Interventional Medicine & Applied Science, Vol. 7 (4), pp. 152–160 (2015) ORIGINAL PAPER

Comparative and evolutionary insights into *CD4* gene across mammalian and avian taxa

NAAZNEEN KHAN*

Department of Biotechnology, Jamia Millia Islamia University, Jamia Nagar, Delhi, India
*Corresponding address: Dr. Naazneen Khan; Jamia Millia Islamia, Jamia Nagar, Delhi-110025, India;
E-mail: naaz.27khan@outlook.com

(Received: July 6, 2015; Revised manuscript received: August 15, 2015; Accepted: August 20, 2015)

Abstract: The present day genetic architecture of a species bears much significance to its closely related species which is due to species-specific differences, shaped by different evolutionary forces across time scale. With the availability of whole genome sequence of several closely related species, it is now possible to infer evolutionary patterns of genes and genomes in specific lineages. To this respect, *CD4* gene, primarily responsible for defensive mechanism in human, is conserved across a few taxa, and thus, comparative genomic studies could be useful for better understanding of host–pathogen biology. Comparative and evolutionary analyses were performed in eleven taxa (10 mammalian and avian) with different statistical algorithms. Phylogenetic inferences revealed recent divergence of human and chimpanzee, and pig was found to be diverged from rest of the taxa significantly. Additionally, gene length, microsatellites, and secondary structures were observed across taxa. The genetic architecture of *CD4* gene and its evolutionary history in different mammalian taxa provide crucial evidence in support of the fact that this gene might have been evolving at a similar rate to other human immune system genes. Future population-based study and structural modeling would unravel the differential ability to interact with HIV virus and influence immune system in humans.

Keywords: comparative, phylogenetic, evolutionary, mammalian, lineages

Introduction

Evolutionary study of genes and genomes is helpful in deciphering the role played by various evolutionary forces including the footprints of natural selection. Such forces change the architecture of genes that are helpful for their adaption in a particular environment. Determination of these changes in functional elements among species that are conserved across evolutionary time is possible with comparative genetic analysis. Comparative genomics offers new insights into genome evolution and the role played by natural selection in molding the sequence evolution [1]. Thus, characterization and comparison of genes among different taxa are helpful in development of new strategies in the battle against human disease. Comparative genomics is believed to be the important aspect of understanding the evolutionary relationship across different taxa. With the advent of high-throughput DNA sequencing facility, it is now easy to compare human genome sequence with several other closely related species to detect any similarities and differences in specific lineages that further helps in understanding disease pathogenesis.

AIDS is considered as one of the three deadly dangerous diseases including T.B. and malaria [2] and remains a major public health issue across the globe. It is a serious disorder of the immune system in which the body's normal defenses against infection breakdown, leaving it vulnerable to a host of life threatening infections, including unusual malignancies. The causative agent of AIDS is the human immunodeficiency virus (HIV), the agent which was caused a pandemic disease that has enormous social, economic, and behavioral impact on individuals, families, communities, and the whole world. AIDS has shattered the global economy with no successful treatment in-sight. The life cycle of HIV virus starts with the high affinity interaction between gp120 envelope protein of HIV virus and "cluster of differentiation" CD4 receptors present on human T-cells [3, 4]. Due to this interaction, another glycoprotein gp41 envelope protein changes its conformation and the virus fuses inside the human cell [5]. As a result of this fusion, massive destruction of CD4 and CD8 cells leads to malfunctioning of immune system.

To this respect, human CD4 molecule is expressed on distinct populations of thymocytes, mature T cells,

macrophages, monocytes, and Langerhans cells [6]. It is a transmembrane glycoprotein of 58 kDa that consists of an extracellular region of 370 amino acids, a transmembrane region of 25 amino acids, and a cytoplasmic tail of 38 amino acids at the C-terminal end. The extracellular portion of CD4 is folded into four distinct domains designated as D1, D2, D3, and D4 [7]. While the N-terminal of the D1 domain shares extensive structural and sequence homology with the variable region of immunoglobulin (Ig) light chains, the other three domains were less closely related to Ig molecules at the level of primary structure but fold similarly to Ig family domains, confirming that CD4 is a member of the Ig like superfamily.

We herewith, attempted to characterize the *CD4* gene in ten mammalian taxa *(Homo sapiens, Pan troglodytes, Macaca mulatta, Canis lupus familiaris, Bos taurus, Sus scrofa, Monodelphis domestica, Oryctolagus cuniculus, Mus musculus, Rattus norvegicus)* and one avian taxon *(Gallus gallus)* in a phylogenetic framework. The data were analyzed keeping the *CD4* gene into account, and evolutionary relationships among the taxa were inferred.

