesting features of these types of genes responsible in human malaria pathogenecity.

Acknowledgement

The authors thank the Indian Council of Medical Research for intramural funding and Ms. Kheerja Agnihotri for initial help in the study.

References

1. Awasthi G, Dash AP, Das A. Characterization and evolutionary analysis of human CD36 gene. *Indian J Med Res* 2008; *129:* 534–41.
2. Awasthi G, Singh S, Dash AP, Das A. Genetic characterization and evolutionary inference of TNF-α with computational analyses. *Braz J Infect Dis* 2008; *12*(5): 374–9.
3. Bustamante CD, Fledel AA, Williamson S, Nielsen R, Todd HM**,** Glanowski S, *et al*. Natural selection on protein-coding genes in the human genome. *Nature* 2005; *437:* 1153–7.
4. Carmel L, Rogozin IB, Wolf YI, Koonin EV. Evolutionarily conserved genes preferentially accumulate introns. *Genome Res* 2007; *17:* 1045–50.
5. Castilho L. The value of DNA analysis for antigens in the Duffy blood group system. *Transfusion* 2007; *47:* 28S–31S.
6. Chaudhuri A, Polyakova J, Zbrzezna, Pogo AO. The coding sequence of Duffy blood group gene in humans and simians: restriction fragment length polymorphism, antibody and malarial parasite specificities, and expression in nonerythroid tissues in Duffy-negative individuals. *Blood* 1995; *85:* 615–21.
7. Fedorov A, Roy S, Fedorova L, Gilbert W. Mystery of intron gain. *Genome Res* 2003; *13:* 2236–41.
8. Hahn MW, Rockman MV, Soranzo N, Goldstein DB, Wray GA. Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* 2004; *167:* 867–77.
9. Hamblin MT, Rienzo AD. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 2000; *66:* 1669–79.
10. Hamblin MT, Thompson EE, Rienzo AD. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* 2002; *70:* 369–83.
11. Haubold B, Wiehe T. Comparative genomics: methods and applications. *Naturwissenschaften* 2004; *91:* 405–21.
12. Hofmann JR, Weber BH. The fact of evolution: implications for science education. *Sci Edu* 2004; *12:* 729–60.
13. Iwamoto S, Omi T, Kajii E, Ikemoto S. Genomic organization of the glycoprotein D gene: Duffy blood group Fya/Fyb alloantigen system is associated with a polymorphism at the 44-amino acid residue. *Blood* 1995; *85:* 622–6.
14. King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. *Science* 1975; *188:* 107–16.
15. Lexer C, Fay MF. Adaptation to environmental stress: a rare or frequent driver of speciation? *J Evol Biol* 2005; *18:* 893–900.
16. Lu ZH, Wang ZX, Horuk R, Hesselgesser J, Yan-chun L, Hadley TJ, Peiper SC. The promiscuous chemokine binding profile of the Duffy antigen/receptor for chemokines is primarily localized to sequences in the amino-terminal domain. *J Biol Chem* 1995; *270:* 26239–45.
17. Marcelo A, Nobrega MA, Pennacchio LA. Comparative genomic analysis as a tool for biological discovery. *J Physiol* 2004; *554:* 31–9.
18. Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB**,** Yang SP, *et al*. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005; *437:* 69–87.
19. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE. Patterns of intron gain and loss in fungi. *PLoS Biol* 2004; *2:* e422.
20. Olsson ML, Hansson C, Avent ND, Akesson IE, Green CA, Daniels GL. A clinically applicable method for determining the three major alleles at the Duffy (FY) blood group locus using polymerase chain reaction with allele-specific primers. *Transfusion* 1998; *38:* 168–73.
21. Pogo AO, Chaudhuri A. The Duffy protein: a malarial and chemokine receptor. *Seminar Hematol* 2000; *37:* 122–9.

22. Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 2003; *19:* 2496–7.

23. Sharpton TJ, Neafsey DE, Galagan JE, Taylor JW. Mechanisms of intron gain and loss in Cryptococcu. *Genome Biol* 2008; *9:* R24.

24. Tournamille C, Colin Y, Cartron JP, Le Van KC. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* 1995a; *10:* 224–8.

25. Tournamille C, Blancher A, Van Kim CL, Pierre GP, Apoil PA, Nakamoto W, Cartron JP, Colin Y. Sequence, evolution and ligand binding properties of mammalian Duffy antigen/receptor for chemokines. *Immunegenetics* 2004; *55:* 682–94.

26. Tournamille C, Le Van KC, Gane P, Cartron JP, Colin C. Molecular basis and PCR-DNA typing of the Fya/fyb blood group polymorphism. *Hum Genet* 1995b; *95:* 407–10.

Corresponding author: Dr Aparup Das, Scientist 'D', Genomics and Bioinformatics Laboratory, National Institute of Malaria Research, Sector 8, Dwarka, New Delhi–110 077, India.
E-mail: aparup@mrcindia.org

Received: 6 April 2009 *Accepted in revised form:* 13 August 2009