

## Characterization & evolutionary analysis of human *CD36* gene

Gauri Awasthi, Aditya P. Dash & Aparup Das

*Evolutionary Genomics & Bioinformatics Laboratory, National Institute of Malaria Research (ICMR)  
New Delhi, India*

Received November 12, 2007

**Background and objectives:** Understanding evolutionary genetic details of immune system genes responsible for infectious diseases is of prime importance concerning disease pathogenicity. Considering malaria as a devastating disease in the world including India, detail evolutionary understanding on human immune system gene is essential. The primary aim of this study was to initiate work on one such gene, the human *CD36* gene responsible in malaria pathogenesis.

**Methods:** DNA sequences of the human *CD36* gene was retrieved from public domain and fine-scale details were characterized. Both comparative and evolutionary analyses were performed with sequences from six other taxa (5 mammalian one avian) where *CD36* homologs are present. Different statistical analyses were also performed.

**Results:** Differential distribution in number and length of exons and introns was detected in *CD36* gene across seven taxa. The CpG islands were also found to be distributed unevenly across the gene and taxa. Neighbour-joining tree was constructed and it was observed that the chimpanzee and human are diverged at the *CD36* gene relatively recently. The chicken, *Gallus gallus* was found to be diverged from rest of the taxa significantly. Also copy number variation was observed across different taxa.

**Interpretation & conclusions:** Comparative genomic study of a human immune system gene *CD36* show relationships among different taxa at the evolutionary level. The information can be of help to study genetic diversity in malaria endemic zones and to correlate it with malaria pathogenicity.

**Key words** *CD36* - comparative genomics - evolution - malaria - phylogenetics

Comparative genomic analyses play an important role in understanding differential organizations of genome across different taxa and in understanding how genomes have evolved over time. With the availability of whole-genome sequence information of an array of taxa from different branches of the tree of life, it has now become easy to compare genes and genomes across different taxa. This approach also helps in finding hitherto unknown genes, regulatory networks, and

regions of the genome that are important in biological functions and thus determining the precise role of evolution in genomes. This whole process should start with computational approach, first in characterization of genes in a particular species of importance and looking at homologous DNA sequences that are conserved across different taxa, followed by comparison of length of genes and introns and exons (in nucleotides), number of introns and exons, distribution of CpG islands, *etc.* Such types

of studies provide comprehensive information that helps in understanding how genes are affected by different evolutionary forces. Further, determination of gene copy number variations is also to some extent responsible for bringing evolutionary changes which accounts for a substantial amount of genetic variation<sup>1</sup>. Current efforts are directed toward a more comprehensive cataloguing and characterization of genes that might provide the basis for determining how genomic diversity impacts biological function, evolution, and common human diseases<sup>2</sup>. This is of special importance, since genetic control and prevention strategies could be undertaken once complete knowledge on character and evolutionary pattern of disease-related genes in genome are at hand.

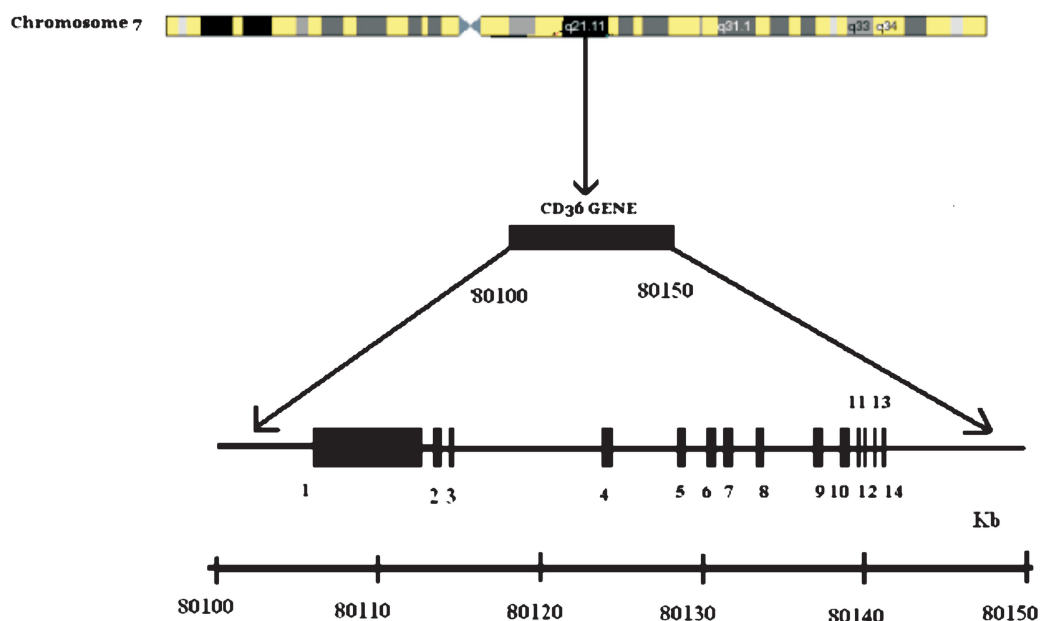
To this respect, the human *CD36* gene, located on chromosome 7 and encoded by 15 coding regions (including one un-translated region)<sup>3</sup> (Fig. 1) is of utmost importance, as it serves as a major receptor for the human malaria parasite *Plasmodium falciparum*<sup>4</sup>. It has been reported that inhibition of the immune response to platelet-mediated clumping of parasite-infected erythrocytes is strongly associated with severe malaria and that *CD36* gene expression is required for such clumping<sup>5</sup>. Moreover it has been reported that a non-sense mutation (T188G) in *CD36* gene is also associated with protection to severe malaria<sup>6</sup>. *CD36* is not only involved in the sequestration of the parasite, but could also play a role in the innate and acquired immune response to malaria infection<sup>7</sup>. This makes

