Studying Macrophage Receptor Evolution and
Function using Bioinformatics

STUDYING MACROPHAGE RECEPTOR EVOLUTION AND FUNCTION USING BIOINFORMATICS

By Nicholas YAP, Hon B.Sc.,

A Thesis Submitted to the School of Graduate Studies in the Partial Fulfilment of the Requirements for the Degree Master of Science, M.Sc.

Abstract

During infection, macrophages are important for the uptake and clearance of microbial products, and producing inflamamtory cytokines. Receptors on the surface of the macrophage bind to bacterial components, and initiate inflammatory pathways. Although it is known that this inflammatory response is mediated through surface receptors, the regulation of this response and the evolution of these receptors remains unclear. This work uses a bioinformatic approach to study the effect of age on lipopolysaccharide (LPS) tolerized macrophages, and the evolution of the class A Scavenger Receptors (cA-SRs).

Previous work has shown that the cA-SRs are evolutionarily related to each other, however the exact nature of this relationship remains unclear because of differences in the functional domains of the proteins. Several of the cA-SRs possess a Scavenger Receptor Cysteine Rich (SRCR) domain, however its exact function and origin remains unknown. Previous studies have shown the SRCR domain is involved in ligand binding in the MAcrophage Receptor with COllagenous structure (MARCO), a member of the cA-SR family. Using recently identified extant species, we show that the SRCR domain has a common origin within the cA-SRs and that several areas within the SRCR domain are under positive selection within MARCO. In addition, recent functional data suggests that MARCO may be co-evolving with the Toll-Like Receptor 2 and CD14 Using a model of 40 non co-evolving proteins, we show that MARCO has indeed co-evolved with TLR2 and CD14. We also identified a polymorphism within the collagenous domain of MARCO that contains a human specific variant. Future experiments to confirm these bioinformatic predictions can further our understanding of macrophage biology and the innate immune system.

Acknowledgements

When I first joined the Golding lab as an undergraduate student, I was a fish out of water. I didn't have any research experience, I didn't take any bioinformatics courses, and I certainly didn't have a background in computer programming. However, despite all of this, Dr. Brian Golding took a chance on me and I am forever grateful. Over the past three years, Dr. Brian Golding has been an incredible mentor to me by teaching me everything from basic computer programming, to writing a scientific paper. Without his mentorship and compassion, I wouldn't be where I am today.

After a few months of being in the Golding lab, I was introduced to Dr. Dawn Bowdish and her lab. Being co-supervised by Dr. Brian Golding and Dr. Dawn Bowdish has been an amazing experience. Dr. Dawn Bowdish truly cares about the success of her students, and goes above and beyond for them. I am thankful for all of her guidance and support both inside and outside the lab.

I want to say thank you to several people who have helped me along the way. Thank you to Dr. Ben Evans and Dr. Jonathon Stone for their continued input and guidance throughout my time at McMaster University. I want to thank Dr. Mihaela Georgescu for her friendship and encouragement. I would also like to thank the McMaster University Biology Department and Medical Sciences Department for giving me this opportunity to be part of this great institution. Finally, I would like to thank the McMaster Biology Department and OGS for funding my research.

Looking back on my time at McMaster, the part I will miss the most is all of my labmates. For three years I have been lucky to call these amazing people my labmates, colleagues, and friends. To the members of the Golding lab, keep those computers up and runni-Segmentation fault. To the members of the Bowdish lab, keep those pipettes, mice, centrifuges, and whatever else you do going. And remember . . . #DawnSwipesLeft

Lastly, I want to thank my family for being so supportive of me. Even though my parents are non-science majors, they still listened to every one of my complaints and struggles as a graduate student. Thanks Mom, Dad and Bev! I couldn't have done it without you.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Nicholas YAP, Hon B.Sc., declare that this thesis titled, "Studying Macrophage Receptor Evolution and Function using Bioinformatics" and the work presented in it are my own. I confirm that:

- Age-Associated Changes of Macrophage Function in LPS Tolerance- RNA extractions from macrophages of mice were conducted by Keith Lee. Bioinformatic analysis conducted by Nicholas Yap. Written by Nicholas Yap, Dr. Dawn Bowdish, and Dr. Brian Golding.

- Co-evolution of MARCO with TLR2 and CD14 and Prediction of Functional Sites - All experiments were conducted by Nicholas Yap and Sohidull. Written by Nicholas Yap, Dr. Dawn Bowdish, and Dr. Brian Golding.

- The Evolution of the Scavenger Receptor Cysteine-Rich Domain of the Class A Scavenger Receptors- All experiments were conducted by Nicholas Yap. Written by Nicholas Yap, Fiona Whelan, Dr. Dawn Bowdish, and Dr. Brian Golding.

Signed:

Date: June 21, 2016

Chapter 1

Introduction

1.1 Introduction

1.1.1 Innate Immunity

The innate immune system is an ancient form of immunity present within both vertebrates and lower organisms. It serves as a fast acting defence mechanism against invading pathogens, by recognizing conserved structures (Murphy, 2011). These conserved structures have been coined Pathogen Associated Molecular Patterns or (PAMPs) and are often microbial components such as the lipopolysaccharides (LPS) or the peptidoglycan cell wall component (Murphy, 2011).

In order to recognize these PAMPs, immune cells express different Pattern Recognition Receptors (PRRs). These receptors are constituitively expressed on the surface of adaptive cells and structural cells to infection such as macrophages, endothelial cells, and Natural Killer (NK) cells (Akira, Uematsu, and Takeuchi, 2006). The recognition of PAMPs by PRRs induces inflammatory cytokines including IL-1, tumor necrosis factor (TNF)-α, and IL-6. These cytokines play a role in regulating cell death, endothelial permeability, and leukocyte recruitment (Akira, Uematsu, and Takeuchi, 2006; Takeuchi and Akira, 2010).

Vertebrates have evolved an additional component of immunity known as adaptive immunity. The adaptive immune response can activate innate immune cells for pathogen destruction or tissue repair and can also recruit additional cells to the site of infection. Unlike the innate immune system, the adaptive immune system consists of T and B cells with antigen specific receptors (Murphy, 2011; Akira, Uematsu, and Takeuchi, 2006). The specificity of these receptors is generated through the Recombination Activating Gene (RAG)-1 and RAG-2 proteins that mediate the alternate splicing of gene segments (Agrawal, Eastman, and Schatz, 1998). The innate and adaptive immune system are linked through professional Antigen Presenting Cells (APCs). Nearly all nucleated cells are able to express small peptides called antigens on their surface,

through the Major Histocompatibility Complex (MHC) (Murphy, 2011). However, only dendritic cells, B cells, and macrophages are able to express co-stimulatory molecules alongside MHC, and are therefore called professional APCs (Murphy, 2011). When PRRs on the surface of professional APCs bind their ligand, they express co-stimulatory molecules alongside the antigen to activate T and B cells.

1.1.2 The Macrophage

Macrophages are a group of innate immune cells known for their plasticity, and versatility. They play a role in various biological processes including tissue remodelling and repair, metabolic function, and phagocytosis (Sica and Mantovani, 2012). Macrophages are highly responsive to their micro-environment and adjust their phenotype and function based on levels of cytokines or bacterial products. Depending on the micro-environment, the function of the macrophages can be skewed towards the classically activated (M1) or alternatively activated (M2) phenotypes. In the presence of interferon (IFN)-γ or microbial products, macrophages develop into classically activated or M1 macrophages (Martinez et al., 2007). These M1 macrophages produce pro-inflammatory cytokines such as IL-1β, IL-15, IL-18, TNF-α, and IL-12 (Martinez et al., 2007). M1 macrophages also increase T cell and B cell activity through increased surface expression of co-stimulatory molecules, which further potentiates the inflammatory response and plays a major role in bacterial killing (Mantovani et al., 2004; Martinez et al., 2007). Conversely, M2 macrophages develop in the presence of IL-4 or IL-13 and are associated with tissue repair and regulating the inflammatory response (Mantovani et al., 2002; Martinez et al., 2007).

1.1.3 Macrophage Surface Receptors

In order to mediate phagocytosis and inflammatory signalling, macrophages possess various receptors on their cell surface. The Scavenger Receptors (SRs) are a group of receptors that bind to various anionic ligands (Gordon, 2002). The SRs consist of multiple classes from A to J, where each class contains its own structure and target ligands. The Class A Scavenger Receptors (cA-SRs) are one class of SRs, and have been shown to be evolutionarily related (Whelan et al., 2012). The cA-SRs are often described as promiscious receptors for their ability to bind multiple ligands such as modified forms of low-density lipoprotein (LDL) as well as various bacterial ligands from *Streptococcus pneumoniae*, and *Mycobacterium tuberculosis* (Goldstein et al., 1979; Arredouani et al., 2004; Bowdish et al., 2009; Whelan et al., 2012; Dorrington and Bowdish, 2013).

Although the cA-SRs are important phagocytic receptors, macrophages require additional surface receptors for mediating an inflammatory response. The Toll-Like Receptors (TLRs) reside on the surface of various APCs, including macrophages, and are important in downstream signalling of NF-κB, a transcription factor involved in the

inflammatory response (Gordon, 2002). CD14 is another receptor preferentially found on the surface of macrophages, and plays a role in binding bacterial ligands such as LPS and lipoteichoic acid (LTA) (Takeda, Kaisho, and Akira, 2003; Nilsen et al., 2008). Compared with TLR2 expressing cells, cells expressing both TLR2 and CD14 have an enhanced response to LTA, a cell wall component of gram-positive bacteria (Nilsen et al., 2008).

Recently, the cA-SRs have been shown to interact with TLR2 and CD14. The MAcrophage Receptor with COllagenous structure (MARCO) is a member of the cA-SR family, and has been shown to enhance the inflammatory response of TLR2 and CD14 to *S. pneumonaie* (Dorrington and Bowdish, 2013). Furthermore this was specific to MARCO, as the Class A Macrophage Scavenger Receptor (SR-A), another member of the cA-SR family, was not shown to enhance this response (Dorrington and Bowdish, 2013). Further investigation on the relationship between these three proteins will further our understanding of the underlying mechanisms between phagocytosis and inflammatory signalling.

In addition, MARCO's ability to enhance the TLR2/CD14 response appears to be mediated by one of its domains known as the Scavenger Receptor Cysteine Rich (SRCR) domain. Several studies have shown that MARCO relies on its SRCR for function and ligand binding. Novakowski et al. (2016) have shown that a variant of MARCO, that lacks the SRCR domain, is unable to cause the enhanced TLR2/CD14 response. Brännström et al. (2002) have also shown that the SRCR domain of MARCO contains a conserved RGR motif that is required for liganding binding. However, although it has been shown that the cA-SRs are evolutionarily related, the origin of the SRCR domain within these proteins remains unclear (Whelan et al., 2012). For example, although MARCO requires its SRCR domain to bind ligands, SR-A contains an SRCR domain, but binds ligands using its collagenous domain (Krieger, 1992). Future investigations on the origin of the SRCR domain are important for understanding its role in ligand binding and signalling within the cA-SRs.

1.1.4   LPS Tolerance in the Elderly

LPS is an outer membrane component of bacteria and a potent inflammatory stimulus. During infection, macrophages phagocytose and clear bacterial components, such as LPS, and induce pro-inflammatory signalling pathways to protect the host. Although this inflammatory response is meant to be protective, the resulting inflammation can be deleterious and may result in toxic shock and sepsis (Greisman and Hornick, 1975). Cells exposed to long term sub-lethal doses of LPS have a protective phenotype known as LPS tolerance or endotoxin tolerance. During this "tolerized" phase, cells are hypo-inflammatory and have a reduced responses to subsequent LPS

stimulation (Fan and Cook, 2004).

Foster, Hargreaves, and Medzhitov (2007) have shown that LPS tolerance is a transcriptional response that is established through nucleosome remodelling and histone modifications. Using a model of LPS tolerance, they show that genes involved in the response have two categories: inducible or non-inducible during secondary exposure (Foster, Hargreaves, and Medzhitov, 2007).. By tolerizing cells, inflammatory genes become non-inducible and produce the typical LPS tolerance response. However, the tolerance response appears to be impaired with age. Previous work has shown that macrophage inflammatory signalling (Stout and Suttles, 2005; Boyd et al., 2012) and ligand binding (Gomez, Boehmer, and Kovacs, 2005) become dysfunctional within elderly individuals. Furthermore, the inflammatory response to LPS is also impaired, as macrophages from elderly individuals and mice produce increased TNF-α and IL-6 when stimulated with LPS (Puchta et al., 2016). Although these studies have characterized the functional differences in elderly individuals, it remains unclear whether the mechanisms involved in establishing tolerance are impaired with age.

1.1.5 Research aims

Bioinformatics is a powerful interdisciplinary tool that combines biology with computational analyses and statistics. Scientists often use it to study large data sets such as evolutionary relationships between proteins, as well as association studies between genetic traits and diseases. Due to its wide range of applications, bioinformatics can be used to further our understanding of macrophage biology. Using this tool, I studied the effects of age on LPS tolerance and attempted to identify genes differentially expressed between old and young macrophages. In addition, my study also focused on the evolution of macrophage surface receptors. Due to the enhanced response of cells expressing MARCO, TLR2, and CD14 to bacterial products, I tested if these three proteins are co-evolving. Furthermore, I investigated the evolution of the SRCR domain within MARCO and its origin within the cA-SR family as a whole. Future chapters will investigate these specific hypotheses.

1.1.6 Specific hypotheses

- Macrophages from old mice have an impaired tolerance response when compared to their young counterparts.

- MARCO has co-evolved with TLR2 and CD14. In addition, the SRCR domain of MARCO contains functional residues important for mediating the enhanced response.

- MARCO shares a closer evolutionary relationship with the SRCR containing cA-SRs and futhermore, the SRCR domain of these proteins has one common origin.

Chapter 2

Age-Associated Changes of Macrophage Function in LPS Tolerance

2.1 Introduction

Macrophages are a class of cells that are derived from either monocytes or embryonic stem cells (Geissmann et al., 2010). Their phagocytic properties were originally identified by Ilya Metchnikoff (1905). He proposed that phagocytes originally evolved as physiological regulators of homeostasis that clear the body of "disharmony" (Tauber, 2003). By studying the process of metamorphosis in tadpoles, Metchnikoff further postulated that phagocytes play a role in an organism's structural and tissue development (Tauber, 2003). These early findings laid the ground work for research in macrophage biology. It is now known that macrophage functions include lipid uptake, and bacterial killing and clearance (Wynn, Chawla, and Pollard, 2013; Motoyoshi, 1998; Babamusta et al., 2006; Linton et al., 1999), and tissue homeostasis (Wynn, Chawla, and Pollard, 2013).

Lipopolysaccharide (LPS) is a major cell wall component of Gram-negative bacteria and an inducer of pro-inflammatory signalling (Fan and Cook, 2004; Greaves and Gordon, 2005). In 1946, it was first observed that a primary low dose of LPS leads to a subdued response to a second, high dose of LPS (Beeson, 1946). This phenomenon has since then been called LPS tolerance or endotoxin tolerance, and is defined as a hypo-inflammatory state with reduced responses to subsequent LPS stimulation following an initial treatment (Fan and Cook, 2004). LPS tolerance is protective in toxic shock and sepsis because it increases survival during prolonged exposure (Greisman and Hornick, 1975).

Although researchers have studied LPS tolerance for over 50 years, the molecular mechanisms behind LPS tolerance have yet to be fully elucidated. During LPS

stimulation, CD14 binds to LPS and mediates the interaction between LPS and TLR4 (Chow et al., 1999). TLR4 then triggers intracellular signalling pathways, via the adapter proteins MyD88 and IRAK, which causes the activation of NF-κB and AP-1 (Fan and Cook, 2004). This results in the production and release of cytokines such as TNF-α, IL-6, and IL-12 as well as reactive oxygen and nitrogen species (Nomura et al., 2000; Sweet and Hume, 1996; Wynn, Chawla, and Pollard, 2013). However, this response must be tightly controlled and regulated since many of the byproducts of the inflammatory response can be detrimental to the health and survival of the host. One mechanism of modulating this response is through regulating gene expression at the transcriptional level. Foster, Hargreaves, and Medzhitov, 2007 divided genes involved in tolerance into tolerizeable (genes not inducible or re-inducible in tolerant macrophages) and non-tolerizeable (inducible genes in tolerant macrophages). The tolerizeable genes generally correspond to genes involved in the inflammatory response, while non-tolerizeable genes are generally anti-microbials. Foster, Hargreaves, and Medzhitov, 2007 showed that nucleosome remodelling and histone modifications altered the expression of tolerizeable and non-tolerizeable genes in tolerant macrophages. Within tolerance induced macrophages, gene products from the initial LPS stimulation were able to negatively regulate tolerizeable genes, and conversely prime non-tolerizeable genes (Foster, Hargreaves, and Medzhitov, 2007). These data suggest that the underlying differences between tolerance and induction are due to changes at the level of transcription.

Although age has been shown to effect macrophage function, it is unclear how age effects the macrophage's ability to induce tolerance. Previous work has shown that macrophage inflammatory signalling (Stout and Suttles, 2005; Boyd et al., 2012) and ligand binding (Gomez, Boehmer, and Kovacs, 2005) become dysfunctional within elderly individuals. As individuals age, both the innate and adaptive immune system decline in function through a process called immunosenescence, leading to higher levels of pro-inflammatory cytokines within the circulation (Franceschi et al., 2000; Giefing-Kröll et al., 2015; Montgomery and Shaw, 2015; Puchta et al., 2016). Within both elderly mice and humans, researchers have found impaired responses to IFN-γ, as well as decreased tumor lysis, IL-1β production, and reactive oxygen species production (Castle, 2000). When comparing the effects of LPS stimulation between young and old individuals, monocytes from the elderly have increased TNF-α and IL-6 (Puchta et al., 2016), as well as decreased IL-β (Bruunsgaard et al., 1999). Counterintuitively, these age-associated changes likely contribute to the increased susceptibility to infections such as influenza (McElhaney et al., 2012; Thompson et al., 2004) and pneumonia (Puchta et al., 2016), as well as increased rates of hospitalization due to infectious disease (Jackson et al., 2003).

Taken together with the differences in tolerance at the transcriptional level, we propose that elderly individuals have impaired tolerance responses due to differences in

gene expression, and that these differences in gene expression cause the increased levels of pro-inflammatory cytokines within the elderly. Using an experimental protocol based on Foster, Hargreaves, and Medzhitov, 2007, we compared the differentially expressed genes between macrophages from young and old mice exposed to LPS. We found no differences in the tolerance response between macrophages from young and old mice, however macrophages from old mice have an impaired induction response to LPS. These data suggest that the increased inflammation within the elderly may be due to the inability to mount a robust inflammatory response to bacterial infections.

2.2   Materials and Methods

2.2.1   Bone Marrow Macrophage Isolation and Differentiation

Bone marrow derived progenitor cells were isolated as in Fei et al. (2016). Briefly, bone marrow was collected from the femurs of three old (18-24 months) and three young (6-8 weeks) C57BL/6 mice. Progenitor cells were cultured in 150mm Fisherbrand petri dishes with 25 mL of Roswell Park Memorial Institute 1640 medium (RPMI-1640), 1% penicillin-streptomycin, 1% L-glutamine, 10% Fetal Bovine Serum (FBS), and 15% L929-cell conditioned medium (LCM). Cells were incubated at 4° C for 7 days and the media changed every 2-3 days. On day 8, mature macrophages were lifted by washing with 15 mL of 4° C phosphate-buffered saline, refrigerating at 4° C for 10 minutes, and manually lifted using a disposable cell lifter.

2.2.2   Endotoxin Stimulation Experiments

To induce LPS tolerance, we used a previously defined protocol Foster, Hargreaves, and Medzhitov, 2007, Briefly, macrophages were stimulated with 10 ng/ μL of LPS or vehicle control (PBS) for 16 hrs, and further stimulated with 100 ng/ μL of LPS for 4 hrs. A control group was stimulated with LPS for 4 hrs. LPS was derived from *Escherichia coli* serotype. Cells were seeded at a concentration of $1x10^6$ cells per well with 3 mL of fresh media. Supernatants were collected and stored at -80° C in Trizol until RNA isolation.

TABLE 2.1: **Treatment Groups of Old and Young Macrophages**

Treatment group	First Exposure (16 hrs)	Second Exposure (4hrs)
LPS tolerance	LPS	LPS
LPS induction	LPS	Vehicle Control
LPS control	Vehicle Control	LPS
Control	Vehicle Control	Vehicle Control

2.2.3 RNA Purification and Sequencing

RNA was isolated for each of the treatment groups as in Fei et al. (2016). RNA was purified using the Qiagen RNA Isolation Kit (RNeasy) and eluted using nuclease free water. The concentration and purity of the samples was tested using NanoVue Spectrophotometer (GE Healthcare) and integrity was measured using the Aligent 2100 Bioanalyzer and the Aligent RNA 6000 Nano Kit. Only samples with RNA Integrity Number (RIN) above 8 were selected. RNA samples were treated using the Human/Mouse/Rat RiboZero Magnetic Kit (Epicentre) and purified using RNAClean XP beads (Agencourt) to remove rRNA. The rRNA-depleted samples were then treated with Turbo DNase (Invitrogen) and purified once using RNAClean XP beads. First strand synthesis was conducted using Superscript III (Invitrogen), while complimentary cDNA was synthesized using RNase H and Klenow fragment of DNA polymerase I (Invitrogen). The generated cDNA was sonicated into 150 base pair fragments using a Covaris S220 Focused-ultrasonicator and deoxyadenosine monophosphate was incorporated into the cDNA fragment using NEBNext dA-Tailing Module (New England Biolabs). The cDNA library was sequenced using the Illumina HiSeq system.

2.2.4 RNAseq Analysis

Single-end reads were obtained for each treatment group using the Illumina HiSeq system. Quality control was performed using Trimmomatic to remove adapter sequences and reads shorter than 75 base pairs (Bolger, Lohse, and Usadel, 2014). FastQC plots were generated to determine the quality of reads and ensure adapters were removed (Andrews et al., 2010). In total, 48 samples were collected, with an average of 1.8 million reads per sample with approximately 97% of reads kept after quality trimming. Reads were then mapped to the *Mus musculus* transcriptome, generated from the NCBI database (Pruitt, Tatusova, and Maglott, 2007) and gene annotation from Ensembl (Flicek et al., 2012) using STAR (Dobin et al., 2013). Mapped reads were counted using HTSeq (Anders, Pyl, and Huber, 2014) and imported into R for analysis using the DESeq2 package (Love, Huber, and Anders, 2014). Using the dataset published in Pena et

al. (2011), we identified genes involved in endotoxin tolerance specifically. Genes were selected as being differentially expressed based on an adjusted p-value. Paired T-tests were done to compare the differences between old and young macrophages. Differentially expressed genes were imported into the DAVID bioinformatic database to identify pathways and Gene Ontology (GO) terms (Huang, Sherman, and Lempicki, 2009). Scripts can be found at `https://github.com/inickyap/Bio720/tree/master`.
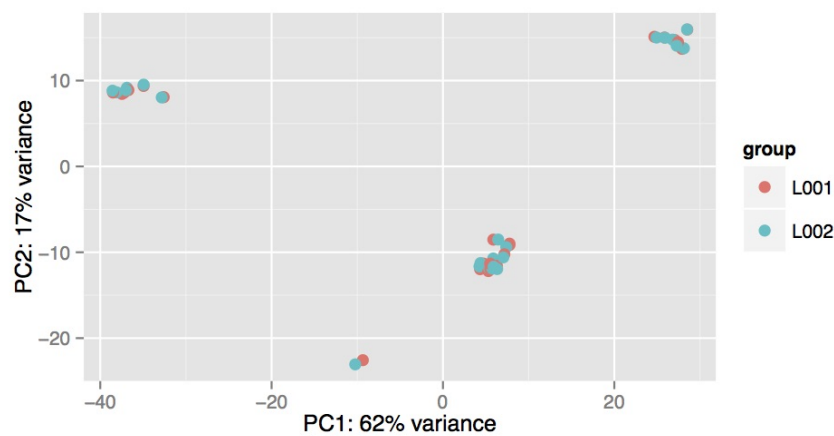
FIGURE 2.1: **PCA Plot of Reads Grouped by Lane.** The 48 samples were run on two independent lanes on the Illumina HiSeq system. In order to test for possible lane effect, we performed a PCA analysis. Red circles correspond to reads ran on lane 1 (L001) and blue circles correspond to reads ran on lane 2 (L002). We observe that there was no significant differences based on Illumina sequencing lanes.

2.3 Results

2.3.1 Differential Expression of Genes Involved in LPS Response

To begin our analysis, we tested for any possible biases in the Illumina sequencing due to lane using DESeq2. Using a PCA analysis, we show that the data does not appear to cluster by lane and observe no biases due to Illumina sequencing (Figure 2.1). Foster, Hargreaves, and Medzhitov, 2007 previously divided genes involved in LPS tolerance into the categories of tolerizeable and non-tolerizeable, which correspond to non-inducible genes and inducible genes in LPS tolerant macrophages respectively (Figure 2.2). Based on the dataset from Foster, Hargreaves, and Medzhitov (2007), we investigated several known tolerizeable and non-tolerizeable genes as controls for our experiment. We then compared the effects of LPS induction and LPS tolerance within

macrophages from young and old mice.

(a)

(b)
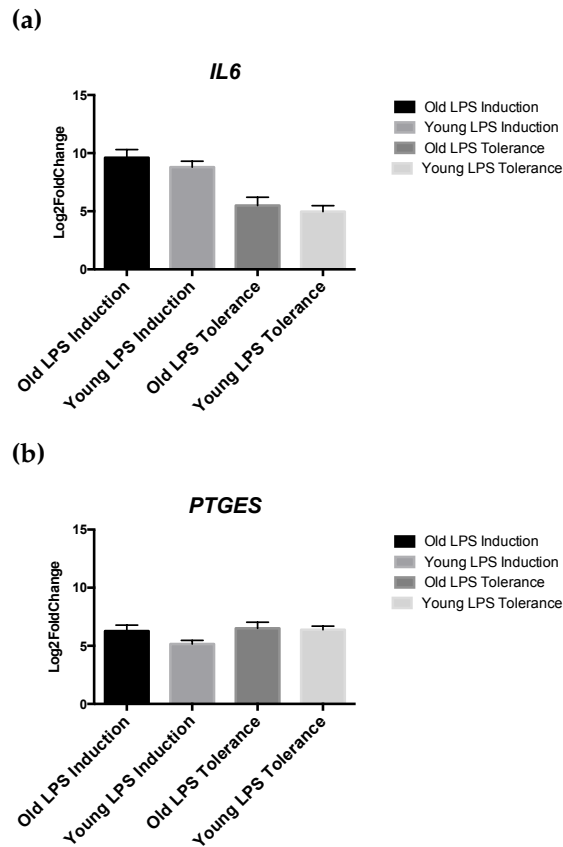
FIGURE 2.2: **Example profile of tolerizeable genes versus non-tolerizeable genes within macrophages from young and old mice.** Within tolerized macrophages, tolerizeable genes such as *IL6* show a decrease in expression or lower log2fold change. Conversely, non-tolerizeable genes such as *PTGES*, show no difference in expression between tolerant and non-tolerant macrophages. Standard error estimate from triplicate values is shown.

First, we studied the effect of LPS induction on macrophages by comparing differentially expressed genes between LPS stimulated macrophages with unstimulated controls. We identified 187 differentially expressed genes specific to young macrophages, 18 genes specific to aged macrophages, and 208 differentially expressed genes common to both groups (Figure 2.3b). For both the tolerizeable and non-tolerizeable genes, we observe an increase in expression after 16 hrs of LPS stimulation in macrophages from young and old mice (Figure 2.4 & 2.5). Macrophages from young mice had higher levels of several tolerizeable genes including *MMP13, Serpine1, IL6, HDC, IL1β,* and *Lipg* compared with their old counterparts. However, *MARCO* was higher in macrophages

from old mice compared to young mice. In the non-tolerizeable genes, *FPR1, PTGES,* and *LCN2* showed no difference with age, while *SAA3* was higher within macrophages from young mice. Taken together at a global level and within specific tolerance related genes, these data suggest that the response to LPS induction is different between the two age groups.

(a) LPS tolerance
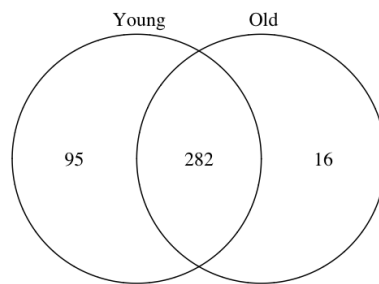
(b) LPS induction
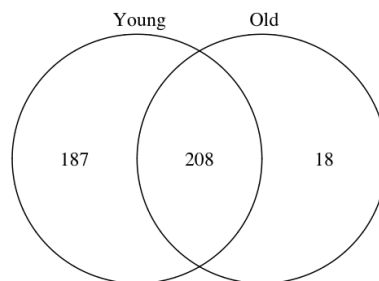
FIGURE 2.3: **Venn diagram of differentially expressed genes in LPS tolerance and LPS induction.**
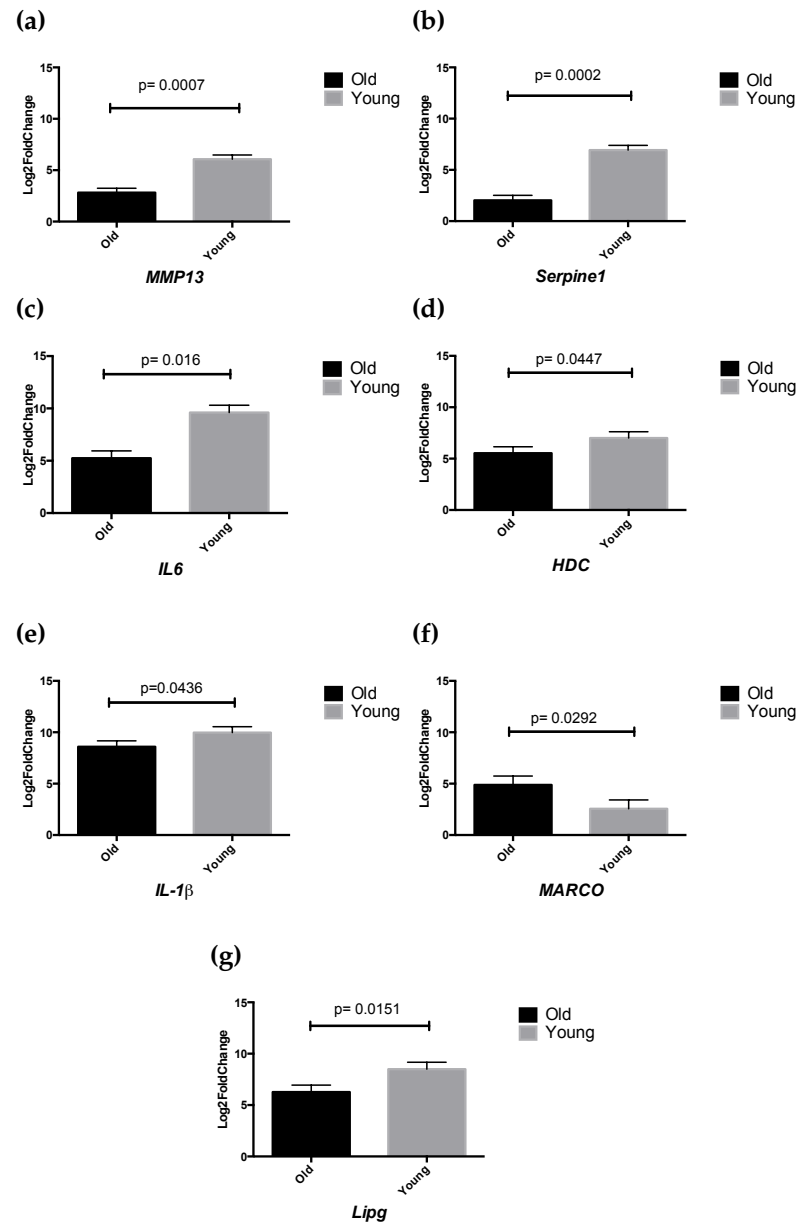
FIGURE 2.4: **Macrophages from young mice have a higher induction of tolerzieable genes than macrophages from old mice.** Log2FoldChange of several genes involved in the LPS response are shown. Standard error estimate from triplicate values is shown.
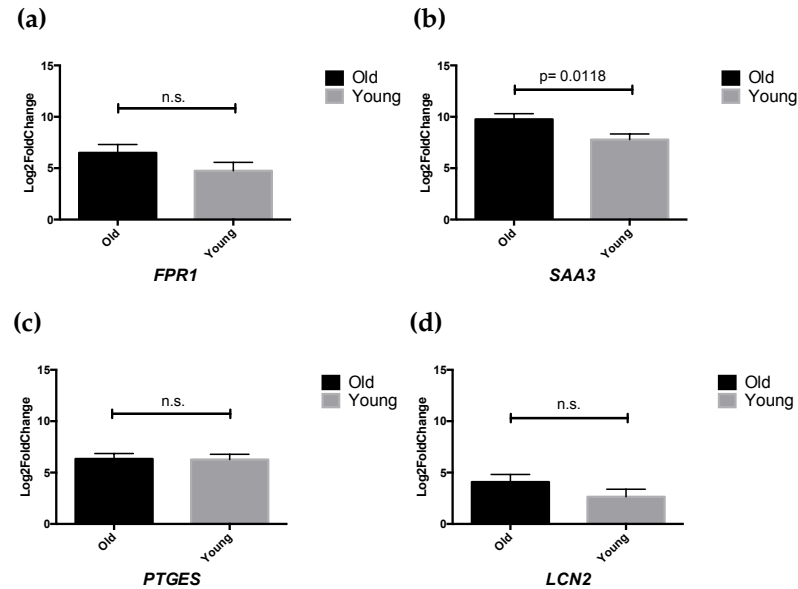
FIGURE 2.5: **Log2FoldChange of Non-tolerizeable Differentialy Expressed Genes in Young and Old during LPS induction.** Standard error estimate from triplicate values is shown.

