Terms and Conditions of Use of Digitised Theses from Trinity College Library Dublin

Copyright statement

All material supplied by Trinity College Library is protected by copyright (under the Copyright and Related Rights Act, 2000 as amended) and other relevant Intellectual Property Rights. By accessing and using a Digitised Thesis from Trinity College Library you acknowledge that all Intellectual Property Rights in any Works supplied are the sole and exclusive property of the copyright and/or other IPR holder. Specific copyright holders may not be explicitly identified. Use of materials from other sources within a thesis should not be construed as a claim over them.

A non-exclusive, non-transferable licence is hereby granted to those using or reproducing, in whole or in part, the material for valid purposes, providing the copyright owners are acknowledged using the normal conventions. Where specific permission to use material is required, this is identified and such permission must be sought from the copyright holder or agency cited.

Liability statement

By using a Digitised Thesis, I accept that Trinity College Dublin bears no legal responsibility for the accuracy, legality or comprehensiveness of materials contained within the thesis, and that Trinity College Dublin accepts no liability for indirect, consequential, or incidental, damages or losses arising from use of the thesis for whatever reason. Information located in a thesis may be subject to specific use constraints, details of which may not be explicitly described. It is the responsibility of potential and actual users to be aware of such constraints and to abide by them. By making use of material from a digitised thesis, you accept these copyright and disclaimer provisions. Where it is brought to the attention of Trinity College Library that there may be a breach of copyright or other restraint, it is the policy to withdraw or take down access to a thesis while the issue is being resolved.

Access Agreement

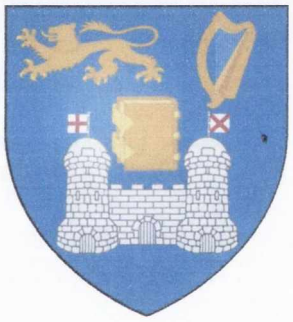
By using a Digitised Thesis from Trinity College Library you are bound by the following Terms & Conditions. Please read them carefully.

I have read and I understand the following statement: All material supplied via a Digitised Thesis from Trinity College Library is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of a thesis is not permitted, except that material may be duplicated by you for your research use or for educational purposes in electronic or print form providing the copyright owners are acknowledged using the normal conventions. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone. This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

A comparative genomics analysis of the vertebrate immune system: genes, pathways and evolution.

by
Paul Cormican
B.*Sc.* M.*Sc.*

A Thesis submitted to
The University of Dublin for the degree of
Doctor of Philosophy



School of Biochemistry and Immunology
Trinity College
University of Dublin
September, 2009

Supervisor: Professor Cliona O'Farrelly

Declaration

This thesis has not been submitted as an exercise for a degree at any other University. Except where other wise stated, the work described here in has been carried out by the author alone. This thesis may be borrowed or copied upon request with the permission of the Librarian, University of Dublin, Trinity College. The copyright belongs jointly to the University of Dublin and Paul Cormican.

Signature of Author: .....

Cormican

ptember 2009

Acknowledgements

I would first of all like to express my thanks to my supervisor Professor Cliona O' Farrelly. She has enthusiastically supported my work from day one while simultaneously displaying a remarkable level of patience. She has also encouraged me to broaden my interests, not just scientifically but in many other areas and because of this I want to promise her here, in writing, that I will one day make it through Ulysses! Cliona, I will always be grateful for the opportunities you have provided me with.

Big thanks too, are due to Dr. Andrew Lloyd, a man who has forgotten more about science than most people will ever know. Andrew's ability and willingness to explain the minute details (God is in the details after all) of our field to me is something I will always appreciate. The fact that he managed to do all this in a fun way was an added bonus.

Thanks too go to all the members of the Comparative Immunology group, both past and present. In particular to Rowan, Sarah, Kieran and Fernando who have toiled in the lab to make my work seem a little less abstract. To everyone else who I have come into contact with over the last few years I wish to say thanks. Though you might not know it, or even have meant to, you have all helped me in some way. To Mary and Sheeona, thanks for basically letting me camp in your living room for about a year. To Aisling, thanks for not changing the radio station on our Friday journeys home. I can't leave out the old (and I do mean old at this stage) St. Vincent's crew who made working there such fun. To Nelly, Jonny, Lydia, Gerry, Shane and countless others, thanks for all those nights in the Merrion and the attempts to play football in the park.

To my friends outside the lab I have to say thanks for so much. To the Castleknock/Firhouse crew, John, James, Katie, Linda and Ted (yes, even Ted) I want to say that it only took about seven years, but you guys have made living in Dublin bearable at last. To James, Sheila, Aoife, Aoife, Amii, Sarah, John, Eleanor and Mike, you guys are all legends. A special thanks to Niall, all

the way back home in Galway. He is always there to go for a pint when I am home and chat about the good old days when we were young and foolish grass cutters.

To my family, Claire, Denise, Micheal and the newer members Sean and Jack, thanks for your support and help over the years. I might not say it much but I do appreciate it. A special thanks to the Kings, without whom I may never have discovered the joys of skiing. To my parents John and Ann, I want to say thanks for all you have done and continue to do for me. This thesis is dedicated to you.

Table of Contents

List of Figures

List of Tables

Publications Arising From this Thesis

1: Higgs R, Cormican P, Cahalane S, Allan B, Lloyd AT, Meade K, James T, Lynn DJ, Babiuk LA, O'farrelly C. Induction of a novel chicken Toll-like receptor following Salmonella enterica serovar Typhimurium infection. Infect Immun. 2006 Mar;74(3):1692-8. PubMed PMID: 16495540; PubMed Central PMCID: PMC1418683

2: Cormican P, Meade KG, Cahalane S, Narciandi F, Chapwanya A, Lloyd AT, O'Farrelly C. Evolution, expression and effectiveness in a cluster of novel bovine beta-defensins. Immunogenetics. 2008 Apr;60(3-4):147-56. Epub 2008 Mar 28. PubMed PMID: 18369613.

3: Meade KG, Cahalane S, Narciandi F, Cormican P, Lloyd AT, O'Farrelly C. Directed alteration of a novel bovine beta-defensin to improve antimicrobial efficacy against methicillin-resistant Staphylococcus aureus (MRSA). Int J Antimicrob Agents. 2008 Nov;32(5):392-7. Epub 2008 Sep 4. PubMed PMID: 18775651.

4: Cormican P, Lloyd AT, Downing T, Connell SJ, Bradley D, O'Farrelly C. The avian Toll-Like receptor pathway--subtle differences amidst general conformity. Dev Comp Immunol. 2009 Sep;33(9):967-73. Epub 2009 Apr 24. PubMed PMID: 19539094.

5: Cormican P, Lloyd AT, O'Farrelly C.
Positive selection in the mammalian CD3 gene family. Manuscript in preparation.

Other Manuscripts:

1: Tim Downing, Paul Cormican, Cliona O'Farrelly, Daniel G. Bradley & Andrew T. Lloyd. Evidence of the adaptive evolution of immune genes in chicken. Accepted BMC Research Notes, September 2009.

Abbreviations

3-D	3-Dimensional
AMP	Antimicrobial Peptide
BBD	Bovine Beta Defensin
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
Bt	Bos taurus
cbd	Canine Beta Defensin
cds	coding sequence
chr	chromosome
d_N	Nonsynonymous substitution rate
d_S	Synonymous substitution rate
DNA	Deoxyribonucleic Acid
EST	Expresed Sequence Tag
Gal	Gallinacin
GG	Gallus gallus
hBD	human Beta Defensin
HMM	Hidden Markov Model
Hs	Homo sapiens
Ig	Immunoglobulin
LAP	Lingual Antimicrobial Peptide
LRR	Leucine Rich Region
ORF	Open Reading Frame
mRNA	messenger RNA
Mm	Mus musculus
MYA	Million Years Ago
PAMP	Pathogen Associated Molecular Pattern
PRR	Pathogen Recognition Receptor
PS	Positively Selected
Rn	Rattus norvegicus
SNP	Single Nucleotide Polymorphism
TIR	Toll/IL-1 Receptor
TLR	Toll Like Receptor
TNF	Tumour Necrosis Factor
UTR	Untranslated Region
ω	d_N / d_S

Summary

All species possess a selection-honed collection of genes whose products function in unison to form a barricade to invading pathogens. The exponential increase in publicly available sequence data has allowed the application of bioinformatics approaches to explore the conservation and variation of immune genes when compared between species. Antimicrobial peptides (AMPs) are important effector molecules of the immune response. Intra and inter genomic analysis indicates that, although AMPs are evolutionarily conserved across species, subgroups of gene lineages exist that are specific to certain species. We applied a Hidden Markov Model (HMM) profile search method to identification of the β-defensin family of AMPs in cattle and characterised the most greatly expanded gene family of any mammal thus far examined.

Recognition of microbial pathogens by Toll-Like-Receptors (TLRs) is an evolutionarily conserved first step in generation of an immune response in all vertebrate species. Homology searching of avian genomes led to the identification of a novel vertebrate TLR, TLR15. This TLR appears to be a diapsid specific variant and is molecularly distinct from other described TLRs. BLAST searches of genes involved in the mammalian TLR pathway against the chicken and zebra finch genomes has allowed for reconstruction of this signalling cascade in birds. Both birds possess an identical but simplified version of the established mammalian pathway with avian specific loss of several pathway intermediates. In addition we have observed that both gene conversion and positive Darwinian selection have shaped the evolution of the avian specific expansion of the TLR2 subfamily. We noted a differing panel of β-defensin genes coded for by each avian species and predicted the presence of amino acid sites that are subject to positive selection and are thus likely to be of functional importance in these genes.

The CD3 family of peptides are fundamental signalling components of the T-cell receptor complex. We note that CD3 subunits have evolved by recurring positive selection in the mammalian lineage with the 3-dimensional spatial

arrangement of positively selected sites providing insight into the form and function of these molecules within the T-cell receptor complex.

1. General Introduction

In every species a means of recognition and elimination of pathogens has developed under the evolutionary pressures generated by interaction with the harmful microbes of its own ecological niche. Whilst the goal of any immune response regardless of species is to protect against pathogens, the means by which this can be achieved varies greatly across the vast evolutionary timescales of life on this planet. Protozoans for example, represent the simplest form of extant free living eukaryote and rely on a process known as phagocytosis for both feeding and pathogen elimination. This relatively simple mechanism provides a sharp contrast to the intricate metazoan defence system of dynamically interacting cells and molecules, the constituents of which must function in unison to generate an appropriate and effective immune response. As a research discipline, immunology has long displayed a strong anthropocentric bias, as most research in the field has naturally focused on the causes and treatment of diseases which principally affect humans. It is because of this focus that much more is known about the make-up and functioning of the immune systems in primates and other mammals than in more distantly related species. Whilst human ailments and conditions may have been the primary focus of much immunological research, the contribution of comparative studies to the understanding of our own immune defences should not be underestimated. Some of the most ground breaking discoveries in the field have derived from comparative studies involving very diverse species, ranging from Metchnikoffs discovery of meschenchymal cell phagocytosis of non-self pathogens in starfish in 1891 to Hoffman's description of Toll receptors as central immune response mediators in insects (YANG 1997). Attempts to relate discoveries from comparative studies back to the human immune system can often lead to a skewed perception of immune defences. This human centred view in immunology has contributed to an inevitable misconception that our own immune system represents the most complex and advanced on the planet and that the systems of other species represent more primitive defences. This misconception ignores the fact that every species alive today has an immune

system whose evolutionary path is the same length as that which has led to our own, and is uniquely suited to providing the appropriate responses to the microbial milieu faced by each organism. Contributing to this misconception is the classical division of immune systems into the innate and adaptive. From the human point of view, those species that possess only components of an innate immune system are viewed as primitive whereas those species (jawed-vertebrates) with an adaptive response similar to our own are considered more highly evolved. The arrival of the genome sequencing era and the application of comparative genomic techniques has to some extent begun to dispel these myths. Comparisons amongst fully sequenced genomes has allowed the evolution of immune systems and their constituent genetic components to be intensively investigated for the first time and shown that all immune systems are to some extent based on differential use of a limited subset of functional protein domains. These functional domains, rather than the gene products in which they are found, can be regarded as the evolutionary units of immune system. It is the different ways in which these domains are fused, split, and shuffled, gained or lost that defines the different immune systems of all species at a molecular level.

1.1 Overview to the study of gene and genome evolution.

Large scale evolution of genomes occurs as a result of structural rearrangement events such as translocations, inversions, transpositions as well as in extreme cases through whole genome duplications. All of these events can lead to an alteration of gene number within a species if they occur in a portion of the genome containing coding genes, and one of the principal findings of large scale comparative genomic studies is that genomes are dynamic and not static with regard to gene content. The most common mechanisms by which gene content can be altered within a genome is through duplication or loss of existing genes and a nomenclature scheme has been devised to accurately define the relationships between genes related by such events (JENSEN 2001). Homologous genes are genes related to each other by descent from a common ancestral gene sequence (FITCH 2000). Within a family of homologous genes,

orthologs are defined as genes in different species that have evolved from a common ancestral gene with the evolution of orthologs reflecting organismal evolution. By contrast, paralogs are homologous sequences that are separated by a gene-duplication rather than a speciation event (SONNHAMMER and KOONIN 2002). Repeated duplications can lead to the presence of multiple paralogous sequences within a genome and the ability to distinguish between orthologs and paralogs is critical to accurately define the evolutionary relationships within a gene family (JENSEN 2001). One means to aid the definition of such evolutionary relationships is to examine the degree of gene order conservation between species. Synteny refers to two or more genes being assigned to the same chromosome in a species (FRAZER *et al.* 2003). Conservation of such gene order between divergent species represents shared synteny for this chromosomal segment and such conservation is one of the most reliable criteria for distinguishing between often highly similar orthologous and paralogous sequences within a genome (FRAZER *et al.* 2003).

1.2 Overview of comparative genomics and the study of immunology

Viewed from the outside, comparative genomics appears to be a research domain in its infancy, even though in reality over a quarter of a century has elapsed since its establishment as a scientific discipline. While genetics has traditionally focused on the link between genes and inheritance within species, comparative genomics is principally concerned with the relationships among the genomes of different species or intra-species strains. The expansion of interest in this field is directly linked to improvements in DNA sequencing technology. These advances have contributed to the availability of an ever increasing number of completely sequenced genomes, providing powerful resources for studying evolutionary changes among organisms. In common with all cellular mechanisms, the immune response of a species is dependent on the underlying gene complement coded in its genome. However unlike other scientific fields such as developmental biology, it has been difficult to carry out immune gene discovery and characterisation by classical means such as cross species polymerase chain reaction (PCR). The rapid rate at which

many immune systems evolve, both in terms of the sequence divergence of individual genes and the rate of gene duplication, loss and retention has meant that, even amongst closely related species, the design of accurate cross-species gene primers has proven virtually impossible in many cases. Comparing the sequences of known immune genes to the genomes of other species by computational means provides a mechanism by which the immune-gene complement of a species can begin to be understood, as distant relationships are often detectable by *in silico* means when they are not *in vitro*.

Comparative genomics provides a powerful resource for gene discovery, but its use in a systems-wide approach in the field of immunology has been undermined by the lack of an accepted definition as to what actually constitutes an immune gene. Most of the annotated immune genes have been identified from studies in vertebrates or more specifically mammals, and a number of different databases have been created which could be used as a "seed" for comparison between species (KELLEY *et al.* 2005), (ORTUTAY and VIHINEN 2009), (LYNN *et al.* 2008). These databases as well as several others have, in the main, been assembled through manual curation of published research literature for genes with known defense characteristics, such as expression in immune cells or tissues, known involvement in established immune pathways or direct interaction with pathogens. The subjective nature of the process of manually reviewing and selecting immune response genes in this manner is clear from the lack of overlap in the content of these databases. In fact these two databases have only one-third of their members in common despite being comparable in overall size. Another means of immune gene identification is through the use of microarray studies to identify genes which display altered expression in response to pathogenic insult (CALVANO *et al.* 2005), (ALIZADEH *et al.* 2000). Specialised immune system cells can be isolated and the genome wide transcriptional responses to pathogenic stimulus determined. The value of such studies is that novel immune-related genes which have not previously been linked to the generation of a defense response can be identified. The problem of high signal-to-noise ratios, endemic in microarray-based analysis

means that immune gene counts derived from such studies are likely to be overestimated. Altered expression of genes involved in normal cellular processes such as energy production and protein synthesis and degradation would be expected as infected cells alter their phenotype to overcome the infection. Using microarrays can erroneously identify these genes as immune genes even though their altered expression is more likely a consequence of, rather than a response to, the infection. The limitations stated above, coupled with the sheer scale and complexity of the immune system has meant that there is no definitive immune gene dataset for use in genome-wide cross species analysis and as such studies claiming to define the "immunome" of any newly sequenced genomes must be evaluated critically.

Considerably more success has come from comparative genomic studies which focus on smaller subsets of immune genes as opposed to an immunome-wide approach. Rather than employing the classical divisions of innate and adaptive to categorize immune genes, it can prove more useful for comparative studies to divide immune-related genes into groups based on their general functions. Three principal divisions are generally considered – (1) A grouping consisting of receptors involved in microbial detection and other cell surface interaction molecules, (2) mediators involved in intracellular immune signalling pathways and (3) effector molecules which directly interact with and inhibit microbial pathogens or influence the generation of an effective immune response by means such as chemotactic recruitment of mobile immune cells or influencing the transcription of other defense genes (SACKTON *et al.* 2007). All three groupings can be considered as inter-dependent, differentiated immune processes that are conserved in the genomes of all species. Whilst this thesis is mainly focused on components that contribute to the immune responses in vertebrates it is clear that, when compared over the massive phylogenetic distances separating the metazoan taxa, immune systems harbour an large amount of variation and diversity in the molecular components in all three groupings. Comparative analysis within each of these categories can provide important insights into both the origin and evolution of immune defense

systems particularly in relation to multi-gene protein families. The evolution of such families allows for the generation of great diversity from a limited number of building blocks with the most "successful" of such expanded families having large sizes within particular species and/or being widely distributed among diverse organisms.