Materials and Methods

In order to perform the detailed characterization and comparative genetic studies, the sequence information on the 3.1 kilo base pair (kbp) nucleotide sequence of human *CD4* gene was retrieved. The orthologous sequences of this gene were searched in the National Centre for Biotechnology Information (NCBI B37.1 assembly) database [8] (accessed 15 Oct 2012) and were retrieved in eleven different taxa: *P. troglodytes* (chimpanzee), *M. mulatta* (rhesus monkey), *M. musculus* (house mouse), *R. norvegicus* (brown rat), *Oryctolagus cuniculus* (hare), *S. scrofa* (pig), *M. domestica* (oppossum), *Bos taurus* (buffalo), *C. lupus familiaris* (dog), and *G. gallus* (fowl). Detailed characterization (accession numbers, start/end of the gene, total gene lengths, chromosomal locations, total number of exons etc.) of the *CD4* gene was done as per the information provided in the NCBI website and is shown in *Table I*. To know whether the size of *CD4* gene had increased due to accumulation of introns, Pearson's correlation coefficient (r) was calculated using Analyse-It, an add-on to the Microsoft Excel software [9]. For all statistical analysis, the P value of <0.05 is considered as level of significance. In order to infer evolutionary relationships among taxa at the *CD4* gene, nucleotide sequences were aligned based on ClustalW algorithm with the help of MegAlign computer program of the DNASTAR software package [10]. The distance matrix generated from the alignment was used to construct a neighbor-joining (NJ) phylogenetic tree using online Phylogeny software [11] to reconstruct and analyze phylogenetic relationship among taxa [12]. Simple sequence repeats (SSRs) also known as microsatellites in the *CD4* gene were screened using the Imperfect Microsatellite Extractor (IMEx) software tool [13]. This tool was used to obtain microsatellites from IMEx using basic mode [14]. Further, the secondary structure of Cd4 protein was estimated with the help of SOPMA software [15, 16] across different taxa.

Results

Characterization of the *CD4* gene in all the 10 mammalian taxa and one avian taxon revealed several interest-

Table I Genetic characterization of the *CD4* gene across eleven different taxa

Species	Chromosome	Gene length	No. of exon	Exon (bp)	Coding region (bp)	Intron (bp)	UTR (bp)
Homo sapiens	12	31,326	10	3103	1377	28,223	1726
Pan troglodytes	12	31,712	10	1553	1377	30,159	176
Macaca mulatta	11	42,004	11	3064	1377	38,940	1687
Canis lupus familiaris	27	46,601	10	1656	1364	44,945	292
Bos taurus	5	16,356	7	2486	1154	13,870	1332
Sus scrofa	5	25,998	11	2709	1374	23,289	1335
Monodelphis domestica	8	8997	7	1152	1152	7845	0
Oryctolagus cuniculus	8	23,491	10	1497	1380	21,994	117
Mus musculus	6	23,517	10	3095	1374	20,422	1721
Rattus norvegicus	4	25,527	10	1749	1374	23,778	375
Gallus gallus	1	12,606	10	1968	1464	10,638	504

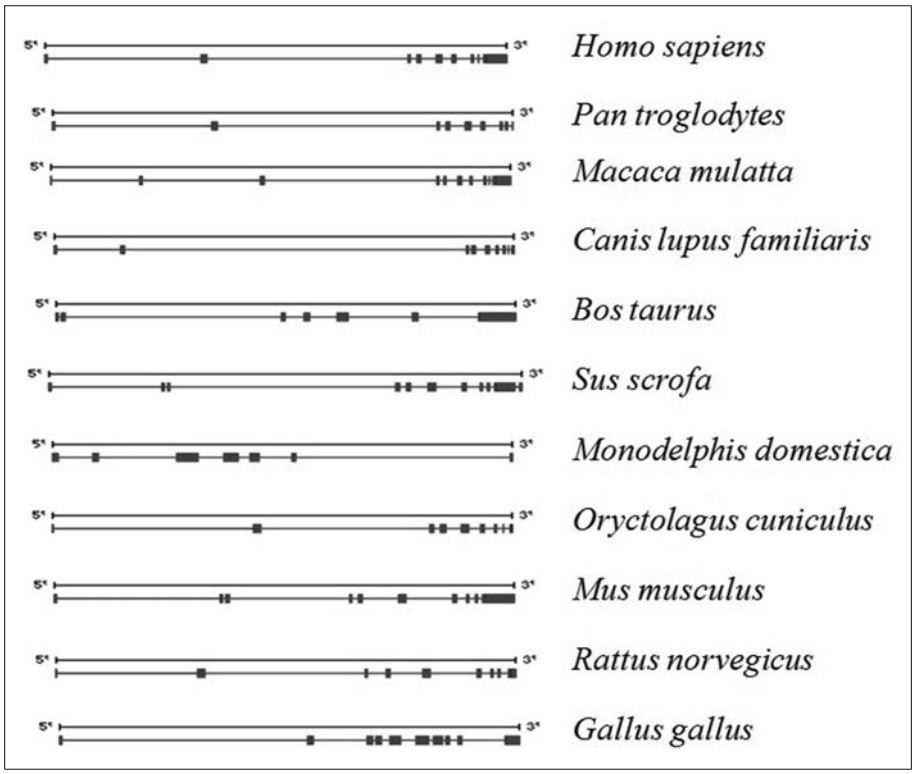

Fig. 1. Schematic representation of *CD4* gene among ten mammalian and avian taxa (■ Coding region, ■ Untranslated region)

ing features. It was observed that the *CD4* gene length varies across taxa with minimum gene length found in *M. domestica* (7845 bp) and maximum in *C. lupus familiaris* (46,601 bp)] *(Table I)*. Similarly, introns were present in all taxa and found to be variable across different taxa. It is intersting to note that introns have constituted a major portion of the *CD4* gene which was found to be maximum in *C. lupus familiaris* (17%) and lowest in

