Morenikeji OB, Thomas BN. 2019. In silico analyses of CD14 molecule reveal significant evolutionary diversity, potentially associated with speciation and variable immune response in mammals. PeerJ 7:e7325
https://doi.org/10.7717/peerj.7325
In silico analyses of CD14 molecule reveals significant evolutionary diversity potentially associated with speciation and variable immune response in mammals

Olanrewaju B Morenikeji ¹, Bolaji N Thomas     Corresponding Author: Bolaji N Thomas  
Email address: bntsbi@rit.edu

¹ Biomedical Sciences, Rochester Institute of Technology, Rochester, NY, United States of America

Cluster differentiation gene (CD14) is a family of monocyte differentiating genes that works in conjunction with lipopolysaccharide binding protein (LBP), forming a complex with TLR4 or LY96 to mediate innate immune response to pathogens. In this paper, we used different computational methods to elucidate the evolution of CD14 gene coding region in 14 mammalian species. Our analyses identified leucine rich repeats (LRRs) as the only significant domain across the CD14 protein of the 14 species, presenting with frequencies ranging from 1-4. Importantly, we found signal peptides located at mutational hotspots demonstrating this gene is conserved across these species. Out of the 10 selected variants analyzed in this study, only 6 were predicted to possess significant deleterious effect. Our predicted protein interactome showed a significant varying protein-protein interaction with CD14 protein across the species. This may be important for drug target and therapeutic manipulation for the treatment of many diseases. We conclude that these results contribute to our understanding of the CD14 molecular evolution, which underlays varying species response to complex disease traits.
In silico analyses of CD14 molecule reveals significant evolutionary diversity potentially associated with speciation and differential immune response in mammals

Olanrewaju B. Morenikeji, Bolaji N. Thomas
Department of Biomedical Sciences, Rochester Institute of Technology, Rochester NY

Olanrewaju B. Morenikeji: obmhst@rit.edu
Bolaji N. Thomas: bntsbi@rit.edu

*Corresponding author:
Dr. Bolaji N. Thomas
Department of Biomedical Sciences
Rochester Institute of Technology
153 Lomb Memorial Drive
Rochester NY 14623
Office: (585) 475-6382
Fax: (585) 475-5809
Email: bntsbi@rit.edu
Abstract

Cluster differentiation gene (CD14) is a family of monocyte differentiating genes that works in conjunction with lipopolysaccharide binding protein (LBP), forming a complex with TLR4 or LY96 to mediate innate immune response to pathogens. In this paper, we used different computational methods to elucidate the evolution of CD14 gene coding region in 14 mammalian species. Our analyses identified leucine rich repeats (LRRs) as the only significant domain across the CD14 protein of the 14 species, presenting with frequencies ranging from 1-4. Importantly, we found signal peptides located at mutational hotspots demonstrating this gene is conserved across these species. Out of the 10 selected variants analyzed in this study, only 6 were predicted to possess significant deleterious effect. Our predicted protein interactome showed a significant varying protein-protein interaction with CD14 protein across the species. This may be important for drug target and therapeutic manipulation for the treatment of many diseases. We conclude that these results contribute to our understanding of the CD14 molecular evolution, which underlays varying species response to complex disease traits.

Keywords: CD14, mammals, species, immune response, evolution, in silico
Introduction

Cluster of differentiation 14 (CD14) gene is a surface differentiation antigen preferentially expressed on mammalian monocytes, neutrophils, macrophages, and plasma cells (Baumann et al., 2010; Tang et al., 2017). It encodes a protein that is important for initiating a robust immune response against microbial pathogens by mediating innate immune response, in concert with several other proteins. It is a co-receptor with Toll-like receptor-4 (TLR4) to activate several intracellular signaling pathways that lead to the synthesis and release of inflammatory cytokines, antimicrobial peptides, chemokines, and other co-stimulatory molecules, which in turn interact with the adaptive immune system (Hartel et al., 2008). Comparative studies have shown that two or more proteins can have common evolutionary origin thereby sharing structural and functional characteristics (Kanduc, 2012). CD14 molecule exists in two forms: soluble (sCD14) or membrane-bound (mCD14) (Panaro et al., 2008; Xue et al., 2012). There are multiple variants of the CD14 molecule that are encoded by the same protein due to alternative splicing and as such has been mapped to varying chromosomal locations in different species. For example, it is mapped to chromosome 5 in humans, chromosome 7 in cattle and chromosome 18 in mouse (Ferrero et al., 1990; Le Beau et al., 1993; Ibeagha-Awemu et al., 2008).

Studies in human, mouse, cattle and sheep have shown that CD14 is significantly involved in innate immunity, playing major roles in susceptibility to tuberculosis, trypanosomosis, malaria and other bacterial infections (Sugawara et al., 2003; Ibeagha-Awemu et al., 2008; Xue et al., 2012; Ojurongbe et al., 2017). Other published reports have shown that there is a higher susceptibility to *Mycobacterium tuberculosis* infection in CD14 transgenic mice compared to the wild type (Reiling et al., 2002; Weiland et al., 2008). Likewise, single nucleotide polymorphisms
(SNPs) in CD14 gene have been associated with higher susceptibility in many disease instances (Oakley et al., 2009; Liu et al., 2012; Xue et al., 2012; Zanoni and Granucci, 2013; Thomas et al., 2015; Xue et al., 2017). In fact, Song et al. (2014) reported how genetic heterozygosity modulate disease resistance and progression in cattle infected with bovine tuberculosis. Furthermore, comparative studies have shown that organism relatedness can be traced through their pattern of genetic divergence (De Donato et al., 2017; Peters et al., 2018).