1.3 Comparative genomics and microbial detection.

The initial step in generation of any immune response is the recognition of foreign antigens and a range of multigene families have been implicated in direct microbial detection. These gene families, collectively referred to as Pathogen Recognition Receptors (PRRs), recognize specific pathogen-associated molecular patterns (PAMPs) (KUMAR *et al.* 2009), that are expressed on the surface of microbes. PRR families, including Toll Like Receptors (TLRs) and NACHT-domain leucine-rich-repeat proteins (NLRs) have been implicated in this process in vertebrates with microbial detection by both these PRR subfamilies being carried out by the same domain type – Leucine Rich Regions (LRRs), (DOYLE and O'NEILL 2006; INOHARA *et al.* 2005). Another important detection domain, the Immunoglobulin Superfamily domain (IGSF), which comprise the antigen binding regions of T-cell receptors (TCR), Major Histocompatability Complex (MHC) proteins and immunoglobulin (Ig) antibodies is not considered here as in many cases they are coded in clusters of interchangeable gene segments rather than complete genes (MAIZELS 2005) providing an excellent example of the primacy of domains over proteins in the immune system discussed earlier. These segments form proteins that are not germ-line encoded and generate B and T cell antigen receptor diversity through rearrangement of immunoglobulin V, D, and J gene fragments in a process known as somatic recombination (MAIZELS 2005). In other cases IGSF domains form part of fully transcribed intact genes and perform a variety of functions, though in vertebrates direct microbial detection does not appear to be one of them.

1.3.1 Toll-Like-Receptors

The Toll-like receptor (TLR) multi-gene family encodes a family of PRRs which play a key role in vertebrate defense against both bacterial and viral pathogens (DOYLE and O'NEILL 2006). TLRs are so named due to homology with the Toll protein in *Drosophila melanogaster*, which was described as a mediator of dorso-ventral patterning in the fruit fly embryo as well as a potent activator of an antifungal response in the adult fly (IMLER and ZHENG 2004). Subsequently, the first Toll-related protein was identified in humans (now denoted TLR4) and expression of this protein on macrophage cells was linked to an increase in cytokine production through activation of the transcription factor NFκB (MEDZHITOV *et al.* 1997). The known TLR family has since expanded to encompass ten functional members in humans (TLR1-TLR10) and twelve members in mice (TLR1-TLR9, TLR11-TLR13) (MEDZHITOV 2007) with similar numbers recently identified in other vertebrates (ROACH *et al.* 2005). The evidence from several knockout studies has supported the proposition that the physiological function of TLRs is the recognition of pathogens and suggested that each TLR recognizes specific subtypes of microbial components (Figure 1.1A) (TAKEDA and AKIRA 2007).

TLRs are transmembrane proteins consisting of the extracellular leucine-rich regions (LRRs) responsible for both microbial recognition and receptor dimerisation and cytoplasmic Toll/interleukin-1 receptor (TIR) domains which activate an intracellular signalling cascade pathway through interaction with a family of adaptor proteins (Figure 1.1B) (O'NEILL and BOWIE 2007). Phylogenetic studies of known vertebrate TLRs supports the division of these genes into 6 relatively distinct families - 1, 3, 4, 5, 7 and 11 (ROACH *et al.* 2005).

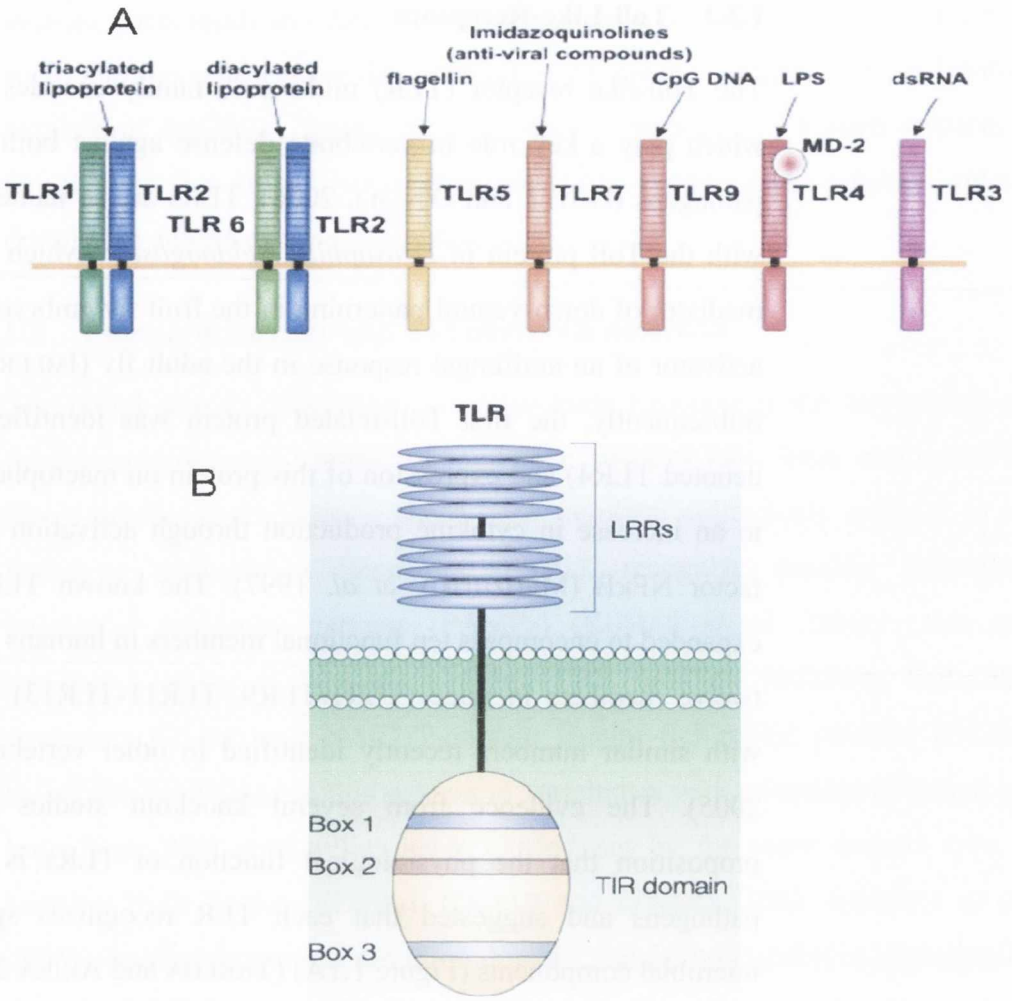


Figure 1-1. A. Different microbial ligands recognised by human TLRs, resulting in TLR pathway activation. Figure is adapted from (Yamamoto et al. 2004). B. Schematic representation of a Toll-like receptor. The extracellular domain contains multiple Leucine Rich Regions. The intracellular TIR domain is linked to a single transmembrane domain and contains three highly conserved regions denoted Boxes 1-3. Figure is adapted from (Akira et al. 2004)

1.3.1.1 TLR 2 family

The TLR2 family includes the mammalian TLR1, TLR2, TLR6 and TLR10 and represents all mammalian TLRs known to be involved in the recognition of bacterial lipopeptides (TAKEDA and AKIRA 2007). Phylogenetic evidence suggests that TLR2 and a TLR1-like gene emerged early in vertebrate evolution following duplication of an ancestral gene (ROACH *et al.* 2005). Later gene duplications in eutherian mammals gave rise to first TLR10 and subsequently TLR1 and TLR6 (KRUITHOF *et al.* 2007). Because of these mammalian specific gene duplications it is clear that any genes related to mammalian TLR1 identified in species that diverged prior to the divergence of eutherians and metatherians (approximately 180 mya) represent TLR1-like genes and should not be considered as true 1:1 orthologs with the mammalian TLR1 gene (TEMPERLEY *et al.* 2008).

A crucial step in lipopeptide detection by the TLR2 family is the formation of a heterodimer between TLR2 and other family members (AKIRA *et al.* 2006). TLR2 has been shown to recognize a wider range of PAMPs than is usual for a TLR, including peptidoglycan and lipoteichoic acid from Gram-positive bacteria, lipoarabinomannan from mycobacteria, glycosylphophatidylinositol anchors from *Trypanosoma cruzi*, zymosan from fungi, and glycolipids from the spirochaete *Treponema maltophilum* (TAKEDA and AKIRA 2007). It is through dimer formation with TLR1, TLR6 and most likely also TLR10 that TLR2 can discriminate between such diverse PAMPs. TLR1 and TLR6, the family members with the highest sequence similarity (69%) are responsible for detection of diacyl and triacyl lipopeptides respectively (OMUETI *et al.* 2005; TAKEUCHI *et al.* 2002; WETZLER 2003). To date, no ligand has been identified for TLR10, however it has been shown to activate the same intracellular pathway as other TLR family members (HASAN *et al.* 2005).

1.3.1.2 TLR 3

TLR3 is an intracellular receptor localised to the endosomal membrane (MATSUMOTO *et al.* 2003) where it has been implicated in recognition of double-stranded RNA (dsRNA). dsRNA is a molecule associated with the replication phase of viral lifecycles and in mice a synthetic dsRNA polyinosine-polycytidylic acid (poly(I:C)) has been shown to induce the expression of the type I anti-viral interferons - interferon α (IFNα) and interferon β (IFNβ) through TLR3 mediated NFκB activation (ALEXOPOULOU *et al.* 2001). Furthermore, TLR3 double negative mice have been shown to be unresponsive to reovirus dsRNA (ALEXOPOULOU *et al.* 2001). Recently human TLR3 has been associated with recognition of *in vitro* transcribed mRNA in dsRNA non-responsive HEK293 cell lines, resulting in the expression of the proinflammatory cytokine IL-8 suggesting that endogenous RNA could be a functional ligand for TLR3 in human (KARIKO *et al.* 2004).

1.3.1.3 TLR4

As previously stated, human TLR4 was the first vertebrate TLR to be described (MEDZHITOV *et al.* 1997), and like the members of the TLR1 family, TLR4 is expressed on the cell surface and has been shown to be involved in lipid recognition. Specifically, experiments using TLR4-expressing HEK293 cells and an NFκB-dependent luciferase reporter assay have shown TLR4 to be the main receptor for bacterial LPS (CHOW *et al.* 1999). Support for the LPS specificity of TLR4 was provided by experiments using two LPS-unresponsive mouse strains. C3H/HeJ mice which have a proline to histidine mutation at the TLR4 residue 712, and C57BL/10ScCr strains which are homozygous for a null mutation in the TLR4 signalling domain were independently shown to be unable to initiate a response to LPS infusion (POLTORAK *et al.* 1998; QURESHI *et al.* 1999). Immune cells derived from TLR4 knockout mice also displayed no response to LPS (HOSHINO *et al.* 1999). It was these studies that provided the first irrefutable links between TLRs and vertebrate immunity.

Uniquely amongst mammalian TLRs, TLR4 requires the presence of at least three non-TLR accessory proteins in order to initiate LPS-induced signalling. In mammals, LPS is initially bound by circulating LPS Binding Protein (LPB) and subsequently transferred to CD14, a protein found secreted in serum or expressed on the surface of macrophages as a glycophosphoinositol (GPI)-linked peptide (WRIGHT *et al.* 1990). CD14 is thought to transfer the bound LPS to a cell surface complex consisting of TLR4 and the small accessory protein MD-2 (DA SILVA CORREIA *et al.* 2001). Loss of, or mutation in any of these components has been shown to contribute to LPS hyporesponsiveness in mammals. In mice, MD-2 (NAGAI *et al.* 2002) and CD14 (HAZIOT *et al.* 1996) deficiency reduces LPS responsiveness considerably while in humans an MD-2 polymorphism causes almost complete loss of lipopolysaccharide induced signalling (HAMANN *et al.* 2004). In addition to LPS, TLR4 has been shown to recognize the plant toxin Taxol (KAWASAKI *et al.* 2001) as well as several endogenous ligands such as fibrinogen and the Heat Shock Proteins HSP60 and HSP70. The ability to recognize HSPs may play a role in the inflammatory response to necrotic cells, though any TLR4 mediated response seen in experiments such as this may yet be shown to have been caused by LPS contamination of the HSP preparations (OHASHI *et al.* 2000).

1.3.1.4 TLR 5

TLR5 is unusual among mammalian TLRs in that it recognizes a protein ligand. Bacterial mobility is derived from the movement of short-rod like structures known as flagella. Flagellin is a critical protein component of these bacterial flagella in both Gram-positive and Gram-negative bacteria (HAYASHI *et al.* 2001). Mutation studies have proposed the region 386-407 in TLR5 as the flagellin binding motif (MIZEL *et al.* 2003), while the TLR5 recognition site in flagellin has been mapped to a 13 amino acid region that is critical for flagellar locomotion (SMITH *et al.* 2003). Recently, human pathogens such as *Campylobacter jejuni* and *Helicobacter pylori* have been shown to evade recognition by TLR5 by mutating key residues in the TLR5 recognition site in flagellin while compensatory mutations elsewhere in the protein allow these

pathogens to maintain mobility (ANDERSEN-NISSEN *et al.* 2005). TLR5 has been shown to induce TNF-α in mouse macrophages in an NFκB-dependent manner (HAYASHI *et al.* 2001) while an increased production of IL-6 and IL-8 has been detected in human corneal epithelia (ZHANG *et al.* 2003). A common stop codon polymorphism in the ligand-binding domain of TLR5 which results in a truncated protein without a transmembrane or intracellular signalling domain has been shown to be associated with susceptibility to pneumonia caused by the flagellated bacterium *Legionella pneumophila* (HAWN *et al.* 2003).

1.3.1.5 TLR 7 family

The TLR7 family comprising TLRs 7, 8 and 9 are all intracellular, endosomally expressed proteins responsible for recognition of nucleic acids and haem motifs (BARTON *et al.* 2006; ROACH *et al.* 2005). TLR7 is the best characterized family member and was first identified as a receptor for imidazoquinolone compounds such as imiquimod, which stimulate an antiviral response in dendritic cells and macrophages by inducing synthesis of IFN-α and pro-inflammatory cytokines (STANLEY 2002). This response is absent in TLR7 knock out mice (HEMMI *et al.* 2000). Subsequently, single strand RNA (ssRNA) was identified as the natural ligand for TLR7 (HEIL *et al.* 2004) and TLR7 has been implicated in detection of both the human parechovirus 1 (HPEV1) and the influenza virus (MELCHJORSEN *et al.* 2005) (LUND *et al.* 2004).

TLR8 is both sequentially and structurally similar to TLR7 and in humans both genes are located on the X chromosome. Despite the relatively ancient gene-duplication event that separates the two genes (ROACH *et al.* 2005), a low level of sequence divergence has meant that both TLR7 and TLR8 recognise similar ligands. Imidazoquinolones and ssRNA are both bound by TLR8 in humans though this affinity is absent in mouse (AKIRA *et al.* 2006). The retention of both TLR7 and TLR8 in most species thus far studied could indicate a level of

functional redundancy between these two TLRs, though recent evidence suggests that in mice at least, TLR8 may be activated in a ligand specific manner by simultaneous stimulation with the imidazoquinolines and polyT oligodeoxynucleotides (GORDEN *et al.* 2006).

TLR9 recognizes CpG DNA motifs that are present in bacterial and viral genomes (BARTON *et al.* 2006). In contrast to vertebrates, microbial CpG motifs are unmethylated and can thus be differentiated from the host's own DNA. Expression of human TLR9 confers cellular responsiveness to synthetic CpG containing oligodeoxynucleotides in HEK293 cells (CHUANG *et al.* 2002), a response that is not observed in TLR9 knockout mice (HEMMI *et al.* 2000). TLR9 also appears to be capable of differentiation between the alternative forms of microbial CpG DNA – CpG-A and CpG-B. These structurally different CpG types elicit a differing cytokine response following binding by TLR9. CpG-B is a potent inducer of inflammatory cytokines such as IL-12 and TNF-α whereas CpG-A recognition induces dendritic cells to produce IFN-α (HONDA *et al.* 2005).

1.3.1.6 TLR 11 family

The TLR11 family consisting of TLRs 11-13 and 21-23 represent the most diverse TLRs when compared among species and also the most sparsely represented TLR family when inter-species repertoires are compared (ROACH *et al.* 2005). In humans for example, this family is represented by a single TLR11 pseudogene whilst mice retain functional copies of TLRs 11, 12 and 13. Little is known about the specific ligands for this family though in mice TLR11 has been shown to recognize uropathogenic bacteria, most likely through recognition of the actin binding protein Profilin (YAROVINSKY *et al.* 2005).

1.3.2 TLRs – genomic comparisons

One of the critical outcomes resulting from the comparative analysis of diverse sequenced genomes was the discovery that gene families involved in pathogen recognition represent evolutionarily labile subsets of genes when compared between species. Both the TIR and LRR domains which constitute the functionally active regions of TLRs have been identified in genome searches of organisms separated by the earliest divergences within the animal kingdom. However, the single TLR gene identified in the basal cnidarian, *Nematostella vectensis* represents the earliest diverged animal lineage in which an accepted TLR gene, consisting of an LRR, transmembrane and TIR domain has been identified (PUTNAM *et al.* 2007). This gene points to an origin of the modern TLRs in the metazoan ancestor of more than 800mya, prior to the separation of bilaterians and cnidarians (LEULIER and LEMAITRE 2008). The contribution made by these early ancestral TLRs to development or immunity has not been determined.

The next major divergence within metazoans involved the split leading to the protostomal and deuterostomal superphyla. This divergence point also represents a critical split in relation to the structural organization of animal TLRs. Protostomes including Ecdysozoa (arthropods and nematodes), Platyzoa (helminthes and rotifers) and Lophotrochozoa (annelids and molluscs) almost exclusively code for mccTLR proteins, which contain multiple cysteine clusters in their extracellular region (IMLER and HOFFMANN 2001). In contrast the TLR genes found in the deuterostomes (chordates and echinoderms) are almost entirely of the sccTLR subtype, as they are characterized by the presence of a single cysteine cluster between the final LRR domain and the transmembrane region (IMLER and ZHENG 2004). A small number of exceptions to the protostomal and deuterostomal TLR structural split have been identified, indicating that the last common ancestor (LCA) of these two superphyla most likely coded for both subtypes of TLRs which have been differentially expanded through gene duplication and gene loss events in both the deuetrostome and protostome lineages (LEULIER and LEMAITRE 2008).