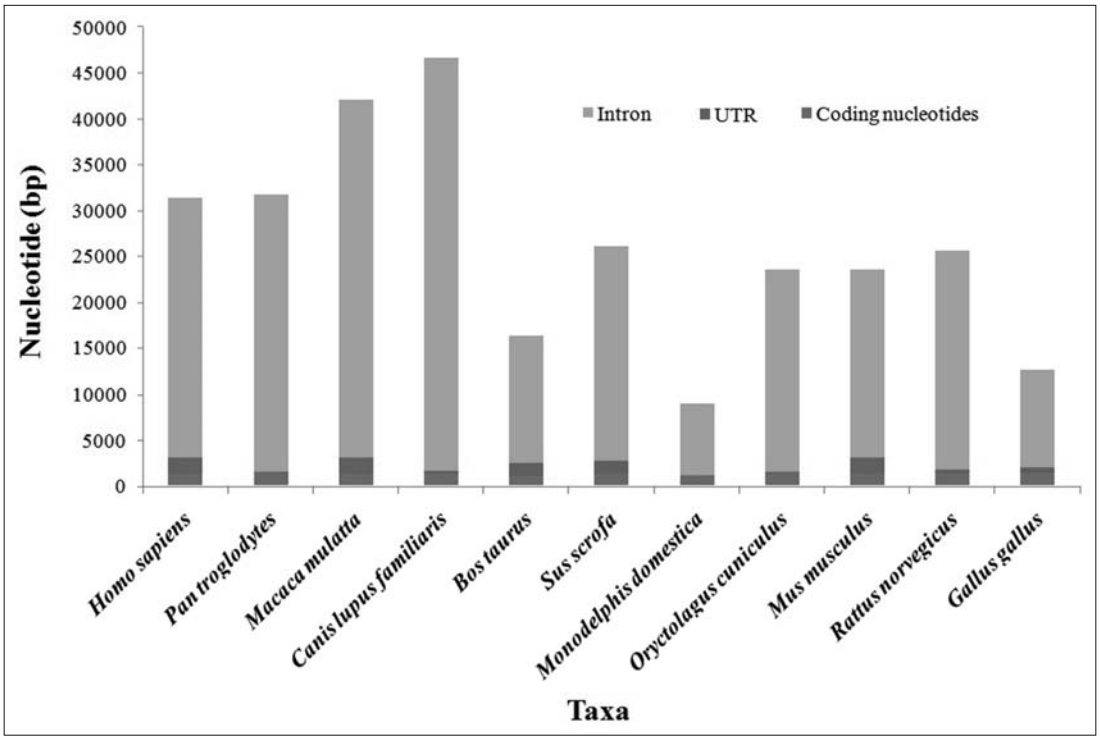

Fig. 2. Comparison on the intron, coding nucleotides, and UTR of the *CD4* gene across different taxa