Several sequence-based methods and tools have been developed to glean evolutionary information in related species via amino acid sequence variation and conservation of homologous proteins through multiple sequence alignment (MSA) (Hepp et al., 2015, Peters et al., 2018). Similarly, other computational methods are available to identify SNP variation within and between amino acid sequences in multiple species, possibly affecting the stability and functionality of such proteins (Ng and Henikoff, 2006; Yue and Moult 2006; Hepp et al., 2015).

Many of these tools can predict the effect of SNP occurrence in protein sequences to determine whether they are disease related, deleterious or neutral. Comparative genomics therefore is a powerful tool to elucidate variants and effects among species in order to detect diseases associated with variations. Variations in amino acid sequence have the ability to alter protein structure and functions like ligand binding, protein folding, impaired intracellular transport and reduced stability (Zeron-Medina et al., 2013; Morisseau et al., 2014; Valastyan and Lindquist 2014).

Due to the significance of CD14 gene in several disease cases in humans and other species, in addition to its considerable involvement in innate immunity, we hypothesize that there might be
evolutionary patterns of similarity and diversification that occurred during speciation, which is important for comparative immune and disease studies in different species. To this end, we carried out a detailed comparative study of CD14 protein in 14 mammalian species to elucidate the evolutionary basis for conserved regions, active sites and mutational hotspots, which could lead to novel disease phenotypes. In addition, we examine the diversification in CD14 protein interactions within and across the species, which could be explored for therapeutic development or drug design.

**Materials and Methods**

*Sequence retrieval and multiple sequence alignment*

Complete CD14 amino acid sequences of 13 mammals were retrieved from UniProtKB/Swiss-Prot (https://www.uniprot.org/uniprot/?query=CD14&sort=score) database. The sequences were retrieved for human (P08571), rat (Q63691), mouse (P10810), cattle (Q95122), rabbit (Q28680), monkey (B3Y6B8), gorilla (G3R4C0), sheep (W5QJA2), horse (F6VK89), pig (A7BG66), buffalo (A0A2R4SDF9), chimpanzee (B3Y6B4) and yak (L8I9P7). The amino acid sequence for goat (ABE68725.1) was retrieved from GenBank. We performed sequence alignment with the Multalin software (http://multalin.toulouse.inra.fr/multalin/), which does a simultaneous alignment of biological sequences with hierarchical clustering. To examine similarity between the sequences, we used SIAS (Sequence Identity And Similarity, http://imed.med.ucm.es/Tools/sias.html) with default BLOSUM62 scoring matrices.

Evolutionary tree was constructed from the sampled species through Phylogeny.fr (http://www.phylogeny.fr/index.cgi) online program.
Comparative physicochemical properties of amino acid sequence in the CD14 molecule

The biochemical properties of the amino acids from the 14 mammalian species were computed with ProtParam (www.expasy.org/protparam/). The following properties were computed for each sequence: aliphatic index, which defines the relative volume of a protein occupied by alanine, valine, isoleucine, and leucine; instability index, which estimates the protein stability based on the amino acid composition; protein net charge, which can be positive, negative or neutral based on the amino acid composition in the protein; molecular weight; grand average of hydropathicity (GRAVY), which determines the hydrophobicity of a protein from the aliphatic side chain; and isoelectric point (pI), which is the pH at which the protein net charge is equal to zero.

Functional analysis, motif scanning and prediction of signal peptides

We performed functional analysis on the protein sequences in order to classify them into superfamilies, predict domains, repeats and find important sites that may be relevant in evolution. We scanned for motif signatures among amino acid sequences with the combined use of ScanProsite (https://prosite.expasy.org/) (Sigrist et al., 2010) and InterPro, an online program that analyzes protein sequences and classification (https://www.ebi.ac.uk/interpro/). The HAMAP profiles, PROSITE patterns, Pfam global models and PROSITE profiles were all included in the search. Sequence logo of the identified conserved domain in the CD14 protein among the 14 mammalian species was constructed with WebLogo (http://weblogo.berkeley.edu/logo.cgi), to show the graphical view of the region containing the conserved amino acid among the species. Furthermore, we predicted the cleavage sites and the presence of signal peptides in CD14 protein from the 14 mammalian species using SignalP 5.0 server (http://www.cbs.dtu.dk/services/SignalP/), which uses recurrent neural network
architecture and deep convolution to classify signal peptides into lipoprotein signal peptides, secretory signal peptides or Tat signal peptides. In order to gain a better understanding of the localization of the protein in each species, we predicted subcellular localizations of CD14 protein using Neural Networks algorithm on DeepLoc-1.0 server (http://www.cbs.dtu.dk/services/DeepLoc/), and the construction of the subcellular pathway hierarchical tree.