Most of our knowledge of protosomal TLR evolution is derived from the studies of the arthropods, *Drosophila melanogaster*, *Apis mellifera* and *Aedes aegyptis,* and the nematode *Caenorhabditis elegans*. Comparison of sequenced arthropod genomes indicates that insects have between five and twelve TLRs (LEULIER and LEMAITRE 2008). Of the nine drosophila TLRs identified, only one, Toll-1, has been conclusively shown to contribute to the immune response in the fruit fly, while the differing embryonic expression patterns of the other eight TLRs has been suggested as evidence that these genes play a role in development rather than defence (LEMAITRE and AUSUBEL 2008). A single gene coding for a TLR (TOL-1) was identified in the *Caenorhabditis elegans* genome (PUJOL *et al.* 2001). Initially this protein was believed to function as part of the worm nervous system, where it would recognize pathogens and allow the worm to avoid them. Recently Tol-1 has been shown to directly contribute to the immune response by altering expression of immune effector genes when in the presence of Gram negative but not Gram-positive bacteria (TENOR and ABALLAY 2008).

TLR genes in deuterostomes are believed to function solely as pathogen recognition receptors and not to play a role in embryonic development (HIBINO *et al.* 2006). Within the deuterostome superphylum, the main phylogenetic division separates chordates from echinoderm (BOURLAT *et al.* 2006). To date, the echinoderm sea urchin, *Strongylocentrotus purpuratus* is the only non-vertebrate deuterosomal genome to have been fully sequenced and is proposed to provide a valuable insight into the immune responses in the animal kingdom in the absence of an adaptive immune system (HIBINO *et al.* 2006). In sea urchins, the TLR family consists of 222 individual genes, representing by far the greatest expansion of the family in any species yet analysed. Phylogenetic analysis indicates that of these genes, 211 represent a species-specific expansion as these genes are more similar to each other than to TLRs in other species (HIBINO *et al.* 2006). These 211 genes show remarkable diversity in the primary amino acid sequences of their LRR containing ectodomains possibly

indicating an expanded pathogen recognition capability when compared to protostomal organisms. Of the remaining 11 sea urchin TLR genes, 3 have been shown to be similar in structure to the protostome TLRs described above. Three others are unusual in that they contain introns and the remaining 5 have a non-characteristic shortening of the extracellular region (HIBINO *et al.* 2006). Until another echinoderm genome is sequenced the analysis of the expanded sea urchin TLR family will be difficult to resolve.

The TLR species repertoire is far more extensively studied in the chordate phylum than in any other. In the subphylum Cephalochordata, the genome of the Amphioxus (*Branchiostoma floridae*) possesses 42 TLR genes (HUANG *et al.* 2008). The closely related urochordates, *Ciona savignyi* and *Ciona intestinalis,* showed notable differences in their TLR repertoire coding for 7 and 3 TLR genes respectively (ROACH *et al.* 2005). Sequence divergence by amino acid replacement and insertions and deletions, particularly in the LRR containing ectodomain mean that no direct one-to-one orthology can be determined to any of the true vertebrate TLRs discussed below.

Vertebrates as a subphylum have been in existence for over 600 million years. Despite this timespan there is remarkably little variation in the number of TLR genes coded for by each vertebrate species with sequenced genomes coding for between 9 and 19 genes (ROACH *et al.* 2005). One possible explanation for this is the emergence in these species of an adaptive immune response which may have complemented the TLRs as primary immune sentinels in this lineage. Indeed, increasing evidence suggesting adaptive immune system activation to be an important consequence of TLR mediated immune responses could support that view. Similar gene counts are, of course, a crude measure of biological similarity. There is no reason why two genomes should not encode differing sets of TLR coding genes, but still have similar overall totals and in fact each vertebrate genome sequenced to date appears to code for different and presumably species-specific repertoire of TLRs.

The six principal families of vertebrate TLR (described above) appear to have arisen very early in vertebrate evolution and with a few notable exceptions, at least one member of each of these families is represented in all the vertebrate genomes thus far examined (ROACH *et al.* 2005). The structural similarity of the PAMPs recognized by members within each TLR family suggests a degree of functional redundancy for pathogen recognition and could explain why one member of each family is often sufficient for survival. In spite of this, species or clade-specific expansions of individual TLR families have been observed. Within the TLR1 family, TLR2 is ubiquitously represented in vertebrates, whilst the TLR1-like genes appear to have undergone differential, independent expansions in the mammals, birds, amphibians and fishes (ROACH *et al.* 2005). The TLR3, TLR4 and TLR5 families are also universally represented with the exception of TLR4 which appears to have been lost in the Green spotted puffer fish (LEULIER and LEMAITRE 2008). Gene duplication events early in vertebrate evolution expanded the TLR7 family to three members – TLR7, TLR8 and TLR9. Subsequently TLR7 has been lost in the *Tetraodon nigroviridis* genome, TLR8 has been independently lost from the chicken and rat genomes and a TLR9 gene is also absent in the chicken genome (LEULIER and LEMAITRE 2008; TEMPERLEY *et al.* 2008). TLR8 appears to have been duplicated in zebrafish (ROACH *et al.* 2005). The TLR11 family has been shown to be the most sparsely represented receptor subset, when vertebrate TLR repertoires are examined. A functional, non-pseudogenised TLR11 has been identified in eutherian mouse (ZHANG *et al.* 2004), the marsupial opossum and the amphibian xenopus genomes (ROACH *et al.* 2005). TLR12, the most recently diverged member of this family appears to be mammalian specific and to date has only been identified in the mouse, rat and opossum genomes. TLR13 identification has been restricted to rodents and xenopus and most likely lost from the genomes of all other sequenced vertebrates (ROACH *et al.* 2005). The other TLR11 family members – TLRs 21-23 are predominantly found in fish species, though an ortholog of TLR21 has been claimed for chicken (TEMPERLEY *et al.* 2008). No mammalian species has been shown to utilize any of the members of this TLR21-23 subfamily for pathogen recognition.

Prior to the work carried out for this thesis, only two potentially novel TLRs had been identified in any vertebrate species. A gene – denoted TLR14 - has been identified in the genomes of the fugu and tetraodon fish species and the amphibian xenopus (ROACH *et al.* 2005). This gene also appears to have been the progenitor gene for a subsequent duplication leading to TLR18 in fugu. A second novel xenopus gene – TLR16 - appears to be a distantly related member of the TLR11 family (ROACH *et al.* 2005). Though each of these TLRs has been provisionally assigned to one of the established TLR families described above, sufficient sequence diversification has occurred to prevent accurate resolution of orthology for these genes with any of the better established family members. The designation of these TLRs as novel PRRs is based on the supposition that the level of sequence divergence observed could permit the recognition of a different class of PAMP to the other family members. As little functional work has been conducted on TLRs outside of human, mice and drosophila this view is open to challenge.

1.4 Comparative genomics and immune signalling pathways

Recognition of invading microbes represents only the initial step in the generation of an immune response. Ultimately a pathogen specific profile of immune proteins is generated to respond to the invader. This profile of proteins can range from antimicrobial peptides which can directly attack and kill bacterial cells, to cytokines and co-stimulatory molecules that can orchestrate both the cellular and humoral immune responses (O'NEILL 2008). Expression of most of these proteins is tightly regulated and controlled by a small number of transcription factors. In all cases a complex multi-protein signalling cascade links the activated ligand-bound PRR with the appropriate transcription factor. The best characterized and most conserved of these cascades is the pathway utilized by both Toll-Like-Receptors and the Interleukin-1 receptor.

1.4.1 Toll-Like Receptor Signalling pathway

The initial signalling event following the formation of ligand induced TLR hetero or homodimers is the recruitment of TIR domain containing adaptor molecules to the intracellular side of the cell membrane (Figure 2). With the exception of TLR3, all TLRs can recruit the TIR domain-containing adaptor molecule myeloid differentiation factor 88 (MyD88) (MEDZHITOV *et al.* 1998). MyD88 proteins consist of a C-terminal TIR domain and an N-terminal death domain. Interaction between the TIR domains of the TLR and the adaptor protein recruits members of the IL-1R-associated kinase (IRAK) family to the membrane bound complex (TAKEDA and AKIRA 2004). Members of the IRAK family contain a serine-threonine kinase domain as well as a death domain which interacts with the death domain of the adaptor proteins. Four members of this family have been described in mammals- IRAK-1, IRAK-2, IRAK-M and IRAK-4, (LI *et al.* 2002; MUZIO *et al.* 1997; SUZUKI *et al.* 2002; WESCHE *et al.* 1999) , with knockout studies in mice indicating that there exists a functional redundancy within the family (THOMAS *et al.* 1999). IRAK4 is first recruited to the TLR/MyD88 complex, followed by IRAK1. Both IRAK-1 and 4 can undergo autophosphorylation during signal transduction leading to dissociation of IRAK-1 and -4 from MyD88. IRAK-2 is associated with a MyD88 independent pathway of TLR signalling (see below). IRAK-M is now thought to down regulate TLR signalling, by preventing this dissociation (KOBAYASHI *et al.* 2002). Another negative regulator of the TLR pathway is the Toll interacting protein (Tollip). This protein is found complexed with IRAK (ZHANG and GHOSH 2002) and must dissociate before IRAK1 can interact with a central protein in the TLR pathway known as Tumor Necrosis Factor receptor-associated factor-6 TRAF6 (MUZIO *et al.* 1998).

Activated TRAF6 can lead to the initiation of the inhibitor of Nuclear Factor Kappa B kinase (IKK) via either the "evolutionarily conserved signalling intermediate in Toll pathways" (ECSIT) protein and mitogen-activated protein kinase/ERK kinase kinase-1 (MEKK-1) (KOPP *et al.* 1999), or via recruitment

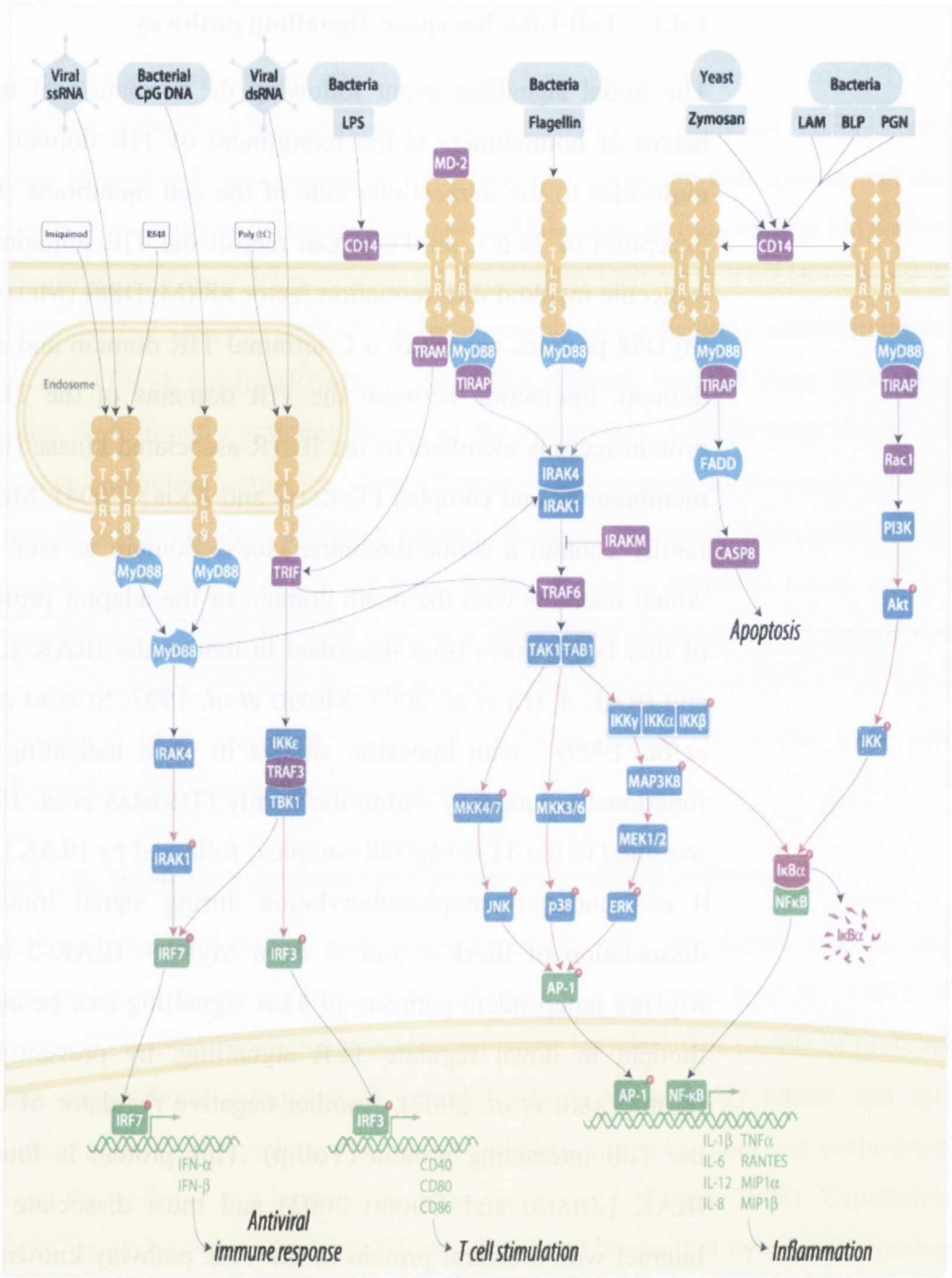


Figure 1-2. Schematic of the mammalian TLR signalling pathway. The details of the pathway are described in the main text. Figure adapted from Invitrogen website (http://www.invitrogen.com).

of TGF-β-activated protein kinase-1 (TAK1), with associated TAK1-binding proteins (TAB1 and TAB2) (NINOMIYA-TSUJI *et al.* 1999) and the ubiquitinylating factors ubiquitin-conjugating enzyme E2 variant 1 (UEV1A) and ubiquitin-conjugating enzyme 13 (UBC13). Once activated, the IKK proteins phosphorylate and thus target for degradation the Inhibitor of nuclear factor Kappa B (IκB), allowing the NFκB transcription activator to translocate to the nucleus and initiate expression of particular immune response genes.

The type of signalling pathway utilized and the transcription factor activated is ultimately dependent on which one of a number of TIR-domain containing adaptor protein is recruited to the cell surface. A second adaptor protein, MyD88 adaptor-like (Mal) (O'NEILL 2002), also known as TIR domain containing adaptor protein (TIRAP) was found to mediate the MyD88 dependent pathway response of both TLR2 and TLR4 by acting as a bridging molecule between MyD88 and both these TLRs (YAMAMOTO *et al.* 2002). Subsequent homology searches of publicly available resources led to identification of previously unknown TIR domain containing proteins (O'NEILL *et al.* 2003). Two of these proteins have been shown to be involved in the MyD88 independent pathway, which is initiated principally following TLR3 and TLR4 activation. The recruitment of one of these proteins, TIR-domain-containing adaptor protein inducing IFN-β (TRIF or TICAM-1) has been implicated in increased activity of the IFN-β promoter IRF-3 and as such forms a vital link in the mammalian response to viral infections initiated by TLR3 recognition of double stranded viral RNA (OSHIUMI *et al.* 2003). The adaptor molecule, TRIF-related adaptor molecule (TRAM) has been shown to be involved in the MyD88 independent signalling by TLR4 receptors, but not by TLR3 (YAMAMOTO *et al.* 2003), while a further adaptor peptide Sterile α and armadillo motifs (SARM), has as yet not undergone sufficient functional characterisation to accurately determine its probable interactions (MCGETTRICK and O'NEILL 2004). The MyD88 independent pathway has also been shown to activate NFκB, but with comparatively decreased effectiveness (KAWAI *et al.* 1999).

Recent years has seen a considerable increase in the number of genes implicated in TLR signal transduction. It would be naive to consider our current perspective as complete and in reality additional genes will most likely be identified which contribute to this signalling cascade. Furthermore systems-biology approaches will provide a more 3-dimensional perspective of signalling pathways such as the TLR pathway and should provide a clearer picture of the protein-protein interactions within the pathway and in between pathway crosstalk.

1.4.2 TLR signalling pathway - genomic comparisons

Naturally, pathways as complex as the TLR pathway described above do not arise spontaneously. Complex signalling pathways tend to develop and evolve through recruitment of existing proteins, based on whether the chemical action they can perform is appropriate within the constraints imposed by existing components (JENSEN 1976). When viewed conceptually, the TLR pathway, in common with most signalling cascades, can be considered as a series of tightly controlled protein phosphorylation events carried out by differentiated families of kinase domain containing proteins. Approximately 2% of all eukaryotic genes possess a kinase domain and across a broad species range, they help propagate cellular signals in processes as diverse as immunity, cell-cycle regulation, neuronal functions and morphogenesis (SCHEEFF and BOURNE 2005). The principle subdivisions within the kinase superfamily are observed across all eukaryotes indicating that rapid divergence and specialization of the different kinase subfamilies occurred very early in the eukaryotic phylogeny (SCHEEFF and BOURNE 2005). Subsequently they appear to have been differentially recruited to function in the various signalling pathways, often in a lineage specific manner.

Comparative genomic studies have identified orthologs for several individual TLR signalling components in the earliest diverged eukaryotic lineages, including multiple members of the IRAK and TRAF families (ORTUTAY *et al.*

2007). No functional studies have been carried out to show whether these genes function in unison to form a pathway involved in immune responses. The absence of canonical TLRs and transcription factors of known involvement in immune related processes at this taxonomic level would suggest that any signalling pathway composed of these proteins in early eukaryotes would be significantly different from the well characterized bilaterian model.