Next, we studied the effect of LPS tolerance on gene expression at a global level. We identified 95 differentially expressed genes specific to young macrophages, 18 specific to old macrophages, and 282 differentially expressed genes shared between the two groups (Figure 2.3a. See also Table B.1 & B.2). Of the 95 genes specific to young macrophages, 68 were down-regulated in tolerance including Rel, CXCL1, and TRAF1 (Table B.1). In addition, young macrophages from young mice also showed an increase in expression of IRAK3, SMAD6, and TGF-β inducer, which are associated with anti-inflammatory responses.

We then identifed differentially expressed genes between tolerized macrophages during LPS stimulation and non-tolerant macrophages exposed to LPS. Based on the tolerizeable and non-tolerizeable gene profile from Foster, Hargreaves, and Medzhitov (2007), we show several tolerizeable genes are down-regulated in tolerized macrophages including *IL6, Serpine1, HDC, IL-1β*, and *MMP13* (Figure 2.4). Conversely, we observe that several non-tolerizeable genes such as *FPR1, LCN2, and PTGES* show an increase in expression in tolerant macrophages (Figure 2.5). When comparing the differences between macrophages from young and old mice, we found no difference in any of the tolerizeable genes between the two age groups (Figure 2.6a - 2.6e). Within the non-tolerizeable genes, we found no difference between the age groups as well (Figure 2.7a - 2.7d). Based on these findings, we were unable to detect differences between the age groups.
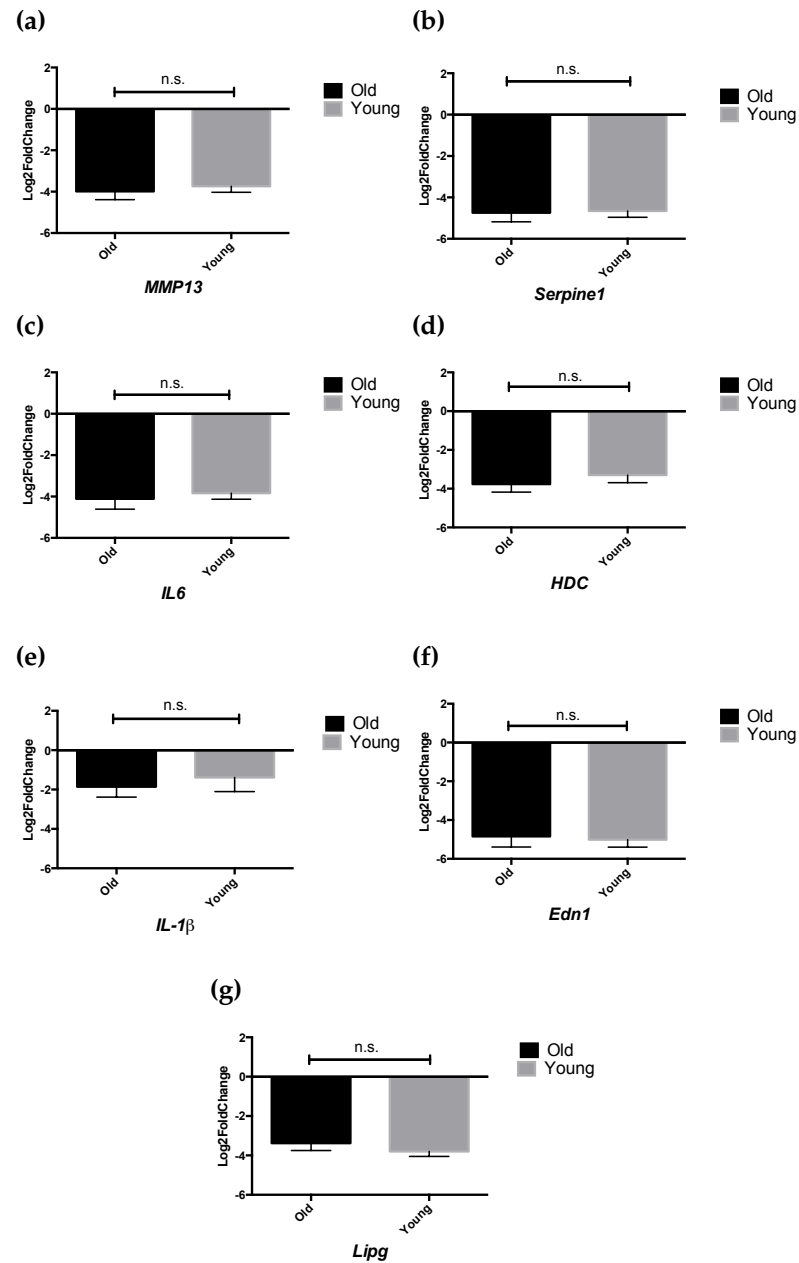
FIGURE 2.6: **Macrophages from young and old tolerized mice show no difference in gene expression levels.** All seven genes show a consistent decrease in expression in both young and old mice. Standard error estimate from triplicate values is shown.

2.3.2 Gene Ontology and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Analysis

To determine GO terms and KEGG pathways associated with all differentially expressed genes between macrophages from young and old mice, we used the DAVID bioinformatic database. In LPS tolerant macrophages from young and old mice, we found genes involved in response to external stimulus, immune system processes, as well as others were down-regulated. Interestingly, we found that several processes involved in metabolism were up-regulated including monosaccharide metabolic process, cellular nitrogen compound metabolic process, alcohol metabolic process, as well as others (Figure 2.8 & 2.9). When comparing the biological processes of differentially expressed genes in LPS induction we found only up-regulated biological processes for macrophages from both young and old mice (Figure 2.10). We found both age groups had similar up-regulated biological processes including immune response, positive regulation of leukocyte activation, immune system processes, as well as others.
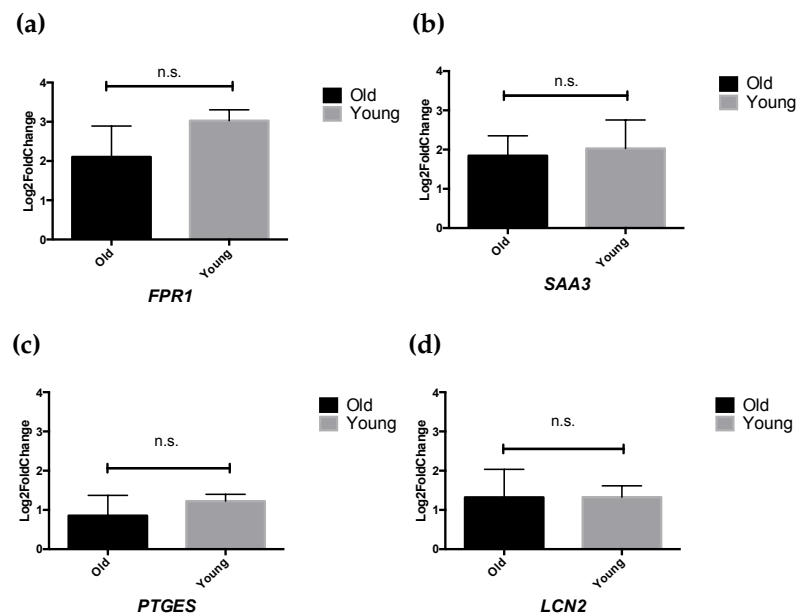
FIGURE 2.7: **Log2FoldChange of non-tolerizeable in young and old tolerized macrophages.** All four genes were up-regulated in macrophages from young and old mice, however there were no differences between age groups. Standard error estimate from triplicate values is shown.

(a)

(b)
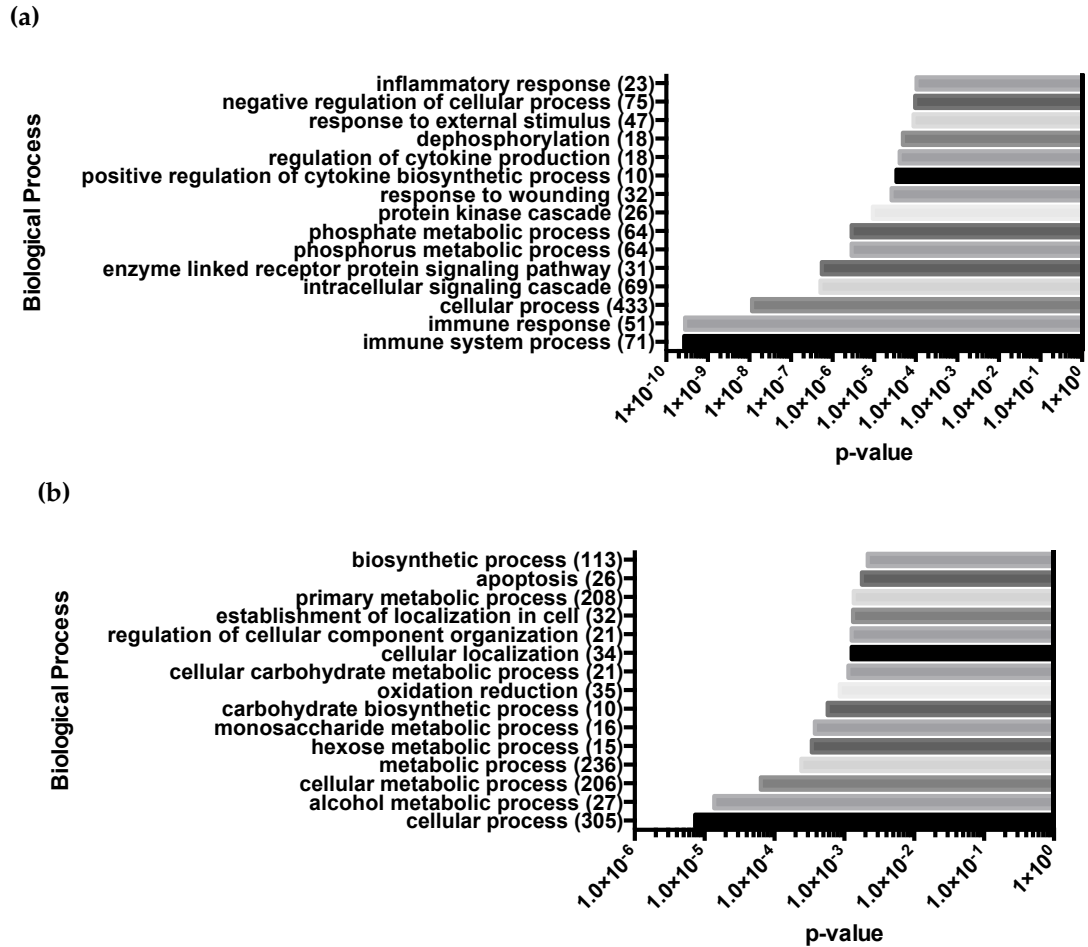
FIGURE 2.8: **Biological processes of differentially expressed genes in LPS tolerance in young macrophages.** Top fifteen biological processes by p-value are shown. Biological processes within LPS tolerized macrophages that are up-regulated are shown in (a) and processes down-regulated are shown in (b). Number of differentially expressed genes associated with each biological process are shown in brackets.

(a)

(b)
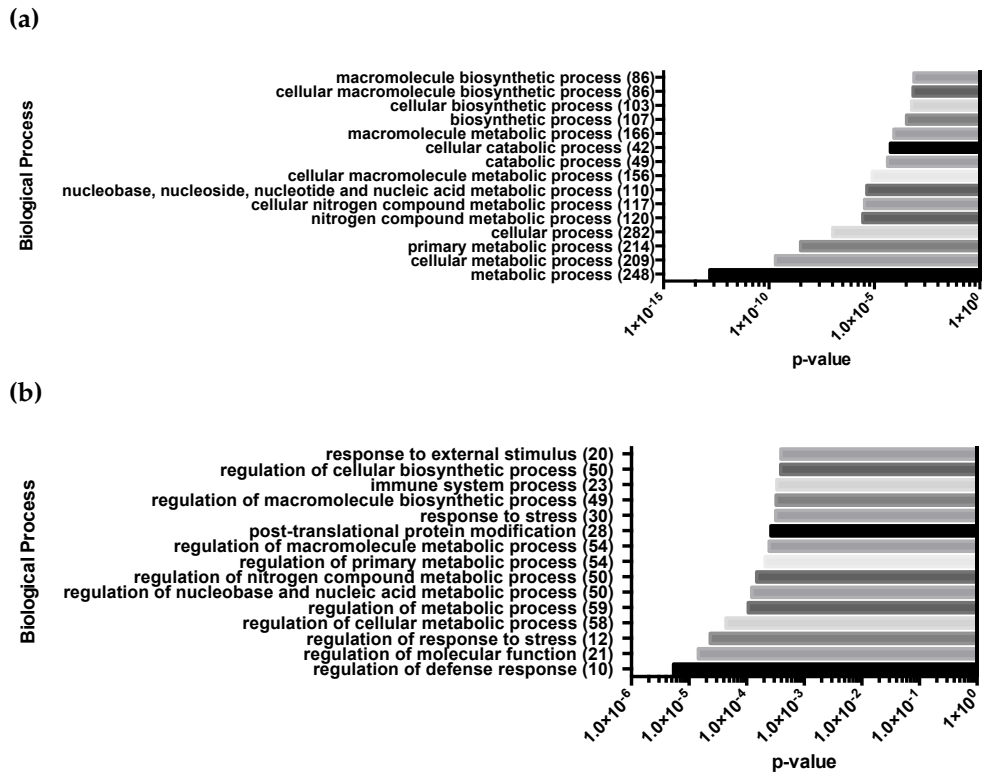
FIGURE 2.9: **Biological processes of differentially expressed genes in LPS tolerance in old macrophages.** Top fifteen biological processes by p-value are shown. Biological processes within LPS tolerized macrophages that are up-regulated are shown in (a) and processes down-regulated are shown in (b). Number of differentially expressed genes associated with each biological process are shown in brackets.

17

(a)

(b)
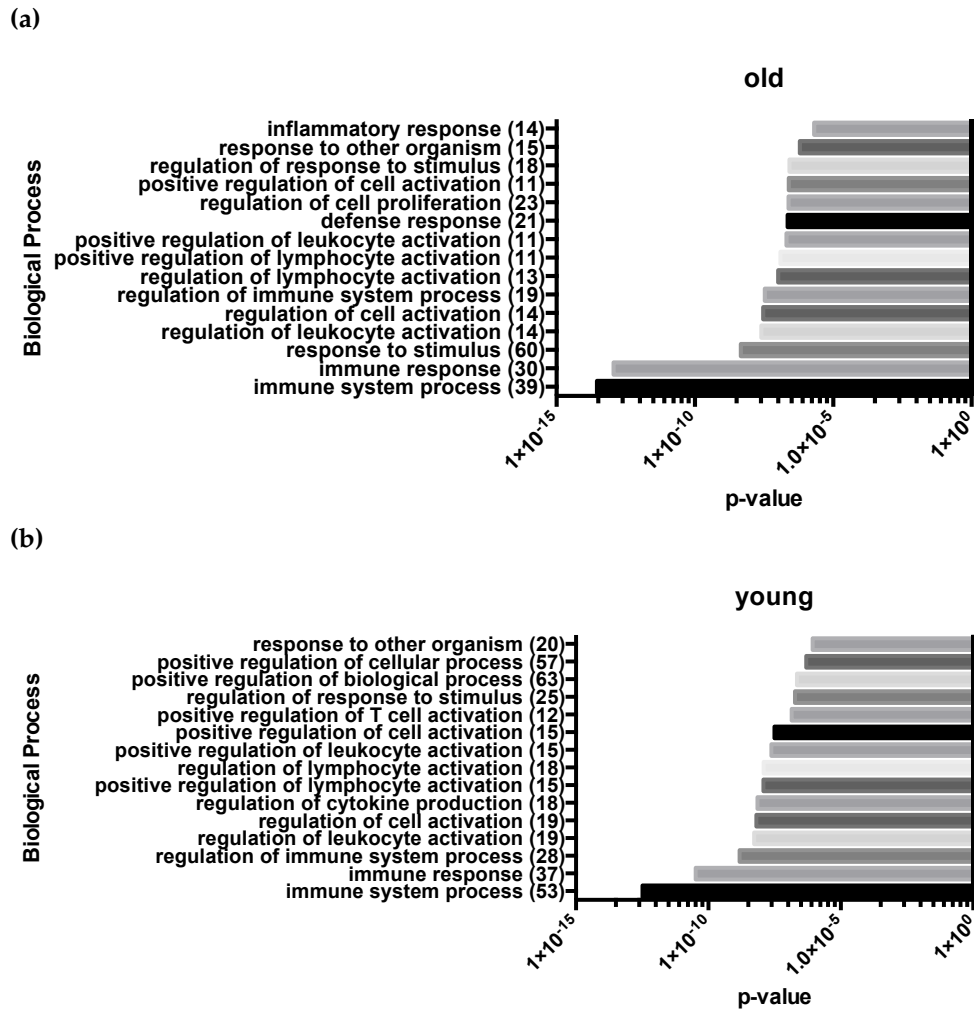
FIGURE 2.10: **Biological processes of differentially up-regulated genes in LPS induction** Top fifteen biological processes by p-value are shown. Biological processes of differentially expressed genes after LPS induction in macrophages from old mice are shown in (a) and genes from macrophages of young mice are shown in (b). Number of differentially expressed genes associated with each biological process are shown in brackets.

2.4 Discussion

Macrophages are important for mounting a robust inflammatory response to infections, however this response must be regulated in order to prevent cytotoxicity and septic shock. The ability for cells to induce a tolerance response, or a reduced response to subsequent stimuli, has been characterized over the past several years. Foster,

Hargreaves, and Medzhitov (2007) has shown that several genes are differentially expressed within tolerance due to histone modifications. They classified genes that are non-inducible in tolerance as tolerizeable and genes that remain inducible as non-tolerizeable. However, due to the decline of innate immune response with age, we chose to study the effect of age on LPS induction and LPS tolerance.

We compared the differences in gene expression during the LPS response using macrophages from young and old mice. Our data suggest that the LPS tolerance response is not impaired with age because there was no difference between young and old mice at the transcriptional level. However, here we show that age plays a role in the induction to LPS. We identified 187 genes that were differentially expressed in young macrophages in LPS induction but were absent from macrophages of old mice. In addition, macrophages from young mice had significantly higher levels of several tolerizeable and non-tolerizeable genes when compared with their old counterparts.

We also identified biological processes associated with differentially expressed genes. We found several metabolic processes were up-regulated in tolerized macrophages from both age groups, while various immune related processes were down-regulated. Previous findings have shown that during LPS stimulation, macrophages rely heavily on glycolysis (Rodríguez-Prados et al., 2010). Fei et al. (2016) have shown that macrophages from old mice have an impaired ability to switch to glycolysis during LPS stimulation, and conclude that old macrophages have a delayed metabolic response. Our data suggests that there are differences in the metabolic response in LPS tolerance between the two age groups, however we did not find differences in the metabolic response in LPS stimulation.

Chapter 3

Co-evolution of MARCO with TLR2 and CD14 and Prediction of Functional Sites

3.1 Introduction

The innate immune system uses many cell surface receptors to bind and internalize pathogens in order to protect the host. These receptors have overlapping target ligands, and initiate phagocytosis. Many of the microbial ligands of these receptors have conserved structures or hydrophobicity. These common features of microbial ligands are defined as Pathogen-Associated Molecular Patterns (PAMPs), while the receptors that recognize patterns are called Pattern Recognition Receptors (PRRs).

The MAcrophage Receptor with COllagenous structure (MARCO), is a PRR commonly found on the surface of macrophages. MARCO contains several protein domains including an expansive collagenous domain, and a c-terminal Scavenger Receptor Cysteine-Rich (SRCR) domain. Like other cA-SRs, MARCO is a promiscuous receptor, binding modified forms of low-density lipoprotein (LDL) (Chen et al., 2006) as well as the bacterial cell wall component LPS (Sankala et al., 2002). In addition, MARCO plays an important role in bacterial clearance as MARCO-deficient mice have increased mortality in pneumococcal infections, due to excessive inflammation within the lungs (Arredouani et al., 2004).

When comparing MARCO to other members of the cA-SR family there are several functional differences. The Class A Macrophage Scavenger Receptor (SR-A), also known as Macrophage Scavenger Receptor 1 (MSR1), is another member of the cA-SR family and has been shown to be evolutionarily related to MARCO (Whelan et al., 2012; Yap et al., 2015). The two proteins share a similar protein structure with both proteins containing a collagenous domain and an SRCR domain (Figure 3.1). However, unlike MARCO, SR-A's ligand binding region may reside within its collagenous

domain. Experiments of SR-A have shown its isoform SR-AII possesses a truncated SRCR domain, but still binds bacterial ligands with through the use of its collagenous domain (Krieger, 1992). This is surprising since the two proteins have been shown to be evolutionarily related, and share similar motifs within their SRCR domains (Whelan et al., 2012; Yap et al., 2015).
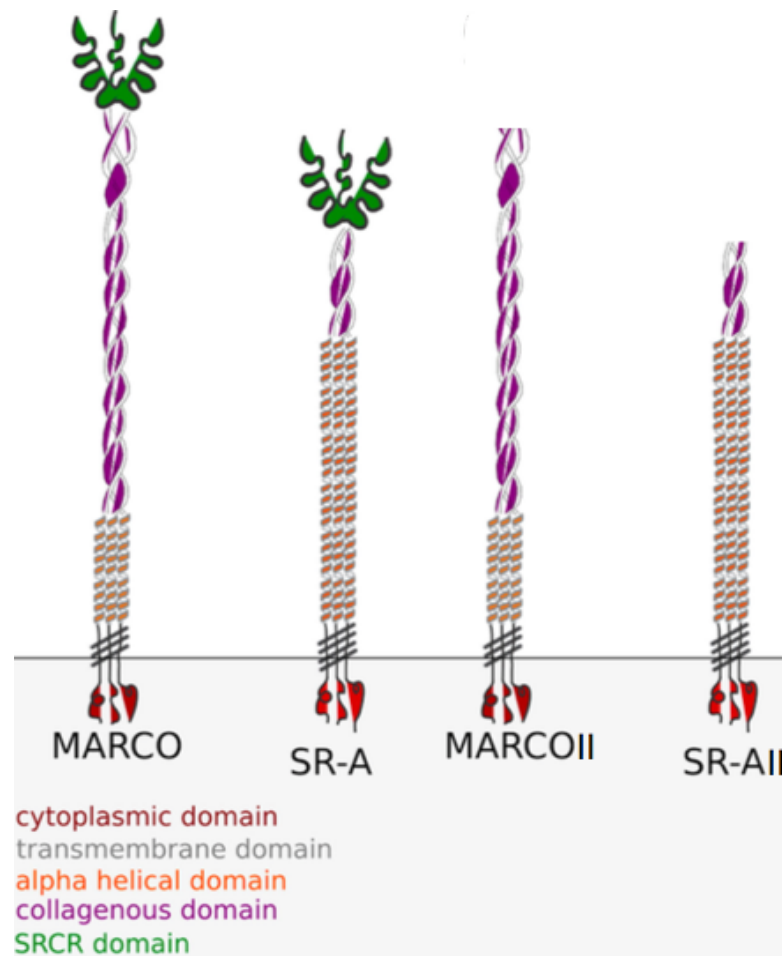
FIGURE 3.1: **Structure of MARCO and SR-A.** The two proteins have similar protein domains including a collagenous domain and an SRCR domain. MARCOII and SR-AII are splice variants of each protein that lack their respective SRCR domains. Adapted from Yap et al. (2015).

In contrast to SR-A, the SRCR domain of MARCO has been shown to be essential in binding bacterial ligands. Novakowski et al. (2016) identified an alternate splice variant of MARCO named MARCOII, that lacks the SRCR domain and is unable to bind traditional MARCO ligands (Figure 3.1). Brännström et al. (2002) have shown that MARCO binds ligands through the use of a positively charged RGR motif within

its SRCR domain. Furthermore, several conserved motifs within the SRCR domain have been identified. Whelan et al. (2012) identified a conserved GRAEVYY motif encompassing the RGR motif found by Brännström et al. (2002), and Yap et al. (2015) identified a WGTICDD motif that is conserved within the SRCR domains of several of the cA-SRs. In a recent study, Lee et al. (2016) used a new method of identifying conserved patterns within proteins called Aligned Pattern Clusters (APC), in order to discover conserved regions within the cA-SRs. They identified several conserved motifs within the SRCR domains of MARCO, including the GRAEVYY and WGTICDD motifs, that were absent from SR-A. These data suggest that although SR-A, and MARCO are evolutionarily related, they possess different functional motifs.

MARCO has been shown to enhance the inflammatory response to bacterial pathogens in conjunction with the Toll-Like Receptor 2 (TLR2), and CD14. TLR2 binds to monomers of bacterial ligands such as lipoteichoic acid (LTA) (Nilsen et al., 2008), while CD14 is required to present these individual monomers from the bacterial surface or from complex polymers. Dorrington and Bowdish (2013) have shown that cells transfected with MARCO, TLR2, and CD14 had higher levels of NF-κB production than cells transfected with TLR2 and CD14 alone when exposed to *Streptococcus pneumoniae*. This was specific to MARCO since cells transfected with SR-A, TLR2, and CD14 did not have an enhanced response (Dorrington and Bowdish, 2013). Similarily, Bowdish et al. (2009) showed that cells co-transfected with MARCO, TLR2, and CD14 have increased NF-κB signalling when exposed to bacterial cell wall components of *Mycobacterium tuberculosis* (Mtb).

Population studies have shown that several of these bacterial ligands have co-evolved with human populations. In particular, 5 different lineages of Mtb have been identified that correspond to different geographical regions, suggesting selective pressure for resistance against different Mtb strains (Perrin, 2015). This selective pressure may have also influenced the frequencies of polymorphisms within the TLR2 gene in different ethnic groups (Perrin, 2015). *S. pneumoniae* is another ancient pathogen that is thought to have co-evolved with humans. Studies have shown that *S. pneumoniae* has retained genes for capsule production and other virulence factors, while other closely-related Streptoccoci species have lost these characteristics (Kilian et al., 2008). In addition, Kilian et al. (2008) trace the origin of modern pneumococcus-lineages to a pneumococcus-like ancestor common between apes and hominoids.

Since both *S. pneumoniae* and Mtb have co-evolved with humans, we hypothesize TLR2, CD14, and MARCO have co-evolved in respose to selective pressure to survive these infections. In addition, due to the compositional differences between MARCO and SR-A, we propose that the functional domains of MARCO may be under selective

pressure to bind microbial ligands. Here we show that TLR2, CD14 and MARCO appear to have co-evolved when compared to a control of 40 proteins not known to be co-evolving. By comparing phylogenies of MARCO, TLR2, and CD14 to 40 unrelated proteins, we found evidence of co-evolution. Using bioinformatics, we attempted to identify critical motifs within MARCO for binding bacteria and enhancing the TLR2/ CD14 response. We identified a Single Nucleotide Polymorphism (SNP) within MARCO's collagenous domain of interest that is specific to *Homo sapiens* as well as a site under positive selection within MARCO's SRCR domain.

3.2 Methods

3.2.1 Testing Co-evolution using Correlation of Branch Lengths

In order to test for co-evolution, phylogenetic trees for 40 non-immune related proteins were used to generate a model of co-evolving proteins. These trees were compared with the phylogenetic trees of MARCO, TLR2, and CD14. Trees for each protein were generated using sequences from *Homo sapiens, Pan troglodytes, Sus scrofa, Mus musculus* and *Bos taurus*, with a pre-defined tree topology. Sequences ranged from 140 amino acids long to 1950 amino acids long, with the majority of sequencing ranging from 300 to 500 amino acids. MrBayes was used to construct phylogenetic trees for the different proteins (Ronquist and Huelsenbeck, 2003). MrBayes was run for 1 million generations using the Whelan and Goldman (WAG) model for estimating amino acid replacement (Whelan and Goldman, 2001). Correlation coefficients were generated through pairwise comparisons of branch lengths from each tree using the ape and vegan packages in R (Paradis, Claude, and Strimmer, 2004; Oksanen et al., 2007). Correlation coefficients for MARCO, TLR2, and CD14 were compared to a model distribution of the 40 non-immune related proteins. To control for pairwise comparisons, 1000 permutations were performed and correlation coefficients were generated and ranked. Branch lengths from the phylogenetic trees were used to calculate the substitution rates for each protein. Scripts can be found at https://github.com/inickyap/Bio720/tree/master.

3.2.2 The 1000 Genomes Project and the Great Apes Genome Project SNP Data Analysis

SNP data for the chromosomal region of MARCO was analyzed from the 1000 Genomes Project (Consortium et al., 2010) and the Great Apes Genome Project (Prado-Martinez et al., 2013) using PERL scripts. SNPS from the Neanderthal and Denisovan genomes were accessed from the Max Planck Institute for Evolutionary Anthropology (Noonan et al., 2006; Meyer et al., 2012). SNPs present within exons of MARCO were characterized as being non-synonymous or synonymous substitutions using NCBI's

dbSNP database (Sherry et al., 2001). Genomes from the Great Apes Genome Project included *Pan troglodytes, Pan paniscus, Pongo pygmaeus,* and *Gorilla gorilla*.