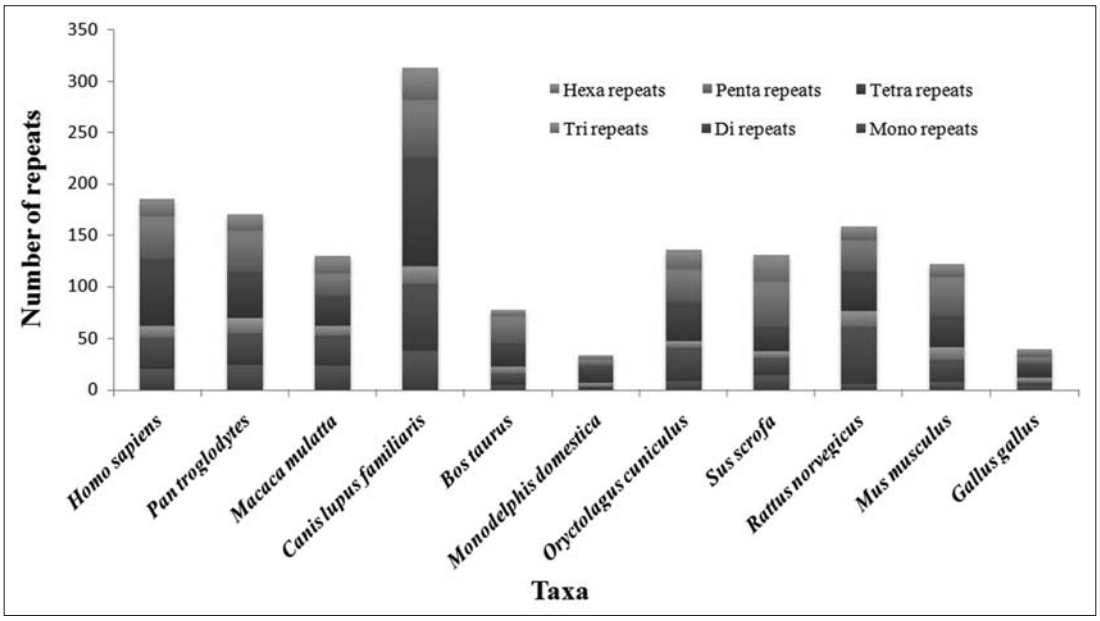

Fig. 3. Comparison of SSRs density in *CD4* gene across eleven taxa

M. domestica (3%). The contribution of intron in *CD4* gene was further confirmed with the positive Pearson's correlation of coefficient ($r = 1$, $p < 0.00001$). Additionally, the number of exon also varied across the taxa; 11 *(M. mulatta, S. scrofa)*, 10 *(H. sapiens, P. troglodytes, M. musculus, R. norvegicus, O. cuniculus, G. gallus, C. lupus familiaris)*, and 7 (*B. taurus* and *M. domestica*). Furthermore, a schematic representation of coding region and untranslated region (UTR) of *CD4* gene across eleven taxa was shown in *Fig. 1*. It was clear from the figure that coding region and the UTR are of varied lengths across taxa. While the the coding nucleotides were highest in *G. gallus* (1464 bp) and lowest in *M. domestica* (1152 bp), UTR was found to be maximum in *H. sapiens* (1726) and absent in *M. domestica*. A comparative analysis of the *CD4* intron, UTR, and coding nucleotides was computed and shown in *Fig. 2*. Furthermore, the occurrence and densities of SSR types were determined in *CD4* gene across 11 taxa *(Fig. 3)*. The overall density of the SSRs was found to be highest in *C. lupus familiaris* with a total of 314 repeats, whereas *M. domestica* contains lowest number (34) of repeats. It is interesting to note that *H. sapiens*, *P. troglodytes*, and *M. mulatta* contain equal number (170) of microsatellite repeats. Among all repeats, the tetra and penta repeats were largely distributed in *CD4* gene across all taxa *(Fig. 3)*.

To understand the evolutionary relationships among all the 11 taxa, phylogenetic tree of *CD4* gene was constructed *(Fig. 4)*. It was clear from the figure that *P. troglodytes* and *H. sapiens* were closely related at the *CD4* gene and belonged to a single clade as are *O. cuniculus* and *R. norvegicus*. Further, it was clear that *M. mulatta* has recently been diverged from the *H. sapiens–P. troglodytes* clade and *M. musculus* has recently been diverged

Table II Percentage of α helix, extended strands, β turns, random coils in eleven different taxa

Species	α Helix (Hh)	Extended strand (Ee)	β Turn (Tt)	Random coil (Cc)
H. sapiens	(118) 19.22%	(107) 17.43%	(70) 11.4%	(319) 51.95%
P. troglodytes	(59) 18.73%	(88) 27.94%	(25) 7.94%	(143) 45.4%
M. mulatta	(49) 13.39%	(113) 30.87%	(27) 7.38%	(177) 48.36%
Canis lupus familiaris	(29) 8.73%	(64) 19.28%	(42) 12.65%	(197) 59.34%
Bos taurus	(53) 8.77%	(130) 21.52%	(59) 9.77%	(362) 59.93%
Sus scrofa	(286) 10.58%	(439) 16.24%	(295) 10.91%	(1683) 62.26%
Monodelphis domestica	(258) 12.3%	(348) 16.6%	(235) 11.2%	(1257) 59.9%
Oryctolagus cuniculus	(72) 22.93%	(90) 28.66%	(33) 10.51%	(119) 37.9%
M. musculus	(75) 11.83%	(161) 25.39%	(59) 9.31%	(339) 53.47%
R. norvegicus	(34) 9.94%	(75) 21.93%	(31) 9.06%	(202) 59.06%
G. gallus	(17) 4.09%	(69) 16.59%	(18) 4.33%	(312) 75%