Sequence comparison studies of extant cnidarians provide insight into the early evolution of the canonical TLR pathway. Homology searches of the genome sequence of the sea anemone *Nematostella vectensis* identified orthologs for most of the principal intracellular mediators involved in the TLR pathway including a single TIR-domain containing adaptor (MyD88) and unambiguous homologs of mammalian transcription factors NF-κB and IRF3 (MILLER *et al.* 2007). In addition, TLR signalling via the Jun N-terminal kinase (JNK) and p38 mitogen-activated protein kinase (MAPK) pathways kinase pathway utilizes the ECSIT adaptor protein. The *Nematostella vectensis* genome possesses a single ECSIT protein, indicating that this variation of NF-kB activation may be considerably more ancient than previously thought. The unexpected finding of so many TLR pathway components in the sea anemone genome is not sufficient to determine if the pathway is carrying out a homologous role in cnidarians and bilaterians, though the presence of a number of downstream, classically NF-kB-mediated effectors could indicate that such homologous functionality is likely (MILLER *et al.* 2007). A second cnidarian species *Hydra magnipapillata* provides a contrast to this extensive conservation of the *Nematostella vectensis* TLR pathway (MILLER *et al.* 2007). Whilst many of the pathway components are conserved in Hydra, no TLR or NF-kB genes were identified in the sequenced genome. Despite this, expression of downstream effector molecules of the TLR pathway, including Antimicrobial peptides (AMPs) has been detected in Hydra suggesting that hydrazoans may possess an analogous rather than homologous immune signalling cascade to other cnidarian species (HEMMRICH *et al.* 2007).

The protostomal species with sequenced genomes, *D. melanogaster* and *C. elegans*, display a similar pattern of within phylum variation as to the make up of the TLR signalling pathway. The toll pathway was initially characterized in *Drosophila* and suggested to play a key role in the establishment of the dorso-ventral axis of the *Drosophila* embryo (BELVIN and ANDERSON 1996). Subsequent studies involving knockouts of individual components of the fly pathway implicated it in defense against both Gram-positive bacterial and fungal infections (LEMAITRE *et al.* 1996). To date, Toll signalling has only been directly implicated in establishing polarity in fruitfly species suggesting this is most likely a secondary role developed within the arthropod lineage (MINAKHINA and STEWARD 2006). Unfortunately, the initial discovery of this pathway in drosophila and the inevitable comparison to the mammalian cascade has led to the misconception among immunologists that the fruitfly pathway represents the progenitor pathway from which the apparently more extensive mammalian one has evolved. However, as described above, the presence of essentially complete signal transduction pathways in both cnidarians and bilaterians suggests that the integration of novel genes into these systems was largely complete in the eumetazoan ancestor (HEMMRICH *et al.* 2007) and that the fruitfly pathway is simply a selection-honed, taxon-specific variant of an already established signalling theme.

Initial experiments using genetic screening techniques identified several vital components of the drosophila Toll pathway including Pelle an adaptor molecule, Tube (an IRAK family kinase), dTRAF2 (a homolog of the known mammalian TRAF6), dECSIT, Cactus (the *Drosophila* homolog of IκB), as well as two NF-kB relate transcription factors Dorsal and Dif (BELVIN and ANDERSON 1996; TAUSZIG-DELAMASURE *et al.* 2002). Identification of various deficiency mutants using genetic screening is likely to miss pathway components where functional redundancy within multi-gene families can mask the effect of these mutations. However, the sequencing of the first drosophila genome (ADAMS *et al.* 2000) allowed for identification of several other TLR pathway molecules in the fruit fly including a drosophila homolog of the

MyD88 adaptor protein which directly links the membrane expressed TOLL protein to the already known downstream mediators (HORNG and MEDZHITOV 2001; SUN *et al.* 2002; TAUSZIG-DELAMASURE *et al.* 2002) , as well as two TAK1-like genes and two further members of the TRAF family of adaptor proteins (MANNING *et al.* 2002). Unlike in mammals, a second pathway, the Immune Deficiency (IMD) pathway, mediates the response to Gram-negative bacteria in drosophila. This pathway shares some similarities with the mammalian TNF-R pathway but is not as well characterized as the TLR cascade. Mutations in the IMD death domain containing adaptor protein inhibited the flies' ability to overcome Gram-negative bacterial infections and activation of the IMD pathway ultimately leads to the activation of the NFκB-like transcription factor Relish (STOVEN *et al.* 2000). Two principal mechanisms are required for Relish activation. In one, a complex consisting of a DREDD caspase and the Fas-associated death domain (FADD) homolog BG4 (HU and YANG 2000) cleaved Relish in the absence of a proteasome (LEULIER *et al.* 2000). Mutation in either protein severely inhibits the expression of genes responsible for antibacterial activity (LEULIER *et al.* 2002; NAITZA *et al.* 2002). The second mechanism implicated in Relish activation involves the Drosophila homologs of the mammalian IKK complex. IKKβ/ird5 and IKKγ/Kenny, which form the *Drosophila* equivalent of the mammalian IKK signalosome (RUTSCHMANN *et al.* 2000). These kinases are themselves activated by dTAK1 (homolog of mammalian TAK1) and mutations in any of these kinases result in a phenotype similar to IMD and Relish mutants. Once phosphorylated Relish is translocated to the nucleus where is induces expression of effector genes specific to Gram-negative infections (TZOU *et al.* 2002) (see below).

Investigation of the immune signalling pathways of another protostomal species, the nematode *Caenorhabditis elegans* has only recently begun. Initial experiments suggested that the single *C. elegans* TLR (TOL-1) was not involved in direct defence responses (PUJOL *et al.* 2001). The absence of a worm ortholog for the NF-kB transcription factor was seen as evidence of

TLR-independent mediated response to bacteria. Recent analysis, however, has suggested that TOL-1 does play a role in the nematode response to *E.coli* (TENOR and ABALLAY 2008). Comparative studies identifying TIR-1, TRF-1, PIK-1 and IκB-1, homologs of the mammalian downstream signal transduction components (SARM), TNF receptor-associated factor 1 (TRAF1), interleukin 1 receptor associated kinase (IRAK) and inhibitor of NF-κB (IκB), respectively (PUJOL *et al.* 2001). Double mutant studies of these components of the canonical TLR pathway show that TRF-1, but not IKB-1, was required for the immune effects of TOL-1, suggesting the TLR pathway in nematodes has evolved to utilize different downstream mediators following the loss of any NF-κB homolog from the worm genome.

Within the deuterostomes, most of the variation observed in the TLR pathway is derived from species-specific expansions of the TIR-domain containing adaptor proteins, suggesting variations in TLR induced responses in deutrostomes is derived from differential use of a limited subset of such proteins. In the sea urchin, the massive expansion of TLRs themselves is paralleled by a modest increase in the family of intracellular adaptors which can interact with these receptors (HIBINO *et al.* 2006). The sea urchin genome codes for a single MyD88 protein as well as 3 cytoplasmic proteins with a MyD88-like domain. The presence of 14 SARM or SARM-like genes indicates an expansion of this adaptor family has also occurred. The presence of homologs for most other components of the mammalian TLR pathway indicates that at the taxonomic level the functioning of the TLR pathway most likely closely mirrors the mammalian cascade to a significant degree (HIBINO *et al.* 2006). The genome of the cephalochordate Amphioxus encodes four MYD88-like, 10 SARM1-like, one TIRAP-like, and one TICAM2-like gene as well as 24 members of an extended TRAF protein family (HUANG *et al.* 2008). At least one ortholog corresponding to each of the known 6 vertebrate TRAFs is recognized in the amphioxus genome. In contrast to this, most of the TLR pathway immune-related transcription factors and the kinases responsible for

their activation are present in comparable numbers to those found in vertebrates (HUANG *et al.* 2008).

Considerably less variation in respect to pathway components is observed within the vertebrate lineage. Early in vertebrate evolution and prior the major vertebrate radiation, whole genome duplication events created 4 paralogous chromosome segments, two of which have been found to contain the TICAM1 and TICAM2 adaptor proteins across in all mammalian species (SULLIVAN *et al.* 2007). The gene coding for TICAM2 was subsequently independently lost in teleosts, amphibians and birds (SULLIVAN *et al.* 2007) indicating that the IFN derived anti-viral response may be subtly different in these lineages than in mammals (SULLIVAN *et al.* 2007). Within amniotes, gene duplication events led to the emergence of the IRAK-2 and TRAF1 genes (ORTUTAY *et al.* 2007), the retention of which is most likely mediated by the necessity for a more highly specific kinase activity to modulate the downstream events associated with the each different vertebrate TLR adaptors.

Overall comparisons of completely sequenced genomes have shown remarkable conservation of the fundamental pathway components required for TLR signalling. As yet it is impossible to rule out the possibility that a simplified form of the pathway may even predate the appearance of recognized modern TLRs themselves as the presence of a pathway in two species is not always indicative of homologous function. For example the vertebrate cytokine signalling cascade which uses members of the Janus kinase family is involved in development rather than defense in drosophila (ORTUTAY *et al.* 2007). What emerges from the comparative genomic approaches is a deeper appreciation of the acquisition of complexity in the evolution of even this one immune signalling pathway. Prior to this thesis a systematic analysis of lineage-specific evolution within the TLR pathway had been carried out using genomic data for members of the teleost (PURCELL *et al.* 2006) and mammalian lineages (KANEHISA and GOTO 2000), while this thesis is a first effort to carry out similar analysis using whole genome comparison techniques within class Aves.

1.5 Comparative genomics and vertebrate immune system effector molecules

The transcriptional response to activation of immune signalling pathways like the TLR cascade described above, varies considerably depending on the taxonomic level being considered. In vertebrates, the best characterized taxon grouping, a variety of proteins capable of directing the immune response are induced including cytokines, chemokines, prostaglandins, acute phase proteins and complement proteins (Janeway: Immunology 2007). The coordinated action of these molecules regulate and direct the immune and inflammatory responses, mainly through recruitment and activation of immune specific cells such as neutrophils and macrophages (Janeway: Immunology 2007). These differing classes of response protein have each arisen at a particular point along the evolutionary tree of life and been incorporated into the already existing immune defenses. For example the origin of cytokine and chemokine families of proteins can be traced to the base of the chordate phylum, before they underwent massive expansion and diversification in all vertebrate lineages (ORTUTAY *et al.* 2007) whilst homologs of components which make up the vertebrate complement system have been identified in more primordial lineages (MILLER *et al.* 2007). In the midst of all this variation in immune effector molecules, antimicrobial peptides (AMPs) are unique in that they are a universal feature of the defense systems of virtually all forms of life including bacteria, fungi, plants and animals (HANCOCK and LEHRER 1998). This grouping of several molecule subtypes with diverse evolutionary origins, represents a potent primary defensive barrier in all eukaryotes whilst prokaryotes almost exclusively use them to kill off same-species competitors for nutrients in their environment (HANCOCK and SAHL 2006).

Accurate classification of AMPs is hindered by the huge primary sequence diversity observed even amongst closely related subtypes. Despite this, all AMPs share a number of important characteristics. Typically these peptides are between 12 and 100 amino acids in length and possess an overall net positive charge (usually between +2 and +9). A high proportion (approx 30%) of

hydrophobic residues contributes to an overall amphipathic state for almost all known AMPs. The majority of AMPs are derived from larger precursor proteins and are post translationally modified to cleave the active peptide from the signal and prepropeptide regions (SEMPLE *et al.* 2006). The direct antimicrobial activity of most AMPs is attributed to the formation of pores following interaction of cationic, positively charged AMPs with the negatively charged phospholipids found in the outer membrane of prokaryotic but not eukaryotic cell membranes (KAGAN *et al.* 1990; SATCHELL *et al.* 2003). There is no clear picture as to the exact sequence of events that lead to pore formation and osmotic lysis of bacterial cells by AMPs (POWERS and HANCOCK 2003). Several models have been proposed for this process with most based around the amphipathic nature of the AMPs, where hydrophobic residues can interact with the membrane lipids and the hydrophilic residues then form the lumen of the transmembrane pore (BROGDEN 2005). Recent evidence suggests that AMPs may also induce bacterial killing via disruption of intracellular processes within microbes including inhibition of protein synthesis, inhibition of enzymatic activity and inhibition of cell wall synthesis (BROGDEN 2005).

As a consequence of the limited sequence similarity among differing AMPs, they are generally classified based on their secondary structure. Three principal groupings are considered. Group 1 includes linear peptides without cysteine residues that generally fold into an amphipathic α-helical structure. These include cecropins, magainins and cathelicidins. Group 2 consists of peptides with a rigid structure consisting of two to four β-strands which are stabilised by disulfide bridges between highly conserved cysteine residues. Defensins, which constitute a significant proportion of the group 2 members are discussed in greater detail below. Group 3 consists of peptides with a predominance of one or more specific amino acids such as the histidine-rich histatins derived from human saliva and the tryptophan-rich indolicidin bovine peptide (SELSTED *et al.* 1992). Comparison of the repertoires of AMPs indicates that, species possess a combination of multiple peptides from several of these structural classes described above. However group 2 peptides such as

defensins represent the one AMP structural type, almost universally maintained across all taxa.

1.5.1 Defensin Antimicrobial peptides

The defensin antimicrobial peptides are a family of small proteins characterized by the presence of six conserved cysteine residues in their active domain (LEHRER and GANZ 2002). Disulphide bridges formed between these residues stabilize the beta sheet structure of these proteins and the presence and spacing of these cysteine amino acids define the principal subdivisions observed within the defensin family (REHAUME and HANCOCK 2008). Most of the functional data pertaining to defensin activity is limited to a relatively small number of genes from an equally small number of species but in general, individual defensin AMPs tend to possess a very specific activity against a limited number of microbes (HIGGS *et al.* 2005). However the combined activity of all the family members present in a species can provide protection against both Gram-positive and Gram-negative bacteria, mycobacteria, fungi and some membrane-enclosed viruses (LEHRER and GANZ 2002).

1.5.2 Defensins – genomic comparison

As described above, comparative genomic analysis has been successful in identification of immune receptors and immune pathway components, even in comparisons attempted between distantly related species. In contrast, several important characteristics hinder the identification in genomic comparisons of antimicrobials such as defensins. The relatively small size of defensin genes coupled with the their rapid rate of duplication and diversification (HOLLOX *et al.* 2008) has meant that homology searches based solely on primary sequence similarity such as BLAST (ALTSCHUL *et al.* 1997) often have only limited potential to detect the presence of these genes in a genome of interest. This problem may be partially overcome by limiting comparisons to closely related species where conservation of synteny can contribute to accurate ortholog identification. However, the sparse representation of sequenced genomes within most other phyla has meant that a comprehensive cross-species genomic

analysis is only possible between members of the more densely sampled vertebrate phylogeny.

All vertebrate defensins possess a primary sequence signature which distinguishes them from other cysteine-containing β-sheet peptides found outside the phylum, although the overall fold adopted by these proteins is very similar (HIBINO *et al.* 2006). By contrast defensins identified in non-vertebrate animals possess a differing sequence signature in relation to the spacing of the conserved cysteine residues that stabilize the β-sheet backbone within the active peptide. Defensins in vertebrate species are further classified into three families, namely alpha, beta, and theta-defensins, based on modification of the vertebrate specific spacing pattern of the six cysteine residues (GANZ 2003)

A β-defensins like sequence most likely represents the ancestral gene from which all vertebrate defensins have descended (XIAO *et al.* 2004). β-defensins contain the motif C-X_6-C-X_4-C-X_9-C-X_6-C-C in their mature peptide region (LEHRER and GANZ 2002) with cysteine bridging occurring as C1-C5, C2-C4, C3-C6 (Figure 1.3). A second class of vertebrate defensin, denoted α-defensins, contain the motif C-X-C-X_4-C-X_9-C-X_9-C-C and form cysteine bridges between C1-C6, C2-C4 and C3-C5 (LEHRER and GANZ 2002). A third type of defensin, the theta or mini-defensin, has so far only been identified in Rhesus macaque (*Macaca mulatta*) called Rhesus θ-defensin 1 (RTD-1) (TANG *et al.* 1999). RTD-1 is an 18-amino acid cyclic peptide that is actually encoded by two separate truncated α-defensin-like precursors, RTD-1a and RTD-1b.

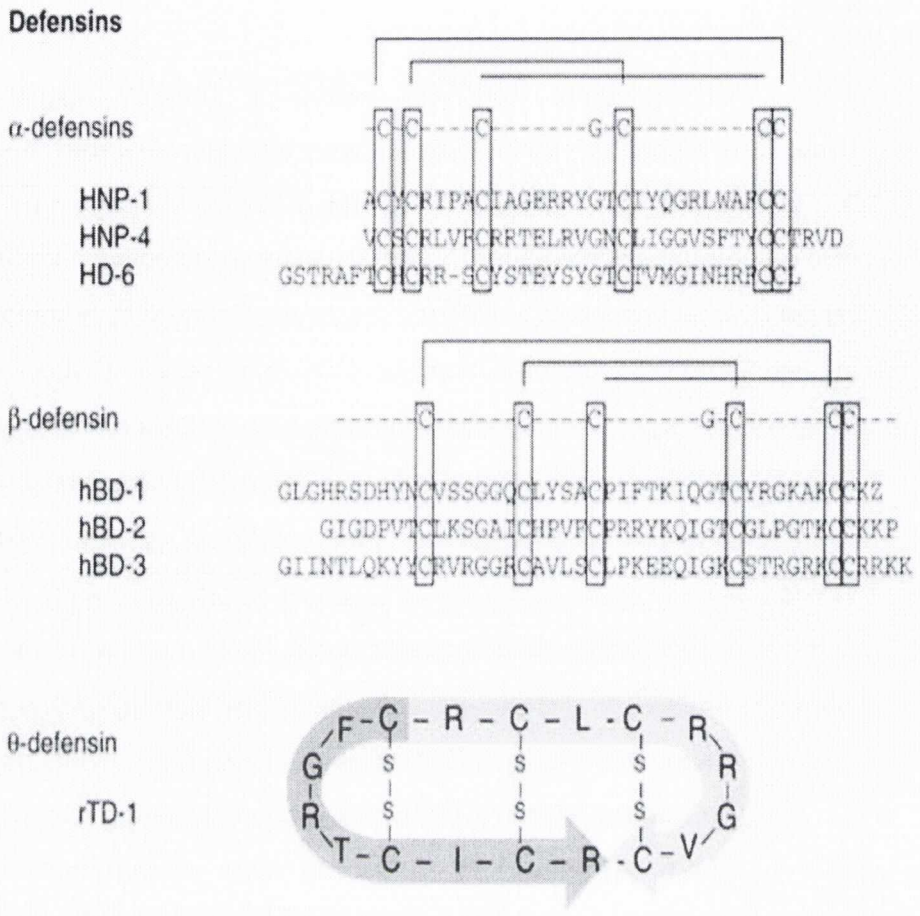


Figure 1-3. Sequence configuration of α-defensins, β-defensins and θ-defensin showing the order of cysteine bridge associations formed in each family subtype.