3.3 Results

3.3.1 Co-evolution of MARCO, TLR2, and CD14

Since previous work has shown that MARCO enhances the inflammatory response of TLR2 and CD14 to bacterial pathogens, we propose that these three proteins have co-evolved with each other. In order to test our hypothesis, we generated phylogenetic trees for 40 non-immune related proteins from 5 species using a pre-defined tree topology and calculated correlation coefficients through pairwise comparisons. Correlation coefficients for MARCO, TLR2, and CD14 were compared with our random data set. We found that TLR2 and CD14 had a very high correlation coefficient of 0.988, which was above 95 percent of the correlations from our distribution of 40 proteins (Figure 3.2). This observation is consistent with data demonstrating that TLR2 and CD14 work co-operatively to bind PAMPs. The correlation coefficient between MARCO and CD14 was 0.929 and was also above the 95th percentile, however the correlation coefficient for MARCO and TLR2 was 0.897, just short of the 95th percentile (correlation coefficient above 0.898) (Figure 3.2).

We also performed a permutation test of 1000 random shuffles of the 40 non-immune related proteins. Correlation coefficients were generated by randomly shuffling the proteins and ranking the correlations. The correlation coefficients between MARCO, TLR2, and CD14 were also included in the ranking. Out of 1000 random shuffles, TLR2 and CD14 ranked in the top three correlation coefficients 999 times. MARCO and CD14 ranked in the top three correlation coefficients 639 times, while MARCO and TLR2 ranked in the top three correlation coefficients 440 times. These data suggest that MARCO, TLR2 and CD14 are co-evolving. Several pairwise comparisons at the extreme ends of the distribution may also be of interest. These can be found at https://github.com/inickyap/Bio720/tree/master.

Furthermore, we calculated the total branch lengths for each phylogenetic tree as a measurement of the rate of evolution. We found that MARCO had the longest branch lengths out of all the proteins, indicating a higher number of substitutions per amino acid (Figure 3.3). TLR2 and CD14 had very similar substitution rates that were also slightly above average.
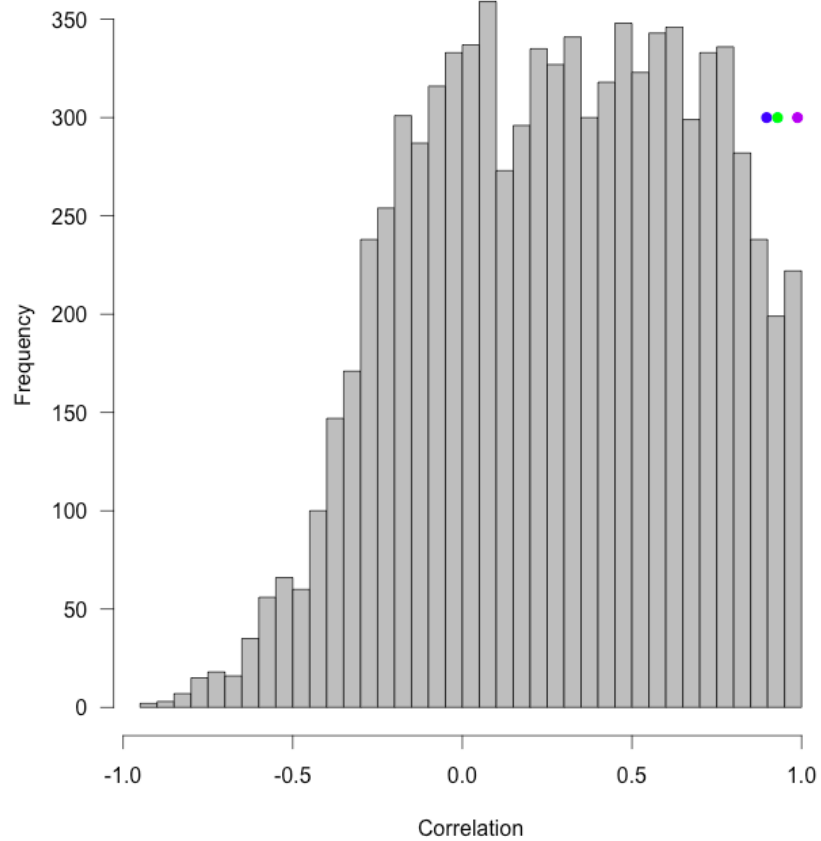
FIGURE 3.2: **Plot of correlation coefficients generated from pairwise comparisons of phylogenies of 40 non-immune proteins.** Correlations for MARCO and TLR2 (blue), MARCO and CD14 (green), and TLR2 and CD14 (purple) are shown, in comparison to our distribution. Bars represent the number of comparisons with a given correlation coefficient.
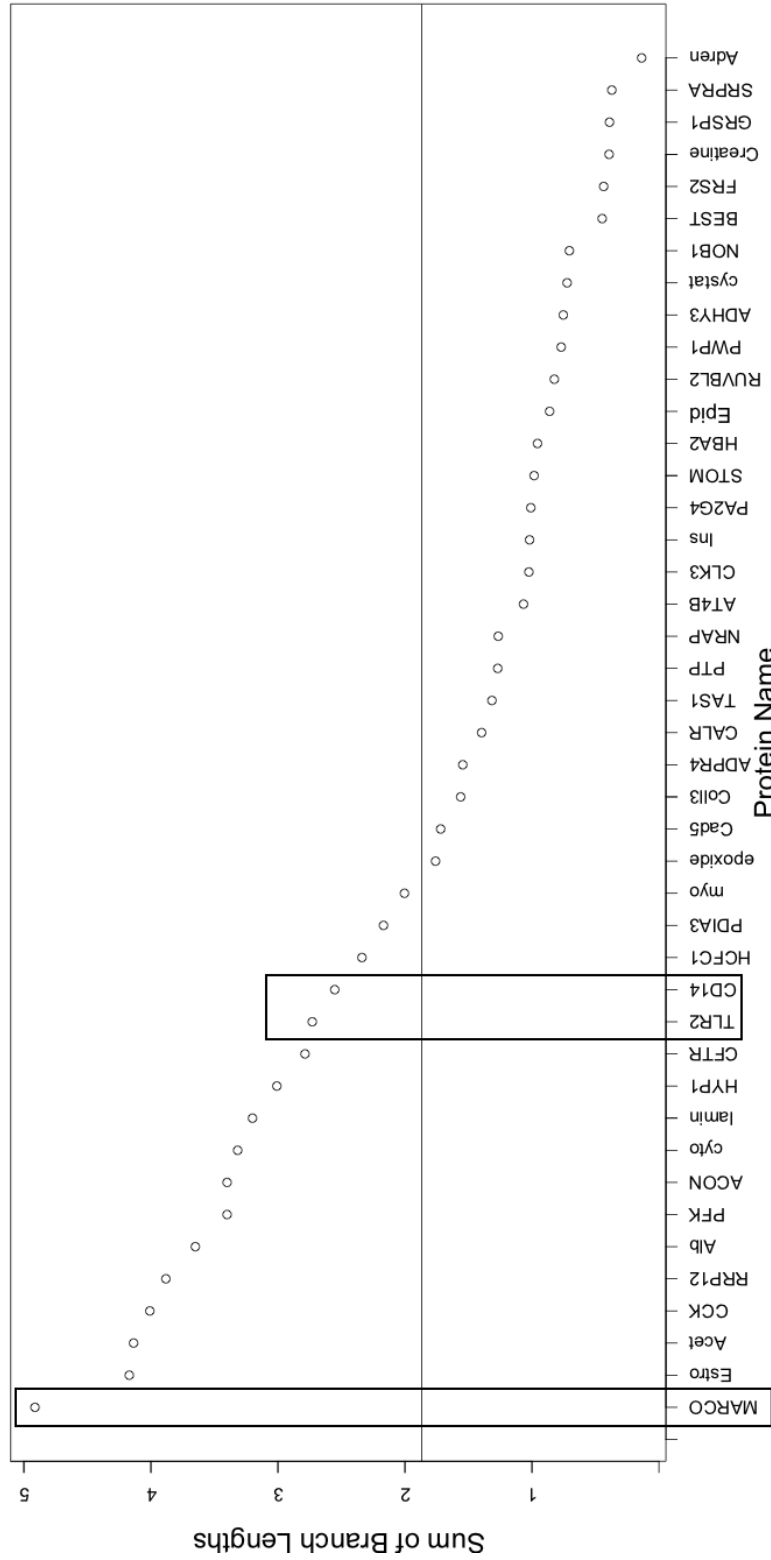
FIGURE 3.3: **Sum of the branch lengths of phylogenetic trees.** Total branch lengths for 40 proteins as well as MARCO, TLR2, and CD14 are shown. Sum of branch lengths represents number of substitutions per site. Black horizontal line shows the average total branch length (1.87). (See Appendix Table B.2 for full protein names.)

3.3.2 Identifying Potential Functional Sites within MARCO

Novakowski et al. (2016) and Dorrington and Bowdish (2013) have shown that the SRCR domain of MARCO is required for enhancing the TLR2 and CD14 response to *S. pneumoniae*. Since MARCO is important for binding two human adapted pathogens, *S. pneumoniae* and Mtb, we attempted to identify functional sites within MARCO. Using PAML, we previously identified three residues within the SRCR domain of MARCO under positive selection, positions 442, 452, and 477. This was done by using the codeML function within PAML, and testing for sites under positive selection within humans, from a codon alignment of MARCO. Within the MARCO protein, position 452 was of particular interest because of its close proximity to the previously defined GRAEVYY (Whelan et al., 2012) and WGTICDD motifs (Yap et al., 2015). Within humans, position 452 encodes a glutamine residue (Q), while most mammals possess an aspartic acid residue (D), and non-human primates vary (Figure 3.4).
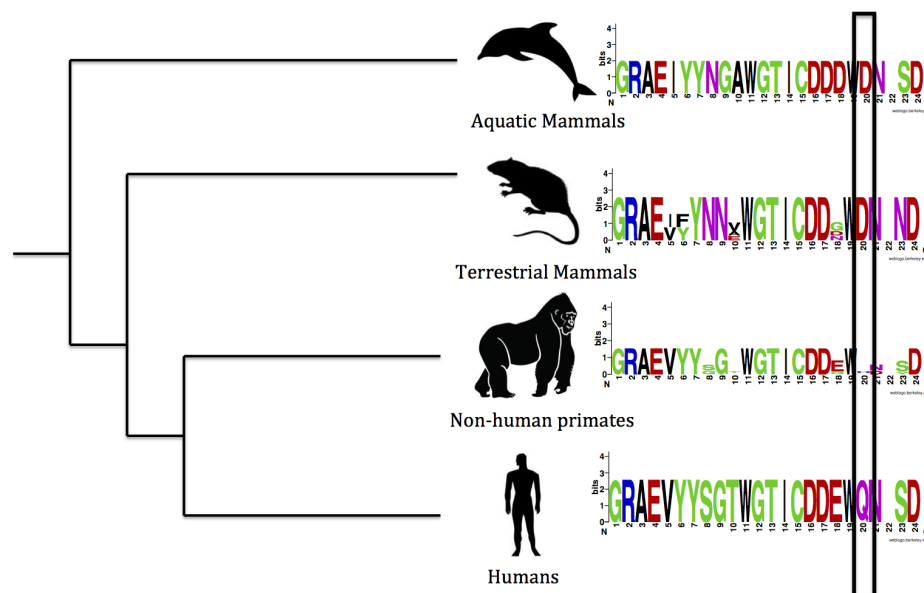
FIGURE 3.4: **Partial alignment of MARCO's SRCR domain around position 452 of humans.** Previous studies have identified position 452 (Q) within humans as under positive selection. Position 452 is downstream of the previously identified GRAEVYY and WGTICDD conserved motifs within the SRCR domain of MARCO (Yap et al., 2015).

Since our previous analysis has shown that infectious diseases have affected the evolution of MARCO, TLR2 and CD14 we propose that MARCO polymorphisms have also been affected by infectious diseases. Using the 1000 Genomes Project we aimed to identify functional SNPs within the exons of the entire MARCO gene. Although the intronic regions of MARCO have been shown to play a role in splicing (Novakowski et al., 2016), and gene expression (Bowdish et al., 2013), we chose to focus on the coding region of MARCO because they would directly impact bacterial binding. We identified SNPs mapped to MARCO using the human MARCO gene as a reference. At amino acid position 282 of MARCO, humans possess a phenylalanine residue. We identified a SNP, rs6761637, which causes a non-synonymous substitution at amino acid 282 from phenylalanine to serine (F282S). rs6761637 has a global global allele frequency of 12% and was found at a high frequency in Asian (13%) and African (29%) populations compared to European (4%) and North American populations (6%). Interestingly, amino acid 282 lies within the collagenous domain of MARCO, which has not been shown to be required for ligand binding.

We also searched for differences in MARCO between humans and other hominins. Using data from the Great Apes Genome Project, we identified SNPs within *Pan troglodytes, Pan paniscus, Pongo pygmaeus,* and *Gorilla gorilla* mapped to the human MARCO gene (Figure 3.5). When we looked at position 282 within apes, all of the different ape species possessed a serine residue (Figure 3.6). We also identified SNPs from Neanderthals and Denisova mapped to the human MARCO gene. Interstingly, both Neanderthals and Denisova possess the serine residue as well.

These data suggests that there are two MARCO variants for amino acid 282. Modern day humans possess a phenylalanine residue while non-human primates as well as Neanderthals and Denisovia possess a serine residue. Furthermore, the identification of rs6761637 shows that modern humans possess both variants. Therefore, we will refer to the primary variant within modern humans as F282, and the ancestral variant as S282.
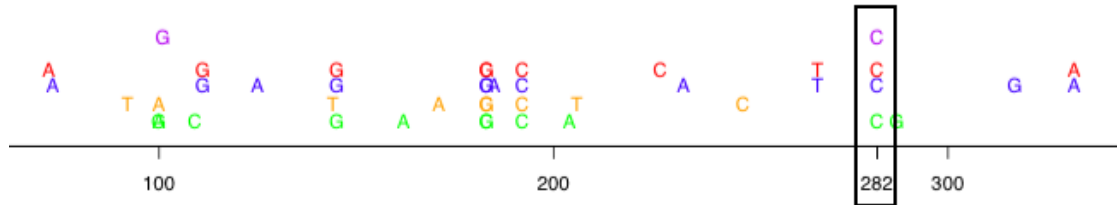
FIGURE 3.5: **Non-synonymous SNPs from The Great Apes Genome Project and Neanderthals mapped onto the human MARCO gene.** SNPs mapped to positions in MARCO exons are shown. Alternate alleles for positions are shown at corresponding amio acids. Scale bar shows amino acid positions of MARCO. SNPs in codons between amino acids 50-350 are shown. Position 282 (outlined in black box) is of particular interest since at this site, all great apes as well as Neanderthals and Denisova possess a serine residue at this site. *Pan troglodytes* (blue), *Gorilla gorilla* (green), *Pan paniscus* (red), *Pongo pygmaeus* (yellow), and Neanderthals and Denisova (purple) are shown. (See Table B.1 for full list of SNPs).
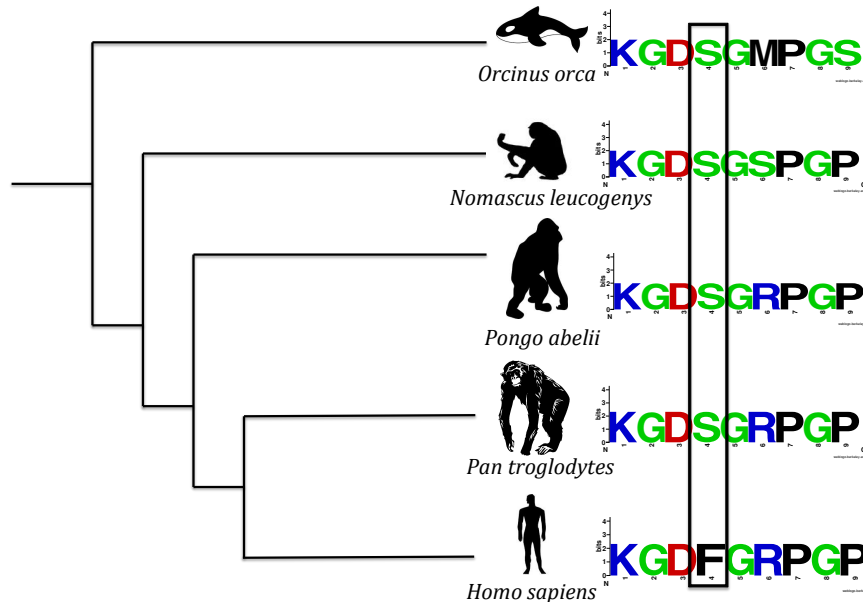
FIGURE 3.6: **Partial alignment of MARCO around position 282 across multiple species.** Within various mammals including *Orcinus orca*, *Nomascus leucogenys, Pongo abellli*, and *Pan troglodytes*, position 282 is exclusively a serine residue. However, the primary variant in humans is a phenylalanine residue at position 282, but humans also have a nonsynonymous substitution of F282S. *Orcinus orca* is included as an outgroup.

3.4 Discussion

The innate immune system is an ancient form of immunity that acts as the first defence against bacterial infection. Ancient microbial ligands have been shown to influence the evolution of human immune receptors in what has been called a molecular arms race. Based on these observations, we propose that the innate immune receptors MARCO, TLR2, and CD14 have co-evolved with each other in order to protect the host against infection. Using bioinformatics, we performed pairwise comparisons to calculate correlation coefficients between MARCO, TLR2, and CD14. We show that MARCO, TLR2, and CD14 have co-evolved with each other, when compared with our model of 40 random proteins. These data are consistent with previous work by Dorrington and Bowdish (2013) and Novakowski et al. (2016) that have shown MARCO enhances the TL2/ CD14 response to *S. pneumoniae* through the use of the SRCR domain. In addition, we previously identified three sites within the SRCR domain of MARCO under positive selection (Yap et al., 2015). Future studies will be

required to determine if these sites alter MARCO's ability to enhance the inflammatory response of TLR2 and CD14 to microbial products.

Human adapted pathogens have also been shown to effect the frequency of different variants of human immune genes. Ioana et al. (2012) speculate that Mtb has influenced the frequencies of three polymorphisms in TLR2. They investigated a guanine to adenosine at polymorphism at amino acid 2258 within TLR2, and found the SNP was specific to European populations and is associated with increased bacterial infections including Mtb (Ioana et al., 2012). Since we show that MARCO, TLR2 and CD14 are co-evolving, we attempted to identify polymorphisms within MARCO encoding non-synonymous mutations. Using SNP data from The Great Apes Genome Project and The 1000 Genomes Project, we found that humans possess two different variants at position 282. Modern humans possess both variants, F282 and S282 respectively. Conversely, non-human primates as well as Neanderthals and Denisova exclusively possess the ancestral S282 variant. Interestingly, we identified a SNP within modern humans that causes a non-synonymous substitution F282S. Therefore, although 88% of modern humans possess the F282 variant, 12% of individuals possess the ancestral S282 variant. These data suggest that modern humans may have experienced selective pressure for the primary F282 variant.

The role of the F282S SNP has been investigated in other studies. Thomsen et al. (2012) attempted to identify variants within MARCO that correlated with Chronic Obstructive Pulmonary Disorder (COPD) within a Danish population. However, they found no correlation between COPD and the F282S mutation. Ma et al. (2011) studied MARCO SNPs associated with pulmonary tuberculosis within Chinese Han population. Consistent with our hypothesis that Mtb may be influencing the evolution of MARCO, Ma et al. (2011) found that having the ancestral S282 variant was associated with increased risk to tuberculosis, but only as part of a haplotype group. Taken together, these data suggest that selective pressure from infectious diseases, such as Mtb and *S. pneumoniae* may have effected MARCO's evolution. Future experiments to test the functional significance of the F282S mutation as well as the positively selected sites within the SRCR domain are required.

Chapter 4

The Evolution of the Scavenger Receptor Cysteine-Rich Domain of the Class A Scavenger Receptors

4.1 Introduction

The Scavenger Receptors (SRs) are a group of Pattern Recognition Receptors (PRRs), which were originally defined for their ability to bind forms of low-density lipoprotein (LDL) and are subdivided into 8 classes (A-H) (Goldstein et al., 1979; Canton, Neculai, and Grinstein, 2013). These receptors are extracellular glycoproteins which mediate phagocytosis of negatively charged ligands (Martínez et al., 2011). This binding ability was later refined to include host-modified ligands such as oxidized LDL (ox-LDL), acetylated LDL (acLDL), and various bacterial ligands including *Streptococcus pneumoniae*, *Escherichia coli*, (Peiser et al., 2000), and *Mycobacterium tuberculosis* (Bowdish et al., 2009).

The class A Scavenger Receptors (cA-SRs), one of eight classes of SRs, are membrane-associated phagocytic receptors which reside on the surface of immune cells (Murphy, 2011). The cA-SR family consists of five members: the Scavenger Receptor class A (SR-A) (Peiser et al., 2000), Macrophage Associated Receptor with CollagenOus structure (MARCO) (Elomaa et al., 1995), SCAvenger Receptor class A member 3 (SCARA3) or Cellular Stress Response 1 (CSR1) (Han, Tokino, and Nakamura, 1998), SCAvenger Receptor class A member 4 (SCARA4) or Scavenger Receptor with C-type Lectin domain (SRCL) (Nakamura et al., 2006), and SCAvenger Receptor class A member 5 (SCARA5) (Jiang et al., 2006). Despite forming a protein family, the 5 cA-SR proteins differ from each other in a few key ways. First, the 5 receptors are expressed differentially on immune cells. For example, it has been shown in mice, that SR-A is restricted to specific myeloid lineages; however SCARA5 is expressed exclusively on epithelial cells (Jiang et al., 2006). Further, there are also differences in domain structure between the cA-SRs. MARCO, SR-A, and SCARA5 differ from SCARA3 and SCARA4 in that they possess a Scavenger Receptor Cysteine-Rich (SRCR) domain (Figure 4.1). The SRCR domain is replaced by a C-type lectin domain in SCARA4, while SCARA3

terminates at the collagenous domain. Functionally, MARCO and SR-A both possess SRCR domains, but do not recognize the same ligands. For example, one study identified the surface proteins of *Neisseria meningitidis* and showed MARCO and SR-A were able to bind different target proteins (Plüddemann et al., 2008). MARCO has also been shown to play a functional role in binding *Mycobacterium tuberculosis*. Polymorphisms within MARCO have been shown to be associated with altered susceptibility to tuberculosis in a Gambian population, whereas no relation was found between infection and polymorphisms in SR-A (Bowdish et al., 2009). In addition, MARCO also plays a direct role in host defence during *Streptococcus pneumoniae* infection. Using an infection model in mice, MARCO has been shown to be important for cytokine and chemokine production in response to *Streptococcus pneumoniae* infection; however SR-A knock-out mice do not show impaired killing of the bacterium (Dorrington and Bowdish, 2013). In addition, MARCO is thought to play a role in antigen presentation and/or antigen transfer to dendritic cells and thereby generating T cell tolerance (Getts et al., 2012). These suggest an important role for MARCO in host defence, while SR-A is primarily involved in the clearance of modified lipids (Bowdish et al., 2009).

Outside of the SRCR domain the five ca-SRs share a similar domain structure, with each protein possessing a cytoplasmic domain, transmembrane domain, and a collagenous domain (Figure 4.1). All of the cA-SRs possess a cytoplasmic domain, a transmembrane domain, and an alpha helical domain, but they differ in the length of their collagenous domain and at their terminal regions. Despite different ligand binding domains, these receptors share similar ligand binding properties. For instance, all of the receptors except SCARA3 have been shown to bind Gram negative and Gram positive bacteria (Canton, Neculai, and Grinstein, 2013). In addition, SR-A, MARCO and SCARA4 all bind ox-LDL despite SCARA4 lacking a SRCR domain (Canton, Neculai, and Grinstein, 2013). SR-A and MARCO also share several bacterial binding capabilities including *Escherichia coli* and *Staphylococcus aureus* (Martínez et al., 2011; Palecanda et al., 1999; Elomaa et al., 1995).

The SRCR domain is an evolutionarily conserved 90-110 amino acid long domain that is characterized by 6-8 cysteine residues (Martínez et al., 2011). Within the genome of *Strongylocentrotus purpuratus* (purple sea urchin), over 1200 SRCR domains were identified, often with several domains found in tandem repeats (Sodergren et al., 2006). It has been hypothesized that these multiple SRCR domains play a role in cell adhesion, a role shared by some SRCR domains found in vertebrate proteins (Bowdish and Gordon, 2009). SRCR domains are classified into two categories: type A domains which possess six cysteine residues encoded by multiple exons, and type B domains which contain eight cysteine residues encoded by a single exon (Martínez et al., 2011). These cysteine residues are thought to bind intracellularly, creating three and four disfulide bridges in class A and class B respectively (Resnick et al., 1996). Three of the cA-SRs possess type A SRCR domains, sharing 6 conserved cysteine residues with other type A SRCR domains. The SRCR domain has been experimentally shown to be required for bacterial binding by utilizing a positively clustered RGR motif within MARCO (Brännström et al., 2002). However, in SR-A, experiments using its isoform, SR-AII,

which possesses a truncated SRCR domain, indicate that the SRCR domain is not necessary for the binding of bacterial ligands (Krieger, 1992).
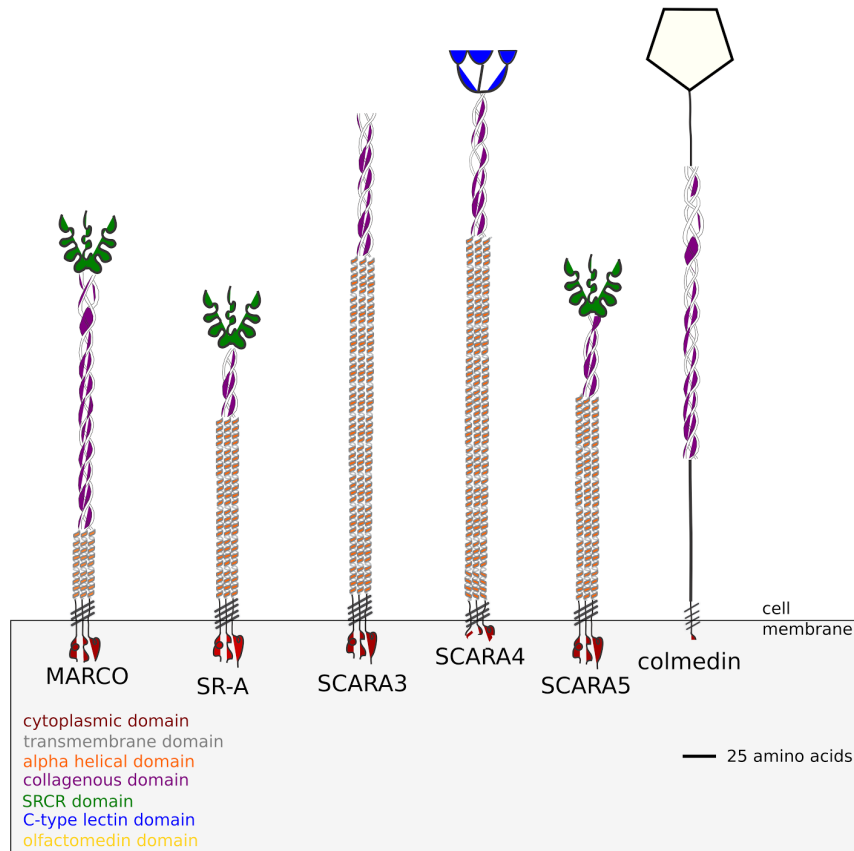
FIGURE 4.1: **Domain structure of the five class A Scavenger Receptors based on the protein sequences obtained from the *Homo sapiens* genome.** SCARA3 terminates at its collagenous domain while SCARA4 possesses a C-type Lectin domain. SCARA5, MARCO, and SR-A all possess a terminal SRCR domain. Colmedin is a transmembrane protein with a collagenous and olfactomedin domain found in *Strongylocentrotus purpuratus* which has been included as an outgroup in this study.

Previous work has hypothesized that the cA-SR family members were created through multiple duplication events of an ancestral gene (Whelan et al., 2012). Of the SRCR-containing receptors, SCARA5 and SR-A have been shown to be more closely related to each other than to MARCO. Analyses have shown that the SRCR-containing cA-SRs diverged from SCARA3 and SCARA4, perhaps as early as within the genomes of teleost fish (Whelan et al., 2012). However, the domain structure of the ancestral receptor remains unresolved. Furthermore, MARCO's relationship to the other cA-SRs is unclear as it contains an SRCR domain that shares functional similarity with SR-A,

but appears to share a common ancestor with SCARA3 and SCARA4 (Whelan et al., 2012). With the recently published genomes of *Petromyzon marinus* (sea lamprey) and *Callorhinchus milii* (ghost shark), we can now study the evolution of the cA-SRs before the divergence of teleost fish. In this study, we test the hypothesis that MARCO, SR-A, and SCARA5 share a common ancestor containing the SRCR domain using various phylogenetic approaches. In addition, we reconstruct a hypothetical ancestral SRCR domain and analyze the evolution of two motifs within the SRCR domain. We also test the hypothesis that MARCO's SRCR domain is under different selective pressure when compared to that of SCARA5 and SR-A due to its direct role in host defence. These data will provide new insight on the origin of the SRCR domain and also its role in ligand binding within the Class A Scavenger Receptors.

4.2 Material & Methods

4.2.1 Gathering Nucleotide and Amino Acid Sequence Data

Amino acid sequences of the 5 cA-SRs were searched for using the National Center for Biotechnology Information (NCBI) and the ENSEMBL databases (Flicek et al., 2013). Full-length, partial, and predicted amino acid sequences for all receptors were included in phylogenetic analyses. Amino acid sequences were gathered from a diverse set of species, with as many representatives as possible from fish, birds, and mammals (accessed March 2014). The total number of sequences for each protein was 40 MARCO, 25 SR-A, 40 SCARA5, 40 SCARA4, and 40 SCARA3 (Table C.1). In addition, PFAM (Protein Families Database) (Bateman et al., 2004) and TMHMM (TransMembrane Hidden Markov Model) (Krogh et al., 2001) were utilized to characterize the domains present within each of the protein sequences . Protein alignments were done using MAFFT (Katoh, Asimenos, and Toh, 2009) due to the numerous collagenous domains of the cA-SRs.

4.2.2 Phylogenetic Analysis

In order to study the evolutionary history of the receptors and their SRCR domains, Bayesian phylogenetic analyses were performed using the software MrBayes (Ronquist and Huelsenbeck, 2003). Several evolutionary models were studied to determine the model best fit to the data set. Each of our data sets was run under a 'mixed' evolutionary model in MrBayes for 1 million generations. PROTTEST (Abascal, Zardoya, and Posada, 2005) was also utilized to search for the model that maximizes the posterior probabilities of the phylogenetic tree, and confirmed the results from MrBayes (Table 4.1). These results differed for the SRCR-containing proteins; however due to the inability to implement the Le and Gascuel (LG) model in MrBayes, we ran our analysis using a Whelan and Goldman (WAG) model.

Analysis of the SRCR domain of SR-A, MARCO, and SCARA5 was carried out using the WAG model with invariable (I) sites and gamma (G) distributed rates for 10 million generations. A combined tree of all 5 cA-SRs was performed using a Jones Thorton and Taylor (JTT) model with an IG distribution for 15 million generations and displayed with midpoint rooting. Finally, an analysis of all 5 cA-SRs with an outgroup was performed in MrBayes using a JTT model, with invariable sites, and with a gamma distribution for 20 million generations. All MrBayes output trees were visualized using TRACER (Rambaut and Drummond, 2012) to ensure convergence. Trees were visualized in FigTree (Rambaut and Drummond, 2009).