the *CD36* gene very important as human malaria is concerned, and demands detailed fine-scale genetic evolutionary understanding.

We therefore conducted a study for detailed characterization of the *CD36* gene and to establish evolutionary relationships among different taxa following computational genomic approaches. Specifically, we determined various aspects of introns particularly the length, number and relationship to CpG islands across different taxa, *etc.* Evolutionary relationships based on *CD36* gene were inferred among all the studied taxa through phylogenetic analysis. We also determined number of copies of *CD36* gene in each individual taxon and compared its variation across these different taxa.

### Material & Methods

Nucleotide sequences of *CD36* gene of seven different taxa *Mus musculus* (GenBank Acc. No.NM\_007643.3), *Rattus norvegicus* (GenBank Acc. No.NM\_031561.2), *Pan troglodytes* (GenBank Acc. No.XM\_519573.2), *Macaca mulatta* (GenBank Acc. No.NM\_001032913.1), *Homo sapiens* (GenBank Acc. No.NM\_001032913.1), *Canis familiaris* (GenBank Acc. No. XM\_533140.2) and *Gallus gallus* (GenBank Acc. No.NM\_001030731.1) were downloaded from National Center for Biotechnology Information (NCBI) database (<http://www.ncbi.nlm.nih.gov/>) during August/September 2007. Out of seven taxa considered, six were



**Fig. 1.** Location of human *CD36* gene at the locus 7q21.11 on the chromosome 7 showing 14 exons (excluding the untranslated region) (Fig. not in scale).

mammalian and one avian (*G. gallus*). Characterization (total gene length, intron and exon number and length, ratio of coding nucleotides to total gene size, details of exons) of *CD36* gene was done following information at the NCBI website. For sequence analysis, different statistical and web resources were used. Multiple sequence alignments and construction of neighbour-joining phylogenetic tree were performed using MegAlign, a part of Lasergene (DNASTAR software, Madison, USA, <http://www.dnastar.com>). Individual branch lengths were calculated using phylogeny option in the statistical program VEGAZZ (downloaded from <http://www.vegazz.net>). Consensus sequences of *CD36* gene were searched in individual taxa. For this, the 'BLAST' option at the NCBI website was used. The hits showing more than 80 per cent homologous DNA sequences were considered as copies of *CD36* gene in those particular taxa. Pearson's correlation coefficients were calculated using 'Analyze-it' (<http://www.analyze-it.com/>) software, an add-on to the MS Excel software. For all statistical analysis,  $P < 0.05$  was considered as level of significance. Determination of CpG islands in the introns of *CD36* genes was conducted using the website <http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/cpgplot.html>. This program defines, by default, a CpG island as a region where the GC level is over 50 per cent, the calculated observed/expected (O/E) CpG ratio is over 0.6 and these two conditions hold for a minimum of 200 continuous nucleotide bases. Short and long introns are defined according to description provided by Gazave and co-workers<sup>8</sup>, where it was suggested that introns of more than 1029 bp nucleotides be considered as long and less than 1029 bp as short introns. Similar guidelines were followed in the present analysis.