Comparative genomic studies searching within vertebrates for defensin motifs corresponding to those described above indicate that teleosts possess only β-defensins genes (Zou *et al.* 2007). A total of seven β-defensins genes were identified in searches of the zebrafish *Danio rerio* genome, with the pufferfish, *Takifugu rubripes* (3 genes) and *Tetraodon nigroviridis* (2 genes) genomes all possessing a β-defensin family (Zou *et al.* 2007). All seven genes identified in these searches share a gene structure unique to fish defensin genes. In fish, the six conserved cysteines span two exons with four cysteines being present in exon 2 and the last two contiguous cysteines coded in a third exon. This is in contrast to all other vertebrate defensins where the six cysteines are located

within a single exon. Whilst it is not possible to determine direct orthology to β-defensins genes in other vertebrates syntenic mapping of neighbouring genes in zebrafish matches those adjacent to the cluster in defensin locus on human chromosome 8 (denoted synteny group A) (WHITTINGTON *et al.* 2008) indicating that this might be the primary locus from which all other vertebrate defensin expansions have subsequently derived (see below).

The chicken defensin gene repertoire was one of the first to be completely characterised (LYNN *et al.* 2004; XIAO *et al.* 2004). A significant avian-specific expansion at the same syntenic locus (group A) as in zebrafish was identified, resulting in a cluster of 14 genes located in an 86 kb region in chromosome 3 (XIAO *et al.* 2004). As with the teleost lineage, no α-defensin genes were present in this taxon (LYNN *et al.* 2004). Recently the sequencing of the monotreme platypus (*Ornithorhynchus anatinus*) (WARREN *et al.* 2008) and the marsupial opossum (*Monodelphis domestica*) genomes (MIKKELSEN *et al.* 2007) has filled the gap in our knowledge relating to the considerably different defensin gene repertoires observed in avian and mammalian species. In platypus a differential expansion of defensin synteny group A to the one observed in chicken was identified. Six beta- and four alpha-defensin genes were predicted from genomic sequences including a single alpha-defensin displaying the sequence characteristic of both the alpha and beta defensin families. Until recently, α-defensins were believed to be restricted to primates and glires (rabbits, rodents) (SEMPLE *et al.* 2006) as no evidence of their existence had been found in either the artiodactyls or carnivore genomic sequences (PATIL *et al.* 2005; YOUNT *et al.* 1999). The recent availability of a large number completely sequenced mammalian genomes has shown that α-defensins are widely but sparsely distributed in that order (LYNN and BRADLEY 2007). Identification of homologous sequences in the platypus genome suggests a considerably older point of origin for the α-defensin lineage than previously thought and suggests that they have been differentially lost and expanded along individual branches of the mammalian lineages. Interestingly the platypus genome was also found to code for 3 defensin-like genes which

have diversified in their function and now form the main peptide components of platypus venom (WARREN *et al.* 2008). The opossum genome shows that defensin genes are arranged in a number of discrete clusters on mammalian chromosomes since before the meta and eutherian split, 200 mya. Three separate clusters totaling 37 genes (including a single α-defensin) were identified in the opossum. These are divided up into the established mammalian syntenic clusters A and D and a progenitor of the syntenic clusters B and C which subsequently split in therian mammals (MIKKELSEN *et al.* 2007).

As mammalian genomes have been sequenced, they have been thoroughly searched to quantify their defensin repertoires. The dog, human, mouse, and rat genomes have 43, 39, 52, and 43 beta-defensin genes respectively (PATIL *et al.* 2005) with the human and mouse genomes coding for 6 and 7 α-defensins respectively. The clustering of mammalian defensins into 4 syntenic groups (five in human and chimp following the split of syntenic group D in the family hominidea) is maintained across all eutherians. As a contrast to this overall synteny conservation, in each of the species so far analysed, novel species-specific defensin genes have been identified. The evolution of mammalian defensins most likely followed the standard model of gene family expansions, with duplication of an existing gene, followed by mutation, drift and natural selection based on the needs of the individual species (ZHANG *et al.* 2005). It is worth noting that the expansion of defensin gene numbers and the huge amount of sequence diversity observed between genes could be linked to an expanded functional role played by these peptides in mammals. Besides the direct antimicrobial activity of mammalian defensins, a receptor-mediated chemotactic action has been identified for the human β-defensin HBD3, though the molecular nature of the interaction between this defensin and the CCR6 chemokine receptor are not well understood (YANG *et al.* 1999). In addition mammalian defensins have been shown to contribute to the maturation of spermatozoa following secretion from epididymal epithelial cells (DACHEUX *et al.* 2003) as well as contributing to coat colour determinations in dogs and owls

through interaction with 7-transmembrane receptors on the cell surface. In effect, evolution appears to have utilized the basic defensin structural motif as a solution to several differing problems (YOUNT and YEAMAN 2004).

1.6 General techniques used in comparative genomic analysis

Bioinformatics is a broad, all encompassing term used to define the application of computers in biological research. In particular, pattern recognition - a task computers are well suited for - has been one of the key tools of bioinformatics since the infancy of the field, though the techniques and methods used have matured considerably, driven principally by the explosion in the amount of sequence data available for analysis. Two of the principal methods used for this purpose are described below.

1.6.1 BLAST

Similarity searching is one of the central tools of bioinformatics and BLAST (Basic Local Alignment Search Tool) is undoubtedly the most widely used algorithm for this purpose. BLAST identifies homologous sequences using a heuristic method by comparing a query sequence to all the sequences in a database in a pairwise manner. Many of the publicly available sequence databases contain millions of sequences composed of billions of bases or amino acids and as a consequence, search methods using entire sequence lengths are not feasible due to processing speed and memory requirements. To maximise efficiency BLAST finds the best local, rather than global, alignments between query and target sequences. To achieve this, the query and database sequences are broken up into "words" – the ideal length of which depends on whether DNA or proteins are being analysed. Initial steps attempt to align the query words of length "W" to the target sequences. Matching word alignments which satisfy a threshold value, "T" are referred to as High-Scoring-Segment-Pairs (HSPs). The score used is calculated using a scoring matrix such as the default BLOSUM62 used in protein-protein searches. The algorithm is designed to find all the common HSPs between a query and hit sequence. Subsequently, each HSP alignment is extended in both directions. Following

each extension the quality of the new alignment is recalculated using the same scoring matrix and a score is assigned. Each alignment extension results in either an increased or decreased score, depending on the quality of the newly added matches. When the alignment score drops below a second predefined threshold, "S", the extension of the alignment stops (ALTSCHUL *et al.* 1997), ensuring only regions of high quality alignment are included in the final BLAST results.

The final results reported for each search are ranked based on both the calculated bit-score and a reported expect value (E-value), defined as the number of alignment matches of similar quality that one would expect to find by chance. To calculate the E-value both the length of the query sequence and the size of the database being searched are considered, with a lower E-value indicating a higher confidence that the alignment between the query and hit sequence is a result of homology rather than chance.

All BLAST programs use the method described above, though there are slight variations in the parameters employed by each one. For example BLASTP, the program used to search a protein database with a protein query sequence uses a default word size of 3 amino acids. By contrast, the BLASTN program used to search DNA sequences against nucleotide databases has a default word size of 11 bases. This parameter can be altered for more tailored sequence specific searches. In addition, parameters such as the gap-opening and gap-extension penalties can be modified to tailor for problems associated with finding distantly related sequences. The price of introducing gaps in the alignment, and cost of every extension past the initial opening gap is an important property of any homology search using BLAST. A high gap costs will result in alignments with fewer gaps and is less likely to detect homologs which have diverged substantially as a result of insertion and deletions in their sequences.

1.6.2 Hidden Markov Models

As described above homology search tools such as BLAST, use scoring matrices to weight matches between non-identical residues in a comparison between two sequences. Such matrices, including the widely used Dayhoff mutation (DM) and BLOSUM tables, are based on the observed sequence diversity amongst related globular proteins (BARKER *et al.* 1978; HENIKOFF and HENIKOFF 1992). These matrices are essentially used to consider the likelihood of an observed substitution between two residues in query and subject sequences. Whilst these matrices allow for information relating to the relative mutability rate of one residue for another across a wide range of proteins to be exploited in similarity searching they do not take into account the relative importance of each residue in a protein as defined by its degree conservation amongst closely related homologs. A Hidden Markov Model (HMM) is a computational structure, which can exploit such position specific information derived from the alignment of multiple homologous protein sequences. Certain amino acid residues in a protein are more critical for correct structure and function, and are as a consequence more highly conserved than other less critical positions. The sensitivity of HMMs for detection of distant relatives based on the alignment of a set of known gene family members comes from the generation of a "fingerprint" which is unique to that set of sequences that can be used to query a database of unknown proteins (MADERA and GOUGH 2002). HMMs are widely used to model linear problems in a wide range of fields including speech and handwriting recognition. The use of HMMs in comparative sequence analysis involves 2 principle steps – (1) Training the model and (2) using the model to detect distant homologies.

Training the model involves capturing the patterns inherent in an alignment of sequences and to accomplish this, a probability distribution for each of the 20 amino acids at every site in an alignment is determined. Generation of a HMM requires that each position in an alignment be assigned to one of a match, delete or insert states. Match states emit a residue based on the probabilities of amino acids in the underlying alignment. A second probability distribution

governs the choice of successor state. Gaps in alignments caused by insertions and deletions are accounted for by assigning each state in a HMM both an "insert" and "delete" state as well as the match state. If an insert states appears it allows insertions relative to the match states in the consensus sequence while a delete state skips a column in the alignment and essentially corresponds to opening a gap in the final state sequence that defines the model created to distinguish the underlying multiple sequence alignment. At the end of this process we have a 'hidden' state sequence that we do not observe and a symbol sequence that we do observe, thus the name Hidden Markov Model. The calculated amino acid probability distribution and state transition probabilities are unique properties of the input sequences and allow for the same mathematical framework to be employed to search for different protein families. Given a trained HMM model any sequence can be searched and the probability that the HMM would generate the test sequence can be calculated. A high probability indicates that the test sequence most likely belongs to the same family as those used to create the HMM.

1.6.3 Positive Selection Analysis

The production of functional proteins is critically dependent on the underlying DNA, as these nucleotide sequences act a template for a cells protein production machinery to translate into amino acids (AA). As more DNA sequence data has accumulated it has allowed for the molecular evolution of these proteins to be studied both within and between species, by comparing and contrasting the changes that have occurred in the encoding DNA. Although such analysis can be carried out on an individual nucleotide basis, the most common methodology employed is to treat the codon as the unit of evolution within protein-coding genes.

In codon evolution synonymous mutations result in an different codon but do not alter the associated amino acid and therefore do not affect the protein sequence. Such mutations are regarded as neutral and their rate of fixation within a sequence is solely dependent on the rate by which they are produced

by mutation (ANISIMOVA and LIBERLES 2007). In contrast, nonsynonymous mutations result in a change in the amino acid coded for by a codon and result in the generation of an altered protein sequence for that gene. Multiple sequence alignment of related sequences allow us to calculate the substitution rates for a gene; d_S, the number of synonymous substitutions per synonymous site and d_N, the number of nonsynonymous substitutions per nonsynonymous site (MUSE and GAUT 1994; YANG *et al.* 1994). The selection pressures influencing the evolution of a protein coding gene can be examined through calculation of a synonymous/nonsynonymous rate ratio, $\omega=(d_N / d_S)$.

If nonsynonymous sites are evolving in a neutral manner, then substitutions at such sites should have a rates of fixation equal to the synonymous substitutions and result in an ω ratio of 1. Such nonsynonymous changes are predicted to have no effect on the fitness of a protein. In most genes, however, nonsynonymous mutations are deleterious and are fixed at a lower rate than synonymous changes and the ω ratio will be < 1. In contrast, ω values of > 1 are indicative of an excess of nonsynonymous substitutions when compared to synonymous changes and considered as evidence for positive selection. By this criteria such amino acid changing substitutions are considered beneficial and are favoured by natural selection (ANISIMOVA and LIBERLES 2007).

Early studies involving estimation of these rates were carried out in an ad hoc pairwise manner between sequences. Such methods have proven unsatisfactory as the ω ratio was estimated as an average over the entire sequence as well as over the entire evolutionary time separating the sequences. By ignoring the biological consideration that many codons within a sequence are under purifying selection as well as ignoring the phylogeny linking the sequences being analysed, ω ratios calculated in this manner may be biased. To over come these limitations, estimates of these rates are often derived in a maximum likelihood framework (YANG 1997). Such methods use advanced probabilistic codon substitution models to estimate a number of parameters from alignments of related sequences. As well as estimation of ω ratios such methods can be

used to investigate other parameters influencing coding sequence evolution including variation in synonymous substitution rates, codon usage variation and more recently selection for alteration of physicochemical properties within a protein (YANG 1997).

1.7 Thesis objectives

Understanding the molecular mechanisms underlying the immune systems of other species has three direct implications for humans. Firstly, such comparative knowledge allows us to chart directly the evolution of the immune system especially with regard to the gain and loss of components and can provide us with valuable information as to the form and function of our own immune mechanisms. Secondly, many of these other species including the domesticated species, cow and chicken represent economically important agricultural animals, intensive faming of which underpins much of the multi-billion dollar food industry worldwide. Parasitic and infectious diseases cost the beef and chicken industry billions of dollars annually through animal deaths and rearing inefficiencies. A clearer understanding of the immune components and mechanisms of each of these species would allow targeted, disease-specific treatments for economically important disease to be developed which could replace the widespread, non-disease-specific use of antibiotics which can lead to the emergence of tough, drug-resistant bacteria. These bacteria can survive and spread easily, due to the intensive farming conditions utilised in modern large scale animal production and can eventually be ingested by the consumer and thus can directly impact on human health. Thirdly, these other species can act as reservoirs for zoonotic diseases which can be transferred to humans. Classical examples such as TB in cattle have been supplemented in recent years by well documented viral infections which have crossed the species barrier. Two recent examples include the outbreaks of avian influenza in Asia between 1997 and the present, notably in Europe in 2007, as well as the current worldwide H1N1 "swine-flu" epidemic. For these reasons, there is now greater interest comparative studies of the immune systems in other species beyond human.

The specific aims of this thesis are to:

- Systematically analyse the avian TLR gene repertoire and compare and contrast it with that of the established mammalian gene set.
- Examine the avian TLR pathway with regards to gene gain and gene loss of its constitutive components as well as investigation of the mechanisms underlying the evolution of pathway components in avian species.
- Characterise the bovine β-defensin AMP genes with respect to sequence, gene structure and genomic organisation and investigate species-specific modifications of this gene family in the bovine lineage.
- Examine the CD3 gene family for evidence of the selective pressures which have influenced the evolution of these important molecules in mammals.

2. Identification and characterization of TLR15 - a novel Vertebrate Toll-Like Receptor

Abstract

Toll-like receptors (TLRs) are a group of highly conserved molecules that initiate the innate immune response to pathogens by recognizing structural motifs expressed by microbes. Throughout the vertebrate lineage, these proteins have evolved diverse ligand specificities, providing a broad-spectrum recognition mechanism to the extensive array of pathogens encountered. Within vertebrates, variation in the numbers and subtypes of TLRs expressed by each species have been observed with the repertoire of TLRs in a species presumably honed by natural selection to provide an effective response to infection. We have identified a novel TLR, TLR15, by bioinformatic analysis of the chicken and zebra finch genomes, which is distinct from any known vertebrate TLR and appears to be specific to diapsids. The gene for TLR15 gene codes for an archetypal vertebrate gene containing highly conserved TIR and transmembrane domains as well as a distinctive arrangement of extracellular leucine-rich regions.

2.1 Introduction

The initiation of any immune response is reliant upon the efficient initial detection of invading pathogens by the class of molecules collectively termed pathogen recognition receptors (PRRs). Detection of conserved microbial structures known as Pathogen Associated Molecular Patterns (PAMPs) results in the activation of intracellular signalling cascades and ultimately induces the expression of a range of immune response molecules under the control of transcription factors such as NF-κB and IRF3 (WERLING and JUNGI 2003). The best characterized and most highly conserved family of PRRs is the Toll-Like Receptors (TLRs). TLRs were initially identified in vertebrates through similarity with the transmembrane Toll protein in *Drosophila* (MEDZHITOV *et al.* 1997), which regulates early embryonic development as well as mediating innate immune responses to fungal infection (HASHIMOTO *et al.* 1988; LEMAITRE *et al.* 1996). Since then, members of this family of molecules have been identified in species encompassing a timespan of at least 600 million years ranging from the cnidarian sea anemones to mammals (ROACH *et al.* 2005).