In the reconstruction of a tree for the SRCR domains of MARCO, SCARA5, and SR-A, the ninth (1042-1142) and tenth repeat (1153-1255) of the *Geodia cydonium* (sea sponge) SRCR-containing protein (GCSRCR) (NCBI ID: CAA75175.1) were used as outgroups. The GCSRCR protein's ninth repeat has been shown to share sequence similarity to both MARCO and SR-A's SRCR domains (Pancer et al., 1997). We also investigated the tenth repeat of the protein because of a nearby alternative splice site (Pancer et al., 1997). Finally, a phylogenetic tree of all five cA-SRs was constructed with a Colmedin protein sequence from *Strongylocentrotus purpuratus* as the outgroup (NCBI ID: NP_001073014.1). Colmedin is a membrane spanning protein containing a transmembrane domain, multiple collagenous domains, and an olfactomedin (Tomarev and Nakaya, 2009) (Figure 4.1) which plays a role in the sea urchin's innate immune system by aiding in the formation of clots where the skin of the organism has been pierced (Smith et al., 2010). Due to the similar domain structure to the cA-SRs, Colmedin is a suitable outgroup to the scavenger receptor family.

Predicted ancestral SRCR domain sequences were reconstructed using the FastML webserver (Ashkenazy et al., 2012). The MrBayes generated phylogenetic tree of SCARA5, SR-A, and MARCO was used to reconstruct the ancestral SRCR domains of these proteins and compared.

4.2.3   Motif Evolution within the SRCR domain

Consensus sequences of SCARA5 and MARCO were generated for mammals, birds, fish, and other species using Jalview (Clamp et al., 2004) and represented as logos using WebLogo (Crooks et al., 2004). We focused our analysis on two motifs within these proteins; MARCO contains a RGRAEVYY motif (amino acids 440-488 in *Mus musculus*) and a WGTICDD motif (amino acids 452-458 in *Mus musculus*) of interest; SCARA5 contains a EGRVEVYH motif (position 399-406 in *Mus musculus* and a WGTVCDD motif (position 410-416 in *Mus musculus*. We included the RGRAEVYY motif in our analysis due to its known functional role in ligand binding within MARCO (Brännström et al., 2002). WebLogos for MARCO were made using 1 sequence from *Petromyzon marinus* (sea lamprey), 1 sequence from *Callorhinchus milii* (ghost shark), 23 sequences from various mammals, 7 sequences from reptiles, amphibians and birds, and 8 sequences from fish. WebLogos for SCARA5 were made using 1 sequence from *Petromyzon marinus*

TABLE 4.1: **Analysis of the best model for each data set of protein alignment based on PROTTEST and running each receptor in MrBayes for 1 million generations under a "mixed" model.**

Receptors used for analysis	Model predicted from PROTTEST /MrBayes	Support (Log likelihood)
SRCR-containing proteins (MARCO, SR-A, SCARA5)	LG + G / WAG + IG*	-6950.09
All 5 cA-SRs (MARCO, SCARA3, SCARA4, SCARA5, SR-A)	JTT + G*	-82497.07

* WAG= Whelan and Goldman model with invariable (I) sites and gamma (G) distributed rates, LG +G = Le and Gascuel model with gamma distributed rates, JTT + G= Jones Thorton and Taylor model with gamma distributed rates.

(sea lamprey), 26 sequences from various mammals, 6 sequences from reptiles, birds and amphibians, and 7 fish sequences.

4.2.4 Differences in selective pressure within the SRCR domain

Using the Phylogenetic Analysis by Maximum Likelihood (PAML) package, we tested whether MARCO is under a different selective pressure from SCARA5 and SR-A (Yang, 2007). Using codeml (**yang2000**), we analyzed only the SRCR domain of MARCO to determine any sites under positive selection. We restricted our analysis to include only the SRCR domains from MARCO, SR-A, and SCARA5. To generate our phylogenetic tree, primate sequences were used with *Mus musculus* as the outgroup. MARCO, SR-A, and SCARA5 sequences from *Mus musculus, Homo sapiens, Pan troglodytes,* and *Gorilla gorilla* were used. Protein alignments were performed using MAFFT (Katoh, Asimenos, and Toh, 2009) and were subsequently transformed into codon alignments using Pal2nal (Suyama, Torrents, and Bork, 2006). We used a branch-site model to allow the ratio of non-synonymous substitutions (dN) to synononymous substitutions (dS) to vary along the branches of the tree and the codon sites. We performed a Likelihood Ratio Test (ΔLRT) to determine the significance for the alternative model compared to a null model.

4.3 Results

4.3.1 MARCO Shares a More Recent Common Ancestor with the SRCR-Containing cA-SRs than with SCARA3 and SCARA4

The evolutionary history of the five cA-SRs has been studied previously (Whelan et al., 2012). In this previous study, it was hypothesized that a single gene duplication event created SCARA5 and SR-A, while MARCO's relationship to the cA-SRs was left unclear (Whelan et al., 2012). We hypothesize that SCARA3 and SCARA4 were generated through one duplication event while MARCO may have been generated from a duplication event of a SCARA5/SR-A precursor.

To test this hypothesis, our analysis includes additional protein sequences from divergent taxa and includes sequences from more diverse species. We are able to expand upon this previous work due to recent genome sequencing projects including that of the sea lamprey and ghost shark. This allowed for our analysis to include in total, 40 MARCO, 25 SR-A, 40 SCARA5, 40 SCARA4, and 40 SCARA3 protein sequences for this analysis. MARCO was found in birds, reptiles, fish, and mammals and a partial sequence was found in *Petromyzon marinus* (sea lamprey). The domains of MARCO varied amongst species, with different numbers of collagen repeats found across different taxa. Within the SRCR domain of MARCO, the RGRAEVYY motif was highly conserved across mammals but less conserved in birds, reptiles, and fish. The SRCR domain of *Petromyzon marinus* however did not possess the conserved RGRAEVYY motif characteristic of MARCO. SR-A was found in mammals exclusively, except for a sequence found in *Xenopus tropicalis* (western clawed frog). We included it in our analysis and found that the domains of SR-A were highly conserved in both the collagenous and SRCR regions of mammals and the *Xenopus tropicalis* sequence, 70% and 78% conservation respectively. SCARA5 was found in birds, reptiles, fish, mammals, and as well in *Petromyzon marinus*. The collagenous and SRCR domains of SCARA5 were fairly conserved across the 40 species with 69% and 75% conservation respectively.

Using Bayesian phylogenetics, we generated a new phylogenetic tree of the cA-SR family. Here we show using midpoint rooting that the non-SRCR-containing proteins are more closely related, while the SRCR-containing proteins branch together (Figure C.1). Within the SRCR-containing proteins, SCARA5 and SR-A cluster together while MARCO appears to have diverged from them before early teleost fish and possibly before the sea lamprey. We constructed a second phylogenetic tree with the addition of the colmedin sequence as an outgroup. Colmedin was used as it possesses a transmembrane domain and multiple collagenous domains (Figure 4.2; C). Using colmedin as an outgroup, MARCO, SCARA5, and SR-A still cluster together while the non-SRCR-containing receptors form their own branch (Figure 4.2). Due to the uncertainty regarding the *Xenopus tropicalis* SR-A sequence, we repeated our analysis excluding the sequence and found no difference in our findings (data not shown). These data suggest

that MARCO shares a more recent common ancestor with SR-A and SCARA5 than with SCARA3 and SCARA4.

4.3.2 Ancestral Reconstruction Shows Conservation of Functional Motifs Within MARCO and SCARA5, and Reveal a Common Origin for the SRCR Domain Within the Class A Scavenger Receptors

Our current knowledge of functional motifs within the SRCR domain is limited to the RGRAEVYY motif within MARCO, which contains a positive cluster essential for ligand binding (Brännström et al., 2002). Although we have shown that it is most likely that MARCO, SCARA5, and SR-A share a common origin of the SRCR domain, SCARA5 and SR-A lack the RGRAEVYY motif. We chose to examine whether this motif is specific to MARCO, or if there are similar motifs within the SRCR domains of SR-A, and SCARA5. Furthermore, we wanted to determine if there are other conserved motifs between the SRCR domains present in these three receptors. Using FastML (Ashkenazy et al., 2012), we reconstructed the ancestral SRCR domains of MARCO, SR-A, and SCARA5. We focused our analyses on the SRCR domains of MARCO and SCARA5, since SR-A is present primarily in mammals.

Within MARCO's SRCR domain, amino acids 440-448 (in *Mus musculus*) contains the RGRAEVYY motif. The motif is highly conserved within mammals, but is less conserved in fish, where it takes the form of QGRVEVFH (Figure 4.3). These two motifs are homologous between mammals and fish, but differences in selective pressure may have changed the content of the motif throughout evolution. We reconstructed the ancestral SRCR domain of MARCO that predates the sea lamprey to analyze the domain's original form. Based on our ancestral reconstruction, the ancestral version of this motif was a EGRVEIFH motif.

We also studied the SRCR domain of SCARA5 across various species to compare with MARCO. Based on our multiple sequence alignment, SCARA5 contains a similar motif to RGRAEVYY. Across all the different species, SCARA5 contains a highly conserved EGRVEVYH motif, where only the first glutamic acid (E), tyrosine (Y), and histidine (H) are somewhat variable (Figure 4.3). We also constructed an ancestral SCARA5 sequence which contained a motif of the form EGRVEVFH. Interestingly the MARCO motif, QGRVEVKH, within fish resembles the EGRVEVFH motif of SCARA5. These suggest that the RGRAEVYY motif is specific to mammalian MARCO proteins, and may be under selective pressure due to its role in bacterial binding. In addition, the ancestral motif to both SCARA5 and MARCO most likely took the form of a EGRVEVFH motif (Figure 4.3).
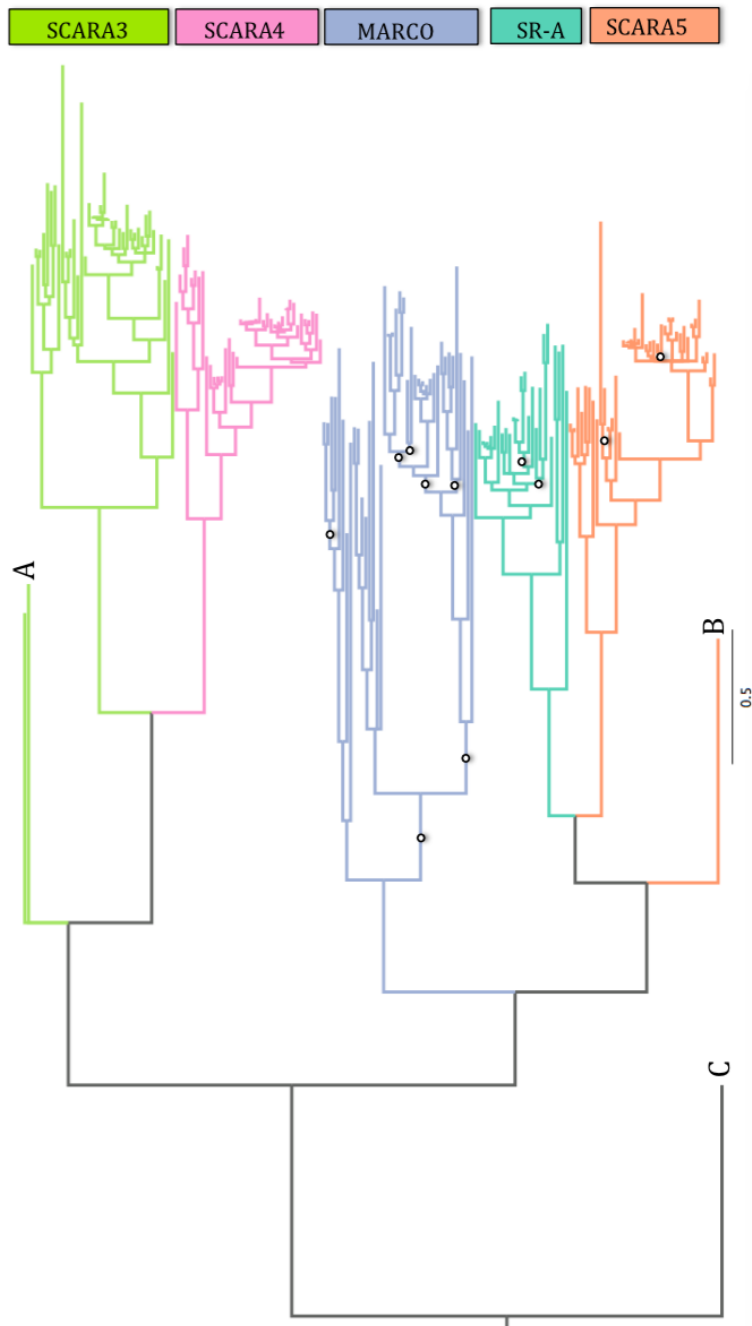
FIGURE 4.2: **Phylogeny of all five Class A Scavengers using colmedin as an outgroup.** MARCO branches with SCARA5 and SR-A after rooting on this outgroup. Posterior probabilities for branches with less than 0.7 confidence are shown with open circles. Scale bar denotes number of substitutions per site. See Table C.1 for complete sequence list and accession numbers. SCARA3 in sea lamprey (*Petromyzon marinus*) and SCARA3 in southern platyfish (*Xiphophorus maculatus*) are labeled as A and label B shows the sea lamprey sequence of SCARA5. These are shown due to their long branching pattern. Label C shows the colmedin sequence.

We also identified another highly conserved motif consisting of WGTICDD in MARCO at amino acids 452-458 in *Mus musculus*. This newly identified motif is highly conserved across the SRCR domains of MARCO, SR-A, and SCARA5, and we hypothesize that the motif may have a functional role due its proximity to a conserved cysteine residue (C1) within the SRCR domain. In contrast to the RGRAEVYY motif, the WGTICDD motif within MARCO is highly conserved across all the different taxa examined (Figure 4.4). The isoleucine residue, position 445 of MARCO in *Mus musculus*, is the only variable site which is replaced by a valine (V) in some fish species. Based on our multiple sequence alignment, SCARA5 possesses a homologous, highly conserved, WGTVCDD (Figure 4.4). Interestingly, within fish, the sea lamprey, and the ghost shark, MARCO's WGTICDD motif resembles the WGTVCDD motif of SCARA5. Taken together with our previous finding, both the RGRAEVYY motif and the WGTICDD motif within MARCO resemble those within SCARA5; however the two motifs may be under different selective pressures. These suggest that the two motifs were found in the ancestral SRCR domain and most likely resembled a EGRVEVFH and WGTVCDD motif, respectively.

4.3.3   Evidence of Positive Selection within the SRCR domain of MARCO

MARCO's SRCR has been shown to play a direct role in ligand binding and host immunity defence (Bowdish et al., 2009). Previous work has shown the importance of arginine residues (R) within the SRCR domain (Brännström et al., 2002), as well as the RGRAEVYY motif (Elomaa et al., 1995). Due to MARCO's direct role in binding various bacteria including *Streptococcus pneumoniae* (Arredouani et al., 2004) and *Escherichia coli* (Palecanda et al., 1999), and an uncertain role in SCARA5, we hypothesize that MARCO's SRCR domain is under positive selection.

We tested for positive selection within MARCO using a branch-site model, where the ratio of non-synonymous substitutions (dN) to synonymous substitutions (dS) is free to vary along both the branches and the sites of the phylogeny. Using this model, we found evidence for positive selection along several sites within MARCO's SRCR domain. We confirmed our results with a Likelihood Ratio Test (ΔLRT) at a 95% significance level. The sites identified as under positive selection were 442, 452, and 477 (Table 4.2). In humans, these sites correspond to tryptophan, glutamine, and valine respectively. Sites 442 and 452 are of particular interest because of their close proximity to the RGRAEVYY and WGTICDD motifs of MARCO (431- 438, 442-448 in mouse respectively) (Figure 4.5). Tryptophan 442 corresponds to the first residue of the WGTICDD motif, and could be an indication of positive selection acting on this motif. Due to the high conservation of this motif across all of the SRCR domains, this suggests a potential biological function for T442 and Q452. We also identified position 477 as possibly being under positive selection. Although position 477 is not within close proximity to any of the known motifs, the site had a relatively high BEB score and may have some uncharacterized biological function.
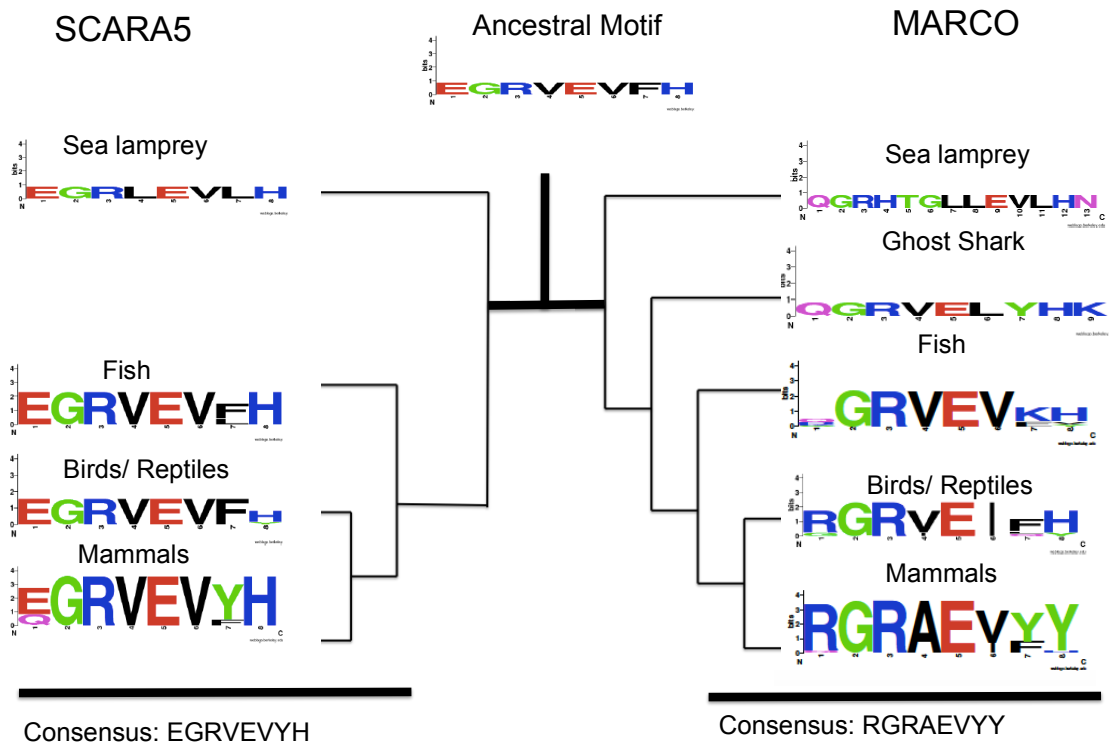
FIGURE 4.3: **Evolution of a EGRVEVYH motif within SCARA5 and MARCO's RGRAEVYY motif among different taxa.** Taxa groups shown include mammals, birds and reptiles, fish, ghost shark, and sea lamprey. The ancestral sequence predicted from FastML is shown as a weblogo. This SCARA5 motif is highly conserved across taxa but MARCO's motif is less conserved outside of mammals.
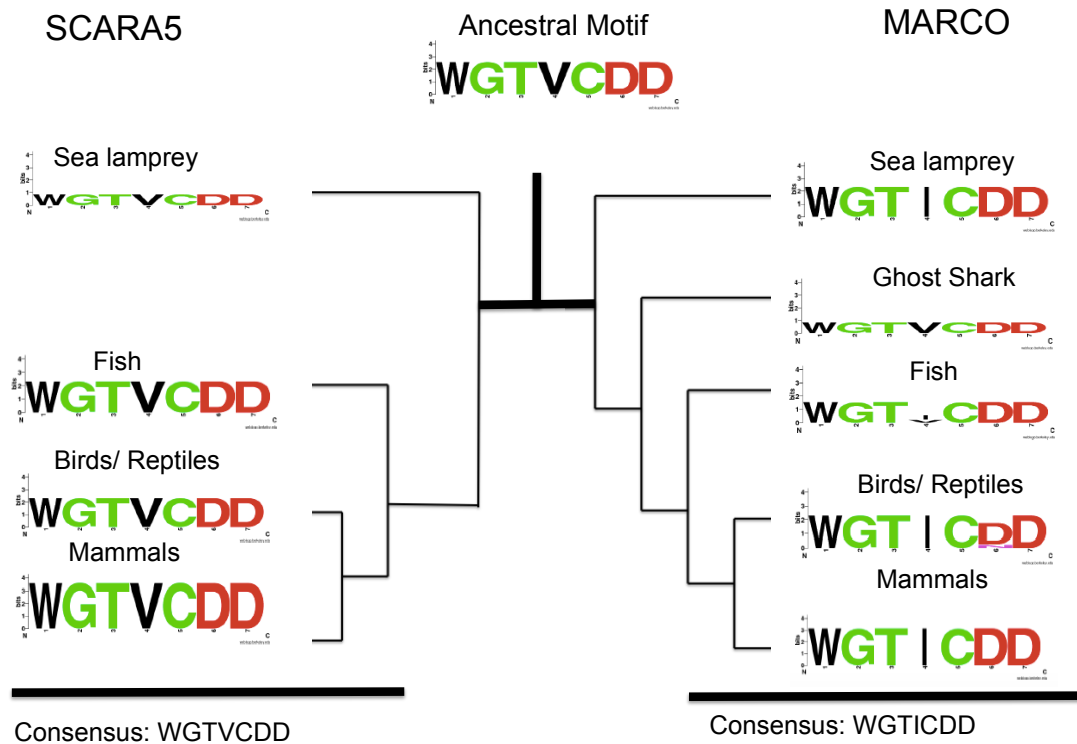
FIGURE 4.4: **Analysis of SCARA5's WGTVCDD motif and MARCO's WGTICDD motif within different taxa.** Taxa groups shown include mammals, birds and reptiles, fish, ghost shark, and sea lamprey. The ancestral sequence predicted from FastML is shown. The WGTVCDD motif is conserved between the ancestral proteins and within both SCARA5 and MARCO except at the valine residue (site 4 in the motif).

TABLE 4.2: **Identified sites putatively under positive selection within MARCO's SRCR domain from PAML**

Site identified with respect to full length human MARCO	Bayes Empirical Bayes (BEB) score
W 442	0.700
Q 452	0.976
V 477	0.861
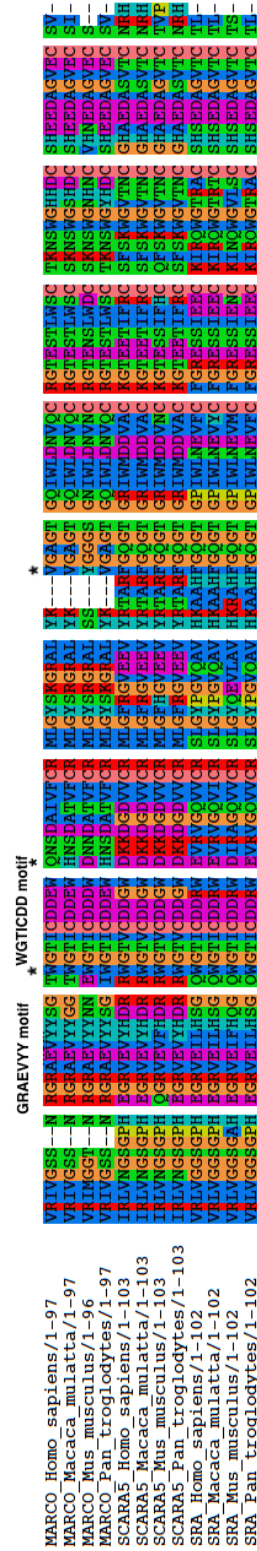
FIGURE 4.5: Alignment of the SRCR domains studied for positive selection test in PAML. The GRAEVYY motif and WGTICDD motif, found in MARCO, are labeled. Stars denote positions 442, 452, and 477 as sites identified as under positive selection from PAML.

4.4   Discussion

The Class A Scavenger Receptor family is a diverse group of Pattern Recognition Receptors involved in innate immunity. Previous work has suggested that the five family members were generated through multiple duplication events, however several questions remained unanswered. It was unclear whether MARCO shared a more recent common ancestor with the SRCR-containing receptors (SCARA5 and SR-A), or with the non-SRCR-containing receptors (SCARA3 and SCARA4) (Whelan et al., 2012). In addition, the ancestral domains were also unresolved because of the uncertainty regarding MARCO's relationship to the other cA-SRs (Whelan et al., 2012). Here, we present new phylogenetic data to resolve these uncertainties within the scavenger receptor family.

Using Bayesian methods, we generated a new phylogenetic tree of all five cA-SRs. Our phylogenetic tree shows that MARCO shares a more recent common ancestor with SCARA5 and SR-A than with SCARA3 and SCARA4. The addition of an outgroup sequence, colmedin from *Strongylocentrotus purpuratus*, also shows the same relationship. This suggests that an ancestral cA-SR containing an SRCR domain was duplicated to produce an ancestral MARCO and a SCARA5/SR-A like precursor. Following this duplication event, the SCARA5/SR-A like precursor underwent a duplication event to produce modern SCARA5 and SR-A sequences. Based on the available sequence data, it still remains unknown if the ancestral gene to all five cA-SRs lacked or contained the SRCR domain. It is possible that the ancestral gene terminated at its collagenous domain and resembled SCARA3 and later acquired the SRCR domain in the MARCO/SCARA5/SR-A precursor. Equally likely, the ancestral gene may have contained the SRCR domain and it was lost in a SCARA3/SCARA4 precursor. Additional sequence data are required to fully uncover the origin of these proteins.

To investigate the origin of the SRCR domains, we utilized Bayesian methods to construct a phylogenetic tree and subsequently reconstructed ancestral SRCR domains. Based on our analysis, MARCO shares a recent common ancestor with SCARA5 and SR-A. Using FastML we reconstructed the SRCR domains at ancestral nodes between SCARA5, MARCO, and SR-A. We focused on two motifs within SCARA5 and MARCO; the RGRAEVYY motif within MARCO and a downstream WGTICDD motif. SCARA5 possesses two motifs similar to these, as it contains a EGRVEVYH motif and a WGTVCDD motif. Based on our multiple sequence alignment, the RGRAEVYY motif is specific to MARCO while the WGTVCDD motif is shared between the three SRCR domains. Furthermore, the ancestral motif to all three cA-SR SRCR domains resembled SCARA5's EGRVEVYH and WGTVCDD motifs. This suggests that MARCO's SRCR domain originally resembled SCARA5, and may have undergone purifying selection. Recent data has shown the EGR-E residues of MARCO and SR-A may have a role in amyloid-β peptide internalization and surface trafficking (Tsay et al., 2016). Since SCARA5 and the ancestral domain lack the RGRAEVYY motif, we hypothesize the ancestral SRCR domain did not play a functional role in bacterial binding due to the functional importance of the RGRAEVYY motif within MARCO. In invertebrate

species, the SRCR domains are thought to play a role in cellular recognition as opposed to ligand binding (Bowdish and Gordon, 2009). This theory may also apply to the original SRCR domain of the cA-SRs.

Given MARCO's role in bacterial binding and clearance within the immune system, we hypothesized that its SRCR domain may be under positive selection. Here we have shown that MARCO has several sites under positive selection including positions 442, 452, and 477. Two of the sites are adjacent to the highly conserved WGTVCDD motif within the SRCR domain, and may have a biological function. However, we did not detect positive selection acting on the RGRAEVYY motif, despite the domain having a known role in ligand binding. Future experiments will look to identify the functional relevance of the positively selected sites and the WGTVCDD motif. We hypothesize that the WGTVCDD motif may have a structural role within the SRCR domain due to the inclusion of a conserved cysteine within the motif. Studying the WGTICDD motif will further our understanding of the SRCR domain and its biological relevance in MARCO's ability to facilitate bacterial binding, phagocytosis, and induction of T-cell tolerance.

Chapter 5

Discussion

5.1 Discussion

Macrophage biology and function have been extensively characterized using various *in vitro* and *in vivo* models. These models have been used to study various components of macrophage biology including LPS tolerance, predictions of functional motifs within receptors, as well as the immune response to various bacterial infections such as *S. pneumonaie*. However, these models can be limited by the cost of the experiments or sample size. Bioinformatics is an easily accessible tool that that can circumvent these problems. This work utilizes a bioinformatic approach to investigate macrophage biology and function.

Using a previously established model of LPS tolerance in Foster, Hargreaves, and Medzhitov (2007) and Fei et al. (2016), I studied the effect of age on the tolerance response. Using macrophages derived from old and young mice, I analyzed differences in gene expression between the two age groups in LPS induction and LPS tolerance. Surprisingly, there was no difference in genes involved in the tolerance response between the two age groups. However, macrophages from young mice had a higher induction of LPS genes than their old counterparts, except for MARCO. These findings suggest that the ability to induce tolerance is not impaired with age, however the induction of the response to bacterial products is dysfunctional. These data can be used to guide future experiments to determine the mechanisms in establishing tolerance.

Studies from the Bowdish lab have shown that MARCO enhances the response of TLR2 and CD14 to bacterial ligands such as *S. pneumoniae* (Dorrington and Bowdish, 2013; Novakowski et al., 2016). In addition, this enhancement is specific to MARCO and relies on the presence of its SRCR domain (Dorrington and Bowdish, 2013; Novakowski et al., 2016). Based on these observations, I propose that MARCO, TLR2, and CD14 have co-evolved in order to bind human adapted pathogens such as *S. pneumoniae*. The correlation coefficients between MARCO, TLR2, and CD14 were compared to a model of 40 non co-evolving proteins. This analysis showed that MARCO, TLR2,

and CD14 have co-evolved. In addition, I identified potential functional SNPs within MARCO. Using data from The 1000 Genomes Project, The Great Apes Project, and the Max Planck Institute for Evolutionary Anthropology, I searched for SNPs within MARCO across various species. This lead to the identification of a SNP causing a non-synonymous substitution at amino acid position 282 within humans. Combined with data from the Great Apes Project, position 282 within humans appears to have two variants: the primary variant F282, and the ancestral variant S282. The primary variant F282 is specific to modern humans, whereas the ancestral variant S282 is common between hominids. These data suggest that the primary F282 variant may have been under a selective pressure within modern humans, and suggests a potential functional role. Since the F282 primary variant is specific to humans, it may have a role in binding a human specific pathogen or ligand. Future studies to determine the functional role of F282 and S282, as well as the positive selection sites are ongoing.

Previous work has shown that the SRCR domain of MARCO contains residues and motifs important for ligand binding (Brännström et al., 2002; Whelan et al., 2012; Novakowski et al., 2016; Lee et al., 2016). Therefore, I identified additional areas for ligand binding within MARCO by searching for sites under positive selection. Using the program PAML, three residues within the SRCR domain of MARCO were identified. One of these sites falls within close proximity to two previously identified functional motifs within the SRCR domain, and may have a role in ligand binding. Future experiments such as site-directed mutagenesis, can be used to test these predictions.

In addition, although the cA-SRs are evolutionarily related, the origin of the SRCR domain within these proteins remains unclear. Although SR-A possess an SRCR domain, it does not require the domain to bind ligands (Krieger, 1992). Using bioinformatics, I studied the evolution of the SRCR domain within the cA-SRs and its origin. The ancient evolution of the cA-SR families was conducted using the genomes from various organisms ranging from coelacanth, and elephant shark to birds, mammals, and others. By generating phylogenetic trees, I show that the SRCR domains of MARCO, SR-A and SCARA5 originated from one common origin. Furthermore, I investigated two motifs within the SRCR domain: the GRAEVYY motif and the WGTICDD motif. The GRAEVYY motif has been shown to contain residues required for ligand binding (Brännström et al., 2002). When we compare this domain in MARCO proteins, the motif resembles the SCARA5 and SR-A form within more divergent species such as the coelacanth and sea lamprey. Since SR-A is known to bind ligands through the use of its collagenous domain, the SRCR domain of MARCO may have evolved to specifically bind bacterial ligands. These findings suggest that the SRCR domain is an ancient part of the innate immune system.