**Prediction analysis of amino acid substitution**

The effect of the amino acid substitution was predicted using the combination of SIFT (Sorting Intolerance from Tolerance), PANTHER (Protein ANalysis THrough Evolutionary Relationship) and PROVEAN (Protein Variation Effect Analyzer). Briefly, we used human CD14 amino acid sequence to query the multiple sequence alignment of other mammalian species in this study using SIFT, which predict the tolerance or deleterious effect of substitutions for each position in the query sequence. Any position with probability less than 0.05 is classified as deleterious, as previously described (Bendl et al., 2014; Choi and Chan 2015). We selected a total of 10 variants from the mutational hotspots as predicted by SIFT and further estimate the likelihood of the selected variants and their effects on protein function through PROVEAN and PANTHER.

**Prediction of protein interactome with CD14 protein in different species**

In order to establish specific interaction of the CD14 protein with other molecules as a result of biochemical events during speciation, we used the retrieved CD14 amino acid sequence from each mammalian species to predict its association with other protein groups and generate different networks using STRING, a database that predicts protein-protein interactions (https://string-
db.org/). This is important in order to examine the diversity shaped by evolution in the association of CD14 gene with other molecules in different organisms. Venn diagrams were constructed for the comparison and visualization of overlapping protein-protein interaction (PPI) among different species using two web-based applications

(http://bioinformatics.psb.ugent.be/software/details/Venn-Diagrams and
http://bioinfogp.cnb.csic.es/tools/venny/)

Results

Comparative analysis and sequence evolutionary trace

In this study, we examined the evolutionary pattern of CD14 protein sequences in 14 mammalian species. The alignment is conserved within two groups separated into ruminants and non-ruminants. The multiple sequence alignment identified leucine (L), aspartic acid (D), lysine (K), glutamic acid (E), valine (V), glycine (G), serine (S) and asparagine (N) as evolutionarily conserved amino acid residues, while others like proline (P), glutamine (Q), methionine (M), alanine (A), phenylalanine (F), isoleucine (I) and threonine (T) were evolutionarily varied. The CD14 protein sequence demonstrates significant variability in both percentage identity and similarity across the 14 species, despite the common evolutionary origin (Figure 1, 2). The percentage identity of CD14 protein in monkey, gorilla, chimpanzee and human was similar while gorilla shares the closest identity with human (Table 1). Among the ruminants, cattle and yak share the closet similarity compared to buffalo, sheep and goat, although the phylogenetic tree suggests that goat is distantly related. While mouse and rat cluster with the same origin, the analysis shows that they share less identity (7.4%) and similarity (13.4%). Rabbit, horse and pig are distantly apart from other species, as they do not share high conservation (Table 1, Figure 2).
In all, the sequence of CD14 protein in goat and horse share the least identity (6.7% and 6.9% for goat and horse respectively) and similarity (9.9% and 13.2% for goat and horse respectively) with human.

**Physicochemical properties at the CD14 promoter region**

ProtParam tool (www.expasy.org/protparam/) was used to compute the physical and chemical properties of CD14 amino acid sequences among the 14 species (Table 2). The aliphatic index is generally high for all species showing that the protein is thermally stable. A higher instability index was observed in the CD14 molecule of rabbit, pig and monkey (53.0, 46.8 and 45.1 respectively), indicating that the protein is less stable and hydrophobic amino acids, such as leucine, valine, serine and asparagine, occupy majority of the sequence, providing higher tolerance against diseases. The lowest instability index is observed in horse (33.5) and goat (35.1) showing that the protein is more stable in these species. The CD14 protein in goat also has the lowest aliphatic index (99.7) while mouse has the highest (107.7). We observed a closer range of molecular weight among the species in this study, although gorilla, monkey, human, chimpanzee and rat had the higher molecular weight with close range (Table 2). Negative net charge, indicative that the protein is more basic than acidic, ranged from -9 as found in mouse and rat to +4 as found in goat. Goat, horse and gorilla has higher isoelectric point (Ip) indicating that CD14 molecule is highly basic in these species than others. The GRAVY values obtained were generally positive and higher in ruminants than non- ruminants, suggesting the proteins are more hydrophobic, which enhances oligomerization and higher binding capability to different proteins.
Characterization of functional motifs and prediction of signal peptides

The CD14 amino acid sequences of the 14 mammalian species in this study were individually scanned for matches against the InterPro and PROSITE collection of protein signature databases. We found one domain (Leucine-rich repeat (LRR), PS51450) with varying frequency across the 14 species (Figure 3). Comparison of the predicted intra-domain features show one LRR domain in human, two each in gorilla, chimpanzee, monkey, horse and pig, three each in cattle, sheep, buffalo, yak, and mouse, with the highest number (4) found in rat. Figure 4 shows the multiple sequence alignment of the homology of LRR domain across the 14 species, showing that leucine, aspartic acid, serine and asparagine are 100% conserved in this region. The sequence logo built from the multiple sequence alignment of the domain is displayed in Figure 5, with the logo showing the relative frequencies of each conserved amino acid and their position in the LRR domain. The domain homology reveals that there is significant conservation of most amino acids in this region.