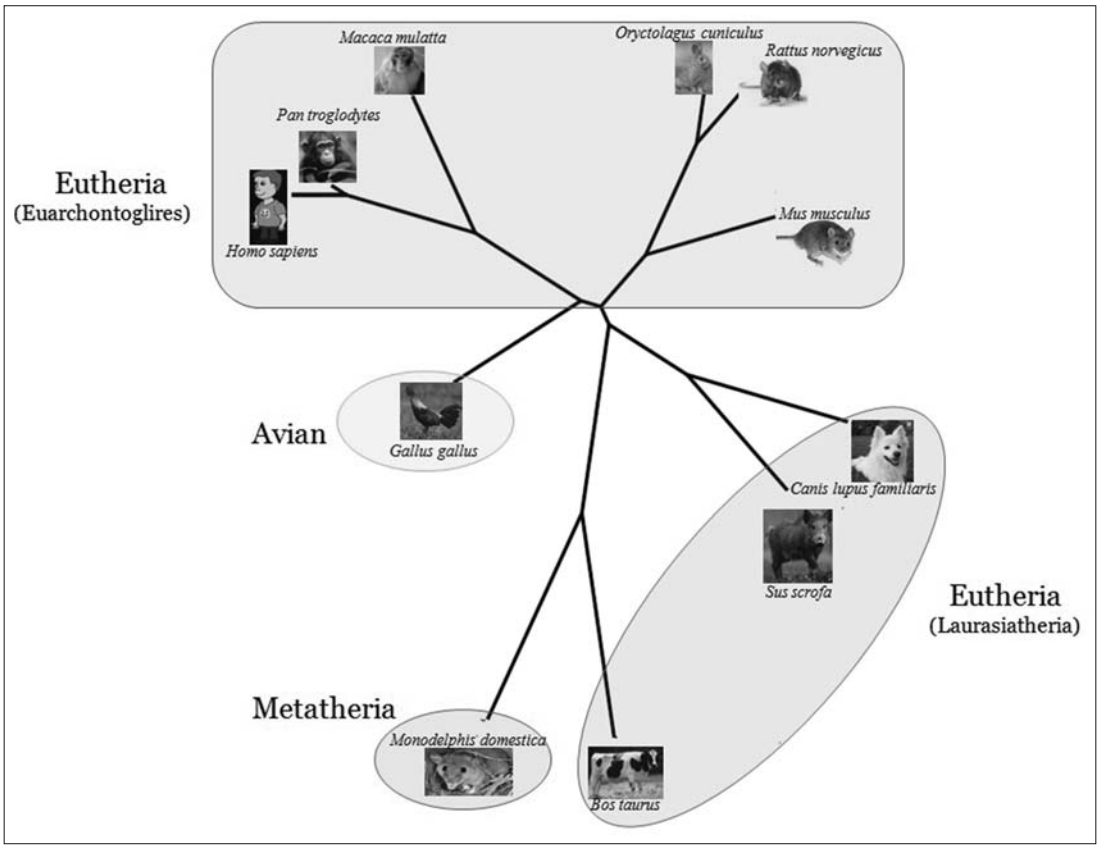

Fig. 4. Comparison on the intron, coding nucleotides, and UTR of the *CD4* gene across different taxa

from *O. cuniculus–R. norvegicus* clade. Likewise, *M. domestica* and *B. taurus* occupy a single clade and *C. lupus familiaris* and *S. scrofa* occupy another clade. Interestingly, *G. gallus* occupies an entirely separate branch, acting as an outgroup. The strength of internal branches was found to be almost absolute for almost all the clades after 1000 simulations.

Apart from genomic and phylogenomics inferences, secondary structure of cd4 protein was determined using SOPMA *(Fig. 5)*. Comparing the secondary structure of cd4 protein, it was observed that human cd4 protein structure has 19.22% α helix, 11.4% β strand, and 51.95% random coils showing a small change in comparison to *P. troglodytes* (18.73% α helix, 27.94% β strand, 45.4% random coils) as shown in *Table II*.

Discussion

With the availability of high-throughput technologies such as sequencing, a large number of genome sequences of different organisms are becoming available rapidly. These increasing amounts of genomic data have revolutionized the field of genomics, especially comparative genomics. Comparative genomics thus is considered as a powerful tool for understanding the evolutionary role in shaping the genetic patterns in closely related species [17]. Acquiring comparative genomic approaches, the present study attempts to infer evolutionary status of the *CD4* gene involved in HIV pathogenicity. We herewith report the systematic computational study of *CD4* gene in human, mammalian, and avian taxa. We found several characteristic patterns of this gene, *viz.*, location, length of the gene, length of exon, intron and UTR, and number of exon and intron varied considerably across the taxa. The length of the whole *CD4* gene was found to be maximum in *C. lupus familiaris* and minimum in *M. domestica*. Additionally, *C. lupus familiaris* was also found to have the longest intron whereas, and *M. domestica* was found to have shortest intron. Moreover, a significant proportion of *CD4* gene in all taxa *(H. sapiens, P. troglodytes, M. mulatta, M. musculus, C. lupus familiaris, M. domestica, G. gallus, B. taurus, R. norvegicus, O. cuniculus, S. scrofa)* was constituted by the introns. The observed feature was supported by the fact that the higher eukaryotes carry a large burden of junk DNA in their genomes [18], which act as the raw material for natural selection and appear to have been conserved by purifying natural selection [19]. This fact was substantiated by the significant positive Pearson's correlation coefficient (r) between the total gene length and intronic DNA. This signifies sufficient increase in size of *CD4* gene due to accumulation of non coding nucleotides, a feature similar to the human drug metabolizing gene