Appendix A

Chapter 2 Supplementary Data

TABLE A.1: **Differentially Expressed Genes Involved In Tolerance In Young Macrophages**

Gene Name	BaseMean	Log2FoldChange	lfcSE	Stat	p-value	padj
1700113A16Rik	1.919	2.737	0.912	3.002	2.682E-03	1.389E-02
2210009G21Rik	77.115	1.084	0.26	4.173	3.010E-05	2.596E-04
4831426I19Rik	35.934	-2.165	0.372	-5.813	6.130E-09	9.380E-08
4930578N16Rik	3.835	2.127	0.818	2.6	9.333E-03	3.930E-02
5730508B09Rik	112.798	-2.295	0.221	-10.393	2.660E-25	1.840E-23
9030425E11Rik	63.273	0.866	0.254	3.414	6.393E-04	4.000E-03
A230050P20Rik	10.542	-2.078	0.511	-4.068	4.740E-05	3.895E-04
Acsl1	3814.117	0.4	0.127	3.157	1.593E-03	8.850E-03
Aebp2	216.525	-1.161	0.224	-5.182	2.200E-07	2.800E-06
Bcl3	21.98	-0.975	0.386	-2.527	1.149E-02	4.629E-02
Ccl22	27.223	-1.046	0.385	-2.716	6.601E-03	2.954E-02
Cd200	173.035	-1.626	0.21	-7.733	1.050E-14	3.090E-13
Cd38	233.294	2.574	0.236	10.886	1.340E-27	1.080E-25
Col18a1	180.396	-1.78	0.279	-6.373	1.850E-10	3.390E-09
Cp	180.755	-0.939	0.224	-4.188	2.810E-05	2.457E-04
Ctnnal1	2.337	-2.075	0.815	-2.546	1.089E-02	4.439E-02
Cxcl1	130.387	-2.817	0.252	-11.198	4.150E-29	3.650E-27
D17Wsu92e	244.898	-1.236	0.166	-7.445	9.710E-14	2.610E-12
Daam1	134.408	-1.054	0.253	-4.174	2.990E-05	2.589E-04
Dgka	21.192	-2.636	0.451	-5.848	4.990E-09	7.700E-08
Dst	720.508	0.597	0.204	2.927	3.422E-03	1.702E-02
Dusp2	26.966	-4.348	0.439	-9.896	4.320E-23	2.660E-21
E330016A19Rik	34.844	-1.657	0.303	-5.468	4.560E-08	6.320E-07
Etv3	169.622	-1.719	0.222	-7.757	8.720E-15	2.600E-13
F10	471.151	-0.687	0.164	-4.199	2.680E-05	2.347E-04
Fancc	21.035	-1.141	0.375	-3.038	2.378E-03	1.255E-02
Fkbp5	12.905	-1.563	0.447	-3.495	4.741E-04	3.075E-03
Fos	165.73	0.87	0.231	3.767	1.653E-04	1.199E-03
Gbp4	123.244	-1.183	0.232	-5.091	3.560E-07	4.420E-06

Glipr2	30.38	-0.957	0.32	-2.995	2.746E-03	1.417E-02
Gyk	178.308	-1.353	0.244	-5.55	2.850E-08	4.070E-07
H2-T24	193.983	0.794	0.183	4.327	1.510E-05	1.401E-04
Hbegf	4.976	-1.794	0.69	-2.602	9.280E-03	3.911E-02
Homer1	45.375	-1.445	0.3	-4.821	1.430E-06	1.630E-05
Id3	90.396	-0.808	0.308	-2.62	8.783E-03	3.732E-02
Igsf6	1162.51	-0.625	0.172	-3.625	2.894E-04	1.980E-03
Igsf9	8.976	-1.576	0.506	-3.115	1.840E-03	1.005E-02
Iqsec2	28.93	-1.071	0.358	-2.989	2.799E-03	1.439E-02
Irak3	242.654	0.462	0.172	2.69	7.156E-03	3.152E-02
Lcn2	54.912	1.326	0.289	4.581	4.620E-06	4.710E-05
Lmna	665.441	-0.399	0.157	-2.55	1.076E-02	4.396E-02
Lrrfip1	306.914	-1.081	0.154	-7.01	2.390E-12	5.490E-11
Ly6a	24.031	2.033	0.438	4.641	3.470E-06	3.660E-05
Mafk	106.408	-2.265	0.218	-10.4	2.480E-25	1.730E-23
Met	659.734	-0.645	0.197	-3.27	1.077E-03	6.302E-03
Mpp7	17.87	-1.053	0.4	-2.632	8.487E-03	3.631E-02
Ms4a4c	474.594	1.603	0.286	5.6	2.140E-08	3.060E-07
Ms4a6d	1482.429	0.951	0.222	4.289	1.790E-05	1.634E-04
Mthfd2	94.662	0.917	0.228	4.015	5.930E-05	4.749E-04
Mxd1	100.124	-1.554	0.215	-7.229	4.860E-13	1.190E-11
Myo10	449.769	-1.643	0.244	-6.737	1.620E-11	3.350E-10
Myst3	146.768	-0.945	0.242	-3.911	9.200E-05	7.047E-04
Ncoa7	83.455	-0.801	0.28	-2.858	4.260E-03	2.044E-02
Notch1	113.743	-1.206	0.285	-4.236	2.270E-05	2.025E-04
Nupr1	69.563	-1.032	0.216	-4.775	1.790E-06	1.990E-05
Odc1	66.079	-1.143	0.249	-4.593	4.370E-06	4.490E-05
Orm1	1.927	2.942	0.899	3.272	1.069E-03	6.261E-03
Otud1	39.943	-1.931	0.291	-6.646	3.020E-11	6.040E-10
Phldb1	89.603	-2.686	0.318	-8.451	2.880E-17	1.080E-15
Pik3ap1	607.869	-0.596	0.164	-3.639	2.733E-04	1.878E-03
Pla2g4a	101.875	-0.996	0.244	-4.089	4.330E-05	3.612E-04
Pou3f1	11.629	-1.261	0.503	-2.509	1.212E-02	4.805E-02
Ppp3cc	25.276	-1.646	0.341	-4.833	1.350E-06	1.540E-05
Prpf4	50.399	-1.09	0.257	-4.246	2.180E-05	1.947E-04
Ptges	273.455	1.226	0.171	7.176	7.180E-13	1.730E-11
Ptx3	14.675	-1.895	0.451	-4.201	2.660E-05	2.332E-04
Rel	60.059	-1.545	0.339	-4.563	5.050E-06	5.100E-05
Rnf125	0.93	-2.8	0.932	-3.004	2.662E-03	1.380E-02
Sdccag8	68.08	0.774	0.289	2.682	7.311E-03	3.214E-02
Serpina3g	22.472	-1.324	0.394	-3.362	7.725E-04	4.711E-03
Slamf9	136.507	0.653	0.241	2.712	6.696E-03	2.984E-02
Slc12a4	150.517	-1.385	0.199	-6.971	3.160E-12	7.090E-11
Slc13a3	50.736	1.92	0.29	6.625	3.480E-11	6.900E-10
Slc16a1	26.767	-1.208	0.326	-3.707	2.096E-04	1.484E-03
Slc31a1	381.42	0.549	0.139	3.942	8.070E-05	6.240E-04
Slco3a1	104.663	-1.125	0.267	-4.212	2.530E-05	2.225E-04
Slfn4	247.947	1.142	0.255	4.483	7.350E-06	7.210E-05
Smad6	26.425	2.306	0.38	6.063	1.330E-09	2.200E-08
Snx10	640.936	-1.393	0.229	-6.073	1.260E-09	2.090E-08

Spata13	202.318	-1.211	0.192	-6.315	2.700E-10	4.820E-09
Sphk1	4.595	-2.663	0.759	-3.507	4.535E-04	2.961E-03
Spic	22.302	1.729	0.409	4.232	2.320E-05	2.054E-04
St6galnac4	39.232	-0.849	0.291	-2.92	3.498E-03	1.732E-02
Tal1	26.664	-1.142	0.348	-3.281	1.033E-03	6.083E-03
Tank	413.526	-1.043	0.252	-4.134	3.560E-05	3.034E-04
Tgfbi	2226.515	2.03	0.14	14.534	7.410E-48	1.740E-45
Tle3	24.151	-1.68	0.393	-4.275	1.910E-05	1.729E-04
Tmco3	269.031	-0.817	0.141	-5.812	6.180E-09	9.440E-08
Tpbg	5.843	-1.545	0.583	-2.648	8.101E-03	3.490E-02
Traf1	406.025	-0.538	0.196	-2.747	6.020E-03	2.734E-02
Trim21	405.57	-2.141	0.171	-12.509	6.680E-36	8.380E-34
Trim26	112.78	-1.604	0.233	-6.891	5.550E-12	1.210E-10
Wdr59	18.181	-1.26	0.373	-3.378	7.312E-04	4.498E-03

TABLE A.2: **Differentially Expressed Genes Involved In Tolerance In Old Macrophages**

Gene Name	BaseMean	Log2FoldChange	lfcSE	Stat	p-value	padj
9330175E14Rik	7.922	-2.151	0.543	-3.963	7.410E-05	9.457E-04
Arid5b	67.63	-1.031	0.304	-3.396	6.841E-04	5.979E-03
Chd1	222.072	-0.902	0.345	-2.614	8.950E-03	4.439E-02
Cnn3	1.706	-3.961	0.948	-4.179	2.930E-05	4.374E-04
Eng	58.792	-1.587	0.4	-3.969	7.230E-05	9.277E-04
Errfi1	28.209	-1.441	0.426	-3.378	7.296E-04	6.262E-03
Fmnl2	113.595	-0.879	0.197	-4.471	7.800E-06	1.324E-04
Gja1	11.829	-2.479	0.766	-3.235	1.215E-03	9.303E-03
Gosr1	115.693	-0.6	0.229	-2.616	8.885E-03	4.415E-02
Gspt1	266.234	-0.57	0.199	-2.872	4.082E-03	2.397E-02
Gsta3	4.262	2.657	0.876	3.034	2.417E-03	1.602E-02
Il1b	2847.482	-1.858	0.525	-3.537	4.041E-04	3.891E-03
Irf2	28.751	-0.834	0.307	-2.717	6.591E-03	3.501E-02
Sertad3	22.056	-1.111	0.332	-3.348	8.126E-04	6.791E-03
Slc25a37	117.083	-0.613	0.193	-3.169	1.528E-03	1.108E-02
Zdhhc18	34.664	-1.177	0.34	-3.462	5.366E-04	4.935E-03

TABLE A.3: **Differentially Expressed Genes During Single-dose Short Stimulation In Young Macrophages**

Gene Name	BaseMean	Log2FoldChange	lfcSE	Stat	p-value	padj

2510009E07Rik	67.614	-0.964	0.261	-3.694	2.211E-04	1.050E-03
A230050P20Rik	10.542	1.309	0.45	2.909	3.621E-03	1.268E-02
Aebp2	216.525	2.441	0.227	10.762	5.220E-27	1.360E-25
App	705.675	0.892	0.212	4.199	2.680E-05	1.527E-04
Ccl22	27.223	5.147	0.586	8.777	1.680E-18	2.910E-17
Ccl8	3.624	2.805	0.863	3.252	1.144E-03	4.591E-03
Col18a1	180.396	2.134	0.275	7.76	8.490E-15	1.210E-13
Cxcl1	130.387	8.467	0.577	14.68	8.620E-49	4.820E-47
Dusp2	26.966	4.786	0.44	10.89	1.290E-27	3.470E-26
E330016A19Rik	34.844	2.774	0.322	8.609	7.390E-18	1.250E-16
Elk3	90.119	1.003	0.233	4.313	1.610E-05	9.490E-05
Etf1	359.964	0.942	0.212	4.448	8.660E-06	5.330E-05
Etv3	169.622	1.113	0.21	5.308	1.110E-07	8.920E-07
F11r	8.802	2.588	0.544	4.757	1.970E-06	1.340E-05
Fabp7	33.492	1.031	0.34	3.032	2.431E-03	8.965E-03
Fancc	21.035	0.992	0.354	2.804	5.054E-03	1.689E-02
Fkbp5	12.905	2.198	0.452	4.865	1.150E-06	8.080E-06
Fos	165.73	-0.647	0.227	-2.855	4.307E-03	1.474E-02
Foxp1	103.904	0.992	0.186	5.339	9.370E-08	7.590E-07
Gng4	3.333	1.851	0.762	2.43	1.512E-02	4.272E-02
H2-D1	2316.508	0.599	0.161	3.711	2.063E-04	9.870E-04
Hbegf	4.976	1.765	0.66	2.673	7.520E-03	2.371E-02
Lmna	665.441	-0.517	0.152	-3.404	6.630E-04	2.799E-03
Lrrfip1	306.914	0.522	0.146	3.575	3.501E-04	1.581E-03
Luzp1	158.134	0.74	0.191	3.877	1.059E-04	5.408E-04
Mmd	32.404	0.881	0.305	2.888	3.878E-03	1.344E-02
Mpp7	17.87	1.117	0.383	2.916	3.544E-03	1.244E-02
Noc4l	23.534	1.448	0.331	4.374	1.220E-05	7.300E-05
Pou3f1	11.629	2.051	0.51	4.023	5.750E-05	3.088E-04
Prkrip1	25.198	1.07	0.425	2.519	1.177E-02	3.454E-02
Psme4	338.142	0.733	0.207	3.551	3.841E-04	1.721E-03
S100a10	109.147	0.593	0.217	2.728	6.366E-03	2.053E-02
Sash1	526.838	0.69	0.233	2.966	3.016E-03	1.084E-02
Sema4c	14.051	1.932	0.433	4.463	8.080E-06	5.000E-05
Slc16a3	122.797	1.586	0.301	5.276	1.320E-07	1.050E-06
Slc31a1	381.42	2.013	0.152	13.217	7.010E-40	2.920E-38
Slc8a1	200.738	-0.439	0.17	-2.581	9.841E-03	2.972E-02
Smad6	26.425	-1.078	0.388	-2.774	5.531E-03	1.824E-02
Tgfbi	2226.515	0.901	0.143	6.318	2.640E-10	2.770E-09
Tle3	24.151	1.645	0.373	4.404	1.060E-05	6.450E-05
Trim26	112.78	2.159	0.23	9.386	6.250E-21	1.210E-19
Znrf3	12.435	1.153	0.424	2.721	6.511E-03	2.093E-02

TABLE A.4: **Differentially Expressed Genes During Single-dose Short Stimulation In Old Macrophages**

Gene Name	BaseMean	Log2FoldChange	lfcSE	Stat	p-value	padj
Ctnnal1	1.689	2.578	0.929	2.777	5.492E-03	1.891E-02
Eng	58.792	1.098	0.374	2.934	3.346E-03	1.238E-02
Hmgn3	23.876	1.287	0.335	3.846	1.202E-04	6.742E-04
Irf2	28.751	1.147	0.292	3.930	8.483E-05	4.959E-04
Ly6i	14.158	2.463	0.849	2.901	3.720E-03	1.359E-02
Ogfrl1	71.680	0.569	0.193	2.952	3.155E-03	1.178E-02
Pdzk1ip1	4.672	1.295	0.542	2.390	1.685E-02	4.840E-02
Plagl1	2.708	3.345	0.876	3.816	1.354E-04	7.509E-04
Rab4a	6.652	-3.020	0.768	-3.934	8.368E-05	4.899E-04
Spic	18.671	1.915	0.556	3.443	5.744E-04	2.670E-03
Traf3ip2	22.397	0.963	0.394	2.445	1.449E-02	4.277E-02

TABLE A.5: **Differentially Expressed Genes During Single-dose Long Stimulation In Young Macrophages**

Gene Name	BaseMean	Log2FoldChange	lfcSE	Stat	p-value	padj
1200009I06Rik	24.303	4.167	0.566	7.363	1.803E-13	3.610E-12
2010109K11Rik	275.891	0.901	0.311	2.898	3.756E-03	1.369E-02
3110001I22Rik	16.289	2.210	0.346	6.385	1.709E-10	2.463E-09
4932438A13Rik	356.745	1.070	0.257	4.162	3.158E-05	2.028E-04
9330175E14Rik	7.922	2.284	0.509	4.484	7.330E-06	5.382E-05
A630072M18Rik	9.840	1.542	0.555	2.781	5.426E-03	1.874E-02
Adamts4	39.613	7.242	0.753	9.619	6.624E-22	2.644E-20
Aff1	138.906	1.909	0.170	11.227	3.012E-29	1.966E-27
Angpt1	1.293	2.653	0.937	2.830	4.648E-03	1.644E-02
Arhgef3	242.127	3.425	0.381	8.987	2.549E-19	8.362E-18
Arid5a	37.815	2.638	0.282	9.354	8.427E-21	3.131E-19
Arid5b	67.630	0.841	0.284	2.964	3.037E-03	1.139E-02
Armc8	114.200	2.115	0.259	8.175	2.955E-16	7.706E-15
Asb13	16.375	1.613	0.369	4.377	1.201E-05	8.502E-05
Atp11b	79.692	0.665	0.267	2.487	1.288E-02	3.872E-02
Atp9b	228.669	0.714	0.181	3.933	8.385E-05	4.904E-04
B4galt5	265.455	1.975	0.222	8.913	4.957E-19	1.596E-17
BC016423	209.272	2.715	0.343	7.920	2.373E-15	5.629E-14
Bcl3	28.521	2.563	0.603	4.249	2.149E-05	1.443E-04
Bfar	133.741	1.883	0.299	6.306	2.871E-10	4.025E-09
Cacnb3	20.539	4.536	0.685	6.626	3.449E-11	5.392E-10
Camk2d	265.347	1.857	0.167	11.113	1.082E-28	6.878E-27
Casp12	3.726	3.578	0.800	4.471	7.802E-06	5.711E-05
Ccrn4l	34.117	3.893	0.345	11.294	1.400E-29	9.385E-28
Cd200	124.136	2.792	0.806	3.465	5.311E-04	2.501E-03

Cds1	24.825	1.459	0.364	4.010	6.075E-05	3.651E-04
Cenpj	30.652	1.802	0.291	6.195	5.836E-10	7.919E-09
Chd1	222.072	1.216	0.339	3.591	3.300E-04	1.643E-03
Chst11	38.887	2.146	0.393	5.455	4.893E-08	5.203E-07
Cnn3	1.706	4.184	0.944	4.430	9.414E-06	6.771E-05
Col27a1	24.455	5.577	0.896	6.226	4.796E-10	6.580E-09
Ctnnal1	1.689	2.578	0.929	2.777	5.492E-03	1.891E-02
Cycs	3.814	1.980	0.699	2.831	4.639E-03	1.641E-02
D17Wsu92e	273.019	1.141	0.469	2.431	1.505E-02	4.412E-02
D1Ertd622e	116.422	1.499	0.300	4.995	5.891E-07	5.297E-06
Daam1	137.431	2.507	0.374	6.710	1.946E-11	3.154E-10
Dcbld2	251.674	3.519	0.518	6.799	1.056E-11	1.745E-10
Dcp1a	76.432	1.040	0.311	3.340	8.370E-04	3.728E-03
Dgkh	39.938	1.595	0.280	5.691	1.260E-08	1.448E-07
Dll1	4.697	4.453	0.793	5.618	1.930E-08	2.165E-07
Dnaja2	314.121	1.977	0.179	11.053	2.125E-28	1.311E-26
Dock10	899.772	1.506	0.618	2.437	1.481E-02	4.352E-02
Dusp14	1.697	3.929	0.956	4.108	3.983E-05	2.503E-04
Dusp16	113.783	3.833	0.494	7.755	8.867E-15	1.981E-13
Edn1	43.938	7.983	0.771	10.349	4.229E-25	2.138E-23
Eng	58.792	1.098	0.374	2.934	3.346E-03	1.238E-02
Errfi1	28.209	1.908	0.409	4.665	3.080E-06	2.439E-05
Etv6	108.583	2.488	0.277	8.991	2.446E-19	8.045E-18
Fabp3	9.816	2.015	0.613	3.286	1.016E-03	4.430E-03
Fez2	115.693	0.695	0.188	3.690	2.242E-04	1.171E-03
Fmr1	101.016	2.290	0.250	9.173	4.582E-20	1.596E-18
Fnbp4	151.252	1.576	0.252	6.252	4.040E-10	5.580E-09
Frmd4a	38.157	1.037	0.279	3.711	2.063E-04	1.088E-03
Fscn1	20.321	5.777	0.723	7.991	1.336E-15	3.270E-14
Fzd1	129.333	4.405	0.220	19.984	7.586E-89	1.170E-85
Gja1	11.829	3.960	0.778	5.091	3.558E-07	3.303E-06
Gna15	65.626	2.582	0.302	8.556	1.171E-17	3.424E-16
Gnb4	73.245	2.223	0.320	6.944	3.820E-12	6.637E-11
Golga3	285.070	2.006	0.265	7.566	3.859E-14	8.180E-13
Gosr1	115.693	0.851	0.218	3.898	9.708E-05	5.585E-04
Gpr126	39.976	2.531	0.388	6.531	6.515E-11	9.824E-10
Gpsm2	75.800	0.906	0.353	2.568	1.024E-02	3.194E-02
Gypc	16.189	1.372	0.385	3.560	3.712E-04	1.823E-03
H2-Q8	4.878	2.214	0.848	2.611	9.036E-03	2.881E-02
Hivep2	26.865	2.970	0.733	4.051	5.102E-05	3.122E-04
Homer1	38.413	1.064	0.376	2.828	4.684E-03	1.654E-02
Ifi202b	1.202	2.522	0.989	2.550	1.078E-02	3.332E-02
Igsf6	1047.643	2.026	0.575	3.521	4.295E-04	2.074E-03
Igsf9	8.870	2.335	0.820	2.847	4.419E-03	1.571E-02
Il10	13.031	3.887	0.591	6.580	4.716E-11	7.235E-10
Il13ra1	89.754	2.637	0.507	5.203	1.959E-07	1.884E-06
Il15	180.691	3.984	0.212	18.783	1.048E-78	1.077E-75
Iqsec2	29.052	1.626	0.472	3.446	5.681E-04	2.645E-03
Irf2	28.751	1.147	0.292	3.930	8.483E-05	4.959E-04
Itga4	700.914	1.364	0.421	3.244	1.179E-03	5.061E-03

Itgav	778.301	2.988	0.379	7.891	3.001E-15	6.998E-14
Jarid2	71.214	1.430	0.221	6.479	9.241E-11	1.377E-09
Junb	506.102	2.470	0.316	7.809	5.751E-15	1.306E-13
Katna1	97.721	2.236	0.254	8.806	1.301E-18	4.082E-17
Klf7	15.534	3.367	0.467	7.205	5.804E-13	1.085E-11
Kpna3	601.052	3.174	0.621	5.114	3.152E-07	2.943E-06
Kpna4	310.368	1.731	0.129	13.436	3.726E-41	4.942E-39
Kremen1	119.316	2.815	0.385	7.313	2.617E-13	5.141E-12
Ktn1	167.034	1.951	0.229	8.503	1.853E-17	5.290E-16
Lancl2	100.332	1.859	0.441	4.211	2.540E-05	1.672E-04
Lass6	776.048	3.082	0.427	7.215	5.376E-13	1.009E-11
Lck	1.982	3.388	0.883	3.837	1.248E-04	6.977E-04
Lhx2	6.404	4.791	0.736	6.511	7.481E-11	1.121E-09
Lnp	61.207	2.068	0.335	6.182	6.342E-10	8.549E-09
Lphn2	102.220	2.698	0.427	6.315	2.705E-10	3.809E-09
Lrch1	88.323	2.151	0.216	9.956	2.382E-23	1.053E-21
Mafk	111.376	1.666	0.464	3.591	3.293E-04	1.640E-03
Map2k4	146.237	1.570	0.331	4.744	2.100E-06	1.734E-05
Map3k8	136.430	2.353	0.283	8.307	9.838E-17	2.673E-15
Mapkbp1	128.582	3.538	0.301	11.755	6.624E-32	5.139E-30
Minpp1	79.593	0.979	0.183	5.355	8.572E-08	8.848E-07
Mt1	126.189	1.095	0.356	3.072	2.128E-03	8.400E-03
Mxd1	120.224	2.243	0.571	3.927	8.595E-05	5.017E-04
Myadm	231.442	2.400	0.304	7.895	2.909E-15	6.822E-14
Myo10	703.920	2.746	0.674	4.074	4.612E-05	2.850E-04
Myst3	155.437	1.099	0.317	3.463	5.347E-04	2.515E-03
Ncoa7	70.016	2.027	0.452	4.484	7.321E-06	5.382E-05
Nfil3	18.585	1.802	0.580	3.107	1.890E-03	7.602E-03
Nfkb1	838.684	3.501	0.215	16.307	8.795E-60	3.013E-57
Nfxl1	101.897	2.252	0.397	5.677	1.369E-08	1.568E-07
Notch1	135.095	1.633	0.553	2.951	3.163E-03	1.180E-02
Nr3c1	218.799	1.692	0.547	3.096	1.964E-03	7.849E-03
Odc1	67.524	1.769	0.352	5.019	5.201E-07	4.728E-06
Ogfrl1	71.680	0.569	0.193	2.952	3.155E-03	1.178E-02
Otud1	42.589	1.186	0.456	2.601	9.299E-03	2.946E-02
Parp8	88.189	2.024	0.337	6.004	1.924E-09	2.412E-08
Pdzk1ip1	4.672	1.295	0.542	2.390	1.685E-02	4.840E-02
Phc2	89.274	1.238	0.304	4.078	4.547E-05	2.815E-04
Phip	277.936	0.725	0.300	2.419	1.556E-02	4.526E-02
Phldb1	130.657	3.035	0.705	4.305	1.667E-05	1.144E-04
Pik3ap1	715.178	1.575	0.447	3.525	4.243E-04	2.050E-03
Pim1	194.736	4.207	0.450	9.350	8.804E-21	3.251E-19
Plagl1	2.708	3.345	0.876	3.816	1.354E-04	7.509E-04
Plekha2	326.654	0.833	0.143	5.819	5.913E-09	7.027E-08
Plekha4	5.241	3.297	0.797	4.137	3.517E-05	2.235E-04
Plekhf2	127.995	1.657	0.311	5.321	1.033E-07	1.053E-06
Ppap2b	33.084	2.474	0.584	4.233	2.305E-05	1.535E-04
Ppfibp1	337.930	2.274	0.161	14.114	3.104E-45	5.318E-43
Ppm1h	127.425	-0.613	0.199	-3.085	2.037E-03	8.089E-03
Ppm1k	88.804	3.222	0.331	9.746	1.928E-22	7.954E-21

Ppp1r15b	229.380	2.082	0.190	10.975	5.029E-28	3.056E-26
Ppp3cc	19.828	1.936	0.381	5.076	3.848E-07	3.558E-06
Prkx	113.160	1.114	0.265	4.209	2.562E-05	1.684E-04
Prpf4	48.086	0.648	0.242	2.679	7.391E-03	2.425E-02
Ptprj	1568.435	2.578	0.253	10.184	2.328E-24	1.117E-22
Ptx3	11.317	4.471	0.859	5.205	1.945E-07	1.871E-06
Rab32	505.381	1.753	0.637	2.751	5.934E-03	2.017E-02
Rab4a	6.652	-3.020	0.768	-3.934	8.368E-05	4.899E-04
Ranbp2	963.318	1.581	0.569	2.778	5.469E-03	1.886E-02
Rap2c	182.516	3.068	0.306	10.009	1.389E-23	6.275E-22
Rbl1	78.559	1.986	0.195	10.167	2.778E-24	1.328E-22
Rbpms	15.321	1.966	0.748	2.627	8.613E-03	2.773E-02
Rel	74.462	3.111	0.678	4.590	4.427E-06	3.404E-05
Rffl	95.471	2.001	0.183	10.929	8.390E-28	5.000E-26
Rgs1	82.764	1.010	0.344	2.939	3.295E-03	1.221E-02
Rin2	388.699	1.449	0.509	2.848	4.399E-03	1.566E-02
Riok3	440.797	1.061	0.281	3.781	1.561E-04	8.518E-04
Rnd3	155.802	3.199	0.635	5.034	4.811E-07	4.402E-06
Rnf125	0.871	2.578	0.990	2.603	9.232E-03	2.932E-02
Samhd1	1134.018	1.923	0.455	4.231	2.330E-05	1.549E-04
Sdccag3	107.124	1.039	0.328	3.171	1.519E-03	6.303E-03
Sdccag8	62.278	-1.658	0.347	-4.783	1.731E-06	1.449E-05
Sec24b	76.156	2.283	0.275	8.310	9.578E-17	2.608E-15
Sertad1	102.396	1.819	0.260	6.990	2.755E-12	4.862E-11
Sertad3	22.056	0.713	0.294	2.427	1.524E-02	4.460E-02
Sfpq	143.390	1.199	0.313	3.827	1.298E-04	7.232E-04
Sgk3	243.445	1.996	0.318	6.269	3.629E-10	5.035E-09
Slc12a4	161.691	2.245	0.466	4.817	1.454E-06	1.232E-05
Slc16a1	25.819	1.397	0.301	4.643	3.436E-06	2.699E-05
Slc30a4	112.229	1.967	0.194	10.151	3.296E-24	1.570E-22
Slc4a7	320.188	2.535	0.400	6.334	2.383E-10	3.383E-09
Snn	54.913	3.071	0.310	9.898	4.239E-23	1.828E-21
Snx10	603.309	2.304	0.633	3.642	2.708E-04	1.383E-03
Socs7	79.869	1.736	0.216	8.033	9.481E-16	2.358E-14
Spint2	2.626	2.196	0.777	2.828	4.686E-03	1.654E-02
Spsb1	20.545	2.372	0.432	5.488	4.066E-08	4.376E-07
Stat3	357.561	1.483	0.195	7.623	2.473E-14	5.362E-13
Stat5a	178.996	3.194	0.296	10.791	3.780E-27	2.169E-25
Stx6	193.753	2.113	0.199	10.643	1.875E-26	1.047E-24
Tank	409.438	2.845	0.539	5.275	1.328E-07	1.322E-06
Tbc1d1	208.084	1.957	0.307	6.375	1.834E-10	2.636E-09
Tiparp	201.545	3.050	0.492	6.203	5.535E-10	7.528E-09
Tjp2	64.536	1.403	0.231	6.062	1.345E-09	1.726E-08
Tmcc3	69.170	1.050	0.263	3.994	6.496E-05	3.869E-04
Tmem50b	152.873	1.327	0.486	2.728	6.368E-03	2.140E-02
Tnfsf4	15.754	5.167	0.639	8.082	6.365E-16	1.609E-14
Traf3ip2	22.397	0.963	0.394	2.445	1.449E-02	4.277E-02
Trip10	81.041	1.356	0.201	6.754	1.437E-11	2.355E-10
Usp12	182.553	2.768	0.167	16.534	2.093E-61	7.375E-59
Usp25	831.327	2.065	0.503	4.106	4.031E-05	2.530E-04

Usp42	31.469	2.696	0.341	7.896	2.884E-15	6.776E-14
Vcpip1	310.547	1.842	0.530	3.479	5.042E-04	2.383E-03
Vps54	260.205	2.250	0.239	9.412	4.884E-21	1.848E-19
Wdr59	31.902	2.043	0.554	3.688	2.259E-04	1.177E-03
Xrn1	76.649	2.096	0.295	7.110	1.162E-12	2.111E-11
Zc3h7a	298.933	2.391	0.349	6.851	7.334E-12	1.233E-10
Zcchc6	271.058	1.422	0.133	10.703	9.903E-27	5.604E-25
Zdhhc18	34.664	1.894	0.332	5.713	1.108E-08	1.285E-07
Zfp281	139.600	1.874	0.523	3.582	3.411E-04	1.689E-03
Zfp36	214.987	2.342	0.428	5.477	4.338E-08	4.637E-07

TABLE A.6: **Differentially Expressed Genes During Single-dose Long Stimulation In Old Macrophages**

Gene Name	BaseMean	Log2FoldChange	lfcSE	Stat	p-value	padj
4930578N16Rik	3.717	1.973	0.706	2.794	5.200E-03	2.847E-02
4933430I17Rik	6.755	2.351	0.854	2.753	5.900E-03	3.146E-02
Afp	6.133	2.253	0.570	3.951	7.777E-05	7.876E-04
Ccl8	3.026	2.058	0.739	2.787	5.317E-03	2.895E-02
Dgka	25.107	-1.809	0.472	-3.833	1.264E-04	1.212E-03
H2-D1	2162.007	0.679	0.234	2.906	3.661E-03	2.128E-02
Hhex	8.368	-2.098	0.713	-2.941	3.268E-03	1.942E-02
Hipk2	55.378	-1.764	0.354	-4.977	6.471E-07	1.076E-05
Id3	98.437	-1.800	0.400	-4.498	6.872E-06	9.055E-05
Mmd	33.841	-1.112	0.384	-2.897	3.773E-03	2.182E-02
Mtus1	34.247	-1.480	0.332	-4.461	8.174E-06	1.053E-04
Pdlim5	212.982	-1.379	0.409	-3.368	7.570E-04	5.698E-03
Serpinb9	89.937	2.328	0.398	5.847	5.017E-09	1.234E-07
Slc31a1	353.384	1.813	0.600	3.023	2.502E-03	1.559E-02
Tgfbi	1882.320	2.268	0.623	3.640	2.723E-04	2.348E-03
Thbs1	1344.918	2.439	0.574	4.250	2.134E-05	2.515E-04
Trem3	5.795	4.194	0.772	5.435	5.481E-08	1.109E-06

TABLE A.7: **Pathway Enrichment in Old Tolerized Macrophages** Top 10 pathways by p-value and pathways related to innate immunity are shown.