## Results

Characterization in all seven taxa revealed a variable size of *CD36* gene [minimum of 29058 nucleotide base

pair (bp) in *M. mulatta* and maximum of 74804 bp in *H. sapiens*]. A maximum of 15 exons were found (including untranslated regions of exons) in three taxa (*H. sapiens*, *P. troglodytes*, and *C. familiaris*) and a minimum of 12 exons (including untranslated regions of exons) in *M. mulatta* (Table). This gene is located in different chromosomes in different taxa and the exon-intron ratio varies widely across taxa. It was evident that *C. familiaris* had the highest ratio and *H. sapiens* had the lowest, signifying that in *H. sapiens*, the non-coding nucleotides are at abundance. The details of exon-intron structure of *CD36* gene (including the UTRs in blue colour) in each taxa is shown in Fig. 2. The size of *CD36* gene in each taxa and relative size of exon and intron are shown in Fig. 3. The figure revealed the differential size distribution of both exons and introns across all the taxa. It was evident that the non-coding nucleotides constituted a major portion of the total nucleotides of the *CD36* gene and that the size of the gene is somehow dependent on the size of the introns in all taxa. This claim was further substantiated by the observation of statistically significant positive correlation ( $r = 1$ ,  $P < 0.001$ ). Further, the percentage of coding nucleotides in the *CD36* gene varied across taxa; the highest was detected in *M. musculus* (9.3% of the total nucleotides) and lowest in *H. sapiens* (2.9%; Fig. 4).

In order to test the hypothesis that the first intron of a gene is usually bigger in size in comparison to the rest, the size of the first introns of *CD36* gene was determined in each taxon. All the introns of each taxa were plotted separately based on sizes (Fig. 5). The length of first intron was largest for almost all the taxa except *C. familiaris* and *M. musculus*. The second largest intron was found to be the third one, in general.

It was suggested that the first introns contain about ten-fold number of CpG islands<sup>8</sup>. To find if this contention is valid in the *CD36* gene, CpG islands were

**Table.** Properties of *CD36* gene in seven taxa studied

Taxa	Properties of <i>CD36</i> gene					
	Chromosomal location	Total gene length (in bp)	Total exon length (in bp)	No. of exons with UTR	No. of exons without UTR	Intron/Exon ratio
<i>Mus musculus</i>	5	53681	3016	15	12	0.059
<i>Rattus norvegicus</i>	4	53743	2551	14	12	0.049
<i>Pan troglodytes</i>	7	39392	2201	15	12	0.059
<i>Macaca mulatta</i>	3	53681	3016	11	10	0.043
<i>Homo sapiens</i>	7	74804	2048	15	14	0.030
<i>Canis familiaris</i>	18	32945	2267	15	12	0.073
<i>Gallus gallus</i>	1	43316	2331	14	12	0.056