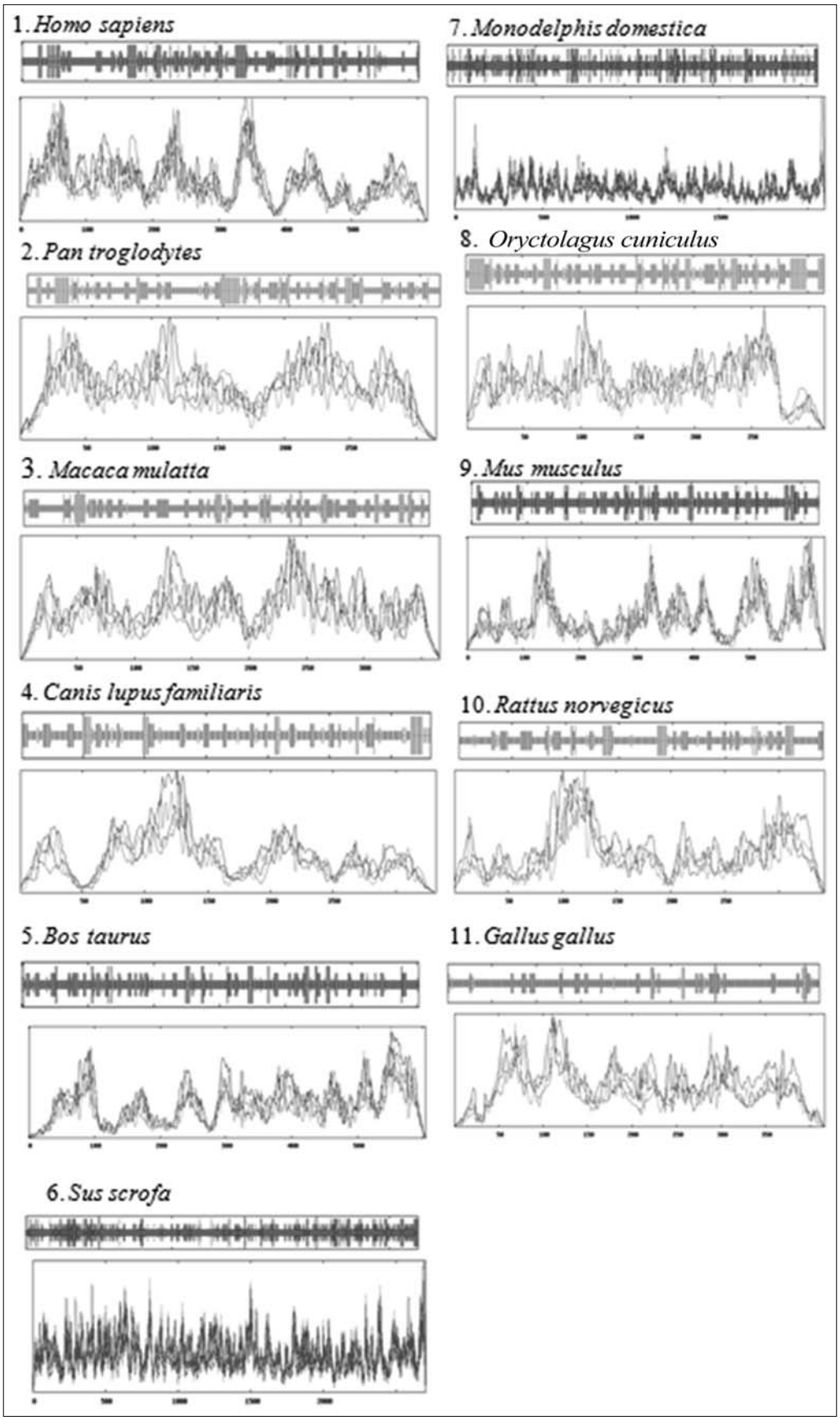

Fig. 5. Comparison of secondary structure of CD4 across different taxa (α Helix [Hh], 3_{10} helix [Gg], Pi helix [Ii], β bridge [Bb], extended strand [Ee], β turn [Tt], bend region [Ss], random coil [Cc])

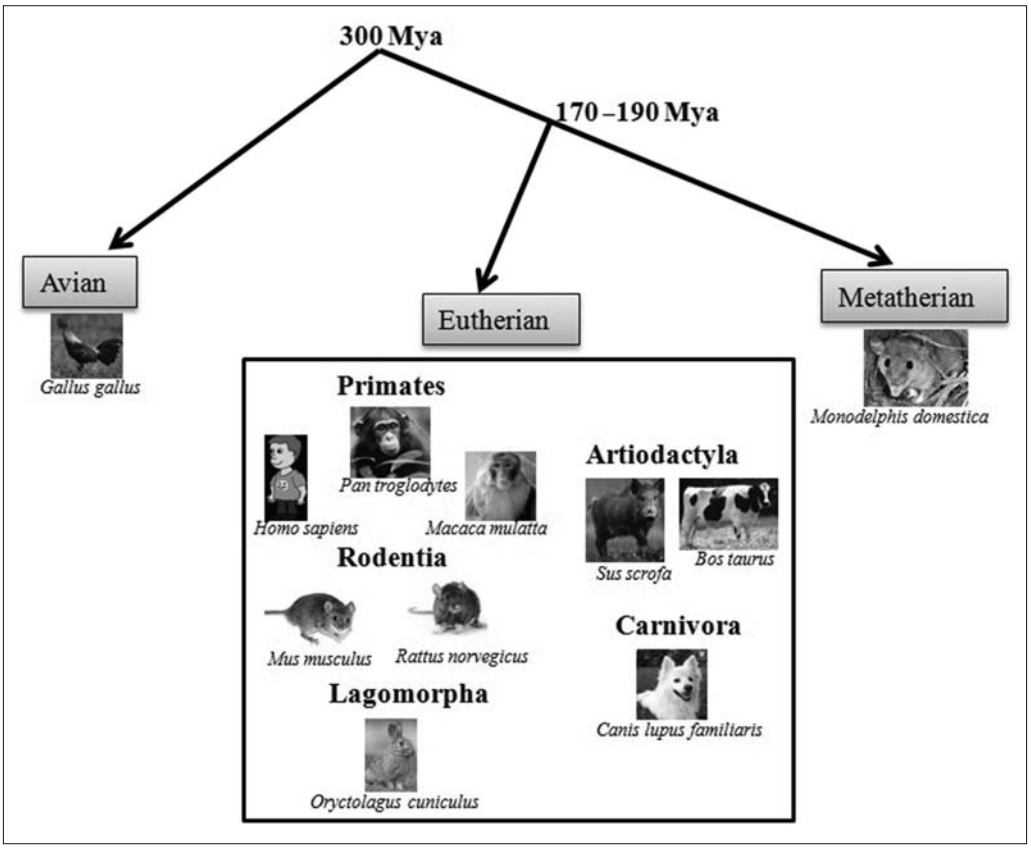

Fig. 6. Phylogenetic splitting topology for mammalian (Eutheria and Metatheria) divergence