KEGG Pathway	Count	P-Value	Fold Enrichment	Bonferroni	Benjamini
Apoptosis	36	1.83E-10	3.153180382	3.29E-08	3.29E-08
Toll-like receptor signaling pathway	37	2.74E-09	2.847948274	4.93E-07	2.46E-07
NOD-like receptor signaling pathway	24	1.40E-06	2.94974939	2.52E-04	5.03E-05
MAPK signaling pathway	62	4.18E-06	1.782835952	7.52E-04	8.36E-05
Chemokine signaling pathway	47	4.98E-06	1.967850211	8.95E-04	8.96E-05
Cytosolic DNA-sensing pathway	21	9.70E-06	2.909525534	0.001745303	1.59E-04
Adipocytokine signaling pathway	22	7.66E-05	2.502150602	0.013687964	0.001147883
RIG-I-like receptor signaling pathway	22	9.74E-05	2.465354269	0.017378841	0.001347678
Jak-STAT signaling pathway	38	9.98E-05	1.905046481	0.017803299	0.001282297
T cell receptor signaling pathway	30	4.66E-04	1.937335404	0.080496608	0.004651456

TABLE A.8: **Pathway Enrichment in Old Short Single-Dose Stimulation Macrophages** Top 10 pathways by p-value and pathways related to innate immunity are shown.

KEGG Pathway	Count	P-Value	Fold Enrichment	Bonferroni	Benjamini
Toll-like receptor signaling pathway	48	9.81E-09	2.225648485	1.85E-06	1.85E-06
NOD-like receptor signaling pathway	33	2.06E-07	2.443277419	3.89E-05	1.30E-05
Apoptosis	40	1.15E-06	2.110528736	2.17E-04	5.42E-05
Neurotrophin signaling pathway	52	3.88E-06	1.83616	7.33E-04	1.47E-04
MAPK signaling pathway	89	7.07E-06	1.541681509	0.001334698	1.91E-04

	Count	P-Value	Fold Enrichment	Bonferroni	Benjamini
Chemokine signaling pathway	65	1.75E-05	1.639428571	0.003310995	4.14E-04
Cytokine-cytokine receptor interaction	79	1.03E-04	1.486236066	0.019238439	0.002156106
Fc gamma R-mediated phagocytosis	38	2.07E-04	1.77995102	0.038462165	0.003012481
RIG-I-like receptor signaling pathway	29	2.27E-04	1.957670588	0.041998673	0.003060031
Endocytosis	64	9.68E-04	1.454384158	0.167263126	0.01070916

TABLE A.9: **Pathway Enrichment in Young Tolerized Macrophages** Top 10 pathways by p-value and pathways related to innate immunity are shown.

KEGG Pathway	Count	P-Value	Fold Enrichment	Bonferroni	Benjamini
Toll-like receptor signaling pathway	49	6.00E-13	2.85429166	1.13E-10	1.13E-10
Chemokine signaling pathway	69	4.44E-11	2.186327241	8.36E-09	4.18E-09
Apoptosis	41	4.12E-10	2.71770346	7.74E-08	2.58E-08
NOD-like receptor signaling pathway	33	6.08E-10	3.069443994	1.14E-07	2.86E-08
Fc gamma R-mediated phagocytosis	39	3.12E-07	2.294964619	5.87E-05	8.39E-06
Neurotrophin signaling pathway	47	4.34E-07	2.084932354	8.16E-05	1.02E-05
MAPK signaling pathway	78	9.12E-07	1.697407794	1.71E-04	1.91E-05
Jak-STAT signaling pathway	50	4.17E-06	1.896984925	7.84E-04	7.13E-05
Cytokine-cytokine receptor interaction	71	4.79E-06	1.678054205	9.00E-04	7.50E-05
T cell receptor signaling pathway	41	8.28E-06	2.003730517	0.001555921	1.20E-04

TABLE A.10: **Pathway Enrichment in Young Short Single-Dose Stimulation Macrophages** Top 10 pathways by p-value and pathways related to innate immunity are shown.

KEGG Pathway	Count	P-Value	Fold Enrichment	Bonferroni	Benjamini
Toll-like receptor signaling pathway	59	1.10E-13	2.488803611	2.09E-11	2.09E-11
NOD-like receptor signaling pathway	41	1.06E-11	2.761633094	2.02E-09	1.01E-09
MAPK signaling pathway	103	4.48E-08	1.623174315	8.51E-06	2.84E-06
Apoptosis	45	5.28E-08	2.160066255	1.00E-05	2.51E-06
Chemokine signaling pathway	76	1.21E-07	1.743877665	2.29E-05	4.59E-06
Fc gamma R-mediated phagocytosis	47	4.52E-07	2.002836943	8.59E-05	1.23E-05
Proteasome	28	7.98E-07	2.487906098	1.52E-04	1.68E-05
Endocytosis	79	1.57E-06	1.633238215	2.99E-04	2.49E-05
Neurotrophin signaling pathway	56	2.26E-06	1.798947486	4.29E-04	3.30E-05
RIG-I-like receptor signaling pathway	34	7.86E-06	2.088064047	0.001492207	9.95E-05

Appendix B

Chapter 3 Supplementary Data

TABLE B.1: **Non-synonomous SNPs in MARCO based on The 1000 Genomes Project.**

Chromosome Position	SNP ID	Ancestral allele	Alternate allele	Amino acid position	Allele frequency
119699906	rs80217020	C	T	11	AF=0.0009;AMR_AF=0.0028; EUR_AF=0.0013
119699932	rs141017045	C	T	19	AF=0.0023;AFR_AF=0.01
119699959	rs148348624	C	T	28	AF=0.0018;AFR_AF=0.01
119726770	rs143848029	C	T	44	AF=0.0018;AMR_AF=0.01
119727706	rs142489484	G	A	72	AF=0.0014;ASN_AF=0.0035; AFR_AF=0.0020
119727771	rs148983889	T	C	94	AF=0.0009;AFR_AF=0.0041
119727793	rs41279766	C	G	101	AF=0.0018;AMR_AF=0.0028; EUR_AF=0.0040
119727815	rs76112551	G	C	109	AF=0.0009;ASN_AF=0.0035
119727909	rs201601376	A	G	140	AF=0.0005;ASN_AF=0.0017
119731929	rs146520914	G	A	161	AF=0.0032;AFR_AF=0.01
119731958	rs140866852	G	A	170	AF=0.0009;ASN_AF=0.0017; EUR_AF=0.0013
119732104	rs190918804	G	A	192	AF=0.0009;AMR_AF=0.01
119732135	rs183062542	G	A	203	AF=0.0009;AMR_AF=0.01
119732139	rs139091970	C	T	204	AF=0.0005;EUR_AF=0.0013
119739063	**rs6761637**	T	C	**282**	AF=0.12;ASN_AF=0.13; AMR_AF=0.06;AFR_AF=0.29; EUR_AF=0.04
119739077	rs145447814	C	T	287	AF=0.0014;AFR_AF=0.01

119739738	rs199883809	C	T	303	AF=0.0027;ASN_AF=0.01
119739754	rs147872741	T	C	308	AF=0.01;AMR_AF=0.0028; AFR_AF=0.04
119739784	rs202033703	A	G	318	AF=0.0005;ASN_AF=0.0017
119739799	rs79247155	G	A	323	AF=0.01;AFR_AF=0.02
119739817	rs61732824	C	T	329	AF=0.04;AMR_AF=0.01; AFR_AF=0.17;EUR_AF=0.0026
119748203	rs201613882	G	A	368	AF=0.0005;ASN_AF=0.0017
119750815	rs199852920	C	T	456	AF=0.0005;ASN_AF=0.0017
119751997	rs34536804	T	C	488	AF=0.01;AMR_AF=0.04; EUR_AF=0.02
119752006	rs61732823	G	A	491	AF=0.01;AMR_AF=0.01; AFR_AF=0.01;EUR_AF=0.02
119752029	rs201104770	C	T	499	AF=0.0005;ASN_AF=0.0017
119752066	rs61732822	G	C	511	AF=0.01;AMR_AF=0.01; EUR_AF=0.01
119752090	rs150890503	C	T	519	AF=0.0009;AMR_AF=0.0028; EUR_AF=0.0013
119752091	rs200590124	G	A	520	AF=0.0009;ASN_AF=0.0017; EUR_AF=0.0013

TABLE B.2: **List of Proteins Used for Calculating Correlation Coefficients.** Sequences from *Mus musculus, Sus scrofa, Bos taurus, Pan troglodytes,* and *Homo sapiens* were used.

Abbreviation	Protein Name
MARCO	Macrophage Receptor with Collagenous structure
TLR2	Toll-like Receptor 2
CD14	Cluster of Differentiation 14
Estro	Estrogen Receptor
Acet	Acetylcholine Receptor Subunit alpha
CCK	Cholecystokinin
RRP12	Ribosomal RNA Processing 12 homolog
Alb	Serum Albumin
PFK	ATP-dependent 6-phosphofructokinase
ACON	Aconitase 1
Cyto	Cytochrome C
Lamin	Lamin A
Hyp1	Hypoxia Up-Regulated Protein 1
CFTR	Cystic Fibrosis Transmembrane conductance Regulator

HCFC1	Host Cell Factor 1
PDIA3	Protein Disulfide-Isomerase A3
Myo	Myoglobin
Epoxide	Epoxide Hydrolase 1
Cad5	Cadherin-5
Coll3	Collagenase 3
ADPR4	ADP-ribosylation Factor 4
TAS1	Thromboxane-A Synthase Isoform 1
PTP	Phospholipid Transfer Protein
NRAP	Nebulin-Related Anchoring Protein
AT4B	AT-rich interactive domain-containing protein 4B
CLK3	Dual Specificity Protein Kinase CLK3
Ins	insulin receptor substrate
PA2G4	Proliferation-Associated Protein 2G4
STOM	Stomatin
Hemo	Hemoglobin Subunit Alpha
Epid	Epidermal Growth Factor Receptor
RUVBL2	RuvB like AAA ATPase 2
PWP1	PWP1 homolog, endonuclein
ADHY3	Alcohol Dehydrogenase Class-3
cystat	Cystatin-C
NOB1	RNA-binding protein NOB1
BEST	Bestrophin 1
FRS2	Fibroblast Growth Factor Receptor Substrate 2
Creatine	Creatine kinase U-type
GRSP1	Golgi Reassembly-Stacking Protein 1
SRPRA	Signal Recognition Particle Receptor Subunit Alpha
Adren	Alpha-1B Adrenergic receptor

Appendix C

Chapter 4 Supplementary Data
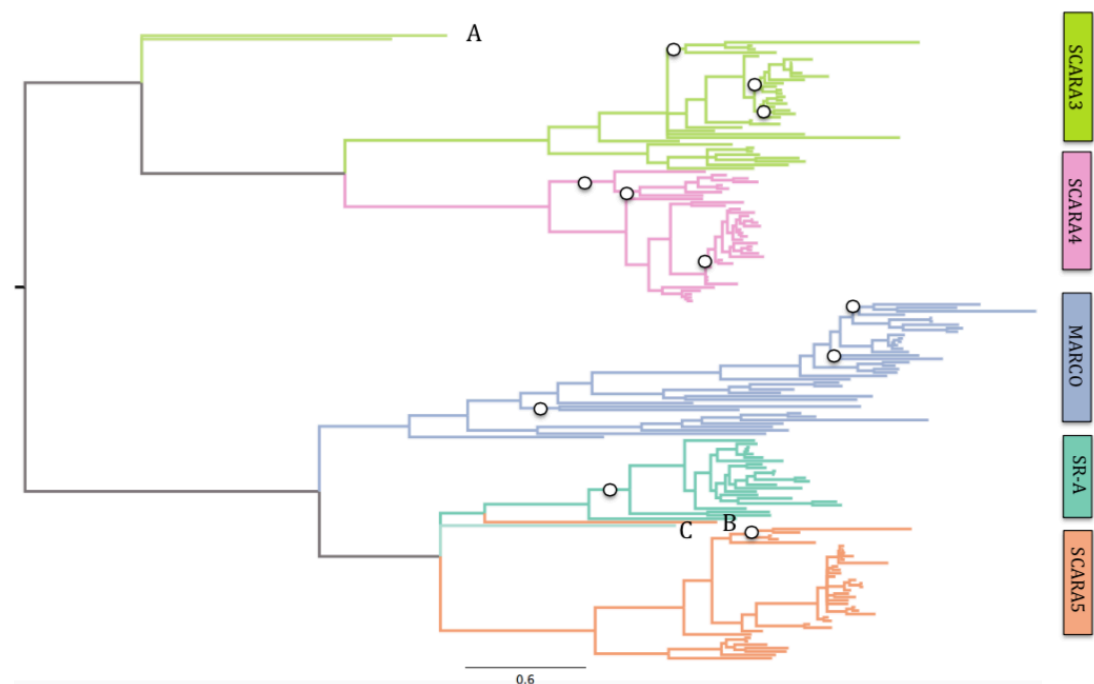
FIGURE C.1: **MrBayes phylogeny of all five Class A Scavenger Receptors using midpoint root.** MARCO branches with SCARA5 and SR-A which suggests a common ancestor between the three proteins. Posterior probabilities less than 0.7 are shown with open circles on their respective branches. Scale bar denotes number of substitutions per site. SCARA3 sequences in sea lamprey (*Petromyzon marinus*) and southern platyfish (*Xiphophorus maculatus*) are denoted as A. The sea lamprey SCARA5 sequence is shown by label B and the western clawed frog (*Xenopus tropicalis*) SR-A sequence is labeled as C. These are shown due to their long branching pattern.
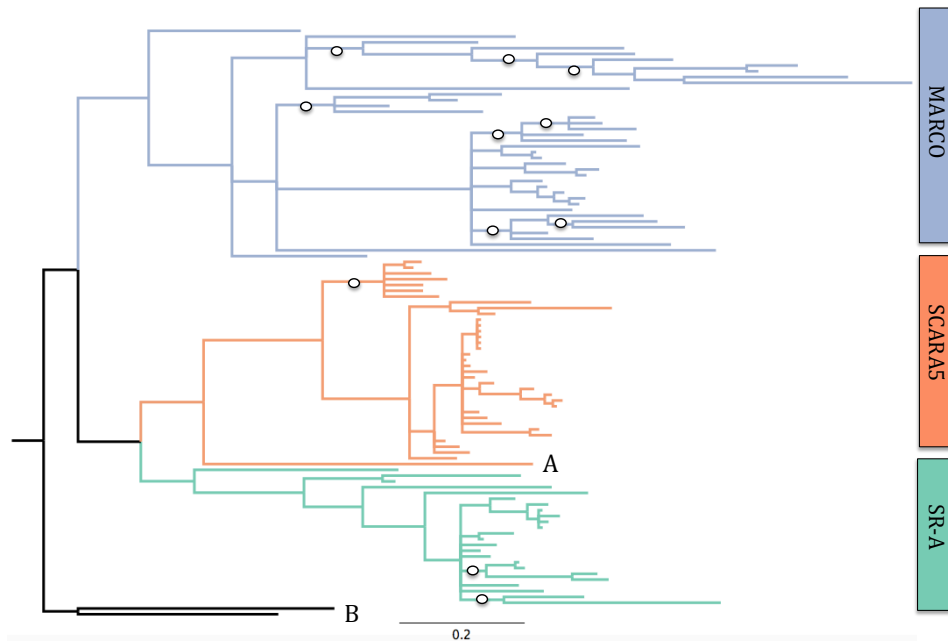
FIGURE C.2: **The evolutionary relationship between the SRCR-containing Class A Scavenger Receptors using the GCSRCR ninth and tenth SRCR domain repeats as outgroups (label B).** The GCSRCR is an SRCR-containing protein found in *Geodia cydonium* (sea sponge). Phylogenetic analysis was performed in MrBayes, with posterior probabilities less than 0.7 are labeled with open circles on their respective branches. Scale bar denotes number of substitutions per site. Label A shows the SCARA5 sequence of sea lamprey (*Petromyzon marinus*).

TABLE C.1: **Accession Numbers of Sequences of the cA-SRs used for Bayesian analysis and PAML**

Accession Numbers	*Organism*	cA-SR
XP_004481938	*Dasypus novemcinctus*	SR-A
XP_002756952	*Callithrix jacchus*	SR-A
XP_848261	*Canis lupus*	SR-A
XP_001488613	*Equus caballus*	SR-A
XP_002928655	*Ailuropoda melanoleuca*	SR-A
EHB08261	*Heterocephalus glaber*	SR-A
XP_003412472	*Loxodonta africana*	SR-A
XP_001097884	*Macaca mulatta*	SR-A
XP_001374095	*Monodelphis domestica*	SR-A
NP_619729	*Homo sapiens*	SR-A
NP_001106797	*Mus musculus*	SR-A
ELK37699	*Myotis davidii*	SR-A
XP_003256747	*Nomascus leucogenys*	SR-A
XP_004277216	*Orcinus orca*	SR-A
XP_001512876	*Ornithorhynchus anatinus*	SR-A
NP_001075717	*Oryctolagus cuniculus*	SR-A
XP_003797724	*Otolemur garnettii*	SR-A
XP_004021785	*Ovis aries*	SR-A
XP_001140701	*Pan troglodytes*	SR-A
NP_001230803	*Sus scrofa*	SR-A
NP_001178868	*Rattus norvegicus*	SR-A
XP_003772876	*Sarcophilus harrisii*	SR-A
NP_001106711	*Bos taurus*	SR-A
XP_004324021	*Tursiops truncatus*	SR-A
ENSXETP00000037776	*Xenopus Tropicalis*	SR-A
XP_004469503	*Dasypus novemcinctus*	MARCO
AFK11537	*Callorhinchus milii*	MARCO
XP_003478655	*Cavia porcellus*	MARCO
NP_990067	*Gallus gallus*	MARCO
EMC87396	*Columba livia*	MARCO
EGV99069	*Cricetulus griseus*	MARCO
XP_533324	*Canis lupus*	MARCO
XP_002917620	*Ailuropoda melanoleuca*	MARCO
XP_003416719	*Loxodonta africana*	MARCO
XP_001083118	*Macaca mulatta*	MARCO
NP_006761	*Homo sapiens*	MARCO
XP_002749572	*Callithrix jacchus*	MARCO

XP_004572010	*Maylandia zebra*	MARCO
XP_003207755	*Meleagris gallopavo*	MARCO
Q9WUB9	*Mesocricetus auratus*	MARCO
XP_001368478	*Monodelphis domestica*	MARCO
NP_034896	*Mus musculus*	MARCO
ELK25702	*Myotis davidii*	MARCO
XP_003456263	*Oreochromis niloticus*	MARCO
XP_003275234	*Nomascus leucogenys*	MARCO
XP_004276557	*Orcinus orca*	MARCO
XP_002712449	*Oryctolagus cuniculus*	MARCO
XP_004086950	*Oryzias latipes*	MARCO
XP_515756	*Pan troglodytes*	MARCO
NP_001243295	*Sus scrofa*	MARCO
XP_002812467	*Pongo abelii*	MARCO
NP_001102481	*Rattus norvegicus*	MARCO
AGH27725	*Sciaenops ocellatus*	MARCO
XP_004004798	*Ovis aries*	MARCO
XP_869408	*Bos taurus*	MARCO
XP_002944607	*Xenopus tropicalis*	MARCO
XP_004175222	*Taeniopygia guttata*	MARCO
ENSLACP00000019789	*Latimeria chalumnae*	MARCO
ENSGACP00000001962	*Gasterosteus aculeatus*	MARCO
ENSDARP00000076803	*Danio rerio*	MARCO
ENSPSIP00000003188	*Pelodiscus sinensis*	MARCO
ENSPMAP00000008064	*Petromyzon marinus*	MARCO
GAFZ01108062	*Anolis Carolinensis*	MARCO
GAIB01020145_1	*Nothobranchius furzeri*	MARCO
XP_004315380	*Tursiops truncatus*	MARCO
EOA93253	*Anas platyrhynchos*	SCARA3
XP_004454217	*Dasypus novemcinctus*	SCARA3
XP_614788	*Bos taurus*	SCARA3
XP_002807525	*Callithrix jacchus*	SCARA3
XP_543225	*Canis lupus*	SCARA3
EMP24884	*Chelonia mydas*	SCARA3
XP_693010	*Danio rerio*	SCARA3
XP_001492909	*Equus caballus*	SCARA3
XP_001234416	*Gallus gallus*	SCARA3
NP_057324	*Homo sapiens*	SCARA3
XP_001110670	*Macaca mulatta*	SCARA3
XP_004550616	*Maylandia zebra*	SCARA3
XP_003204626	*Meleagris gallopavo*	SCARA3
XP_001380024	*Monodelphis domestica*	SCARA3
NP_766192	*Mus musculus*	SCARA3
ELK29387	*Myotis davidii*	SCARA3

XP_003446373	*Oreochromis niloticus*	SCARA3
XP_003272990	*Nomascus leucogenys*	SCARA3
XP_004270781	*Orcinus orca*	SCARA3
XP_002709349	*Oryctolagus cuniculus*	SCARA3
XP_001515573	*Ornithorhynchus anatinus*	SCARA3
XP_004084013	*Oryzias latipes*	SCARA3
XP_003794065	*Otolemur garnettii*	SCARA3
XP_519678	*Pan troglodytes*	SCARA3
XP_002818982	*Pongo abelii*	SCARA3
NP_001102340	*Rattus norvegicus*	SCARA3
XP_003971950	*Takifugu rubripes*	SCARA3
XP_003757833	*Sarcophilus harrisii*	SCARA3
XP_003359089	*Sus scrofa*	SCARA3
ELW62541	*Tupaia chinensis*	SCARA3
XP_002938223	*Xenopus tropicalis*	SCARA3
ENSACAT00000015059	*Anolis carolinensis*	SCARA3
ENSPSIT00000005080	*Pelodiscus sinensis*	SCARA3
ENSGMOT00000004583	*Gadus morhua*	SCARA3
ENSLACT00000011225	*Latimeria chalumnae*	SCARA3
ENSPMAT00000007605	*Petromyzon marinus*	SCARA3
ENSXMAT00000000548	*Xiphophorus maculatus*	SCARA3
ENSGACT00000010151	*Gasterosteus aculeatus*	SCARA3
ENSTRUT00000033803	*Takifugu rubripes*	SCARA3
ENSTNIT00000015378	*Tetraodon nigroviridis*	SCARA3
EOB08780	*Anas platyrhynchos*	SCARA4
XP_003219726	*Anolis carolinensis*	SCARA4
XP_004447445	*Dasypus novemcinctus*	SCARA4
NP_001095313	*Bos taurus*	SCARA4
XP_002757150	*Callithrix jacchus*	SCARA4
XP_849057	*Canis lupus*	SCARA4
XP_003474092	*Cavia porcellus*	SCARA4
XP_003507186	*Cricetulus griseus*	SCARA4
NP_001116312	*Danio rerio*	SCARA4
CBN82070	*Dicentrarchus labrax*	SCARA4
XP_001492967	*Equus caballus*	SCARA4
NP_001034688	*Gallus gallus*	SCARA4
XP_002922671	*Ailuropoda melanoleuca*	SCARA4
NP_569057	*Homo sapiens*	SCARA4
XP_003406821	*Loxodonta africana*	SCARA4
XP_001088438	*Macaca mulatta*	SCARA4
XP_004546924	*Maylandia zebra*	SCARA4
XP_003205037	*Meleagris gallopavo*	SCARA4
XP_001368023	*Monodelphis domestica*	SCARA4
NP_569716	*Mus musculus*	SCARA4

ELK31640	*Myotis davidii*	SCARA4
XP_003439401	*Oreochromis niloticus*	SCARA4
XP_003262063	*Nomascus leucogenys*	SCARA4
ABV44703	*Oncorhynchus mykiss*	SCARA4
XP_004273818	*Orcinus orca*	SCARA4
XP_001508422	*Ornithorhynchus anatinus*	SCARA4
XP_002713638	*Oryctolagus cuniculus*	SCARA4
XP_004081176	*Oryzias latipes*	SCARA4
XP_003784843	*Otolemur garnettii*	SCARA4
XP_524004	*Pan troglodytes*	SCARA4
XP_002828148	*Pongo abelii*	SCARA4
NP_001020892	*Rattus norvegicus*	SCARA4
XP_003968105	*Takifugu rubripes*	SCARA4
XP_002194661	*Taeniopygia guttata*	SCARA4
ELW63801	*Tupaia chinensis*	SCARA4
XP_004319786	*Tursiops truncatus*	SCARA4
XP_004411262	*Odobenus rosmarus*	SCARA4
XP_002934169	*Xenopus tropicalis*	SCARA4
GSTENT10018049001	*Tetraodon nigroviridis*	SCARA4
EOA94326	*Anas platyrhynchos*	SCARA5
XP_002756852	*Callithrix jacchus*	SCARA5
XP_003479729	*Cavia porcellus*	SCARA5
XP_001234366	*Gallus gallus*	SCARA5
EGW14787	*Cricetulus griseus*	SCARA5
XP_543223	*Canis lupus*	SCARA5
XP_002914457	*Ailuropoda melanoleuca*	SCARA5
EHB14244	*Heterocephalus glaber*	SCARA5
NP_776194	*Homo sapiens*	SCARA5
XP_003412507	*Loxodonta africana*	SCARA5
XP_002805343	*Macaca mulatta*	SCARA5
XP_004550682	*Maylandia zebra*	SCARA5
XP_003204623	*Meleagris gallopavo*	SCARA5
XP_001370534	*Monodelphis domestica*	SCARA5
NP_083179	*Mus musculus*	SCARA5
ELK29382	*Myotis davidii*	SCARA5
XP_003456669	*Oreochromis niloticus*	SCARA5
XP_003272963	*Nomascus leucogenys*	SCARA5
XP_004270748	*Orcinus orca*	SCARA5
XP_001507091	*Ornithorhynchus anatinus*	SCARA5
XP_002709496	*Oryctolagus cuniculus*	SCARA5
XP_003794062	*Otolemur garnettii*	SCARA5
XP_519680	*Pan troglodytes*	SCARA5
XP_003132867	*Sus scrofa*	SCARA5
XP_002818988	*Pongo abelii*	SCARA5

NP_001129327	*Rattus norvegicus*	SCARA5
XP_004433598	*Ceratotherium simum*	SCARA5
XP_003757828	*Sarcophilus harrisii*	SCARA5
XP_004004485	*Ovis aries*	SCARA5
NP_001095969	*Bos taurus*	SCARA5
CAG12980	*Tetraodon nigroviridis*	SCARA5
XP_004382241	*Trichechus manatus*	SCARA5
EMP24889	*Chelonia mydas*	SCARA5
XP_002941495	*Xenopus tropicalis*	SCARA5
NP_001025361	*Danio rerio*	SCARA5
ENSACAT00000014115	*Anolis carolinensis*	SCARA5
ENSGMOT00000002420	*Gadus morhua*	SCARA5
ENSPMAT00000006907	*Petromyzon marinus*	SCARA5
ENSXMAT00000006370	*Xiphophorus maculatus*	SCARA5
ENSGACT00000009202	*Gasterosteus aculeatus*	SCARA5

Appendix D

RAG1 is a chimeric protein with two evolutionary origins

Abstract

Through alternate splicing of V(D)J segments, RAG1 and RAG2 generate a diverse repertoire of functional T cell and B cell receptors that allow for the development of specific immune responses to diverse antigens. Previous work has shown functional and structural similarities in the DNA binding and splicing mechanism of the RAG proteins and transposases. Transposases are enzymatic proteins that catalyze the movement of transposons into other parts of the genome and occasionally into other genes. Based on this observation, the RAG transposon hypothesis was proposed where the RAG proteins were originally derived from a transposable element. Phylogenetic analysis suggested that the RAG1 core domain was derived from the transposase family known as Transib. However, the RAG1 amino terminus does not share sequence similarity with any transposases, and contains two domains that are absent from Transib transposases. The RAG1 amino terminus plays an essential role in RAG1 dimer formation, DNA splicing efficiency, and contains an E3 ubiquitin ligase. Therefore, the RAG transposon hypothesis is incomplete without accounting for the origins of the amino terminus. We propose that RAG1 is a chimeric protein where the RAG1 amino terminus shares a different evolutionary origin from the RAG1 core domain. Using Bayesian phylogenetics, we confirm the evolutionary relationship between RAG1 and Transib. We show that the RAG1 amino terminus is distantly related to the TNF Receptor Associated Factor (TRAF) protein family, as both proteins share a ring domain and zinc finger. In addition, we identify an example of a Transib insertion event that generated a chimeric protein, and show that Transib is able to insert into other protein coding genes. Our results show that the RAG transposon was inserted into a protein-coding gene containing the RAG1 amino terminus to generate the full-length RAG1 sequence.

Introduction

RAG1 (Recombination Activating Gene-1) and RAG2 (Recombination Activating Gene-2) are essential for V(D)J recombination. RAG1 and RAG2 function by binding to the Recombination Signal Sequences (RSS) that flank V (variable), D (diversity), and J (joining) gene segments, thereby initiating DNA cleavage and rejoining (Agrawal, Eastman, and Schatz, 1998). The alternative joining of V(D)J gene segments generates a repertoire of diverse antigen receptors essential for the maturation of B cells and T cells (Agrawal, Eastman, and Schatz, 1998). The similarities in DNA cleavage products between RAG1, retroviral integrases and transposons, such as bacteriophage Mu and Tn10 transposition, has led to the current hypothesis of adaptive immunity evolving from a transfer event of the "RAG transposon" into a primitive antigen receptor gene (Agrawal, Eastman, and Schatz, 1998; Thompson, 1995).