Bp, nucleotide base pairs; UTR, un-translated regions

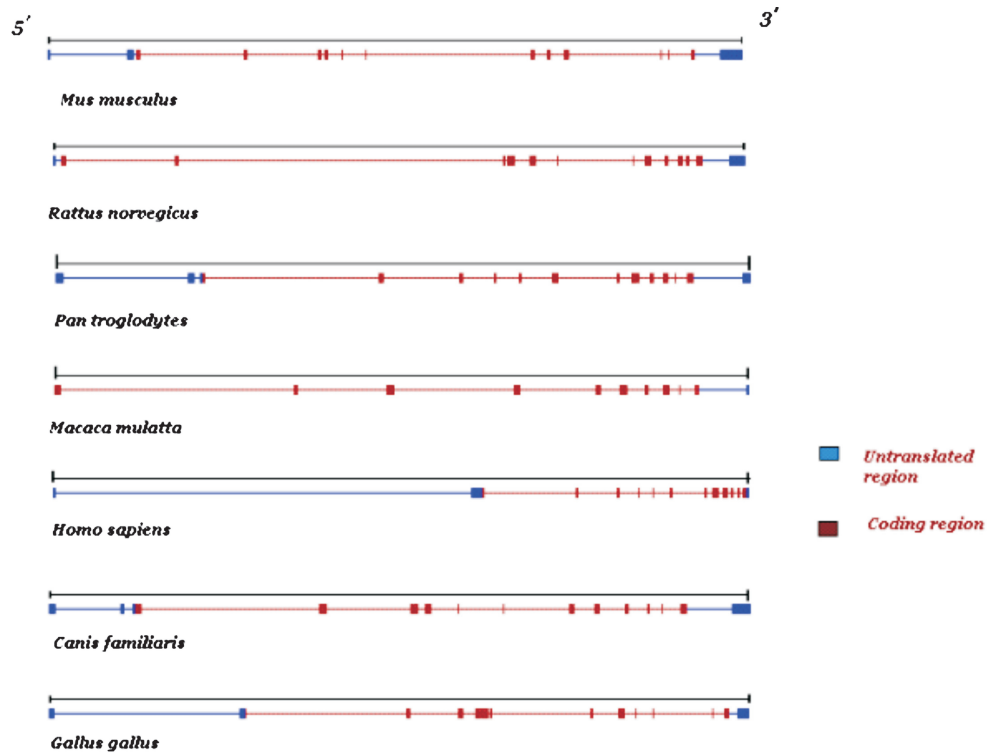


Fig. 2. Characterization of *CD36* gene in different taxa showing the position of different exons (Fig. not in scale).

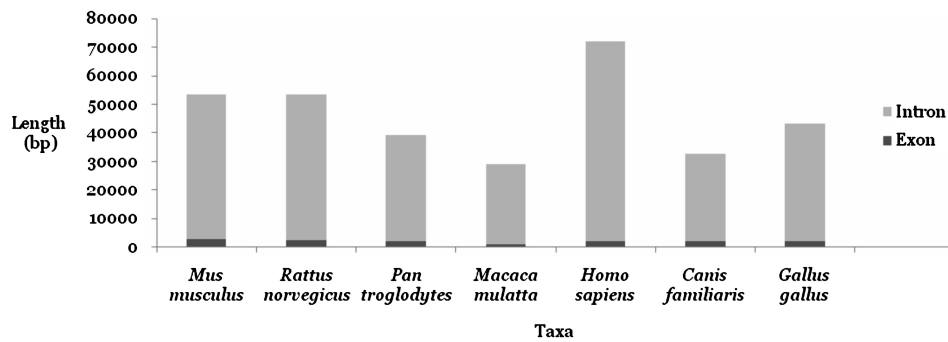


Fig. 3. Relative composition of exon and intron in *CD36* gene among different taxa.

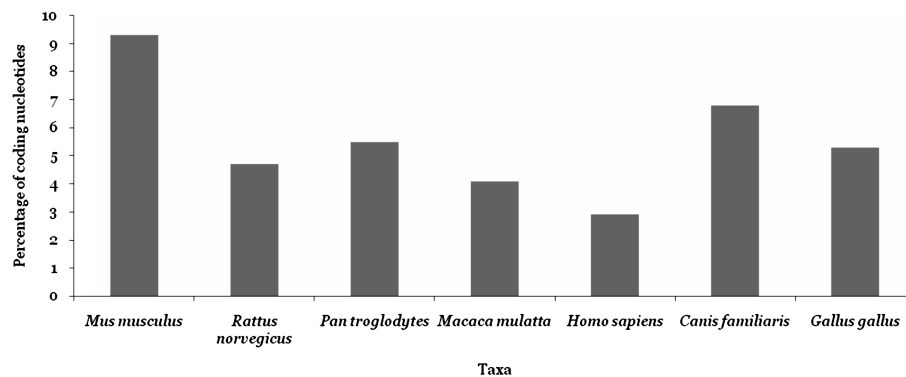


Fig. 4. Mean percentage of coding nucleotides in *CD36* gene across different taxa.