NAT2 [20]. Furthermore, DNA repeat expansion leads to formation of introns, [21] and presence of introns in genes indicates high rate of genome evolution that plays a role in directing transcript and post-transcriptional control of gene expression [22, 23]. Thus, the presence of introns in a gene helps species to evolve at a faster rate [23]. This fact was validated with the results obtained in the present study as a large number of microsatellites of different repeat size were observed in the human *CD4* gene. It is well known fact that microsatellites are present in eukaryotic genomes due to replication slippage which results into high mutation rate [24]. The maximum numbers of microsatellites were observed in *C. lupus familiaris* (314), having maximum gene length and longest intron, whereas *M. domestica* harbors the lowest (34) number of microsatellites. The smaller *CD4* gene size of *M. domestica* compared to other mammalian and avian gene size would logically predict metatherian to contain lower number of microsatellites. Among all repeat types, tetra repeats and penta repeats were found to be highly abundant which was thought to be the characteristic of all vertebrate taxa [24]. Surprisingly, in the present study, the *CD4* coding nucleotides that were found to be maximum in Avian taxa *(G. gallus)* contain the highest *CD4* coding nucleotide which indicates high rate of expression of cd4 protein for defensive mechanism. This pattern was quite different from other human genes, e.g., *NAT2* [20]. Furthermore, it is well-known fact that UTR plays a vital role in gene regulation by playing a crucial role in post translational regulation [25]. Thus, the variation in UTR across eleven taxa signifies that the mechanism of *CD4* gene regulation might be different in different taxa with maximum in *H. sapiens* and absent in *M. domestica*.

In order to understand phylogenetic relationships among different taxa, which can throw insights into the evolutionary paradigm of the *CD4* gene, NJ phylogenetic tree was constructed. It is quite clear that *H. sapiens* and *P. troglodytes* are closely related to each other at *CD4* gene, as *M. musculus* and *R. norvegicus* are. Phylogenetic positions of different taxa based on the *CD4* gene were found to be quite similar for other genes (TNF, IL-10, IL-23a, MHC) [26, 27] that are involved in regulation of human immune system. Thus, it can be predicted that the *CD4* gene might be evolving at a similar rate as to human immune genes.

The observed comparative genomic data can also be discussed in terms of functional aspect of the *CD4* gene. Secondary structure of cd4 protein was predicted from SOPMA server as shown in *Fig. 5*. Comparing the secondary structure of cd4 protein obtained from human and chimpanzee, it was found that the secondary structure was approximately similar in both the taxa maintaining the cd4 protein structure.

The present study focuses on comparative study based on CD4 gene across eleven taxa which can be discussed dividing all eleven taxa into three main groups, viz., Metatherian (marsupials), Eutherian (placental), and Avian (Birds) *(Fig. 6)*. It was reported that metatherian mammal lineage split from eutherians (placental mammals) approximately 170–190 million years ago (Mya) and from avian taxa approximately 300 Mya [28]. Thus, the positioning of metatherian between Eutherian and Avian makes it valuable for evolutionary comparisons [29]. Although metatherian and eutherian have acquired distinct physiological and behavioral features, they still share many ancestral therian characters, most notably lactation using mammary papilla, and the bearing of live young without using a shelled egg. Despite having characteristics of eutherian lineage, the composition of *CD4* gene in metatherian lineage *(M. domestica)* was found to be lower than avian group which further supports the fact that *M. domestica CD4* is "minimal essential" and is much more compact and simple than other mammalian and avian lineage.

Conclusions

In conclusion, the detailed genetic characteristic and evolutionary analysis of human *CD4* gene that is involved in activation of immune system with comparative genomic analyses provide many interesting results. Since comparative genomics provide a powerful approach to better understand regulation of complex biological sciences and offers deeper insights into the way evolutionary forces (natural selection) moulds genes and genomes, such approaches have been proven to be useful in an array of animal taxa, e.g., in *Bubalus bubalis* [30], sponge [31], adenovirus [32], and in several human genes, *viz.*, MS4A and TMEM 176 [33]. With the basic understanding on genetic characterization and evolutionary genomics, the patterns of immune system activation and defensive mechanism in humans would be better visualized. Further work on the diversity patterns of *CD4* gene in different human population would provide additional clues to which extent this gene varies and responds to different evolutionary forces.

* * *

Funding sources: None.

Conflict of interest: None.

Acknowledgements: The author thanks the Indian Council of Medical Research (ICMR) for Junior and Senior Research Fellowships.