Furthermore, we predicted the signal peptides, position and secretory pathway of the CD14 amino acids in the 14 species under consideration. Our analysis show that chimpanzee, gorilla, human and monkey share the same signal peptide (VSA-TT) at the same position (19 and 20), with high likelihood (Table 3). Buffalo, cattle, sheep and yak also share the same signal peptide (VSA-DT) and position (20 and 21), although sheep has a different position (19 and 20). We observed a significant variation for the rest of the species in terms of signal peptides and their positions (Table 3). Interestingly, signal peptide for all the species (Figure 6a), except sheep (Figure 6b), share the same subcellular localization in the neural network.
Mutational analysis of predicted variation

A total of 10 variants were selected from the predicted mutations by SIFT and the effects were tested as deleterious or not in the 14 species with PROVEAN and PANTHER. Our analysis showed that 4 of these variants (D28V, W45H, G62E, L70D) were validated mutations with deleterious effect on all species with 2 others found in few species. These variants cluster in the C-terminus region of CD14 protein between 20 to 100 amino acids. A closer look suggests that mutational effect on the CD14 protein sequence varied from C-terminus to N-terminus with less mutational effect towards the N-terminus (Table 4). The deleterious mutations observed in our study were all at the C-terminus region thus identifying it as a mutational hotspot. Q100G, V301M, L318I, G335T, L357H and G370K mutation spots were neutral for most species. This might mean that CD14 is less conserved in this region because of evolutionary divergence of all species. However, L-H at position 357 showed a deleterious effect in cattle, yak, pig, gorilla, human, monkey, buffalo and chimpanzee, while there is also a deleterious effect of G-K at position 370 of CD14 in rat.

Protein-protein interaction cluster with CD14 gene in different species

In order to deduce protein-protein interactions (PPI) that evolved through speciation due to co-localization, additive genetic interaction, co-expression or repression and physical association with CD14 in the mammalian species under study, we used STRING to build the protein network, based on collection of laboratory experimental results from the database (Figure 7) and segment the gene pool base on our phylogenetic result to build Venn diagrams for each species cluster (Figure 8a, b, c). We could not find any protein network for horse and so was excluded in the analysis. Our result shows that there is significant variation in the CD14 protein interactome
across species (Figure 7). Generally, we found that there were different proteins that clustered with CD14 in all the species. All species had 10 proteins in their cluster except cattle and goat that had 11. Looking at the Venn diagram, rabbit had the highest CD14 PPI that is not shared with others while 3 protein set (CD14, TLR2 and TLR4) is common to members of this group (Figure 8a). Figure 8b shows the ruminant group, including goat, sheep and yak had no unique gene set, meaning the PPI is duplicated in one or two other members of the group. However, cattle has 8 unique PPI while buffalo has 4, that were not shared with others. CD14 and TLR2 are common to all in this group. Likewise, there were 8 unique PPI in human, 6 in gorilla and none in monkey and chimpanzee (Figure 8c).

**Discussion**

Comparative analysis of CD14 protein in this study enhances our understanding of genome plasticity among 14 mammalian species and establishes functional, molecular and structural relationships in different clades that are important in an evolutionary trace. The significant variability in the multiple sequence alignment of CD14 molecule across the species suggests a high evolutionary divergence especially between the ruminant and non-ruminant group. This implies that CD14 amino acid sequence had undergone significant changes during speciation leading to functional and structural modification in different species. Studies have shown that variation in amino acid sequences could impact immunogenicity, immunotolerance and immunoreactivity (Tauber, 2004, Kanduc, 2012; Bendl et al., 2014). However, we found that amino acid residues like leucine (L), glutamic acid (E), lysine (K), valine (V), aspartic acid (D), glycine (G), serine (S) and asparagine (N) are highly conserved, thereby retaining some degree of homology in functional, molecular and structural characteristics. In addition, this reveals the
common origin between the mammalian species before divergent speciation. Based on the percentage identity and similarity, monkey, gorilla and chimpanzee are closer to human in their CD14 amino acid sequence, suggesting a lower degree of variation and may infer some degree of similar CD14 expression during disease condition (Ferrero et al., 1990; Ibeagha-Awemu et al., 2008; Bendl et al., 2014).