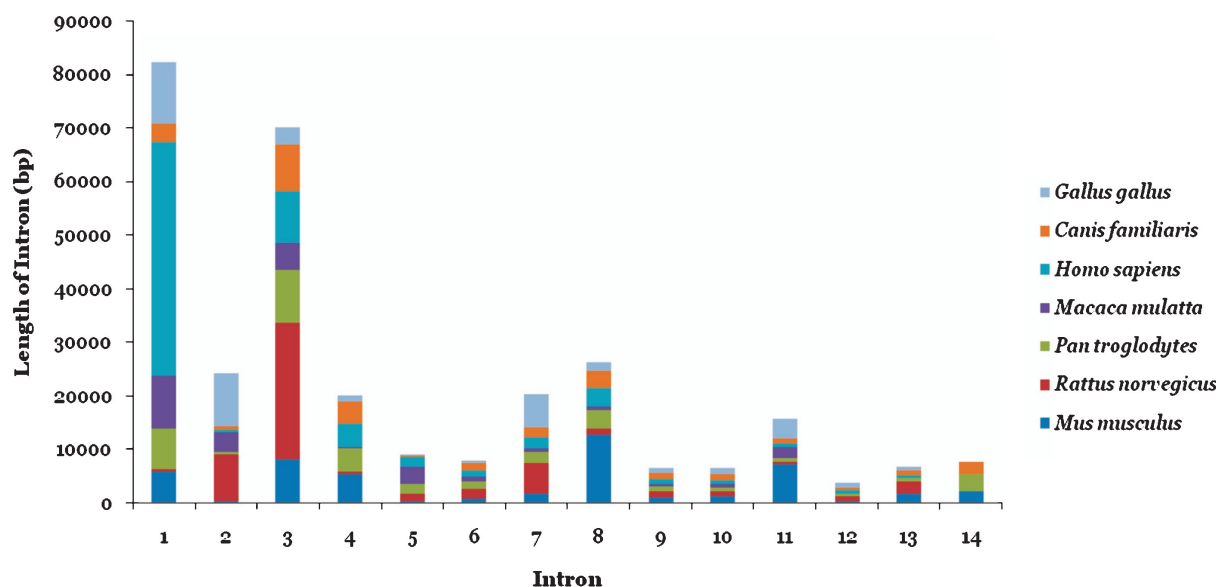


Fig. 5. Distribution of intron length among different taxa.

determined in the first, long and short introns<sup>8</sup> in all the taxa. The figure (Fig. 6) depicts the CpG islands in the first introns separately for all the taxa. The occurrence of CpG islands varied across different taxa for both the long and short introns (Figure not shown). Contrary to the expectation<sup>8</sup>, GC poor regions were found in short introns in most of the taxa (Fig. 6).

With a view to understand phylogenetic relationships among all the seven taxa at *CD36* gene, an un-rooted neighbour joining (NJ) tree was constructed (Fig. 7). The length of each branch is also shown in the NJ tree which is highly variable across taxa. It was clearly seen in the NJ tree that *P. troglodytes* and *H. sapiens* were closely related to each other and belonged to one clade. The avian sequence from *G. gallus* fell in a completely separate branch. Also, *M. musculus* and *M. mulatta* sequences seemed to have diverged from a single common ancestor in recent past.

Since it is known that many immune system genes are present in multiple copies in the genome and *CD36* gene is one of the very important human immune system genes, we were interested in finding whether multiple copies of this gene are present in each taxon. Thus the whole genome of each taxa was scanned separately for extra copy of *CD36* gene. In all the taxa, more than one copy of *CD36* gene was found, except *M. musculus*. However, there seemed to be a great variation in copy number across taxa (Fig. 8). A minimum of only one extra copy of *CD36*

gene was found in *M. musculus*, whereas a maximum of 55 copies were detected in the avian taxa, *G. gallus*. Comparing only the six mammalian taxa, *C. familiaris* seemed to possess the highest number (37 copies) of *CD36* gene. On the basis of copy numbers, *H. sapiens* (11 copies) appeared to bear somehow a close resemblance to *M. mulatta* (10 copies) and *P. troglodytes* (7 copies). Similarly, *M. musculus* (single copy) and *R. norvegicus* (two copies) appeared to be close to each other in this characteristic. Further, in order to understand if the copy numbers in some respect relates to the size of gene, intron and exon, correlation coefficients ( $r$ ) were calculated. In all cases the  $r$  values were found to be negative but not statistically significant (data not shown).

## Discussion

Our study on characterization and comparative analysis of *CD36* gene in different taxa revealed several interesting features, both on basic understanding and evolutionary perspectives of this gene. This gene is known to be responsible, although partly, for human malaria pathogenicity<sup>3-7</sup>. The size of the gene, introns, exons, number of exons and introns in each taxon, size of individual introns and exons and the ratio of coding to the non coding regions varied considerably across taxa. Further, location and exon-intron ratio of this gene also varied across taxa. In humans, the non coding nucleotides are in abundance<sup>9,10</sup>, signifying the fact that human *CD36* gene contains more non coding nucleotides in