All TLR molecules share certain characteristics which distinguish them from other PRRs and surface expressed molecules. They are type I transmembrane receptors consisting of amino-terminal Leucine Rich Repeat (LRR) domain, a transmembrane domain and an approximately 160 amino acid long carboxy-terminal cytoplasmic region called Toll/Interleukin-1 receptor (TIR) domain (BELL *et al.* 2003). The TIR domain is a critical functional unit that is conserved between TLRs, Interleukin-1 (IL-1) receptors and the intracellular adaptors that interact with these two receptor subtypes (FITZGERALD and O'NEILL 2000). Crystal structure analysis of a number of TIR domains indicates a conserved 3-dimensional fold consisting of 5 parallel Beta-sheets, 5 alpha-helical regions and 5 connecting loops which have been implicated in the formation of interactions with other TIR domain containing proteins (TAUSZIG-DELAMASURE *et al.* 2002; XU *et al.* 2000). Whilst the TIR domain tends to be relatively conserved amongst orthologus genes, weaker overall sequence

conservation is observed in comparisons between paralogous TLR proteins, reflecting the differing intracellular signalling mechanisms exploited by the various vertebrate TLRs. In general, overall sequence conservation is confined to three regions within the domain, denoted Box 1, 2 and 3 (FITZGERALD and O'NEILL 2000; O'NEILL 2006). Both Box 1 and 2 have been directly implicated in signal transduction. In particular, mutation of a highly conserved proline residue in box 2, which forms the BB-loop responsible for TLR2 dimerisation with TLR1, has been shown to disrupt signal transduction and render mice resistant to the effects of LPS (POLTORAK *et al.* 1998; XU *et al.* 2000). Conservation in box 3 has not been shown to be vital for signalling transduction as mutational studies have indicated that it does not contribute to known protein-protein interactions (SLACK *et al.* 2000).

The extracellular portion of TLRs is characterised by the presence of 19-25 tandem copies of leucine-rich repeat (LRR) motifs. Each of these LRRs shares a typical consensus sequence consisting of the highly conserved 11-residue pattern - LxxLxLxxNxL – where x can be any amino acid and the leucines can be replaced by any other hydrophobic amino acids (Bella et al. 2008). The asparagine (N) residue is often replaced by a cysteine or threonine residue as all three are capable of forming the hydrogen bonds which make up the "asparagine ladder" responsible for structural stabilization of the protein (MATSUSHIMA *et al.* 2007). This common sequence pattern allows the extracellular domain of TLR proteins to adopt a horseshoe shaped solenoid structure (BELL *et al.* 2005; JIN *et al.* 2007). The inner (concave) surface of the horseshoe consists of beta-strands coded for by the 11 amino acids of the LRR described above. The outer (convex) surface is composed of the subsequent 13 amino acids which have a general consensus of – xxLxxxxFxxLxx (JIN *et al.* 2007) .

Prior to the sequencing of the chicken genome (2004), seven TLRs had been identified in this species. These included two chicken TLR1-like (type 1 and type 2) TLR2 (type 1 and type 2), TLR3 TLR4, TLR5 and TLR7 (BOYD *et al.*

2001; FUKUI *et al.* 2001; IQBAL *et al.* 2005; LEVEQUE *et al.* 2003; LYNN *et al.* 2003; PHILBIN *et al.* 2005; YILMAZ *et al.* 2005). Previous analysis of sequenced fish and amphibian genomes has also shown counterparts for most known mammalian TLRs or TLR subfamilies were present in these species whose divergences significantly predate that of birds and mammals. In addition two novel TLRs, TLR21 and TLR22 were identified in all teleost species examined while the genome of the amphibian *Xenopus tropicalis* coded for TLR14, a receptor later predicted to be a highly diverged member of the TLR2 subfamily (Roach et al. 2005). The sequencing of the chicken and the zebra finch genomes, representing the first available complete avian sequences, provided an opportunity to completely characterize the TLR repertoire possessed by these species and to determine if the avian lineage carried any species- or taxon-specific expansions of the TLR gene family.

2.2 Materials and Methods

Publicly available protein sequences corresponding to the 10 known human TLRs were retrieved from Uniprot (http://www.uniprot.org/) (See Appendix). In addition, the sequences for all known non-primate mammal, fish and chicken TLRs were downloaded. Masked assemblies for each chicken chromosome were downloaded from the University of California Santa Cruz Genome Browser (KENT *et al.* 2002) (http://genome.ucsc.edu). Each chromosome assembly was translated in all six reading frames using a purpose written Perl program. This procedure was repeated for the zebra finch genome following the release of the first sequence draft in October 2008 (http://www.songbirdgenome.org). Each genome was searched using the assembled vertebrate TLR dataset with the TBLASTN program (ALTSCHUL *et al.* 1997). In addition the initial ensembl predicted gene dataset consisting of 28,416 proteins was searched using the BLASTP program for sequences displaying a high degree of similarity to TLR family members. The TIR domain from all known vertebrate TLR sequences were extracted and a multiple sequence alignment of these regions was generated using the T-Coffee program (NOTREDAME *et al.* 2000). This alignment was used to construct a

Hidden Markov Model (HMM) using the HMMER 2.1.1 suite of programs (http://hmmer.wustl.edu/) (EDDY 1998). The HMM profile generated using the TIR domain multiple sequences alignment was then used to search the six frame translations of each chromosome assembly in order to identify any TIR domain containing regions in the genomes of these avian species which may have been sufficiently diverged to have been missed by the initial BLAST searches.

Genomic DNA corresponding to putative TLRs was retrieved using BLAT and used for prediction of intron-exon boundaries using the GenScan program (http://genes.mit.edu/GENSCAN.html) (BURGE and KARLIN 1997). Initial characterization of the functional domain structures of the novel TLRs were predicted using the SMART program (SCHULTZ *et al.* 2000) as well as in-house annotation using motif matching profiles. The proposed avian homologs of known mammalian TLRs were further analyzed by aligning them with homologous sequences from other vertebrate species using the T-COFFEE multiple sequence alignment program (NOTREDAME *et al.* 2000). Phylogenetic analysis of the proteins was carried out using Mega v.4 with a Poisson corrected model (KUMAR *et al.* 2001). 1000 bootstrap replicates were carried out to test the topological stability of each node in the tree.

2.3 Results and Discussion

A combination of BLAST and HMM searches identified two putative TLR like sequences in the chicken genome which were distinct from any of the previously characterised chicken TLR genes. The first of these was located on the unassigned chromosome (ChrUN) in the initial draft of the chicken release and showed a high degree of similarity to the TLR21 gene identified in multiple fish and amphibian species. This TLR was represented by only a partial sequence in the early chicken genome release as it coding sequence transcended one end of a short assembled contig. It was decided to not pursue this chicken TLR gene any further. The full sequence of this gene was ultimately determined through isolation of cDNAs from bursal lymphocytes (CALDWELL *et al.* 2005) and shown to be a direct 1:1 ortholog of the TLR21

genes in fish and amphibians, with the avian gene representing the only amniote species in which it has so far been identified. As well as confirming the presence of all the known chicken TLRs, initial homology searches identified a further putative TLR coding sequence which displayed poor sequence similarity to any other known vertebrate sequence and was named sequentially as TLR15 as its extracellular domain structure appeared distinctly different from all vertebrates TLRs 1-14.

TLR15 was bioinformatically mapped to chromosome 3 (chr3:2925041-2927644) in the February 2004 assembly of the chicken genome. In chicken, this novel TLR is flanked downstream by the genes GPR75 and PSME4 and upstream the Ensembl predicted gene XTP3-B and CHAC2. The TLR15 locus and surrounding chromosomal portion display a high degree of synteny with conserved areas on the human 2 and mouse 11 chromosomes (Figure 2.1).The syntenic regions in human, mouse were analysed for the presence of genes corresponding to the chicken TLR15. No orthologous genes or recognisable pseudogene were identified using either the Genscan gene prediction program (BURGE and KARLIN 1997) or a Hidden Markov Model (EDDY 1998) search strategy based on the TIR domain (Figure 2.1). Subsequently the genome of a second avian species, the songbird zebra finch was sequenced. Using the chicken TLR15 sequence as a "seed" BLAST query (ALTSCHUL *et al.* 1997) a zebra finch ortholog was located in the expected syntenic location indicating that the TLR15 gene is at least as old as the split that separates the galliform and passerine lineages (approximately 90 million years). This split is a basal split in the avian phylogeny and its presence in these two species would suggest that it is likely to be a feature of all extant bird TLR repertoires. As this thesis was being prepared, the genome sequence of the Green anole lizard, *Anolis carolinensis* was made publicly available representing the first reptilian genome to be sequenced. Reptiles, birds and mammals together form the amniotes, a group of tetrapod vertebrates which possess a distinctive embryonic developmental process when compared to other vertebrates. Within this subgroup, birds and reptiles (diapsids) share a more recent common

ancestor with each other than with mammals (synapsids). A search of this reptilian genome identified a partial sequence corresponding to the TIR, transmembrane and C-terminal LRR domains of a TLR with identical syntenic positioning and orientation to that of the chicken and zebra finch TLR15s (Figure 2.1). A gap in the genome sequence prevents identification of the full length sequence for this lizard TLR15 homolog but the conserved synteny and high degree of sequence conservation would confidently indicate that the origin of the TLR15 gene predates the divergence of birds and reptiles approximately 280 million years ago and is most likely a conserved feature of all diapsid species.

The full coding sequence (cds) for the avian TLR15 gene was predicted from the corresponding genomic sequences. The TLR15 cds in both chicken and zebra finch is maintained in a single exon consisting of 2604 and 2622 base pairs respectively, translating to proteins of 868 amino acids in chicken and 874 in zebra finch. Confirmation of the chicken sequence was carried out by Rowan Higgs (Education and Research Centre, SVUH) who PCR amplified gene-specific cDNAs and sequenced the entire coding region using a combination of specific forward and reverse primers. The chicken TLR15 coding sequence was submitted to GenBank (accession number DQ267901). The sequence was in agreement with the bioinformatic prediction, except for synonymous changes at base positions 168 (A to G) and 450 (C to T) and nonsynonymous changes at base positions 926 (A to C) and 1363 (T to G). The nonsynonymous changes result in amino acid changes from glutamic acid to alanine and from leucine to valine, respectively. Both of these amino acids are located in the variable extracellular region of the protein.

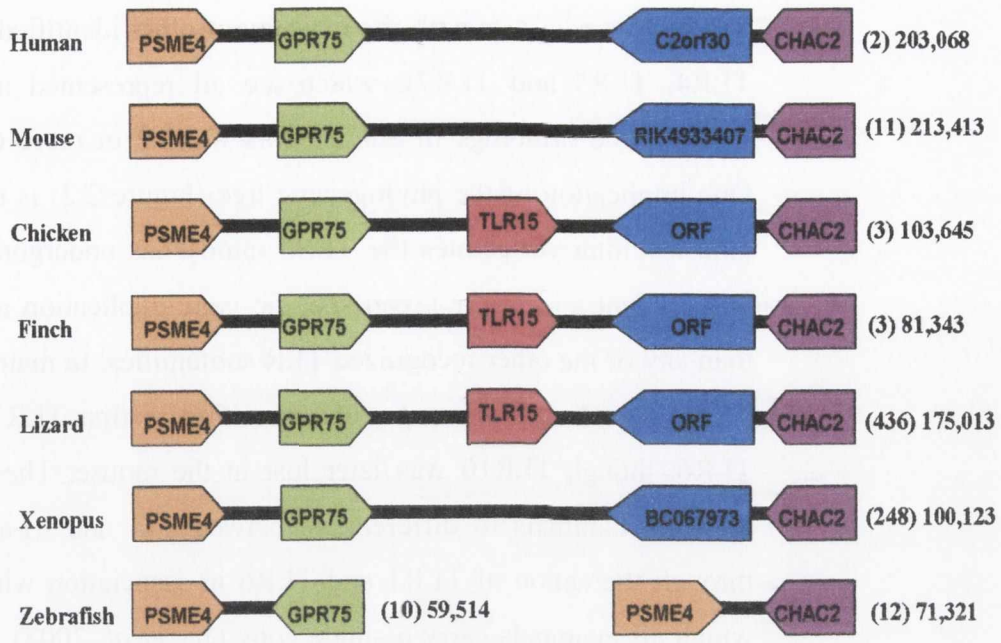


Figure 2-1. Comparative genomic synteny in human, mouse, chicken, zebra finch, anole lizard, xenopus and zebrafish, of genes flanking the TLR15 locus in chicken, finch and lizard species. Figure is not drawn to scale. Chromosome or scaffold numbers are indicated in brackets as well as syntenic span lengths in base pairs. Direction of transcription of each gene is indicated.

Phylogenetic analysis of the TLR family is hindered by the poor sequence conservation observed in the much of the protein sequence. In particular, the extracellular region coding for the LRR domains is virtually unalignable when a wide sample of TLR family members is examined. To overcome this problem phylogenetic reconstruction of the TLR family is more reliably determined using the highly conserved TIR domains. A neighbor-joining tree constructed using all the avian TLRs identified in our genome searches as well as the entire repertoire of genes from mouse, human, xenopus and fugu and the single anole lizard TLR sequence revealed that the TLR15 genes cluster with high bootstrap support as an outgroup to the TLR2 family consisting of TLRs 1,2,6 and 10 (Figure 2.2) but displays poor overall sequence similarity when compared to all the other TLR2 family clade members with the best being 30.1% between chicken TLR2 and TLR15. The low sequence identity, and the relationship pattern observed in the phylogenetic tree, are not sufficient to assign a clear 1:1 orthology, thus TLR15 represents a novel receptor distinct from all known

TLRs. This contrasts with the majority of other identified avian TLRs (TLR3, TLR4, TLR5 and TLR7), which are all represented as highly conserved, recognisable orthologs in amniotes as well as in more diverged vertebrates. One implication of the phylogenetic tree (Figure 2.2) is that, amongst strictly land dwelling vertebrates the TLR2 family has undergone significantly more independent species or taxon specific gene duplication and gene loss events than any of the other recognized TLR subfamilies. In mammals, the archetypal TLR1 gene has duplicated twice, resulting in first TLR10 and subsequently TLR6, though TLR10 was later lost in the mouse. These duplications have allowed mammals to differentiate between di- and tri-acylated lipopeptides through the action of TLR1 and TLR6 in association with TLR2, a gene for which all mammals carry a single copy (JIN *et al.* 2007). In avian species the dynamic and labile nature of the TLR2 family is again evident as both TLR1 and TLR2 have been duplicated, independent of the mammalian events and together with the distantly related TLR15 gene may be maintained to counter the specific pathological challenges faced by birds which differ from those faced by mammals. TLR15 itself, may have evolved to counter specific pathogens for which diapsids are the preferential host. The TLR14 gene, found in amphibian and fish vertebrate species is also a distantly related member of the TLR2 family (ISHII *et al.* 2007) (Figure 2.2) but appears to represent an expansion of the TLR2 family in these lineages which is independent of that in reptiles and birds. TLR14 is maintained in a syntenic position in both amphibian and fish species which differs from that observed for TLR15 in reptiles and birds. Furthermore, the predicted peptide for chicken TLR15 is coded for by a single exon, a feature common to all the amniote members of the TLR1/TLR2/TLR6/TLR10 clade whereas the TLR14 gene is coded for by 5 exons in xenopus and 3 exons in zebrafish and fugu (ISHII *et al.* 2007). This evidence supports the proposal of TLR15 as a novel diapsid-specific TLR.

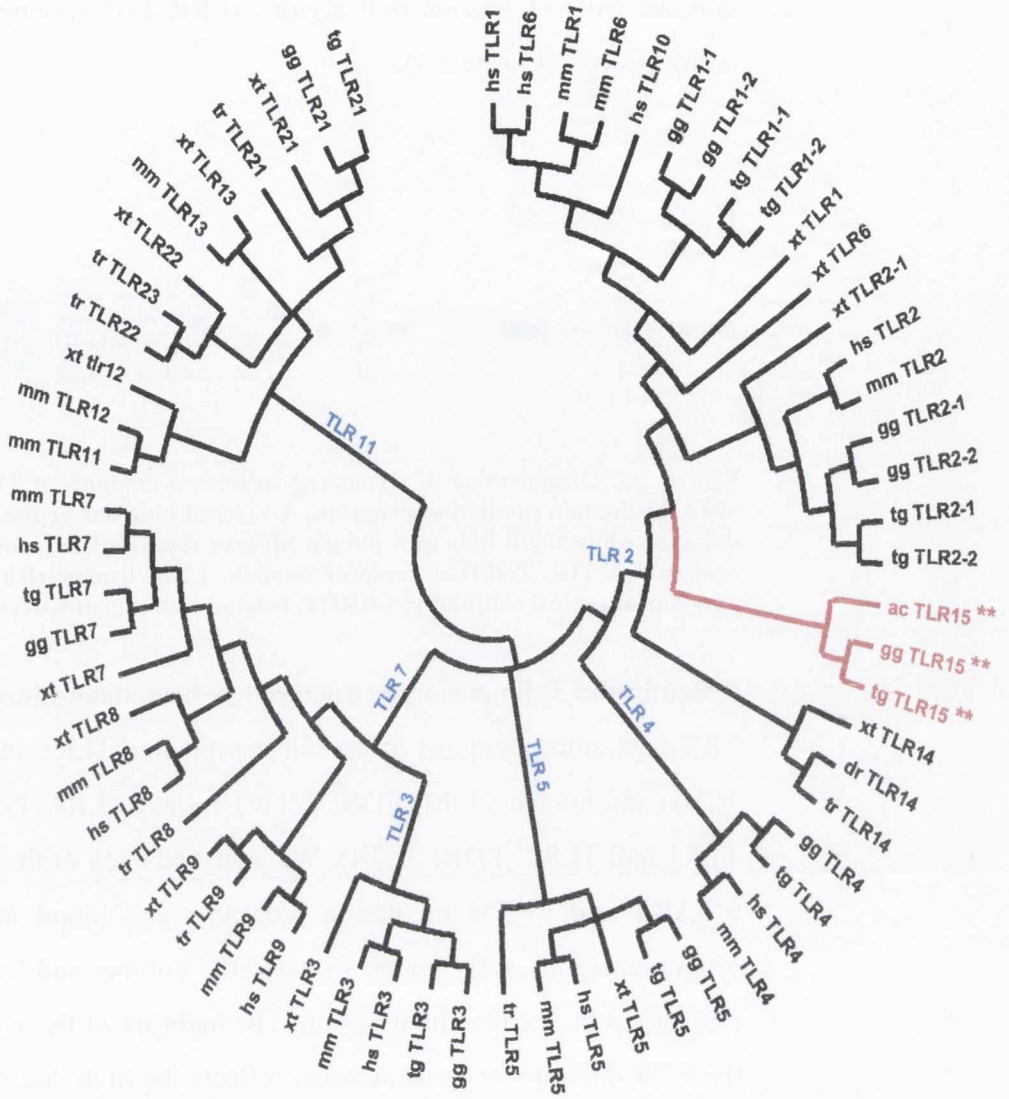


Figure 2-2. Neighbor-joining tree generated from the TIR domain amino acid sequences of human (hs), mouse (mm), chicken (gg), zebra finch (tg), anole lizard (ac), xenopus tropicalis (xt) and fugu (tr) TLR homologs, reconstructed using Mega v.4.0. Branches with less than 50% bootstrap support have been collapsed. Bootstrap analysis is based on multiple resampling of the original data and is the most common method of estimating the degree of confidence in the topology of phylogenetic trees. Branches leading to the TLR15 genes are highlighted in red and the branches leading to each of the accepted TLR subfamilies are labeled in blue.