The first study to analyze the evolutionary relationship between the RAG proteins and a transposable element was conducted by Kapitonov and Jurka (2005), which compared RAG1 with the transposase family known as Transib. They proposed that RAG1 was derived from a Transib-like ancestral gene that inserted into the genome of a Bilaterian or Cnidarian ancestor (Kapitonov and Jurka, 2005). Phylogenetic analyses and multiple sequence alignments of Transib and representative RAG1 proteins revealed sequence simarility at the RAG1 core domain (Kapitonov and Jurka, 2005). In addition, sequence comparisons between the RSS of RAG1 and the inverted terminal repeats (ITR) of Transib showed similar potential DNA cleaving sites. These results suggested an evolutionarily relationship between tranposable elements and RAG1.

Further evidence supporting this theory was recently shown in a study analyzing the functional similarities between a recently isolated Transib protein and RAG1. The RAG1 core domain shares a DDE catalytic triad with various transposases and viral integrases that are essential for DNA nicking and hairpin formation (Landree, Wibbenmeyer, and Roth, 1999). A functionally isolated Transib protein in *Helicoverpa zea*, named Hztransib, also possesses the DDE catalytic triad. Mutations of the triad in both proteins results in decreased or minimal hairpin formation and suggests a similar mechanism of DNA cleavage and rejoining (Hencken, Li, and Craig, 2012; Landree, Wibbenmeyer, and Roth, 1999). These functional similarities further support the theory of a common origin between RAG1 and the Transib transposase.

Despite the functional and evolutionary evidence between RAG1 and transposases, the RAG transposon theory must be expanded. The critical issue of the RAG transposon theory is the lack of functional and sequence similarity between transposases and the RAG1 amino terminus (Dreyfus, 2009). In order to explain this observation, Dreyfus proposes a radically different theory where RAG1 shares a common evolutionary origin with herpes virus recombinase DNA Binding Proteins (DBP) (Dreyfus, 2009).

Dreyfus hypothesizes that an ancestral deuterostome was infected with a primordial herpes virus that was inserted into the genome adjacent to a primordial RAG2 gene, and could coevolve with RAG2 (Dreyfus, 2009). However, the herpes DBP and RAG1 sequences share no sequence similarity.
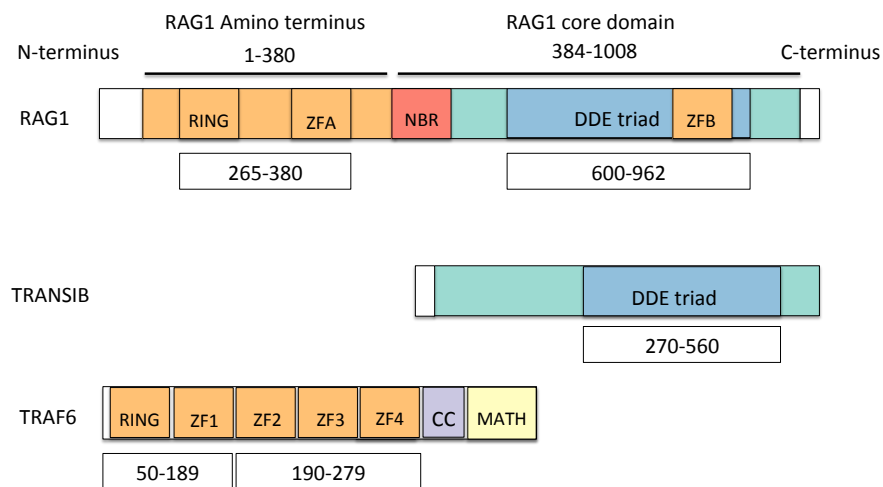
FIGURE D.1: **Protein structure of RAG1.** RAG1 contains a Ring domain (RING), a Nonamer Binding Region (NBR), a DDE catalytic triad, and two Zinc Finger domains (ZFA and ZFB). The Transib transposase contains a DDE catalytic triad and shares similarity with the RAG1 core domain. TRAF6 contains a RING domain, four Zing Finger domains (ZF1-ZF4), a Coiled-Coil domain (CC), and a MATH domain.

In addition, after extensive searching in newly sequenced genomes, the RAG transposon has yet to be found. Although transposases share similarity with RAG1 at the RAG1 core domain, the RAG1 amino terminus contains a zinc finger motif and a ring domain that are both absent in known transposases. The zinc finger motif and ring domain have been experimentally shown to contribute to the stability of RAG1 and play a role in protein-protein interactions (Rodgers et al., 1996). In addition, the removal of these domains results in a decrease in RAG1 function; despite the main catalytic residues residing within the RAG1 core domain (Rodgers et al., 1996).

Although a relationship between RAG1 and Transib has been suggested, the RAG transposon hypothesis must be expanded to account for the RAG1 amino terminus. With more genomes available, we confirm the phylogenetic relationship between

RAG1 and Transib, and furthermore study the RAG1 amino terminus in a phylogenetic context. Given the nature of transposable elements, we present a new theory where the primitive RAG transposon may have inserted into a protein coding gene containing elements of the RAG1 amino terminus. We hypothesize that RAG1 is a chimeric protein; with components of a Transib-like protein and a protein containing the zinc finger motif and ring domain. Using Bayesian analysis, we show that the RAG1 core domain and the RAG1 amino terminus have separate evolutionary origins. When comparing the RAG1 core domain with other transposases, we confirm the evolutionary relationship between RAG1 and Transib. In addition, we show that the RAG1 amino terminus shares an evolutionary origin with other zinc finger containing proteins. Finally, we document an example of a Transib protein insertion event within another unrelated protein within *Aedes aegypti*. Taken together, these data suggest an insertion event of a Transib-like gene into a protein resembling the RAG1 amino terminus, is sufficient to create the full length RAG1 sequence.

Materials and Methods

Transib proteins from Kapitonov and Jurka (2005) were downloaded from the RepBase database (Jurka et al., 2005). The original RAG1 sequences from Kapitonov and Jurka (2005) were used for PSI-BLAST searches to obtain a total of 14 RAG1 sequences from the National Center for BiotechnologyInformation (NCBI) protein database. A second PSI-BLAST search was performed using the RAG1 amino terminus as a search query to discover sequences with simarility to the RAG1 amino terminus. The Tumor necrosis factor Receptor Associated Factor (TRAF) protein family member 6 (TRAF6) was identified as having sequence similarity to the RAG1 amino terminus with E-values over 3×10^{-5} and a coverage ranging from 15-20% . Due to the presence of a zinc finger motif within this protein family, 20 TRAF6, 10 TRAF3 and 10 TRAF2 proteins were downloaded from the NCBI protein database. PSI-BLAST was also performed using Transib sequences as a query and excluding any results denoted as RAG1 from the search. The 'Out At First' (OAF) protein from *Aedes aegypti* had E-values of up to 2×10^{-57} when compared with Transib. Finally, to confirm the evolutionary relationship between RAG1 and transposases, sequences from the hermes, Mu, and Mariner transposases were gathered. Protein sequences were aligned using MAFFT (Katoh et al., 2005).

To select the optimal model for phylogenetic testing, each protein sequence alignment was analyzed by PROTTEST (Abascal, Zardoya, and Posada, 2005). We constructed two phylogenetic trees, using MrBayes Ronquist and Huelsenbeck, 2003, to study the evolutionary relationship of RAG1 at the amino terminus and at the RAG1 core domain. We constructed one tree using RAG1, TRAF6, TRAF2, and TRAF3 sequences; while excluding the RAG1 core domain and the MATH domains of the

TRAF proteins. This tree was run under a Whelan and Goldman (WAG) model. We also constructed a tree using only the RAG1 core domain, Transib, and the transposase sequences under a Jones Thorton and Taylor (JTT) model. Our phylogenetic treess were constructed using MrBayes for 5 million generations. Trees were sampled every 1000 generations with a burn in period of 25%. Convergence was tested with Tracer (Rambaut and Drummond, 2012) and trees were visualized using FigTree (Rambaut and Drummond, 2009).

Results

Several hypotheses have been put forth regarding the origin of adaptive immunity and the RAG genes. The RAG transposon theory proposes that RAG1 originated as part of a transposable element. Kapitonov and Jurka (2005) have shown that RAG1 and the transposase family known as Transib may share a common origin. This was due to conserved functional residues within the RAG1 core domain, as well as similarity between the RSS of RAG1 and ITR of Transib. However, the RAG1 amino terminus does not appear to be derived from a transposase and can not be explained by the current RAG transposon theory. Here we propose that RAG1 is a chimeric protein with separate evolutionary origins for its core domain and amino terminus.

In order to confirm the evolutionary relationship between Transib and the RAG1 core domain, we constructed a phylogenetic tree using the RAG1 core domain, and several transposases (Figure D.2). We include additional RAG1 sequences from more divergent taxa including *Branchiostoma floridae* (Florida lancelet) and *Strongylocentrotus purpuratus* (purple sea urchin), as well as other known transposases. Using MrBayes, we show that RAG1 and Transib are distantly related compared with Mu, Mariner, and hermes transposases. Our results confirm the findings of Kapitonov and Jurka (2005) and suggests a common evolutionary origin of the RAG1 core domain and the Transib transposase.
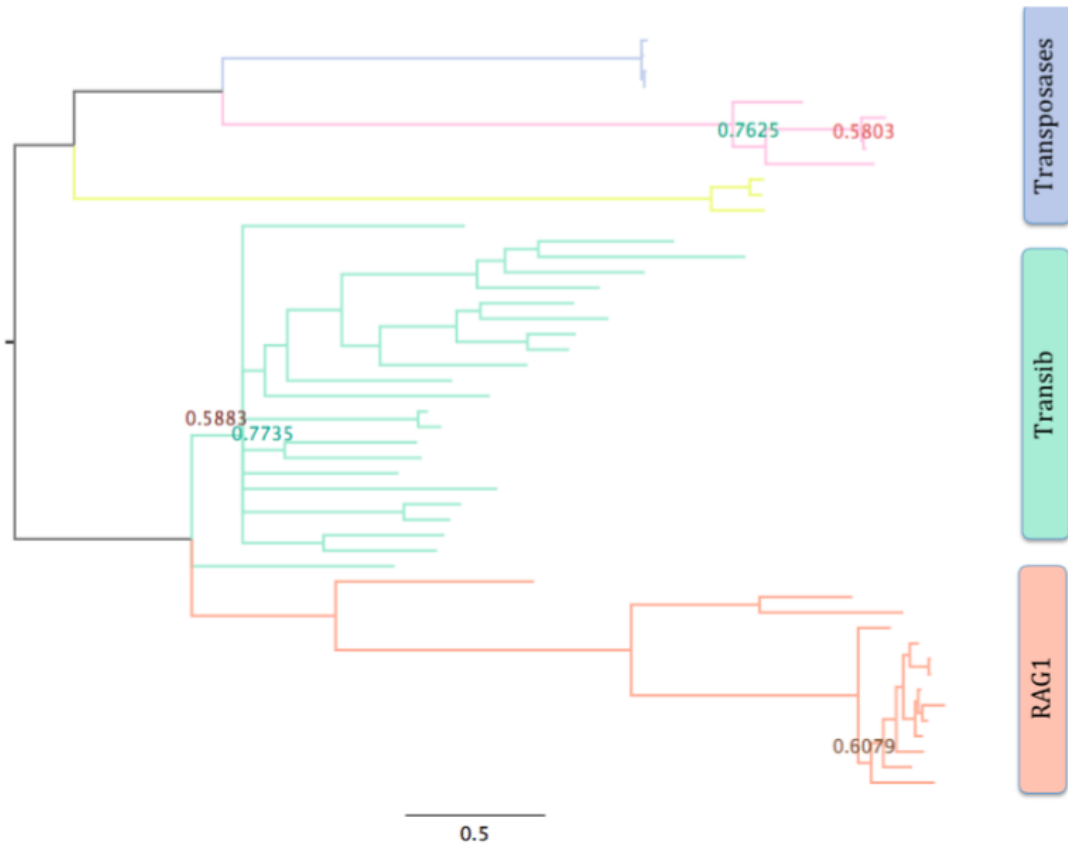
FIGURE D.2: **Phylogenetic tree of the RAG1 core domain and various transposases.** Transposases sequences from the Transib, Mu, hermes, and Mariner families included (green, yellow, pink, and blue respectively). Scale bar denotes the number of substitutions per site. Tree was constructed using mid point rooting. Posterior probabilities above 0.8 are not shown while values below 0.8 are denoted. Transib and RAG1 group together phylogenetically, while the other transposases form their own clade. Our results suggest that the RAG1 core domain and Transib share a distant evolutionary relationship. Taxa sampled are listed in the supplement.

We began studying the evolutionary origin of the RAG1 amino terminus by conducting PSI-BLAST searches and found significant sequence similarity between the RAG1 amino terminus and the TRAF protein family. The TRAF proteins shared similarity with RAG1 at the zinc fingers and ring domains, therefore we gathered 20 TRAF6, 10 TRAF2, and 10 TRAF3 protein sequences and generated a multiple sequence alignment (Figure D.3). The TRAF proteins and RAG1 share several functional residues within these domains including the zinc binding cysteines (Rodgers et al., 1996).
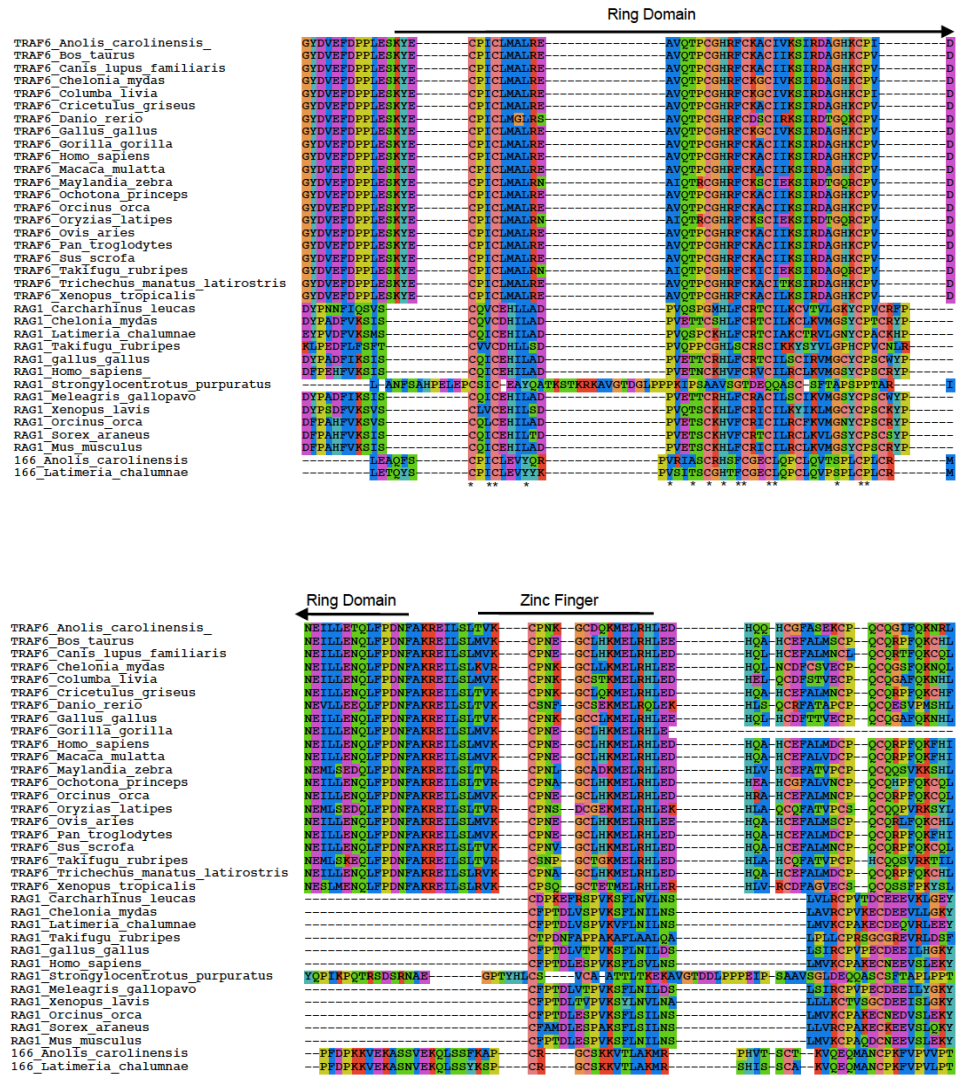
FIGURE D.3: **Multiple sequence alignment of RAG1 and TRAF6 at the zinc finger and ring domain.** Residues that make up the ring domain and zinc finger motif of RAG1 (residues 265-380) are conserved with the ring domain and first zinc finger of TRAF6 (residues 50-189). Within TRAF6, the functional residues involved in ubiquitination reside within an alpha helix region that is absent within RAG1.

We constructed a phylogenetic tree of RAG1 and the TRAF proteins to characterize

the origin of the RAG1 amino terminus (Figure D.4). Using the Ring finger protein 166 as an outgroup, we show that the TRAF proteins share a close evolutionary relationship, with TRAF2 and TRAF3 being closely related. Previous work has shown that TRAF6 is a more ancient member of the TRAF family, while TRAF2 and TRAF3 appeared later in evolution (Zapata, Martínez-García, and Lefebvre, 2007) We also show that the RAG1 amino terminus protein is also distantly related to the TRAF proteins. The RAG1 amino terminus has a ring domain and zinc finger (ZFA) that shares similarity with the ring domain and first zinc finger (ZF1) of TRAF6 (Figure D.3).This suggests that the RAG1 amino terminus originated from a separate protein containing the zinc finger and ring domains.
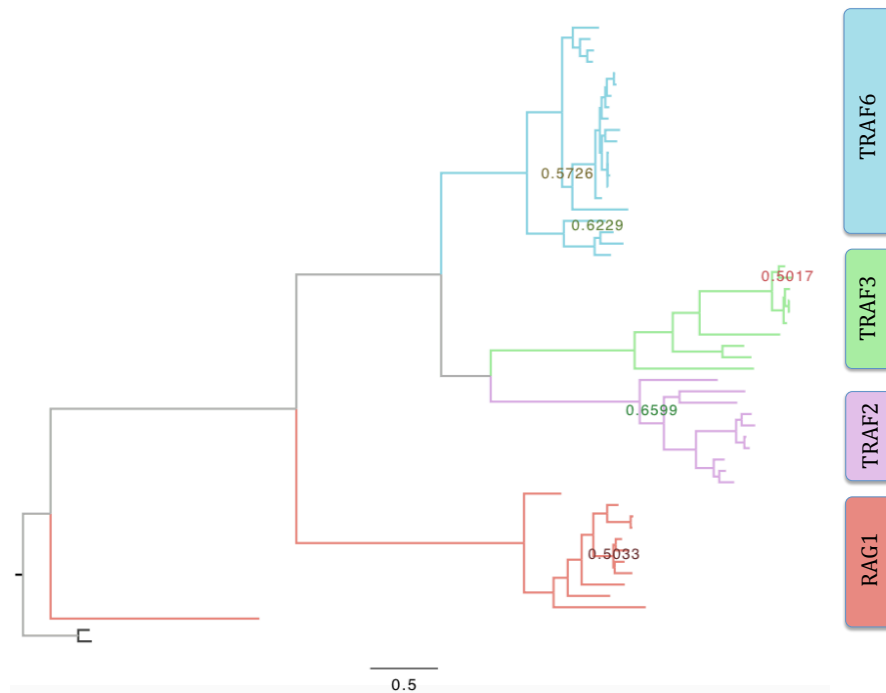
FIGURE D.4: **Phylogenetic tree of RAG1, TRAF2, TRAF3, and TRAF6.** Scale bar denotes the number of substitutions per site. Tree was constructed using Ring finger protein 166 from *Latimeria chalumnae* and *Anolis carolinensis* as outgroups. Posterior probabilities above 0.8 are not shown while values below 0.8 are denoted. TRAF2, TRAF3, and TRAF6 are all closely related, while the RAG1 amino terminus shares a distant evolutionary relationship with the TRAF proteins. Taxa sampled are listed in the supplement.

Finally, we discovered an example of a Transib insertion event resulting in a chimeric protein. Transib1 from *Aedes aegypti* and its respective OAF protein appear to

share a high amount of similarity. A dot plot comparison of Transib from *Drosophila melanogaster* and *Aedes aegypti* were compared with their respective OAF proteins using DOTTER (Sonnhammer and Durbin, 1995) (Figure D.5). The *Aedes aegypti* OAF protein shared sequence similarity with Transib from both *Aedes aegypti* and *Drosophila melanogaster* at the amino terminus, but not at the carboxyl terminus. However, *Aedes aegypti* OAF protein shared similarity to *Drosophila melanogaster* OAF protein at the carboxyl terminus but not at the amino terminus. These data show that a Transib protein was inserted into the OAF protein within *Aedes aegypti*.
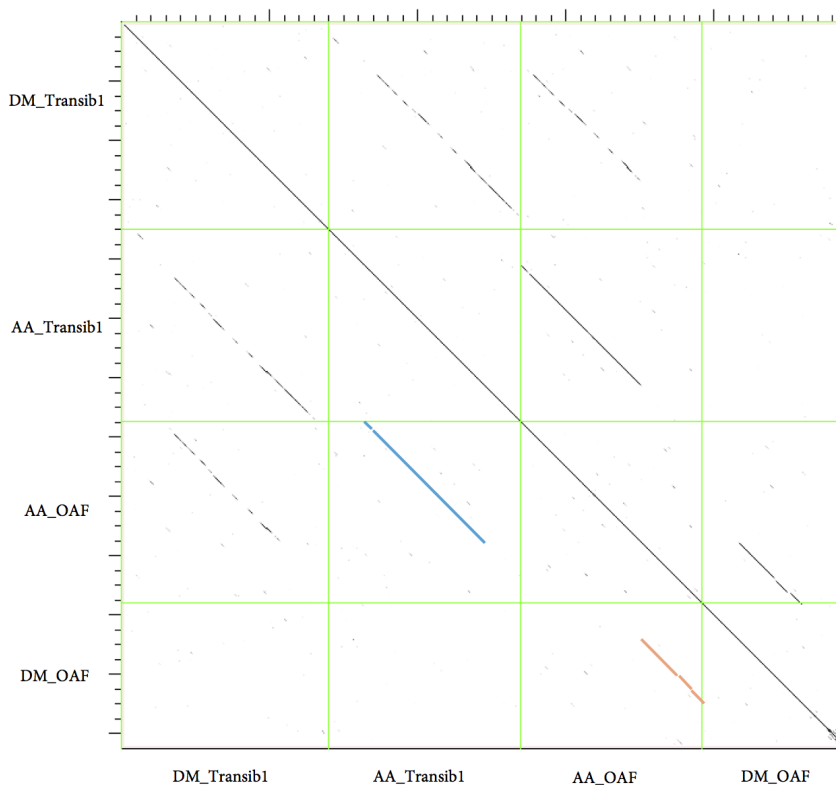
FIGURE D.5: **Dot Plot of Transib1 Versus OAF From *Drosophila melanogaster* and *Aedes aegypti*** Left corner of the graph represents the amino terminus of proteins while the right corner represents the carboxyl terminus. Lines represent strength of conservation between two proteins. DM represents sequences from *Drosophila melanogaster* while AA represents sequences from *Aedes aegypti*. OAF from *Aedes aegypti* shares a high degree of similarity with Transib1 from *Aedes aegypti* and *Drosophila melanogaster* at the amino temrinus (Blue). The OAF of *Aedes aegypti* also shares similarity with OAF from *Drosophila melanogaster* at the carboxyl terminus (Orange).

Discussion

For many years, the dominant hypothesis of the origins of adaptive immunity has been the RAG transposon theory. This theory suggests that adaptive immunity was generated from a tranposition event of a RAG transposon, containing the components of a primordial RAG1, and possibly RAG2, that inserted into the genome of an ancient deuterostome. Functional similarities between RAG mediated recombination and transposases, such as sharing a common catalytic site and DNA cleavage mechanisms (Hencken, Li, and Craig, 2012; Agrawal, Eastman, and Schatz, 1998), initially led researchers to search for the RAG transposon. However, the primordial RAG transposon has not been discovered despite extensive searching, which has raised the questioned of whether the RAG transposon actually exists. Here we present the hypothesis that the RAG1 protein is a chimeric protein; created from the insertion event of a Transib-like transposase into a protein resembling the RAG1 amino terminus.

Work by Kapitonov and Jurka (2005) concluded that RAG1 and Transib were evolutionarily related based on conserved amino acids between Transib and the RAG1 core domain, as well as similarities between the RSS of RAG1 and the ITR of Transib. They reported 14-17 % similarity between Transib and RAG1 at the RAG1 core as well as complete conservation of functional positions between the RSS and ITR (Kapitonov and Jurka, 2005). Based on these results, they proposed that the RAG transposon was either originally a member of the Transib super family or from an unknown family of Transib transposons (Kapitonov and Jurka, 2005).

We confirmed the evolutionary relationship between RAG1 and Transib in order to shed light on the RAG transposon hypothesis. Our analysis includes additional RAG1 sequence as well as transposases from other families. Our phylogenetic analysis shows that RAG1 and Transib share a distant evolutionary relationship when compared to other transposases. This coincides with the hypothesis that the RAG1 core domain was derived from a Transib-like transposase. We also discovered an example of a chimeric protein that originated from a Transib insertion event. The Transib1 and OAF proteins within *Aedes aeygpti* share a high degree of similarity at the amino terminus, but not at the carboxyl terminus. Conversely, the OAF of *Aedes aegypti* shares similarity with other OAF proteins at the carboxyl terminus but not at the amino terminus. This suggests that the Transib1 protein was inserted into the OAF protein at the amino teriminus, and provides an example of how Transib insertion events can generate a chimeric protein.

In order to study the origin of the RAG1 amino terminus, we searched for sequences resembling the zinc finger and ring domain. We identified that the TRAF protein family members share significant sequence similarity with the RAG1 amino terminus. We also show that the RAG1 amino terminus is distantly related to the TRAF protein

family.

Although RAG1 and the TRAF proteins have different roles in the immune system, the two share functional similarities at their ring domains. TRAF6 and RAG1 both form functional homodimers through protein-protein interactions within their ring domains. The TRAF6 amino terminus was shown to form a homodimer through folding of the ring domain and conserved hydrophobic residues (Yin et al., 2009b). This was surprising since the TRAF6 carboxyl terminal forms a trimer, and furthermore improper dimerization resulted in impaired TRAF6 function (Yin et al., 2009b). The RAG1 amino terminus has also been shown to form a homodimer when bound to DNA, where conserved residues within the ring domain and zinc finger motif are critical for homodimer formation (Yin et al., 2009a; De and Rodgers, 2004). In addition, both the RAG1 ring domain and TRAF6 ring domain have been shown to belong to a family of E3 ubiquitin ligases. TRAF6 interacts with signaling molecules such as interleukin-1 (IL-1) or Receptor activator of NF-κB ligand (RANKL), which leads to the production of NF-κB or AP-1 (Lamothe et al., 2008). Previous work has shown that the ring domain and first zinc finger of TRAF6 are essential for NF-κB signaling through ubiquitination of TRAF6 (Lamothe et al., 2008). Mutations of critical residues involved in ubiquitination results in a loss of IKK and TAK1 activation, and in NF-κB activation (Lamothe et al., 2008), where auto-ubiquitination of TRAF6 is mediated through interactions with Ubc13 (Lamothe et al., 2007). Interestingly, the critical residues identified for this process cluster at an alpha helix between the ring domain and first zinc finger of TRAF6 (Lamothe et al., 2007). This region is lost within the RAG1 amino terminus, and may represent a novel innovation specific to the TRAF protein family. However, the RAG1 ring domain has also been shown to be involved with ubiquitination through interactions with UbcH4 and UbcH10, but the ubiquitination target remains unknown (Yurchenko, Xue, and Sadofsky, 2003; De and Rodgers, 2004). Together with our phylogenetic data, the functional similarities between the RAG1 amino terminus and TRAF6, support the possibility of a common origin between the two.

Although we have shown strong evidence supporting a chimeric origin of RAG1, the origin of the RAG2 protein remains unknown. If the RAG1 core domain was inserted into a TRAF6-like gene, it seems unlikely that the full-length RAG1 gene could then translocate to a position upstream of the RAG2 gene. Furthermore, the RAG2 gene shows no similarity with other known transposases and has not been identified in the genomes of sea urchin, lancelet, hydra, and sea anemone (Kapitonov and Jurka, 2005). Kapitonov and Jurka propose that RAG2 may have originated from a separate insertion event, but it remains unclear what the relationship is between the RAG1 transposon and RAG2 (Kapitonov and Jurka, 2005). Our data supports the hypothesis of a separate insertion event, however further analysis is required.

Recently, (Kapitonov and Koonin, 2015) have shown evidence supporting the theory that RAG1 and RAG2 originated from the same transposon termed TransibVDJ. They identify the preliminary evidence for this TransibVDJ gene within the sequenced genome of *Lytechinus variegatus*, which encoded a LVRAG1 and LVRAG2 gene in the same locus. Due to this recent discovey, our hypothesis is flawed since (Kapitonov and Koonin, 2015) have shown RAG1 and RAG2 shared a common evolutionary origin.