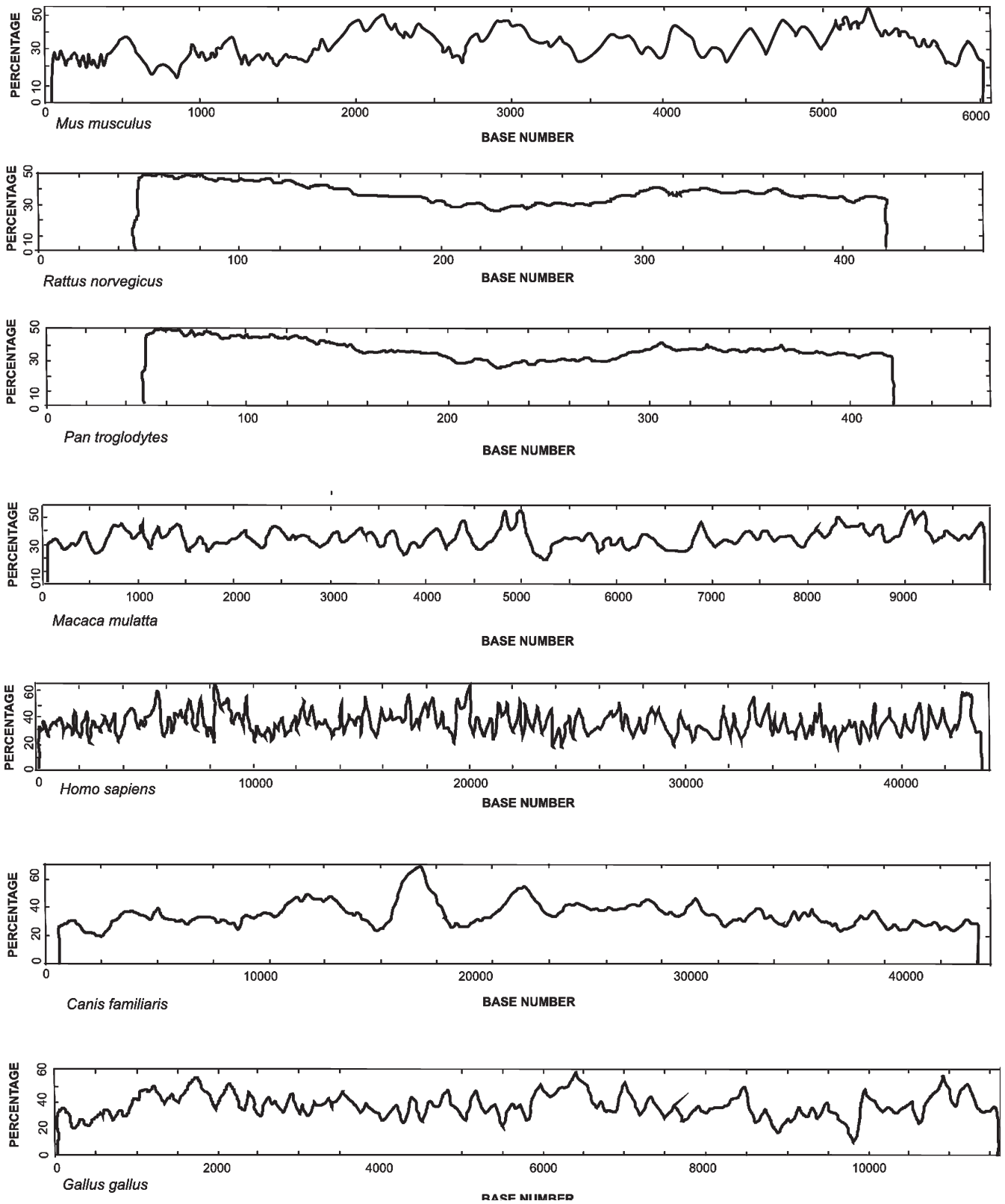


Fig. 6. CpG plots of first introns in different taxa. The peaks above 50 are considered to be the regions containing CpG islands.