Further evidence distinguishing TLR15 from other vertebrate TLRs can be derived from examination of the domain structure of the encoded proteins. Analysis of the domain structure of TLR15 using the SMART prediction tool revealed an archetypal TLR structure comprising a cytoplasmic Toll/IL-1 receptor (TIR) region, a short transmembrane domain, and an extracellular

domain with 11 leucine-rich regions (LRRs) of varying length and sequence compositions (Figure 2.3).

Figure 2-3. Organization of secondary structural domains of TLR15 as predicted using SMART domain prediction program. A vertical blue bar represents the transmembrane domain, while small blue and purple squares represent sequences of low compositional complexity. TIR, Toll/IL-1 receptor domain; LRR, leucine-rich repeat; LRRT, leucine-rich repeat typical subfamily; LRRCT, leucine-rich repeat C-terminal domain.

Recently the 3-dimensional structure has been determined for the extracellular LRR containing domains of several mammalian TLRs, including mouse (PDB: 3CIG) and human TLR3 (PDB: 1ZIW), mouse TLR4 (PDB: 2Z64) and human TLR1 and TLR2 (PDB: 2Z7X). We analysed each of these proteins using both SMART and PFAM prediction programs and found in all cases that these programs significantly underestimated the number and length of LRR domains making up the extracellular region. The inability of these programs to detect all the LRR domains in these proteins reflects the high degree of variability in the sequences that make up the repeating unit of individual LRRs. LRR can be divided up into a Highly Conserved Segment (HCS) based on the first 11 amino acids of the domain and a Variable Segment (VS) composed of a residue stretch of varying length. All LRR domain containing proteins for which a 3-dimensional structure is available have been shown to adopt a solenoid shape, where each LRR represents one turn of the solenoid (BELLA *et al.* 2008). The HCS of each LRR repeat forms the concave (inner) surface of the arc and invariably consists of a short β-strand which forms the continuous β-sheet, characteristic of this surface of these molecules (BELLA *et al.* 2008). The VS of each LRR forms the convex (outer) surface of the molecule and can adopt a variety of secondary structures, depending on sequence composition and length (lrrnb for references). Several secondary structure programs exist including

PORTER (POLLASTRI and MCLYSAGHT 2005) and PSIPRED (JONES 1999) which can predict the presence of β-strand structures within protein sequences. Using both of these programs we were able to accurately predict the correct number of β-sheets corresponding to the HCS of the LRRs in each of the TLR three-dimensional structures listed above (Data not shown). A similar strategy when applied to the novel TLR15 sequences predicts a single Leucine Rich Repeat N-Terminal (LRRNT) domain, 20 individual internal LRRs and one Leucine Rich Repeat C-Terminal (LRRCT) domain in the extracellular region (Table 2.1 and Figure 2.3). In general, the LRRs identified in TLR15 are of two particular subtypes. 14 of the predicted domains – (LRRs 1, 2, 4, 5, 6, 7, 8, 11, 13, 14, 15, 17, 18, 19 and 20) (see Table 2) conform to the typical TLR-LRR consensus sequence (LxxLxLxxNxLxxLxxxxFxxLxx – where x may be any residue, L may be substituted any other hydrophobic amino acid residue and N can be any of N, C, T or S). Several of these LRRs deviate slightly from this TLR-LRR consensus but such deviation is a common feature of these domains in vertebrate TLRs where the repeat can vary in both sequence composition and length (BELL *et al.* 2003). LRRs 9, 10, 12 and 16 more closely resemble proline rich bacterial LRRs (typified by the consensus LxxLxLxxNxLxxLPx(x)LPxx). These bacterial LRRs are a common feature of all members of the TLR2 subfamily of receptors (BELL *et al.* 2003). LRR3 in TLR15 codes for an unusually long motif (99 amino acids in chicken, 106 amino acids in zebra finch). Uniquely among the previously described vertebrate TLRs, all the members of the TLR7 family (TLRs 7, 8, and 9) also contain a single extended LRR motif. In contrast to TLR15, this unusual domain in TLR7 family members is located more centrally (LRR 8 out of 27) (BELL *et al.* 2003)and varies in length from 58 to 73 amino acids. In all cases including TLR15, the downstream LRR (LRR4 in TLR15) is atypical and has an N residue substituted for the first L in the HCS. It has been proposed that these long extended LRR motifs may form an undetermined structure and may act as a cap to the LRRs and allow the members of the TLR7 family to adopt a conformation consisting of two rather than one horseshoe structure (GIBBARD *et al.* 2006). Whilst 3-dimensional modeling of the TLR7 family will resolve

this issue, such a proposition appears unlikely as one of the proposed horseshoe structures would be devoid of a cap (usually provided by the LRRNT and LRRCT domains) and to date no LRR containing protein has been identified that has an "uncapped" LRR domain. In addition the extended LRR in TLR15 is located at the extreme N-terminal end of the proposed peptide and the remaining N-terminal LRRs (1 and 2) would not be sufficient to form any meaningful horseshoe shaped structure.

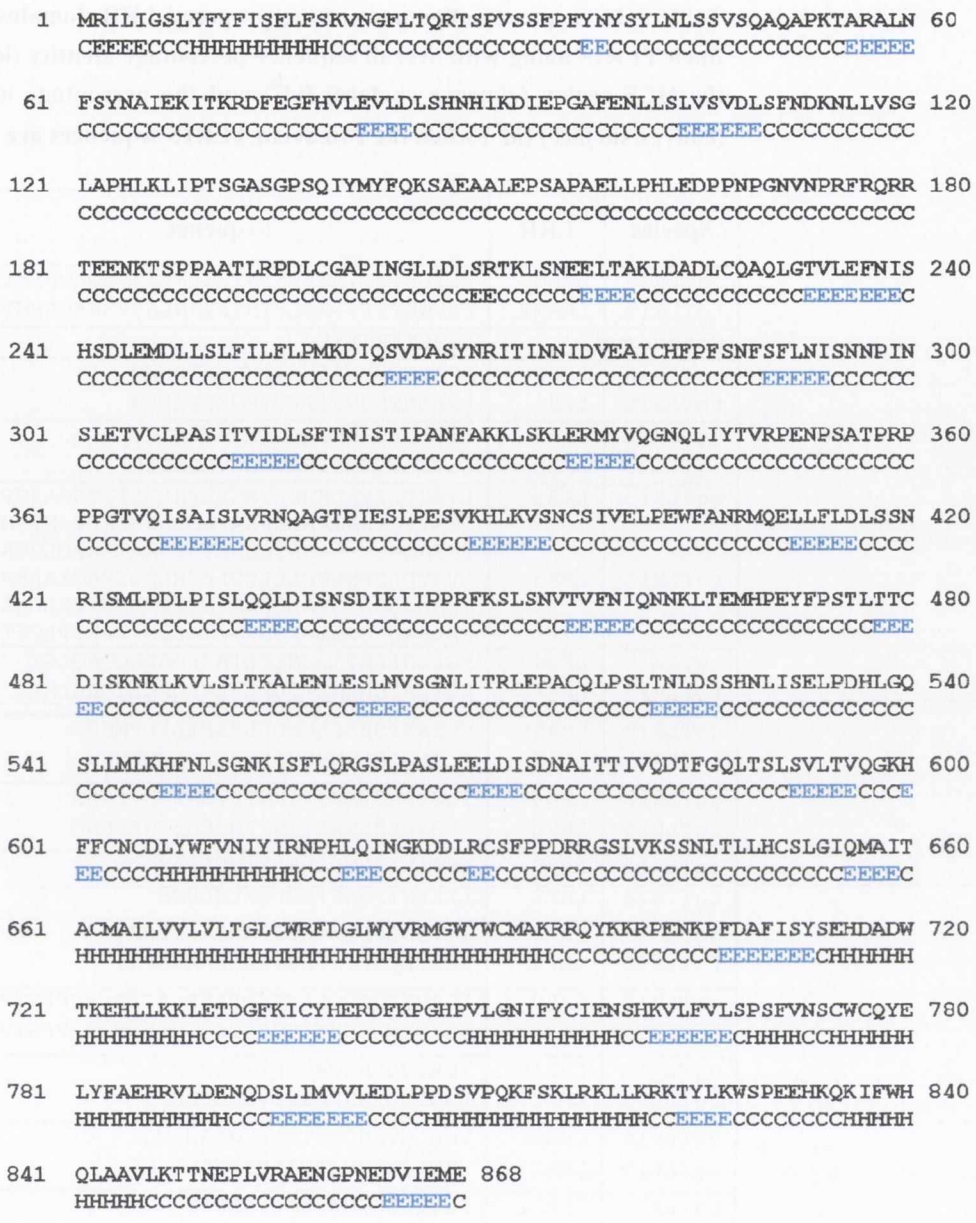


Figure 2-4. The consensus secondary structure prediction for chicken TLR15 by PORTER and PSIpred. Abbreviations used – H, Helix; C, coil; E, β-strand.

Table 2-1. Sequence alignment of the proposed LRR domains from chicken and zebra finch TLR15 along with overall sequence percentage identity (id[a]), percentage identity in the HCS region (concave surface) (id[b]) and the percentage identity in the VS regions (convex surface) (id[c]) when the two avian TLR15 sequences are compared.

Species	LRR	Sequence	% id[a]	% id[b]	% id[c]
ggTLR15	LRRNT	FLTQRTSPVSSFPFYNYSYLNLSSVSQAQAPKT	67		
tgTLR15	LRRNT	FLMWRTPP--AFLVYNYSYSNLSSVSEAQAPKT			
Consensus		LXXLXLXXNXL			
ggTLR15	LRR1	ARALNFSYNAIEKITKRDFEGFHV	74	82	54
tgTLR15	LRR1	ARALNFSHNVIEKVTKGELEGFDW			
ggTLR15	LRR2	LEVLDLSHNHIKDIEPGAFENLLS	63	72	54
tgTLR15	LRR2	LEVLDFSYNQIRAVEPGVFQSLLS			
ggTLR15	LRR3	LVSVDLSFNDKNLLVSGLAPHLKLIPTSGASGPSQ IYMYFQKSAEAALEPSAPAELLPHLEDPPNPGNVN PRFRQ------RRTEENKTSPPAATLRPDLCGAPI	47	91	41
tgTLR15	LRR3	LVSVDLSFNDEKLLLSDLPSHLRLSPAGKAPRSLQ ISRNSGRSSGVALQPSAPAEEPSHSGVLSLLQILS PRLRRSTGNLLRRGERNTTALPMVTPEPTLCGTPI			
ggTLR15	LRR4	NGLLDLSRTKLSNEELTAKLDADLCQAQLGT	52	55	50
tgTLR15	LRR4	NGTLNLSHSNLTQDELVLKLDEDLCQAHLRR			
ggTLR15	LRR5	VLEFNISHSDLEMDLLSLFILFLPMKD	63	45	75
tgTLR15	LRR5	ILELDISHNNVEMDLLSLFSLFFPMEN			
ggTLR15	LRR6	IQSVDASYNRITINNIDVEAICHFPFSN	43	45	44
tgTLR15	LRR6	TLSIDASSNKLTINILNPESFCNFPSHQ			
ggTLR15	LRR7	FSFLNISNNPINSLETVCLPAS	73	82	64
tgTLR15	LRR7	LLFLNISNNPINSLDRLCLPSS			
ggTLR15	LRR8	ITVIDLSFTNISTIPANFAKKLSK	67	82	54
tgTLR15	LRR8	IKEIDLSFTNISQIPLDFAKKLFN			
ggTLR15	LRR9	LERMYVQGNQLIYTVRPENPSATPR--PPPGTVQ	56	73	48
tgTLR15	LRR9	LEKMYVQGNHFIYTAFSESGNTLPTCVPPPGTVH			
ggTLR15	LRR10	ISAISLVRNQAGTPIESLPES	76	73	80
tgTLR15	LRR10	LNALSLVRNKAGTPVESLPEK			
ggTLR15	LRR11	VKHLKVSNCSIVELPEWFANRMQE	75	82	69
tgTLR15	LRR11	VKHLGMSNCSIVELPEWFADTVEE			
ggTLR15	LRR12	LLFLDLSSNRISMLPDLPIS	70	90	44
tgTLR15	LRR12	LLFLDLSSNHISVFPNFPPS			
ggTLR15	LRR13	LQQLDISNSDIKIIPPRFKSLSN	65	73	58
tgTLR15	LRR13	LQHLDISSNDIKVISSSLKSLSN			
ggTLR15	LRR14	VTVFNIQNNKLTEMHPEYFPST	50	54	45
tgTLR15	LRR14	LKIFRIQNNKIMGIHTEFFPSA			
ggTLR15	LRR15	LTTCDISKNKLKVLSLTKALEN	73	64	82
tgTLR15	LRR15	LKKCDFSKNKVKVLSLTSALEK			
ggTLR15	LRR16	LESLNVSGNLITRLEPACQLPS	77	82	73
tgTLR15	LRR16	LEHLNISGNLITRLEPAGHLPA			
ggTLR15	LRR17	LTNLDSSHNLISELPDHLGQSLLM	75	100	54
tgTLR15	LRR17	LTNLDSSHNLIPELPDGFGVSLPG			
ggTLR15	LRR18	LKHFNLSGNKISFLQRGSLPAS	95	91	100
tgTLR15	LRR18	LKYFNLSGNKISFLQRGSLPAS			
ggTLR15	LRR19	LEELDISDNAITTIVQDTFGQLTS	75	91	69
tgTLR15	LRR19	LVELDISDNAITTIVEATFSPLTS			
ggTLR15	LRR20	LSVLTVQGKHFFCNCDLYWFVNIY	79	73	85
tgTLR15	LRR20	LRLLTVQGDHFFCTCDLYWFVNVY			

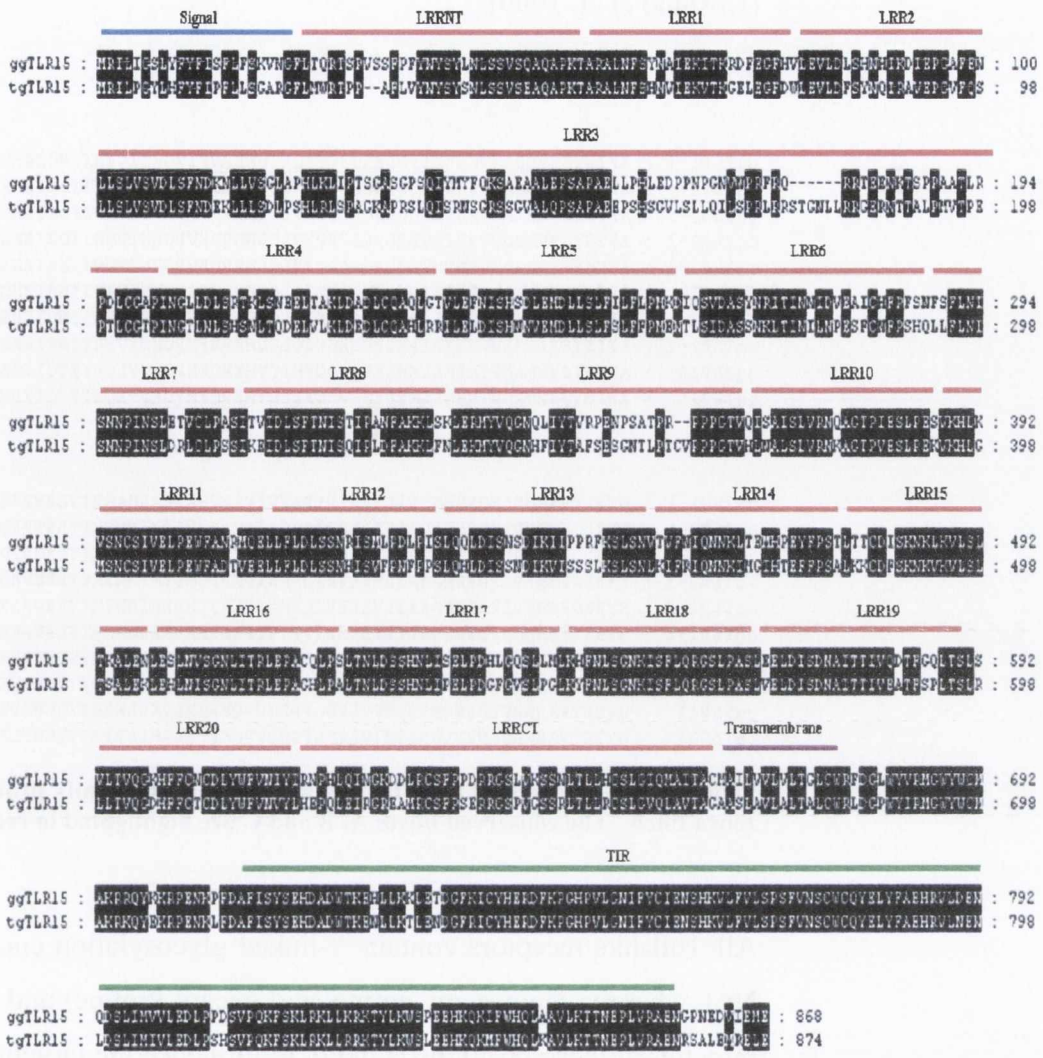


Figure 2-5. Multiple sequence alignment of chicken and zebra finch TLR15 with signal peptide indicated by a blue bar, LRRs indicated by a blue bar, transmembrane domain indicated by a purple bar and the TIR domain indicated by a green bar.