References

1. Caicedo AL, Purugganan MD: Comparative plant genomics: frontiers and prospects. Plant Physiol 138, 545–547 (2005)
2. Gil JP, Gil Berglund E: CYP2C8 and anti-malarial drug efficacy. Pharmacogenomics 8, 187–198 (2007)
3. Clark SJ, Jefferies WA, Barclay AN, Gagnon J, Williams AF: Peptide and nucleotide sequences of rat CD4 (W3/25) antigen: evidence for derivation from a structure with four immunoglobulin-related domains. Proc Natl Acad Sci USA 84, 1649 (1987)
4. Groot F, Kuijpers TW, Berkhout B, de Jong EC: Dendritic cell-mediated HIV-1 transmission to T cells of LAD-1patients is impaired due to the defect in LFA-1. Retrovirology 3, 75 (2006)
5. Stein BS, Gowda SD, Lifson JD, Penhallow RC, Bensch KG, Engleman EG: pH-independent HIV entry into CD4-positive T cells via virus envelope fusion to the plasma membrane. Cell 49(5), 659–668 (1987)
6. DR Littman: The structure of the CD4 and CD8 genes. Annu Rev Immunol 5, 561–584 (1987)
7. Maddon PJ, Dalgleish AG, McDougal JS, Clapham PR, Weiss RA, Axel R: The T4 gene encodes the AIDS virus receptor and is expressed in the immune system and the brain. Cell 47, 333–348 (1986)
8. NCBI: [www.nlm.nih.gov]
9. Pearson's correlation coefficient: [http://analyse-it.com]
10. Alignment algorithm: [www.dnastar.com]
11. Phylogeny software: [www.phylogeny.fr]
12. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard J-F, Guindon S, Lefort V, Lescot M, Claverie J-M, Gascuel O: Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res 36, W465–469 (2008)
13. Microsatellite finder: [http://www.cdfd.org.in/imex]
14. Mudunuri SB, Nagarajaram HA: IMEx: Imperfect Microsatellite Extractor. Bioinformatics 23, 1181–1187 (2007)
15. Protein secondary structure: [http:www.ibcp.fr]
16. Geourjon C, Deleage G: SOPM: A self optimised prediction method for protein secondary structure prediction. Protein Eng 7, 157–164 (1994)
17. Han L, Zhao Z: Comparative analysis of CpG islands in four fish genomes. Comp Funct Genomics 565631 (2008)
18. Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA: Wide distributed non-coding purifying selection in the human genome. Proc Natl Acad Sci USA 104, 12410–12415 (2007)
19. Miller W, Makova KD, Nekrutenko A, Hardison RC: Comparative genomics. Annu Rev Genomics Hum Genet 5, 15–56 (2004)
20. Khan N, Pande V, Das A: Characterization, comparative genomic and evolutionary inferences of a human drug metabolizing (NAT2) gene. Interv Med Appl Sci 3, 65–73 (2011)
21. Shakyawar SK, Joshi BK, Kumar D: SSR repeat dynamics in mitochondrial genomes of five domestic animal species, Bioinformation 4, 158–163 (2009)
22. Ohno S (1970): Heidelberg Pearson's correlation coefficient. Evolution by Gene Duplication. Springer-Verlag, Heidelberg [http://analyse-it.com]
23. Sullivan JC, Reitzel AM: A high percentage of introns in human genes were present early in animal evolution: evidence from the basal metazoan *Nematostella vectensis*. Genome Informatics 17(1), 219–229 (2006)
24. Tóth G, Gáspári Z, Jurka J: Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res 10, 967–981 (2000)
25. Bashirullah A, Cooperstock RL, Lipshitz HW: Spatial and temporal control of RNA stability. Proc Natl Acad Sci USA 98(13), 7025–7028 (2001)
26. Cogswell-Hawkinson AC, McGlaughlin ME, Calisher CH, Adams R, Schountz T: Molecular and phylogenetic characterization of cytokine genes from Seba's short-tailed bat *(Carollia perspicillata)*. Open Immunol J 4(31), 31–39 (2011)
27. Belov K, Sanderson CE, Deakin JE, Wong ES, Assange D, McColl KA, Gout A, de Bono B, Barrow AD, Speed TP, Trowsdale J, Papenfuss AT: Characterization of the opossum immune genome

provides insights into the evolution of the mammalian immune system. Genome Res 17(7), 982–991 (2007)
28. Kumar S, Hedges SB: A molecular timescale for vertebrate evolution. Nature 392, 917–920 (1998)
29. Gentles AJ, Wakefield MJ, Kohany O, Gu W, Batzer MA, Pollock DD, Jurka J: Evolutionary dynamics of transposable elements in the shortailed opossum *Monodelphis domestica*. Genome Res 17(7), 992–1004 (2007)
30. Banerjee P, Gahlawat SK, Joshi J, Sharma U, Tantia MS, Vijh RK: Sequencing, characterization and phylogenetic analysis of TLR genes of *Bubalus bubalis*. DHR Int J Biomed Life Sci (DHR-IJBLS) 3(1), 2278–2301 (2012)
31. Yuen B, Bayes JM, Degnan SM: The characterization of sponge NLRs provides insight into the origin and evolution of this innate immune gene family in animals. Mol Biol Evo 31(1), 106–120 (2014)
32. Rux JJ, Kuser PR, Burnett RM: Structural and phylogenetic analysis of adenovirus hexons by use of high-resolution X-ray crystallographic, molecular modeling and sequence-based methods. J Virol 77(17), 9553–9566 (2003)
33. Zuccolo J, Bau J, Childs SJ, Goss GG, Sensen CW, Deans JP: Phylogenetic analysis of the MS4A and TMEM176 gene families. PLoS One 23, 5(2), e9369 (2010)