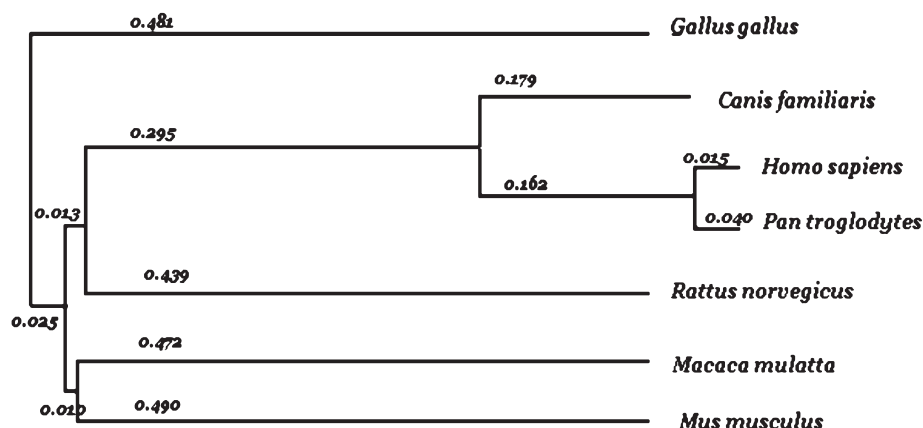


Fig. 7. Phylogenetic status of different taxa at *CD36* gene. Values depict length of each branch leading to a single taxon.

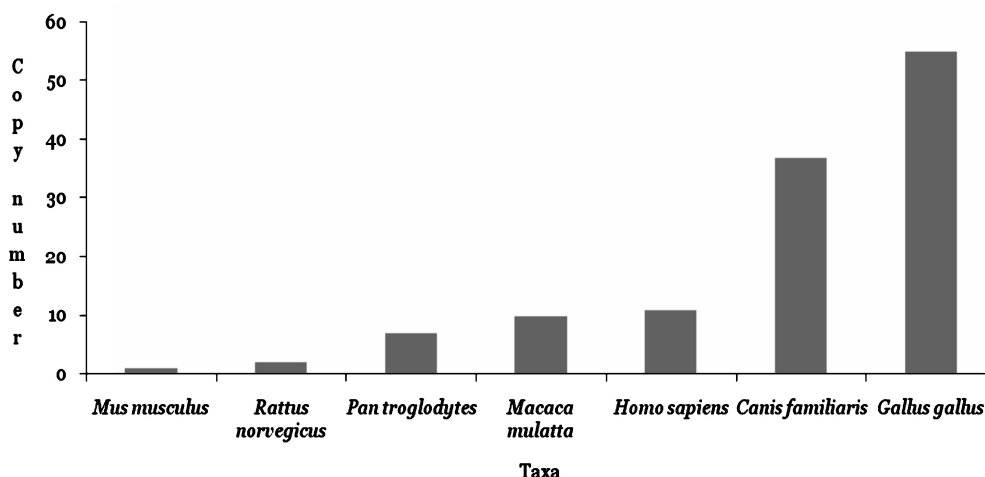


Fig. 8. Number of reported copy numbers of *CD36* gene in different taxa.

comparison to other studied taxa. Although the fact that whether accumulations of non coding DNA has helped in increasing the overall genome size<sup>11</sup> is still debated, our findings, though restricted to a single gene, seemed to corroborate this hypothesis. This is further reflected from the overall percentage of coding nucleotides in each taxa, where *H. sapiens* contained the least percentage of coding regions in *CD36* gene. Further, detection of a statistically significant positive correlation between intron length and gene length signified this fact.

It has been postulated that the first intron of a gene is usually different from rest of the introns both in size and GC content. Usually the first introns are larger and contain substantial amount of GC nucleotides<sup>8</sup>. In the *CD36* gene, a very similar pattern of this intron was observed. Also, in general, a larger size of introns was observed at the third position in almost all the taxa which is a unique observation.

However, in general longer introns were found in almost all the taxa. Further, CpG islands seemed to be abundantly distributed in the first introns of almost all the taxa analysed here. Since we observed a comparatively lower CpG content in short introns, this might explain that short introns harbour a relatively lesser proportion of regulatory elements than long introns<sup>8</sup>. Hence, finding high number of long introns and more CpG islands in the long introns of *CD36* gene suggests presence of more repetitive units<sup>8</sup>. Since it is believed that genes containing short introns are generally highly expressed<sup>12</sup> and we detected long introns in *CD36* genes, reflects the fact that *CD36* might not be a very highly expressed gene<sup>13</sup>. However, to further bolster this fact about *CD36* gene, experimental evidences will be needed.

Phylogenetic analysis at a particular gene throws light on the evolutionary relationships among different

taxa and also evolutionary paradigm of the gene. The phylogenetic analysis could clearly show that *H. sapiens* and *P. troglodytes* are very closely related to each other at *CD36* gene and a recent divergence at this gene from a common ancestor might be certain. These two taxa were previously compared between themselves and amongst other taxa for other genes and it was shown that human and chimpanzee functional DNA were more similar to each other than either is to other apes<sup>14</sup>. Although very identical findings were observed in terms of exon and intron numbers and gene length between *M. musculus* and *R. norvegicus*, at the phylogenetic level they were found to be placed in different branches. Thus, the *CD36* gene seems to have evolved very differently than the rest of the genes across taxa. It became clear from the present study that the *CD36* gene has evolved much to its extent across different taxa that have been studied here.