We also observed that the molecular weight, isoelectric point (Ip), instability index and net charge of CD14 protein for this group of mammals are similar, suggesting a key biochemical and immunological function is retained in these species during evolution (Saha et al., 2013; Ajayi et al., 2018). Of interest, the CD14 sequence in cattle and buffalo were much more conserved than yak, despite their common origin potentially implying that domestication has not affected key biological functions in cattle, and the possibility that buffalo can also be domesticated without loss of immunological function. Furthermore, a higher aliphatic index, net negative charge and GRAVY as shown in the physicochemical properties of CD14 protein in mouse and rat gives an indication of high concentration of alanine, valine, isoleucine and leucine, reported to influence transcription factors, providing higher tolerance against bacterial and viral infections (Korber, 2000; Panaro et al., 2008; Ivanov et al., 2015). This is thought to be an important evolutionary adaptation for these small animals to survive bouts of exposure to diseases in their environment, and may explain the basis for these organisms at times serving as reservoir hosts for many disease pathogens in humans. The general negative net charge of CD14 protein as observed across the species indicates an increasing reactivity and help in its receptor binding mechanism. Therefore, the higher the net charge, the more the reactivity of the protein.
Interestingly, our motif and signal peptide scan found just one domain and one signal peptide in the entire length of CD14 amino acid sequence. The numbers of conserved leucine-rich repeat (LRR) domains vary from species to species. Species with similar number of LRR profile may likely have same immunological implications. This again, is a significant evolutionary signature. CD14 is a co-receptor that bind with LPS, therefore a higher leucine amino acid profile in the molecule may accelerate its binding mechanism to receptor in a significant way because the protein plays a significant regulatory role in initiating a robust innate immune response. Studies have shown that LRR domain is evolutionarily conserved in most of the innate immune related proteins in vertebrates, invertebrates and plants, providing the innate immune defense especially through pathogen-associated molecular patterns (PAMPs) (Aylwin and Ramnik 2011). Some reports also stated that there about 2–45 leucine-rich repeats within the LRR domain, containing up to 30 residues. Classifying our mammalian species under study into ruminants versus non-ruminants, we observed that non-ruminants possess a lower number of LRR domain in their CD14 molecule (one domain in human, three in ruminants and four in rat). Notably, rat again possesses the highest number of LRR domains remarkably traceable to selection pressure across the species. Moreover, the amino acid sequence of this domain is highly conserved for all species under study, and are found towards the C-terminal region of CD14, justifying the fact that amino acid sequence variation that differentiate species are found close to the N-terminal region (Peters et al., 2018).

Our study additionally reveals varying secretory signal peptide sites in the CD14 molecule across the species. Signal peptides have been identified as hydrophobic amino acids, recognized by the signal recognition particle (SRP) in the cytosol of eukaryotic cells. Secretory signal peptide is a
class of signal peptide that allows the export of a protein from the cytosol into the secretory pathway (Nielsen and Krogh, 1998; Park and Kanehisa, 2003; Rivas and Fontanillo, 2010; Sigrist et al., 2010). In this, we found that human, monkey, gorilla and chimpanzee all have the same signal peptide site and position. Cattle, yak, sheep and buffalo also share the same site and position whereas goat did not, confirming why goat is significantly distant to other ruminants in our phylogenetic construction. It is unclear if this is related to disease tolerance when compared to other species. However, we noted in our predicted neural network that the subcellular localization of CD14 protein goes from the extracellular through the intracellular and enters the secretory pathway for all the species, except sheep. In sheep, the subcellular localization begins from the nucleus through the mitochondrion, peroxisomal targeting signal (PTS) and N-terminal sequences before it enters into the secretory pathway. This information may possess potential immunological consequences that will require further analysis and possibly an in-vitro validation.

Of most importance, a higher proportion of the predicted mutations occupying the C-terminal region of CD14 protein show that they are closer to the active site and may have direct structural and functional effects on the protein thereby causing harmful disease phenotype or susceptibility (Malm and Nilssen, 2008). Studies have shown that the leucine-rich repeats at the C-terminal region is required for responses to smooth lipopolysaccharide, whereas the variable region (290 – 375) has been found to be necessary for response to bacterial lipopolysaccharide (Bella et al., 2008; Arnesen, 2011; Xue et al., 2012, 2018). Therefore, variation at this region might be traceable to varied exposure and responses to pathogens in the cause speciation.
We observed a higher proportion of deleterious mutational spots in human, monkey, gorilla and chimpanzee occupying the same loci compared to ruminants and other species. This might suggest that the vital residue conservation at this region is due to selection pressure among these species and has been maintained over time possibly because of their role in evolution, resulting in similar biological and immunological function (Feder and Mitchell-Olds, 2003; De Donato et al., 2017; Peters et al., 2018). Therefore, a perturbation of the amino acid sequence at this region could affect protein folding, ligand binding and other functions, which might be lethal or regarded as disease-causing mutation in all mammals. Understanding the molecular variation in the region could help solve the challenge of Mendelian disease phenotypes. We recommend an in vitro study of this region in CD14 protein sequence to elucidate the molecular mechanism affecting functionality of this region. In all, 3 of these mutations have been characterized and verified in humans to cause disruption of active site and loss of protein activities (Hidam and Debasish, 2018).

Furthermore, we used the STRING database to annotate CD14 protein network with other protein molecules that may have evolved together during speciation. Significantly, we found that CD14 molecule selectively interact with other proteins from species to species. For example, in cattle, CD14 molecule interacts with 8 other proteins, which are not shared with goat, sheep and yak. In a similar vein, buffalo has 4 unique sets of protein that co-express with CD14 protein. Human and gorilla in their group has 8 and 6 genes respectively that uniquely interact with CD14 protein, which are not found in monkey and chimpanzee. These protein interactions are possibly due to the specific molecular or biochemical changes that occur in CD14 protein during selection pressure in different species. This interactome is important to decipher molecular and
biochemical mechanisms shaped by evolution, which may be useful for drug design and therapeutic treatment of many diseases. Several studies have shown that molecular association between chains of different protein molecules is geared by the electrostatic force like hydrophobic effects which define specific bimolecular interaction in different organism (Arkin et al., 2014; De Las and Fontanillo, 2010; Chen et al., 2013). The modulation of this interaction may be useful as putative therapeutic targets for disease treatment in many species. Ivanov et al (2013) have used the interaction of Tirobifan with glycoprotein IIb/IIIa as an inhibitor for cardiovascular drug discovery, likewise the interaction of Maraviroc and CCR5-gp120 for anti-HIV drug.