TABLE D.1: **Accession numbers of sequences used for phylogenetic analysis of RAG1 related proteins and TRAF proteins.**

Protein family	*Organism*	Accession Number
TRAF2	*Alligator mississippiensis*	XP_006267621.1
TRAF2	*Chelonia mydas*	EMP28341.1
TRAF2	*Callorhinchus milii*	XP_007898411.1
TRAF2	*Maylandia zebra*	XP_004574276.1
TRAF2	*Danio rerio*	XP_683631.2
TRAF2	*Anolis carolinensis*	XP_003214697.1
TRAF2	*Canis lupus*	XP_005625129.1
TRAF2	*Sus scrofa*	XP_005652776.1
TRAF2	*Gorilla gorila*	XP_004048994.1
TRAF2	*Homo sapiens*	XP_005266156.1
TRAF2	*Gallus gallus*	CDZ92726.1
TRAF3	*Alligator sinensis*	XP_006018090.1
TRAF3	*Anolis carolinensis*	XP_003224105.2
TRAF3	*Cavia porcellus*	XP_003463146.1
TRAF3	*Gorilla gorilla*	XP_004055781.1
TRAF3	*Homo sapiens*	NP_003291.2
TRAF3	*Mus musculus*	AAC52175.1
TRAF3	*Callorhinchus milii*	XP_007886497.1
TRAF3	*Danio rerio*	NP_001003513.1
TRAF3	*Maylandia zebra*	XP_004569940.1
TRAF3	*Latimeria chalumnae*	XP_006000509.1
TRAF6	*Chelonia mydas*	XP_007055299.1
TRAF6	*Columba livia*	XP_005500297.1
TRAF6	*Gallus gallus*	XP_004941602.1
TRAF6	*Bos taurus*	NP_001029833.1
TRAF6	*Ovis aries*	XP_004016459.1
TRAF6	*Orcinus orca*	XP_004264064.1
TRAF6	*Sus scrofa*	NP_001098756.1
TRAF6	*Canis lupus familiaris*	XP_003432370.1
TRAF6	*Cricetulus griseus*	ERE70467.1
TRAF6	*Ochotona princeps*	XP_004585456.1
TRAF6	*Gorilla gorilla gorilla*	XP_004051013.1
TRAF6	*Homo sapiens*	NP_004611.1
TRAF6	*Macaca mulatta*	NP_001129268.1
TRAF6	*Pan troglodytes*	XP_001154136.1
TRAF6	*Trichechus manatus latirostris*	XP_004369863.1
TRAF6	*Xenopus tropicalis*	NP_001008162.2
TRAF6	*Danio rerio*	AAT37634.1

TRAF6	*Maylandia zebra*	XP_004558127.1
TRAF6	*Takifugu rubripes*	XP_003969671.1
TRAF6	*Oryzias latipes*	XP_004066897.1
RAG1	*Carcharhinus leucas*	XP_003738569
RAG1	*Chelonia mydas*	XP_003738569
RAG1	*Gallus gallus*	P24271
RAG1	*Meleagris gallopavo*	XP_003206426
RAG1	*Homo sapiens*	NP_000439
RAG1	*Mus musculus*	P15919.2
RAG1	*Orcinus orca*	XP_004264065.1
RAG1	*Sorex araneus*	XP_004607968.1
RAG1	*Xenopus lavis*	Q91829
RAG1	*Latimeria chalumnae*	XP_005987022
RAG1	*Takifugu rubripes*	XP_005987022
RAG1	*Strongylocentrotus purpuratus*	AAZ23546
RAG1	*Branchiostoma floridae*	Repbase
RAG1L	*Hydra magnipapillata*	Repbase
Ring protein 166	*Anolis carolinensis*	XP_008120474.1
Ring protein 166	*Latimeria chalumnae*	XP_006008323.1
Hermes	*Bactrocera tryoni*	AAD03082.1
Hermes	*Ceratitis capitata*	AAX13309.1
Hobo	*Drosophila melanogaster*	P12258.1
Hermes	*Mamestra brassicae*	AAL93244.1
Hermes	*Musca domestica*	AAB60236.1
Mariner	*Musca domestica*	AAK54758.1
Mariner	*Drosophila simulans*	AAC16617.1
Mariner	*Drosophila sechellia*	AAC16608.1
Mariner	*Drosophila teissieri*	AAC28261.1
Mu	*Enterobacteria phage mu*	P07636.2
Mu	*Shigella flexneri*	WP_024260740.1
Mu	*Salmonella enterica*	WP_024146401.1
Transib 1	*Aedes aegypti*	Repbase
Transib 1	*Anopheles gambiae*	Repbase
Transib 1	*Culex quinquefasciatus*	Repbase
Transib 1	*Drosophila pseudoobscura*	Repbase
Transib 1	*Drosophila mojavensis*	Repbase
Transib 1	*Drosophila willistoni*	Repbase
Transib 1	*Nasonia vitripennis*	Repbase
Transib 1	*Strongylocentrotus purpratus*	Repbase
Transib 2	*Anopheles gambiae*	Repbase
Transib 2	*Drosophila melanogaster*	Repbase
Transib 2	*Drosophila pseudoobscura*	Repbase
Transib 2	*Aedes aegypti*	Repbase
Transib 2	*Hydra magnipapillata*	Repbase

Transib 3	*Drosophila melanogaster*	Repbase
Transib 3	*Drosophila pseudoobscura*	Repbase
Transib 3	*Hydra magnipapillata*	Repbase
Transib 3	*Anopheles gambiae*	Repbase
Transib 4	*Drosophila pseudoobscura*	Repbase
Transib 4	*Hydra magnipapillata*	Repbase
Transib 4	*Aedes aegypti*	Repbase
Transib 5	*Hydra magnipapillata*	Repbase
Transib 5	*Drosophila melanogaster*	Repbase
Transib 5	*Aedes aegypti*	Repbase

Bibliography

Abascal, F., R. Zardoya, and D. Posada (2005). "ProtTest: selection of best-fit models of protein evolution". In: *Bioinformatics* 21.9, pp. 2104–2105.

Agrawal, A., Q. M. Eastman, and D. G. Schatz (1998). "Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system". In: *Nature* 394.6695, pp. 744–751.

Akira, S., S. Uematsu, and O. Takeuchi (2006). "Pathogen recognition and innate immunity". In: *Cell* 124.4, pp. 783–801.

Anders, S., P. T. Pyl, and W. Huber (2014). "HTSeq–A Python framework to work with high-throughput sequencing data". In: *Bioinformatics*, btu638.

Andrews, S et al. (2010). "FastQC: A quality control tool for high throughput sequence data". In: *Reference Source*.

Arredouani, M., Z. Yang, Y. Ning, G. Qin, R. Soininen, K. Tryggvason, and L. Kobzik (2004). "The scavenger receptor MARCO is required for lung defense against pneumococcal pneumonia and inhaled particles". In: *The Journal of Experimental Medicine* 200.2, pp. 267–272.

Ashkenazy, H., O. Penn, A. Doron-Faigenboim, O. Cohen, G. Cannarozzi, O. Zomer, and T. Pupko (2012). "FastML: a web server for probabilistic reconstruction of ancestral sequences". In: *Nucleic Acids Research* 40.W1, W580–W584.

Babamusta, F., D. L. Rateri, J. J. Moorleghen, D. A. Howatt, X.-A. Li, and A. Daugherty (2006). "Angiotensin II infusion induces site-specific intra-laminar hemorrhage in macrophage colony-stimulating factor-deficient mice". In: *Atherosclerosis* 186.2, pp. 282–290.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, et al. (2004). "The Pfam protein families database". In: *Nucleic Acids Research* 32.suppl 1, pp. D138–D141.

Beeson, P. B. (1946). "Development of tolerance to typhoid bacterial pyrogen and its abolition by reticulo-endothelial blockade." In: *Experimental Biology and Medicine* 61.3, pp. 248–250.

Bolger, A. M., M. Lohse, and B. Usadel (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". In: *Bioinformatics*, btu170.

Bowdish, D. M. and S. Gordon (2009). "Conserved domains of the class A scavenger receptors: evolution and function". In: *Immunological Reviews* 227.1, pp. 19–31.

Bowdish, D. M., K. Sakamoto, M.-J. Kim, M. Kroos, S. Mukhopadhyay, C. A. Leifer, K. Tryggvason, S. Gordon, and D. G. Russell (2009). "MARCO, TLR2, and CD14 are required for macrophage cytokine responses to mycobacterial trehalose dimycolate and *Mycobacterium tuberculosis*". In: *PLoS Pathogens* 5.6, e1000474.

Bowdish, D. M., K. Sakamoto, N. A. Lack, P. C. Hill, G. Sirugo, M. J. Newport, S. Gordon, A. V. Hill, and F. O. Vannberg (2013). "Genetic variants of MARCO are associated with susceptibility to pulmonary tuberculosis in a Gambian population". In: *BioMed Central Medical Genetics* 14.1, p. 47.

Boyd, A. R., P. Shivshankar, S. Jiang, M. T. Berton, and C. J. Orihuela (2012). "Age-related defects in TLR2 signaling diminish the cytokine response by alveolar macrophages during murine pneumococcal pneumonia". In: *Experimental Gerontology* 47.7, pp. 507–518.

Brännström, A., M. Sankala, K. Tryggvason, and T. Pikkarainen (2002). "Arginine residues in domain V have a central role for bacteria-binding activity of macrophage scavenger receptor MARCO". In: *Biochemical and Biophysical Research Communications* 290.5, pp. 1462–1469.

Bruunsgaard, H., K. Andersen-Ranberg, B. Jeune, A. N. Pedersen, P. Skinhøj, and B. K. Pedersen (1999). "A high plasma concentration of TNF-α is associated with dementia in centenarians". In: *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* 54.7, pp. M357–M364.

Canton, J., D. Neculai, and S. Grinstein (2013). "Scavenger receptors in homeostasis and immunity". In: *Nature Reviews Immunology* 13.9, pp. 621–634.

Castle, S. C. (2000). "Clinical relevance of age-related immune dysfunction". In: *Clinical Infectious Diseases* 31.2, pp. 578–585.

Chen, Y., M. Sankala, J. R. Ojala, Y. Sun, A. Tuuttila, D. E. Isenman, K. Tryggvason, and T. Pikkarainen (2006). "A phage display screen and binding studies with acetylated low density lipoprotein provide evidence for the importance of the scavenger receptor cysteine-rich (SRCR) domain in the ligand-binding function of MARCO". In: *Journal of Biological Chemistry* 281.18, pp. 12767–12775.

Chow, J. C., D. W. Young, D. T. Golenbock, W. J. Christ, and F. Gusovsky (1999). "Toll-like receptor-4 mediates lipopolysaccharide-induced signal transduction". In: *Journal of Biological Chemistry* 274.16, pp. 10689–10692.

Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton (2004). "The jalview java alignment editor". In: *Bioinformatics* 20.3, pp. 426–427.

Consortium, . G. P. et al. (2010). "A map of human genome variation from population-scale sequencing". In: *Nature* 467.7319, pp. 1061–1073.

Crooks, G. E., G. Hon, J.-M. Chandonia, and S. E. Brenner (2004). "WebLogo: a sequence logo generator". In: *Genome Research* 14.6, pp. 1188–1190.

De, P. and K. K. Rodgers (2004). "Putting the pieces together: identification and characterization of structural domains in the V (D) J recombination protein RAG1". In: *Immunological Reviews* 200.1, pp. 70–82.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1, pp. 15–21.

Dorrington, M. G. and D. M. Bowdish (2013). "Immunosenescence and novel vaccination strategies for the elderly". In: *Frontiers in Immunology* 4.

Dreyfus, D. H. (2009). "Paleo-immunology: evidence consistent with insertion of a primordial herpes virus-like element in the origins of acquired immunity". In: *PloS One* 4.6, e5778.

Elomaa, O., M. Kangas, C. Sahlberg, J. Tuukkanen, R. Sormunen, A. Liakka, I. Thesleff, G. Kraal, and K. Tryggvason (1995). "Cloning of a novel bacteria-binding receptor structurally related to scavenger receptors and expressed in a subset of macrophages". In: *Cell* 80.4, pp. 603–609.

Fan, H. and J. A. Cook (2004). "Review: Molecular mechanisms of endotoxin tolerance". In: *Journal of Endotoxin Research* 10.2, pp. 71–84.

Fei, F., K. M. Lee, B. E. McCarry, and D. M. Bowdish (2016). "Age-associated metabolic dysregulation in bone marrow-derived macrophages stimulated with lipopolysaccharide". In: *Scientific Reports* 6.

Flicek, P., I. Ahmed, M. R. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, et al. (2012). "Ensembl 2013". In: *Nucleic Acids Research*, gks1236.

— (2013). "Ensembl 2013". In: *Nucleic Acids Research* 41.D1, pp. D48–D55.

Foster, S. L., D. C. Hargreaves, and R. Medzhitov (2007). "Gene-specific control of inflammation by TLR-induced chromatin modifications". In: *Nature* 447.7147, pp. 972–978.

Franceschi, C., M. Bonafè, S. Valensin, F. Olivieri, M. de Luca, E. Ottaviani, and G. de Benedictis (2000). "Inflamm-aging: an evolutionary perspective on immunosenescence". In: *Annals of the New York Academy of Sciences* 908.1, pp. 244–254.

Geissmann, F., M. G. Manz, S. Jung, M. H. Sieweke, M. Merad, and K. Ley (2010). "Development of monocytes, macrophages, and dendritic cells". In: *Science* 327.5966, pp. 656–661.

Getts, D. R., A. J. Martin, D. P. McCarthy, R. L. Terry, Z. N. Hunter, W. T. Yap, M. T. Getts, M. Pleiss, X. Luo, N. J. King, et al. (2012). "Microparticles bearing encephalitogenic peptides induce T-cell tolerance and ameliorate experimental autoimmune encephalomyelitis". In: *Nature Biotechnology* 30.12, pp. 1217–1224.

Giefing-Kröll, C., P. Berger, G. Lepperdinger, and B. Grubeck-Loebenstein (2015). "How sex and age affect immune responses, susceptibility to infections, and response to vaccination". In: *Aging Cell* 14.3, pp. 309–321.

Goldstein, J. L., Y. Ho, S. K. Basu, and M. S. Brown (1979). "Binding site on macrophages that mediates uptake and degradation of acetylated low density lipoprotein, producing massive cholesterol deposition". In: *Proceedings of the National Academy of Sciences* 76.1, pp. 333–337.

Gomez, C. R., E. D. Boehmer, and E. J. Kovacs (2005). "The aging innate immune system". In: *Current Opinion in Immunology* 17.5, pp. 457–462.

Gordon, S. (2002). "Pattern recognition receptors: doubling up for the innate immune response". In: *Cell* 111.7, pp. 927–930.

Greaves, D. R. and S. Gordon (2005). "Thematic review series: the immune system and atherogenesis. Recent insights into the biology of macrophage scavenger receptors". In: *Journal of Lipid Research* 46.1, pp. 11–20.

Greisman, S. and R. Hornick (1975). "The nature of endotoxin tolerance." In: *Transactions of the American Clinical and Climatological Association* 86, p. 43.

Han, H.-J., T. Tokino, and Y. Nakamura (1998). "CSR, a scavenger receptor-like protein with a protective role against cellular damage caused by UV irradiation and oxidative stress". In: *Human Molecular Genetics* 7.6, pp. 1039–1046.

Hencken, C. G., X. Li, and N. L. Craig (2012). "Functional characterization of an active Rag-like transposase". In: *Nature Structural and Molecular Biology* 19.8, pp. 834–836.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2009). "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources". In: *Nature Protocols* 4.1, pp. 44–57.

Ioana, M, B Ferwerda, T. Plantinga, M Stappers, M Oosting, M McCall, A Cimpoeru, F Burada, N Panduru, R Sauerwein, et al. (2012). "Different patterns of Toll-like receptor 2 polymorphisms in populations of various ethnic and geographic origins". In: *Infection and Immunity* 80.5, pp. 1917–1922.

Jackson, L. A., K. M. Neuzil, O. Yu, P. Benson, W. E. Barlow, A. L. Adams, C. A. Hanson, L. D. Mahoney, D. K. Shay, and W. W. Thompson (2003). "Effectiveness of pneumococcal polysaccharide vaccine in older adults". In: *New England Journal of Medicine* 348.18, pp. 1747–1755.

Jiang, Y., P. Oliver, K. E. Davies, and N. Platt (2006). "Identification and characterization of murine SCARA5, a novel class A scavenger receptor that is expressed by populations of epithelial cells". In: *Journal of Biological Chemistry* 281.17, pp. 11834–11845.

Jurka, J., V. V. Kapitonov, A Pavlicek, P Klonowski, O Kohany, and J Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements". In: *Cytogenetic and Genome Research* 110.1-4, pp. 462–467.

Kapitonov, V. V. and J. Jurka (2005). "RAG1 core and V (D) J recombination signal sequences were derived from Transib transposons". In: *PloS Biology* 3.6, e181.

Kapitonov, V. V. and E. V. Koonin (2015). "Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon". In: *Biology Direct* 10.1, pp. 1–8.

Katoh, K., G. Asimenos, and H. Toh (2009). "Multiple alignment of DNA sequences with MAFFT". In: *Bioinformatics for DNA Sequence Analysis*. Springer, pp. 39–64.

Katoh, K., K.-i. Kuma, H. Toh, and T. Miyata (2005). "MAFFT version 5: improvement in accuracy of multiple sequence alignment". In: *Nucleic Acids Research* 33.2, pp. 511–518.

Kilian, M., K. Poulsen, T. Blomqvist, L. S. Håvarstein, M. Bek-Thomsen, H. Tettelin, and U. B. Sørensen (2008). "Evolution of Streptococcus pneumoniae and its close commensal relatives". In: *PloS One* 3.7, e2683.

Krieger, M. (1992). "Molecular flypaper and atherosclerosis: structure of the macrophage scavenger receptor". In: *Trends in Biochemical Sciences* 17.4, pp. 141–146.

Krogh, A., B. Larsson, G. Von Heijne, and E. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes". In: *Journal of Molecular Biology* 305.3, pp. 567–580.

Lamothe, B., A. Besse, A. D. Campos, W. K. Webster, H. Wu, and B. G. Darnay (2007). "Site-specific Lys-63-linked tumor necrosis factor receptor-associated factor 6 auto-ubiquitination is a critical determinant of IκB kinase activation". In: *Journal of Biological Chemistry* 282.6, pp. 4102–4112.

Lamothe, B., A. D. Campos, W. K. Webster, A. Gopinathan, L. Hur, and B. G. Darnay (2008). "The RING domain and first zinc finger of TRAF6 coordinate signaling by interleukin-1, lipopolysaccharide, and RANKL". In: *Journal of Biological Chemistry* 283.36, pp. 24871–24880.

Landree, M. A., J. A. Wibbenmeyer, and D. B. Roth (1999). "Mutational analysis of RAG1 and RAG2 identifies three catalytic amino acids in RAG1 critical for both cleavage steps of V (D) J recombination". In: *Genes and Development* 13.23, pp. 3059–3069.

Lee, E.-S. A., F. J. Whelan, D. M. Bowdish, and A. K. Wong (2016). "Partitioning and Correlating Subgroup Characteristics from Aligned Pattern Clusters". In: *Bioinformatics*, btw211.

Linton, M. F., V. R. Babaev, L. A. Gleaves, and S. Fazio (1999). "A direct role for the macrophage low density lipoprotein receptor in atherosclerotic lesion formation". In: *Journal of Biological Chemistry* 274.27, pp. 19204–19210.

Love, M. I., W. Huber, and S. Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In: *Genome Biology* 15.12, p. 550.

Ma, M.-J., H.-B. Wang, H. Li, J.-H. Yang, Y. Yan, L.-P. Xie, Y.-C. Qi, J.-L. Li, M.-J. Chen, W. Liu, et al. (2011). "Genetic variants in MARCO are associated with the susceptibility to pulmonary tuberculosis in Chinese Han population". In: *PloS One* 6.8, e24069.

Mantovani, A., S. Sozzani, M. Locati, P. Allavena, and A. Sica (2002). "Macrophage polarization: tumor-associated macrophages as a paradigm for polarized M2 mononuclear phagocytes". In: *Trends in immunology* 23.11, pp. 549–555.

Mantovani, A., A. Sica, S. Sozzani, P. Allavena, A. Vecchi, and M. Locati (2004). "The chemokine system in diverse forms of macrophage activation and polarization". In: *Trends in immunology* 25.12, pp. 677–686.

Martinez, F. O., A. Sica, A. Mantovani, and M. Locati (2007). "Macrophage activation and polarization." In: *Frontiers in Bioscience: a Journal and Virtual Library* 13, pp. 453–461.

Martínez, V. G., S. K. Moestrup, U. Holmskov, J. Mollenhauer, and F. Lozano (2011). "The conserved scavenger receptor cysteine-rich superfamily in therapy and diagnosis". In: *Pharmacological Reviews* 63.4, pp. 967–1000.

McElhaney, J. E., X. Zhou, H. K. Talbot, E. Soethout, R. C. Bleackley, D. J. Granville, and G. Pawelec (2012). "The unmet need in the elderly: how immunosenescence, CMV infection, co-morbidities and frailty are a challenge for the development of more effective influenza vaccines". In: *Vaccine* 30.12, pp. 2060–2067.

Metchnikoff, E. (1905). *Immunity in infective diseases*. Oxford University Press.

Meyer, M., M. Kircher, M.-T. Gansauge, H. Li, F. Racimo, S. Mallick, J. G. Schraiber, F. Jay, K. Prüfer, C. de Filippo, et al. (2012). "A high-coverage genome sequence from an archaic Denisovan individual". In: *Science* 338.6104, pp. 222–226.

Montgomery, R. R. and A. C. Shaw (2015). "Paradoxical changes in innate immunity in aging: recent progress and new directions". In: *Journal of Leukocyte Biology* 98.6, pp. 937–943.

Motoyoshi, K. (1998). "Biological activities and clinical application of M-CSF." In: *International Journal of Hematology* 67.2, pp. 109–122.

Murphy, K. (2011). *Janeway's immunobiology*. Garland Science, 1—35.

Nakamura, K., W. Ohya, H. Funakoshi, G. Sakaguchi, A. Kato, M. Takeda, T. Kudo, and T. Nakamura (2006). "Possible role of scavenger receptor SRCL in the clearance of amyloid-β in Alzheimer's disease". In: *Journal of Neuroscience Research* 84.4, pp. 874–890.

Nilsen, N. J., S. Deininger, U. Nonstad, F. Skjeldal, H. Husebye, D. Rodionov, S. von Aulock, T. Hartung, E. Lien, O. Bakke, et al. (2008). "Cellular trafficking of lipoteichoic acid and Toll-like receptor 2 in relation to signaling; role of CD14 and CD36". In: *Journal of Leukocyte Biology* 84.1, pp. 280–291.

Nomura, F., S. Akashi, Y. Sakao, S. Sato, T. Kawai, M. Matsumoto, K. Nakanishi, M. Kimoto, K. Miyake, K. Takeda, et al. (2000). "Cutting edge: endotoxin tolerance in mouse peritoneal macrophages correlates with down-regulation of surface toll-like receptor 4 expression". In: *The Journal of Immunology* 164.7, pp. 3476–3479.

Noonan, J. P., G. Coop, S. Kudaravalli, D. Smith, J. Krause, J. Alessi, F. Chen, D. Platt, S. Pääbo, J. K. Pritchard, et al. (2006). "Sequencing and analysis of Neanderthal genomic DNA". In: *Science* 314.5802, pp. 1113–1118.

Novakowski, K. E., A. Huynh, S. Han, M. G. Dorrington, C. Yin, Z. Tu, P. Pelka, P. Whyte, A. Guarné, K. Sakamoto, et al. (2016). "A naturally occurring transcript variant of MARCO reveals the SRCR domain is critical for function". In: *Immunology and Cell Biology*.

Oksanen, J., R. Kindt, P. Legendre, B. O'Hara, M. H. H. Stevens, M. J. Oksanen, and M. Suggests (2007). "The vegan package". In: *Community Ecology Package* 10.

Palecanda, A., J. Paulauskis, E. Al-Mutairi, A. Imrich, G. Qin, H. Suzuki, T. Kodama, K. Tryggvason, H. Koziel, and L. Kobzik (1999). "Role of the scavenger receptor MARCO in alveolar macrophage binding of unopsonized environmental particles". In: *The Journal of Experimental Medicine* 189.9, pp. 1497–1506.

Pancer, Z., J. Münkner, I. Müller, and W. E. Müller (1997). "A novel member of an ancient superfamily: sponge (*Geodia cydonium*, Porifera) putative protein that features scavenger receptor cysteine-rich repeats". In: *Gene* 193.2, pp. 211–218.

Paradis, E., J. Claude, and K. Strimmer (2004). "APE: analyses of phylogenetics and evolution in R language". In: *Bioinformatics* 20.2, pp. 289–290.

Peiser, L., P. J. Gough, T. Kodama, and S. Gordon (2000). "Macrophage class A scavenger receptor-mediated phagocytosis of *Escherichia coli*: role of cell heterogeneity, microbial strain, and culture conditions *in vitro*". In: *Infection and Immunity* 68.4, pp. 1953–1963.

Pena, O. M., J. Pistolic, D. Raj, C. D. Fjell, and R. E. Hancock (2011). "Endotoxin tolerance represents a distinctive state of alternative polarization (M2) in human mononuclear cells". In: *The Journal of Immunology* 186.12, pp. 7243–7254.

Perrin, P. (2015). "Human and tuberculosis co-evolution: An integrative view". In: *Tuberculosis* 95, S112–S116.

Plüddemann, A., S. Mukhopadhyay, M. Sankala, S. Savino, M. Pizza, R. Rappuoli, K. Tryggvason, and S. Gordon (2008). "SR-A, MARCO and TLRs differentially recognise selected surface proteins from Neisseria meningitidis: an example of fine specificity in microbial ligand recognition by innate immune receptors." In: *Journal of Innate Immunity* 1.2, pp. 153–163.

Prado-Martinez, J., P. H. Sudmant, J. M. Kidd, H. Li, J. L. Kelley, B. Lorente-Galdos, K. R. Veeramah, A. E. Woerner, T. D. O'Connor, G. Santpere, et al. (2013). "Great ape genetic diversity and population history". In: *Nature* 499.7459, pp. 471–475.

Pruitt, K. D., T. Tatusova, and D. R. Maglott (2007). "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins". In: *Nucleic Acids Research* 35.suppl 1, pp. D61–D65.

Puchta, A, A Naidoo, C. Verschoor, D Loukov, N Thevaranjan, T. Mandur, P. Nguyen, M Jordana, M Loeb, Z Xing, et al. (2016). "TNF Drives Monocyte Dysfunction with Age and Results in Impaired Anti-pneumococcal Immunity." In: *PLoS pathogens* 12.1, e1005368–e1005368.

Rambaut, A and A Drummond (2009). "FigTree v1. 3.1". In: *Computer program and documentation distributed by the author at http://tree. bio. ed. ac. uk/software.*

Rambaut, A and A. Drummond (2012). *Tracer, MCMC Trace Analysis Tool, v1. 5.0.*

Resnick, D., J. E. Chatterton, K. Schwartz, H. Slayter, and M. Krieger (1996). "Structures of class A macrophage scavenger receptors electron microscopic study of flexible, multidomain, fibrous proteins and determination of the disulfide bond pattern of the scavenger receptor cysteine-rich domain". In: *Journal of Biological Chemistry* 271.43, pp. 26924–26930.

Rodgers, K. K., Z. Bu, K. G. Fleming, D. G. Schatz, D. M. Engelman, and J. E. Coleman (1996). "A zinc-binding domain involved in the dimerization of RAG1". In: *Journal of Molecular Biology* 260.1, pp. 70–84.

Rodríguez-Prados, J.-C., P. G. Través, J. Cuenca, D. Rico, J. Aragonés, P. Martín-Sanz, M. Cascante, and L. Boscá (2010). "Substrate fate in activated macrophages: a comparison between innate, classic, and alternative activation". In: *The Journal of Immunology* 185.1, pp. 605–614.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models". In: *Bioinformatics* 19.12, pp. 1572–1574.

Sankala, M., A. Brännström, T. Schulthess, U. Bergmann, E. Morgunova, J. Engel, K. Tryggvason, and T. Pikkarainen (2002). "Characterization of recombinant soluble macrophage scavenger receptor MARCO". In: *Journal of Biological Chemistry* 277.36, pp. 33378–33385.

Sherry, S. T., M.-H. Ward, M Kholodov, J Baker, L. Phan, E. M. Smigielski, and K. Sirotkin (2001). "dbSNP: the NCBI database of genetic variation". In: *Nucleic Acids Research* 29.1, pp. 308–311.

Sica, A. and A. Mantovani (2012). "Macrophage plasticity and polarization: *in vivo veritas*". In: *The Journal of Clinical Investigation* 122.3, pp. 787–795.

Smith, L. C., J. Ghosh, K. M. Buckley, L. A. Clow, N. M. Dheilly, T. Haug, J. H. Henson, C. Li, C. M. Lun, A. J. Majeske, et al. (2010). "Echinoderm immunity". In: *Invertebrate Immunity*. Springer, pp. 260–301.

Sodergren, E., G. M. Weinstock, E. H. Davidson, R. A. Cameron, R. A. Gibbs, R. C. Angerer, L. M. Angerer, M. I. Arnone, D. R. Burgess, R. D. Burke, et al. (2006). "The genome of the sea urchin *Strongylocentrotus purpuratus*". In: *Science* 314.5801, pp. 941–952.

Sonnhammer, E. L. and R. Durbin (1995). "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis". In: *Gene* 167.1, GC1–GC10.

Stout, R. D. and J. Suttles (2005). "Immunosenescence and macrophage functional plasticity: dysregulation of macrophage function by age-associated microenvironmental changes". In: *Immunological Reviews* 205.1, pp. 60–71.

Suyama, M., D. Torrents, and P. Bork (2006). "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments". In: *Nucleic Acids Research* 34.suppl 2, W609–W612.

Sweet, M. J. and D. A. Hume (1996). "Endotoxin signal transduction in macrophages." In: *Journal of Leukocyte Biology* 60.1, pp. 8–26.

Takeda, K., T. Kaisho, and S. Akira (2003). "Toll-like receptors". In: *Annual Review of Immunology* 21.1, pp. 335–376.

Takeuchi, O. and S. Akira (2010). "Pattern recognition receptors and inflammation". In: *Cell* 140.6, pp. 805–820.

Tauber, A. I. (2003). "Metchnikoff and the phagocytosis theory". In: *Nature Reviews Molecular Cell Biology* 4.11, pp. 897–901.

Thompson, C. B. (1995). "New insights into V (D) J recombination and its role in the evolution of the immune system". In: *Immunity* 3.5, pp. 531–539.

Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, C. B. Bridges, N. J. Cox, and K. Fukuda (2004). "Influenza-associated hospitalizations in the United States". In: *The Journal of The American Medical Association* 292.11, pp. 1333–1340.

Thomsen, M., B. G. Nordestgaard, L. Kobzik, and M. Dahl (2012). "Genetic variation in the scavenger receptor MARCO and its association with chronic obstructive pulmonary disease and lung infection in 10,604 individuals". In: *Respiration* 85.2, pp. 144–153.

Tomarev, S. I. and N. Nakaya (2009). "Olfactomedin domain-containing proteins: possible mechanisms of action and functions in normal development and pathology". In: *Molecular Neurobiology* 40.2, pp. 122–138.

Tsay, H.-J., Y.-C. Huang, Y.-J. Chen, Y.-H. Lee, S.-M. Hsu, K.-C. Tsai, C.-N. Yang, F.-L. Huang, F.-S. Shie, L.-C. Lee, et al. (2016). "Identifying N-linked glycan moiety and motifs in the cysteine-rich domain critical for N-glycosylation and intracellular trafficking of SR-AI and MARCO". In: *Journal of Biomedical Science* 23.1, p. 1.

Whelan, F. J., C. J. Meehan, G. B. Golding, B. J. McConkey, and D. M. Bowdish (2012). "The evolution of the class A scavenger receptors". In: *BioMed Central Evolutionary Biology* 12.1, p. 227.

Whelan, S. and N. Goldman (2001). "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach". In: *Molecular Biology and Evolution* 18.5, pp. 691–699.

Wynn, T. A., A. Chawla, and J. W. Pollard (2013). "Macrophage biology in development, homeostasis and disease". In: *Nature* 496.7446, pp. 445–455.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood". In: *Molecular Biology and Evolution* 24.8, pp. 1586–1591.

Yap, N. V., F. J. Whelan, D. M. Bowdish, and G. B. Golding (2015). "The evolution of the scavenger receptor cysteine-rich domain of the class A scavenger receptors". In: *Frontiers in Immunology* 6.

Yin, F. F., S. Bailey, C. A. Innis, M. Ciubotaru, S. Kamtekar, T. A. Steitz, and D. G. Schatz (2009a). "Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis". In: *Nature Structural and Molecular Biology* 16.5, pp. 499–508.

Yin, Q., S.-C. Lin, B. Lamothe, M. Lu, Y.-C. Lo, G. Hura, L. Zheng, R. L. Rich, A. D. Campos, D. G. Myszka, et al. (2009b). "E2 interaction and dimerization in the crystal structure of TRAF6". In: *Nature Structural and Molecular Biology* 16.6, pp. 658–666.

Yurchenko, V., Z. Xue, and M. Sadofsky (2003). "The RAG1 N-terminal domain is an E3 ubiquitin ligase". In: *Genes and Development* 17.5, pp. 581–585.

Zapata, J. M., V. Martínez-García, and S. Lefebvre (2007). "Phylogeny of the TRAF/MATH domain". In: *TNF Receptor Associated Factors (TRAFs)*. Springer, pp. 1–24.