A wide range of copies of *CD36* gene was found to be present in different taxa ranging from single to 55 copies. Although no precise information is available to correlate these two characters, the copy number variations found in the seven taxa reflected the fact that these might somehow be related to malaria pathogenesis. For example, the avian taxa (*G. gallus*) had the highest copy numbers of *CD36* gene, inspite of reported malarial incidences. However in dog, *C. familiaris*, malaria incidence has not been reported at all but it also has a high copy number of *CD36* gene. In other five primates (including humans), where malaria incidences are frequently reported, *CD36* gene had very few copies in their respective genomes. However, detail functional analysis (particularly related to dosage dependence) would be necessary to evaluate this contention. Since copy number variations could be related to differential evolutionary paradigm and genetic diversity of the concerned gene<sup>2</sup>, detail genetic diversity studies across taxa and populations of each individual taxon would be necessary to ascertain a definite correlation between the *CD36* gene copy number and malaria pathogenesis.

In conclusion, this study provided new insights and comprehensive understanding of *CD36* gene in different taxa. Considering that *CD36* gene is related to malaria pathogenicity and disease severity, further molecular, evolutionary and genetic diversity studies would be needed to understand this gene and its role in malaria.

## Acknowledgment

The authors thank Dr Surbhi Pal Chaudhary for help in the initial stage of work and Dr U. Sreehari for help in Fig. 6.

## References

1. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, *et al.* Global variation in copy number in the human genome. *Nature* 2006; 444 : 444-54.
2. Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, *et al.* Copy number variation: new insights in genome diversity. *Genome Res* 2006; 16: 949-61.
3. Fernandez-aiz RE, Armesilla AL, Sanchez-Madrid F, Vega MA. Gene encoding the collagen type I and thrombospondin receptor *CD36* is located on chromosome 7q11.2. *Genomics* 1993; 17 : 759-61.
4. Aitman TJ, Cooper LD, Norsworthy PJ, Wahid FN, Gray JK, Curtis BR, *et al.* Malaria susceptibility and *CD36* mutation. *Nature* 2006; 405 : 1015-6.
5. Omi K, Ohashi J, Patarapotikul J, Hananantachai H, Naka I, Looareesuwan S, *et al.* *CD36* polymorphism is associated with protection from cerebral malaria. *Am J Hum Genet* 2003; 72 : 364-74.
6. Pain A, Urban BC, Kai O, Casals-Pascual C, Shafi J, Marsh K, *et al.* A non-sense mutation in *CD36* gene is associated with protection from severe malaria. *Lancet* 2001; 357 : 1502-3.
7. Serghides L, Smith TG, Patel SN, Kain KC. *CD36* and malaria: friends or foes? *Trends Parasitol* 2003; 19 : 461-9.
8. Gazave E, Marque's-Bonet T, Fernando O, Charlesworth B, Navarro A. Patterns and rates of intron divergence between humans and chimpanzees. *Genome Biol* 2007; 8 : R21.
9. Bird CP, Stranger BE, Liu M, Thomas DJ, Ingle CE, Beazley C, *et al.* Fast-evolving non-coding sequences in the human genome. *Genome Biol* 2007; 8 : R118.
10. Ponting CP, Lunter G. Signatures of adaptive evolution within human non-coding sequence. *Hum Mol Genet* 2006; 15 : R170-5.
11. Sironi M, Menozzi G, Comi GP, Bresolin N, Cogliani R, Pozzoli U. Fixation of conserved sequences shapes human intron size and influences transposon-insertion dynamics. *Trends Genet* 2005; 21 : 484-8.
12. Castillo-auiis D CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. Selection for short introns in highly expressed genes. *Nat Genet* 2002; 31 : 415-8.
13. Versteeg R, van Schaik BD, van Batenburg MF, Roos M, Monajemi R, Caron H, *et al.* The human transcriptome map reveals extremes in gene density, intron length, GC content and repeat pattern for domains of highly and weakly expressed genes. *Genome Res* 2003; 13: 1998-2004.
14. Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, *et al.* Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. *Science* 2005; 309 : 1850-4.