As shown earlier, there are variations in the number of the LRR domain among these species, possibly the lesser number of LRR domain in human is supplemented or accounted for by the functionality of other genes in the network (Thakur and Shankar, 2016). From our physicochemical properties, CD14 is classified as hydrophobic across the species due to higher proportion of LRR. The varying degree of LRR among these species is thought to affect the electrostatic force created by the hydrophobic effects of the protein. Published studies have shown that diverse fungal, bacterial, viral and parasite components are sensed by the mammalian LRR domain of proteins like NOD-like receptors and Toll-like receptors (Korber, 2000; Kutay and Guttinger, 2005; Lucchesse et al., 2009; Kamaraj and Purohit 2014). Likewise, about 34 leucine-rich repeat proteins have been associated with diseases in human. Obviously, divergent evolutionary events have shaped the protein-protein interaction of CD14 in different species, which is thought to be significant to varying degrees of disease susceptibility and pathogen selection.
Conclusion

We have used computational methods to gather information on CD14 protein in 14 mammals. Our *in silico* comparison of CD14 amino acid sequences among these species gave molecular evidence of a divergent evolutionary events that occurred during speciation, potentially of significance in modulating innate immune response to pathogenic challenges. Obviously, this gene has been subjected to selection pressure due to sufficient sequence variation we found from one species to another. We identified mutational hotspots with damaging effects in human and other species. In particular, the signal peptides located in these mutational hotspots are possibly of major importance in immunological studies. The variants identified in this study can be further subjected to validation through *in vitro* analysis. Since CD14 molecule is essential in initiating proper immune response to pathogens and the precursor of a robust adaptive immune response, our study highlights the effect of mutations on protein structure and disease outcome, protein-protein interaction that may be essential for drug design, yielding themselves to therapeutic manipulations for treating many diseases. Finally, these results contribute to our understanding of the evolutionary mechanism that underlie species variation in response to complex disease traits.

Acknowledgement

This work was funded by a Laboratory and Faculty Development Award, College of Health Sciences and Technology, Rochester Institute of Technology (BNT). OBM is supported through the American Association of Immunologists Careers in Immunology Fellowship Program. The
funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References


DOI:10.1016/j.chembiol.2014.09.001


Ferrero E, Hsieh C L, Francke U, Goyert SM. 1990. CD14 is a member of the family of leucine-rich proteins and is encoded by a gene syntenic with multiple receptor genes. *Journal of Immunology* 145:331-336


Ng PC, Henikoff S. 2006. Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Human Genetics. 7:61-80. DOI:10.1146/annurev.genom.7.080505.115630


Thakur R and Shankar J. 2016. *In silico* analysis revealed high-risk single nucleotide polymorphisms in human pentraxin-3 gene and their impact on innate immune response


element in KIT ligand influences cancer risk and has undergone natural selection. *Cell.*
Table 1. Percentage identity (similarity) of the CD14 protein across the mammalian species

**Identity:** Minimum: 6.61; Maximum: 99.2; Mean: 23.2603296703297; Standard deviation: 26.568543593553

**Similarity:** Minimum: 9.91; Maximum: 100; Mean: 32.5818367346939; Standard deviation: 30.99414850177
<table>
<thead>
<tr>
<th></th>
<th>Human</th>
<th>Rat</th>
<th>Mouse</th>
<th>Bovine</th>
<th>Rabbit</th>
<th>Goat</th>
<th>Monkey</th>
<th>Gorilla</th>
<th>Sheep</th>
<th>Horse</th>
<th>Pig</th>
<th>Buffalo</th>
<th>Chimp</th>
<th>Yak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>100</td>
<td>23.1 (27.9)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Identity</td>
<td>100</td>
<td>23.1 (27.9)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Similarity</td>
<td>100</td>
<td>23.1 (27.9)</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Identity:** Minimum: 6.61; Maximum: 99.2; Mean: 23.2603296703297; Standard deviation: 26.568543593553

**Similarity:** Minimum: 9.91; Maximum: 100; Mean: 32.5818367346939; Standard deviation: 30.994144850177
Table 2. Physicochemical properties of the CD14 molecule across the mammalian species
Table 2: Physicochemical properties of the CD14 molecule across the mammalian species

<table>
<thead>
<tr>
<th>Species</th>
<th>Amino acids size</th>
<th>Molecular weight (Da)</th>
<th>Isoelectric point</th>
<th>Instability index</th>
<th>Aliphatic index</th>
<th>Net charge</th>
<th>Net charge</th>
<th>GRAVY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimpanzee</td>
<td>375</td>
<td>40135.34</td>
<td>5.92</td>
<td>43.44</td>
<td>104.61</td>
<td>-4</td>
<td>0.113</td>
<td></td>
</tr>
<tr>
<td>Gorilla</td>
<td>375</td>
<td>40005.15</td>
<td>6.10</td>
<td>42.27</td>
<td>102.80</td>
<td>-3</td>
<td>0.094</td>
<td></td>
</tr>
<tr>
<td>Human</td>
<td>375</td>
<td>40076.20</td>
<td>5.84</td>
<td>42.93</td>
<td>101.76</td>
<td>-5</td>
<td>0.083</td>
<td></td>
</tr>
<tr>
<td>Monkey</td>
<td>375</td>
<td>40127.19</td>
<td>5.69</td>
<td>45.10</td>
<td>102.80</td>
<td>-6</td>
<td>0.085</td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>363</td>
<td>38450.27</td>
<td>6.19</td>
<td>33.47</td>
<td>103.06</td>
<td>-3</td>
<td>0.096</td>
<td></td>
</tr>
<tr>
<td>Mouse</td>
<td>366</td>
<td>39203.94</td>
<td>5.08</td>
<td>41.16</td>
<td>107.70</td>
<td>-9</td>
<td>0.051</td>
<td></td>
</tr>
<tr>
<td>Pig</td>
<td>373</td>
<td>39724.01</td>
<td>5.82</td>
<td>46.83</td>
<td>103.40</td>
<td>-4</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>Rabbit</td>
<td>372</td>
<td>39992.29</td>
<td>5.72</td>
<td>52.99</td>
<td>103.33</td>
<td>-5</td>
<td>0.041</td>
<td></td>
</tr>
<tr>
<td>Rat</td>
<td>372</td>
<td>40053.85</td>
<td>5.33</td>
<td>40.19</td>
<td>104.11</td>
<td>-9</td>
<td>0.033</td>
<td></td>
</tr>
<tr>
<td>Buffalo</td>
<td>373</td>
<td>39756.09</td>
<td>5.84</td>
<td>41.49</td>
<td>101.80</td>
<td>-2</td>
<td>0.099</td>
<td></td>
</tr>
<tr>
<td>Cattle</td>
<td>373</td>
<td>39666.79</td>
<td>5.37</td>
<td>41.70</td>
<td>102.06</td>
<td>-5</td>
<td>0.099</td>
<td></td>
</tr>
<tr>
<td>Goat</td>
<td>373</td>
<td>39930.28</td>
<td>8.47</td>
<td>35.07</td>
<td>99.71</td>
<td>+4</td>
<td>0.032</td>
<td></td>
</tr>
<tr>
<td>Sheep</td>
<td>371</td>
<td>39368.43</td>
<td>5.50</td>
<td>40.27</td>
<td>101.54</td>
<td>-5</td>
<td>0.087</td>
<td></td>
</tr>
<tr>
<td>Yak</td>
<td>381</td>
<td>40481.75</td>
<td>5.54</td>
<td>41.63</td>
<td>102.23</td>
<td>-4</td>
<td>0.082</td>
<td></td>
</tr>
</tbody>
</table>
Table 3

Table 3. Prediction of signal peptides and properties of the CD14 molecule in mammalian species
Table 3: Prediction of signal peptides and properties of the CD14 molecule in mammalian species

<table>
<thead>
<tr>
<th>Species</th>
<th>Amino acids size</th>
<th>Cleavage position</th>
<th>Signal site</th>
<th>Probability</th>
<th>Likelihood</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimpanzee</td>
<td>375</td>
<td>19 and 20</td>
<td>VSA-TT</td>
<td>0.9140</td>
<td>0.9991</td>
<td>0.0009</td>
</tr>
<tr>
<td>Gorilla</td>
<td>375</td>
<td>19 and 20</td>
<td>VSA-TT</td>
<td>0.9077</td>
<td>0.9991</td>
<td>0.0009</td>
</tr>
<tr>
<td>Human</td>
<td>375</td>
<td>19 and 20</td>
<td>VSA-TT</td>
<td>0.9142</td>
<td>0.9991</td>
<td>0.0009</td>
</tr>
<tr>
<td>Monkey</td>
<td>375</td>
<td>19 and 20</td>
<td>VSA-TT</td>
<td>0.9142</td>
<td>0.9991</td>
<td>0.0009</td>
</tr>
<tr>
<td>Horse</td>
<td>363</td>
<td>14 and 15</td>
<td>AAT-LE</td>
<td>0.2069</td>
<td>0.675</td>
<td>0.3250</td>
</tr>
<tr>
<td>Mouse</td>
<td>366</td>
<td>17 and 18</td>
<td>ASP-AP</td>
<td>0.4563</td>
<td>0.9991</td>
<td>0.0009</td>
</tr>
<tr>
<td>Pig</td>
<td>373</td>
<td>19 and 20</td>
<td>VSA-AT</td>
<td>0.7699</td>
<td>0.9989</td>
<td>0.0011</td>
</tr>
<tr>
<td>Rabbit</td>
<td>372</td>
<td>19 and 20</td>
<td>AST-DT</td>
<td>0.6574</td>
<td>0.9981</td>
<td>0.0019</td>
</tr>
<tr>
<td>Rat</td>
<td>372</td>
<td>17 and 18</td>
<td>VHA-SP</td>
<td>0.8795</td>
<td>0.9998</td>
<td>0.0002</td>
</tr>
<tr>
<td>Buffalo</td>
<td>373</td>
<td>20 and 21</td>
<td>VSA-DT</td>
<td>0.9712</td>
<td>0.999</td>
<td>0.0010</td>
</tr>
<tr>
<td>Cattle</td>
<td>373</td>
<td>20 and 21</td>
<td>VSA-DT</td>
<td>0.9750</td>
<td>0.9992</td>
<td>0.0008</td>
</tr>
<tr>
<td>Goat</td>
<td>373</td>
<td>20 and 21</td>
<td>VTA-DK</td>
<td>0.9642</td>
<td>0.9991</td>
<td>0.0009</td>
</tr>
<tr>
<td>Sheep</td>
<td>371</td>
<td>19 and 20</td>
<td>VSA-DT</td>
<td>0.9000</td>
<td>0.9453</td>
<td>0.0547</td>
</tr>
<tr>
<td>Yak</td>
<td>381</td>
<td>20 and 21</td>
<td>VSA-DT</td>
<td>0.9752</td>
<td>0.9993</td>
<td>0.0007</td>
</tr>
</tbody>
</table>
Table 4. Prediction of amino acid consequence at the mutational hotspot of CD14 molecules in mammalian species

Prediction (cutoff = -2.5); values above cutoff are considered deleterious; values below cutoff are considered neutral.
<table>
<thead>
<tr>
<th>Species</th>
<th>D28V</th>
<th>W45H</th>
<th>G62E</th>
<th>L70D</th>
<th>Q100G</th>
<th>V301M</th>
<th>L318I</th>
<th>G335T</th>
<th>L357H</th>
<th>G370K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Horse</td>
<td>-3.742</td>
<td>-4.914</td>
<td>-3.513</td>
<td>-3.524</td>
<td>-2.364</td>
<td>-1.896</td>
<td>-1.412</td>
<td>-0.983</td>
<td>-2.054</td>
<td>-2.067</td>
</tr>
<tr>
<td>Rabbit</td>
<td>-2.759</td>
<td>-4.293</td>
<td>-2.910</td>
<td>-4.007</td>
<td>-2.744</td>
<td>-1.969</td>
<td>-1.574</td>
<td>-0.544</td>
<td>-1.865</td>
<td>-2.451</td>
</tr>
<tr>
<td>Rat</td>
<td>-3.478</td>
<td>-4.725</td>
<td>-3.373</td>
<td>-1.058</td>
<td>-2.905</td>
<td>-2.038</td>
<td>-1.351</td>
<td>0.464</td>
<td>-2.497</td>
<td>-2.619</td>
</tr>
<tr>
<td>Sheep</td>
<td>-3.559</td>
<td>-4.952</td>
<td>-4.072</td>
<td>-3.206</td>
<td>-2.312</td>
<td>-1.981</td>
<td>-1.246</td>
<td>-1.376</td>
<td>-2.335</td>
<td>-1.695</td>
</tr>
</tbody>
</table>

1 Table 4 Prediction of amino acid consequence at the mutational hotspot of CD14 molecules in mammalian species

2 Prediction (cutoff= -2.5); values above cutoff are considered deleterious; values below cutoff are considered neutral
Figure 1

Figure 1. Multiple sequence alignment of CD14 promoter regions between mammalian species
131
132
Conensus


133
134


261
262
Conensus


Figure 2

Figure 2: Phylogenetic tree of evolutionary relationships among taxa

The evolutionary history was inferred using the Neighbor-Joining method. The optimal tree with the sum of branch length = 1.48602764. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the p-distance method and are in the units of the number of amino acid differences per site. The analysis involved 14 amino acid sequences. The coding data were translated assuming a standard genetic code table.
Figure 3: Comparison of predicted intra-domain features of CD14 protein

This comparison shows showing leucine-rich repeat (PS51450), which provide additional information about the structure and function of critical amino acids in 14 mammalian species analyzed.
Figure 4: Conserved domain LRR patterns across mammalian species
Figure 5

Figure 5: CD14 protein sequence logo displaying the most conserved domain and the positions of amino acids.

Sequence logo displaying the most conserved domain and the positions of amino acids starting from the N-terminus on the left to C-terminus to the right. The relative frequency of the amino acids is shown on the y-axis.
Fig 6. Hierarchical tree-predicted subcellular localizations of CD14 protein using neural network algorithm

6a: Hierarchical tree for all other mammalian species analyzed

6b: Hierarchical tree for sheep only
Figure 7

Figure 7. The network view of predicted associations for group of proteins with CD14

The network nodes are proteins. The edges represent the predicted functional associations. The thickness of the line indicate the degree of confidence prediction for the interaction.

Red line: fusion evidence
Green line: neighborhood evidence
Blue line: co-occurrence evidence
Purple line: experimental evidence
Yellow line: text mining evidence
Light blue line: database evidence
Black line: co-expression evidence
Figure 8

Figure 8: Venn diagram showing the proportion of intersection and unique genes depicting evolutionary diversity of CD14 molecule

8a: Comparison and visualization of protein interaction with CD14 molecule in pig, rabbit, mouse and rat

8b: Comparison and visualization of protein interaction with CD14 molecule in cattle, yak, sheep, goat and buffalo

8c: Comparison and visualization of protein interaction with CD14 molecule in human, gorilla, chimpanzee and monkey