The TIR domain of the novel TLR15 is highly similar when compared with other members of the avian TLR2 family (Figure 2.6). In particular both Box-1 and Box-2, regions previously shown to be critical for signalling in TLR and associated adaptor proteins are highly conserved and would indicate that all members of this subfamily in birds use the same intracellular TIR domain containing adaptor protein in order to transduce the signal resulting from TLR-ligand interaction. In particular Box-2 of TLR15 carries the highly conserved Proline residue which is maintained in all vertebrate TLRs with the exception

of TLR3, and has been shown to be critical for MyD88 dependent signalling (Hoshino et al. 1999).

```
ggTLR1-1 : AFISYSERDSLWVKNELIPNLEKGEGCIQLCQHERNFIPGKSIVENIINCIEKSYKSIFVLSPNFVQSEWC :  71
ggTLR2-1 : AFVSYSENDSNWVENIMVQQLEQACPPFRLCLHKRDFVPGKWIVDNIIDSIEKSHKTLFVLSEHFVQSEWC :  71
ggTLR2-2 : AFVSYSENDSNWVENIMVQQLEQACPPFRLCLHKRDFVPGKWIVDNIIDSIEKSHKTLFVLSEHFVQSEWC :  71
tgTLR2-1 : AFVSYSENDSDWVENTMVRELEQACPPFRLCLHKRDFVPGKWIVDNIIDSIEKSRKTLFVLSEHFVQSEWC :  71
tgTLR2-2 : AFVSYSENDSDWVENTMVRELEQACPPFRLCLHKRDFVPGKWIVDNIIDSIEKSRKTLFVLSEHFVQSEWC :  71
ggTLR1-2 : AFISYSERDSLWVKNELIPNLEKGEGCIQLCQHERNFIPGKSIVENIINCIEKSYKSIFVLSPNFVQSEWC :  71
tgTLR1-1 : AFISYSERDSLWVKNELIPNLEKGEGCIQLCQHERNFVPGKSIVENIINCIEKSYKSIFVLSPNFVQSEWC :  71
tgTLR1-2 : AFISYSERDSLWVKNELIPNLEKGEGCIQLCQHERNFVPGKSIVENIINCIEKSYKSIFVLSPNFVQSEWC :  71
ggTLR15  : AFISYSEHDADWTKENLLKTLET--DGFKICYHERDFKPGHPVLGNIFYCIENSHKVLFVLSPSFVNSCWC :  69
tgTLR15  : AFISYSEHDADWTKENLLKTLEN--DGFRICYHERDFKPGHPVLGNIFYCIENSHKVLFVLSPSFVNSCWC :  69

ggTLR1-1 : HYELYFAHHRLFSENSNSLILILLEPIPSYVIPARYHKLKALMAKRTYLEWPKERSKHALFWANLRAAVNI : 142
ggTLR2-1 : KYELDFSHFRLFDENNDVAILLLEPIQSQAIPKRFCKLRKIMNTKTYLEWPPDEEQQQMFWENLKAALKS : 142
ggTLR2-2 : KYELDFSHFRLFDENNDVAILLLEPIQSQAIPKRFCKLRKIMNTKTYLEWPPDEEQQQMFWENLKAALKS : 142
tgTLR2-1 : KYELDFSHFRLFDENNDAAILVLLEPIQSKAIPKRFCKLRKIMNTKTYLEWPAGEEQQQVFWENLKGALKS : 142
tgTLR2-2 : KYELDFSHFRLFDENNDAAILVLLEPIQSKAIPKRFCKLRKIMNTKTYLEWPLEEEQQQMFWFNLKIALRS : 142
ggTLR1-2 : HYELYFAHHKLFSENSNSLILILLEPIPPYVIPARYHKLKALMAKRTYLEWPKERSKHALFWANLRAAISI : 142
tgTLR1-1 : HYELYFAHHKLFSENSNSLILILLEPIPPYLIPARYHKLKALMAKRTYMEWPKERSKRALFWANLRAAINI : 142
tgTLR1-2 : HYELYFAHHKLFSENSNSLILILLEPIPPYLIPARYHKLKALMAKRTYMEWPKERSKRALFWANLRAAINI : 142
ggTLR15  : QYELYFAEHRVLDENQDSLIMVVLEDLPPDSVPQKFSKLRKLLKRKTYLKWSPEEHKQKIFWHQLAAVLKT : 140
tgTLR15  : QYELYFAEHRVLNENLDSLIMIVLEDLPSHSVPQKFSKLRKLLRRKTYLKWSLEEHKQKMFWHQLKAVLKT : 140
```

Figure 2-6. Alignment of the TIR domain of all TLR2 family proteins from chicken and zebra finch. The conserved boxes A, B and C are highlighted in red..

All Toll-like receptors contain *N*-linked glycosylation consensus sites (NxS or NxT where x denotes any amino acid except Proline) and the addition of sugar moieties at these positions is likely to influence the targeting of the receptors to the cell surface as well as contributing to microbial pattern recognition (WEBER *et al.* 2004). The avian TLR15 sequences of chicken and zebra finch are predicted to have 14 and 12 potential glycosylation sites respectively. Nine of these potential sites are conserved across both species with the principal differences between the two bird species occurring in LRR 4 (3 potential sites in zebra finch but none chicken) and LRR5 (1 potential site in chicken only) (Figure 2.5). TLRs appear unique among LRR containing proteins in that the ligand-binding site is not located exclusively on the concave surface of the LRR solenoid (BELL *et al.* 2003; HUANG *et al.* 2008). The crystal structures of the human TLR1/TLR2 complex and the mouse TLR3 dimer complex bound to their respective ligands clearly show that residues located in the VS of several

LRRs are critical for receptor-ligand interactions and dimer formation (JIN *et al.* 2007). The location of many of the potential N-linked glycosylation sites on the avian TLR15 proteins indicates that these proteins would most likely possess a conserved localization of critical ligand recognition sites. Eight of the conserved N-linked sites are located in or adjacent to the conserved β-strand of an LRR and similar to that observed in mammalian TLR3. The resulting heavily glycosylated surface would be unlikely to be involved in ligand-receptor interactions as the attached sugars could act as a steric hindrance to any potential binding (CHOE *et al.* 2005)

2.4 Conclusions

In this study we used a combination of BLAST homology and HMM profile searches to uncover a novel vertebrate TLR – designated TLR15. This TLR codes for a protein containing a highly conserved TIR domain as well as an extracellular domain composed of multiple LRR motifs. Phylogenetic analysis indicates that this novel TLR is a distantly related member of the TLR2 subfamily and represents an expansion of this family different to that seen in mammals. TLR15 was named sequentially based on its distinctive extracellular domain structure which differs from any of the known mammalian TLRs (TLR 1-13), and it is also distinctive from TLR14, which had previously been described in *Takifugu rubripes*, *Tetraodon nigroviridis* and *Xenopus tropicalis (ROACH et al. 2005)*. TLR15 may represent a bird and reptile specific receptor that has been either gained in these orders alone or may have been a feature of the last common ancestor of all amniotes and simply lost from the mammalian lineage. It is likely that similarly unique TLR genes are present in other genomes that contribute to species- or class-specific immune defense mechanisms tailored to combat particular pathogens.

As an addition to the analysis described in this thesis, molecular work carried out by Rowan Higgs has shown that in chicken, this novel TLR is expressed in the spleen, bursa, and bone marrow of healthy chickens, suggesting a role for this novel receptor in constitutive host defense (HIGGS *et al.* 2006). Following

in vivo Salmonella enterica serovar Typhimurium infection, quantitative real-time PCR demonstrated significant upregulation of TLR15 in the caecum of infected chickens. Interestingly, similar induction of TLR2 expression following infection was also observed, suggesting a functional relationship between these two TLRs. *In vitro* studies revealed TLR15 upregulation in chicken embryonic fibroblasts stimulated with heat-killed *S. enterica* serovar Typhimurium. These results suggest a role for the TLR in avian defense against bacterial infection.

It is becoming increasingly clear that analysis of the evolutionary relationships among genes can offer significant insights into the different functions of the cognate proteins. Based on the analysis described above the TLR15 most likely evolved and was retained for recognition of a specific lipopeptide ligand distinct from those recognized by either TLR1 or TLR2 (UEMATSU and AKIRA 2008). Further functional characterisation to determine the preferential ligand of this novel TLR will represent a significant step in tracing the evolutionary history and divergence pattern of this immunologically important gene family.

3. The avian TLR Pathway – subtle difference amongst general conformity

Abstract

All jawed vertebrates possess an adaptive immune system which varies greatly in complexity and structure among the different branches of the vertebrate phylogeny. By contrast the more ancient and purportedly more primitive innate immune system, of which the Toll-Like Receptor (TLR) pathway is a major component, is maintained with remarkable consistency across all vertebrates. This pathway is responsible for initial recognition of invading microbial pathogens resulting in transcriptional regulation and stimulation of innate immune functions including expression of effector molecules such as AMPs. Amidst this background of overall conservation, subtle differences in the components that make up this pathway may have important implications for species-specific defense against key pathogens. Here we employ a homology informed method to characterize the TLR pathway in the chicken and recently sequenced zebra finch genomes. In general, the components of the pathway are the same in birds and mammals, although there are clear differences. The TLR receptors show a pattern of gene duplication and gene loss both in individual avian species and birds in general. In particular we observe avian specific duplication of both TLR1 and TLR2 as well as a recent duplication of the TLR7 gene in the zebra finch lineage. We note that both positive selection and gene conversion shape the evolution of the avian specific TLR2 genes. In addition we observe notable differences in the finch AMP repertoire when compared to that of the chicken - the only other well characterised avian species. Bioinformatic analysis revealed no evidence of cathelicidins in the finch genome but find that an additional 10 defensins map to the avian beta defensin cluster on chromosome 3. These findings could provide insight into the differing immune response invoked in individual vertebrate species in response to invasion by particular pathogens from their own to microbiological environment.

3.1 Introduction

Recent years has seen a growth in interest in the molecular components of the vertebrate innate immune system. The application of comparative studies has shown that many innate host defence mechanisms are conserved amongst all vertebrates whilst simultaneously suggesting that at a molecular level, every species has modified these defences, presumably to tailor their immune responses to their particular circumstances. The Toll-Like receptor (TLR) pathway represents a well conserved immune signalling cascade when comparisons are made between even distantly related species (LEULIER and LEMAITRE 2008). The constituent components of the TLR pathway can be divided into three distinct groups; (1) TLRs, the molecules responsible for pathogen recognition, (2) the intracellular signalling components that relay the immune signal to (3) the downstream effector molecules that execute the host response (all reviewed in chapter 1). In spite of this overall conservation of the TLR pathway, differences in the regulation and functioning of TLR signalling between species would be influenced by the presence or absence of any of the intermediate components of the pathway. To date, most studies relating to the TLR pathway have been conducted in mammals and insects, while comparatively little is known about this immune response mechanism in non-mammalian vertebrates. In a previous study, chicken Expressed Sequence Tags (ESTs) were clustered and searched for orthologs of known TLR pathway components (LYNN *et al.* 2003). This work identified at least partial sequence data in chicken for almost half of the pathway proteins known at the time. ESTs represent a valuable resource for gene discovery but are subject to some critical limitations. By definition they are short, single-read sequences derived from expressed mRNA. Current technology limits ESTs to between 200 and 500 nucleotides in length (http://www.ncbi.nlm.nih.gov/projects/dbEST/) with a predicted error rate on base calling of 3% (BOGUSKI *et al.* 1993). In addition EST reads are generated from either the 5' or 3' end of expressed genes and so are sometimes composed solely of untranslated/non-coding regions of the original mRNA transcript. Several limitations can be overcome by clustering of ESTs that are likely to be derived from the same underlying gene. This

technique reduces the redundancy in the EST database, and improves the sequence quality by generating a consensus sequence which represents all ESTs associated with a particular cluster.

Subsequent to the previous study in chicken, a large number of ESTs have become available in a second avian species, the zebra finch as well as high quality draft genome sequences for the chicken (2004) and zebra finch species (http://genome.wustl.edu/genome.cgi?GENOME=Taeniopygia%20guttata). Here we describe the use of *in-silico* techniques to identify and characterise the components of the Toll-Like-Receptor (TLR) pathway present in these two avian species as well as identification of a family of downstream mediators of the immune response whose expression is dependent on the pathway. The ancient divergent point separating chicken (*Galliformes*) and zebra finch (*Passeriformes*) is estimated to have occurred 90-100 million years ago and represents one of the basal splits within the established avian phylogeny (GIBB *et al.* 2007). As a result, these two genomes provide a comprehensive resource in which the TLR pathway can be completely reconstituted for the avian order as well as providing insight into the nature of any avian-clade or avian-species specific gene duplications or losses. In addition, the availability of two avian genome resources allows for investigation of the selective pressures that have influenced the adaptation of the different TLR pathway genes in avian lineages.

3.2 Materials and Methods

A total of 66,719 zebra finch EST sequences were downloaded from Genbank. Before clustering a number of pre-processing steps were performed on the EST dataset. Repeat sequences were masked in RepeatMasker (www.repeatmasker.org). The dataset was then cleaned of vector contaminant sequences using SeqClean (www.tigr.org). ESTs were clustered using the TGICL software (PERTEA *et al.* 2003) based on the criteria that they must overlap by at least 30 base pairs at a greater than 95% identity. Phrap (Green P. 1996) (http://www.phrap.org) was used to assemble the predicted clusters and a consensus sequence was generated for each cluster using Craw (CHOU and

BURKE 1999). ESTScan was used to identify the potential coding regions within the generated clusters and corrects for errors resulting from frame-shifts occurring during the EST generation process. Finally the predicted protein corresponding to the predicted coding region was generated using the TRANSEQ program from EMBOSS (www.emboss.org).

A dataset of proteins with known involvement in the human TLR pathway was created using the KEGG reference database (http://www.genome.jp/kegg/) and the uniprot database (http://www.uniprot.org/). Additional genes involved in TLR signalling were derived from literature searches of published research, resulting in a dataset of 63 genes with which to search the avian genomes. A dataset of all previously identified avian TLRs was also assembled (See Appendix). A dataset of AMPs was also created using known mammalian and avian proteins. The draft versions of the sequenced and assembled chicken and zebra finch genomes were downloaded from the GoldenPath (http://genome.ucsc.edu/) and searched using the TBLASTN program (ALTSCHUL *et al.* 1997). The 16,713 Ensembl predicted chicken gene set as well as the predicted peptide assemblies generated by the clustering of the zebra finch ESTs (see above) were also searched for sequences displaying a high degree of similarity to mammalian TLR components. Putative avian genes were extracted and aligned with orthologs from other species using the T-Coffee program (NOTREDAME *et al.* 2000). Phylogenetic trees were constructed using MEGA 4.0 (TAMURA *et al.* 2007) in order to classify the avian homologs accurately, particularly in the case of multi-gene families where ortholog identification by BLAST similarity alone may be unreliable. Chromosomal location and strand orientation of the identified avian genes were determined using the BLAST Like Alignment Tool (BLAT) (KENT 2002) . This tool was also used to determine the degree of syntenic gene order conservation among the avian species and in comparison with other vertebrates.

Protein evolution is driven by selective forces acting on different sites in the protein as well as the accumulation of neutral changes through genetic drift.

One of the most stringent methods for detecting positive selection at the protein level is comparison of the rate of nonsynonymous substitutions (d_N) to the rate of synonymous substitution (d_S) using the ratio between these rates ($\omega = d_N/d_S$). If nonsynonymous changes are selectively neutral they will be fixed at the same rate as synonymous mutations such that $\omega = 1$. If nonsynonymous mutations are deleterious, they will be removed more frequently than synonymous ones, yielding $\omega < 1$; correspondingly, if nonsynonymous changes are selectively advantageous then $\omega > 1$ (ANISIMOVA *et al.* 2002). The PAML 3.15 suite of programs was used to test for positive selection affecting gene families showing evidence of either gene gain or gene loss in one or both of the chicken and zebra-finch genomes (YANG 1997). We applied the maximum likelihood based branch-specific and branch-site specific models to test for episodic selection following gene duplication affecting specific branches in birds. For this analysis, model M0 assumes a constant ω across all sites and branches in a phylogeny. This was compared by likelihood ratio test (LRT) to a second model, the two-ratio model - this allows ω to vary only on a post duplication branch. The LRT compares twice the difference of the log-likelihoods of the two models to a χ^2 distribution with N-1 degrees of freedom where N is the number of branches in the phylogeny (NIELSEN and YANG 1998). DAMBE was used to detect any problems relating to saturation of synonymous substitutions that could result from the presence of long branches in the gene phylogeny (XIA and XIE 2001). For branch-site specific analysis, variable model A (Ma) is compared to a neutral null model (Ma1) in order to identify particular amino acid residues that have undergone adaptive changes along known lineages of interest.

In the analysis of the avian TLR2 duplicate genes, detection of likely borders of gene conversion events were determined by visual inspection of the alignment of the chicken and zebra finch TLR2-1 and TLR2-2 protein sequences and confirmed using the Geneconv program (SAWYER 1989). The crystal structure of the extracellular domains of human TLR1 and TLR2 were

downloaded from the protein databank (PBD ID. 2Z7X) and visualized using the pymol molecular viewer program (http://www.pymol.org).

In the analysis of the defensin gene family, models M1a, M2a, M7 and M8 are used to conservatively test for sites that have been subject to selection across all the sequences sampled from both bird species. The neutral model M1a assumes two site classes with one fixed at $\omega = 1$. This is compared to model M2a, which adds a third site class with a variable ω estimated from the data. Neutral model M7 (beta) is compared with model M8 (beta & ω), where $0 \geq \omega \leq 1$ for M7 by the assumption of a beta distribution B (p, q). M8 has an additional class of sites with proportions and ω estimated from the data such that ω can be greater than 1. The significance of the differences between the nested models was estimated by LRTs with two degrees of freedom (df) for both comparisons. Where the variable model is significantly more likely, a Bayes empirical Bayes (BEB) approach is used to calculate the probability that each site has $\omega > 1$ for models M2a and M8. Only sites with > 95% probability of being positively selected are reported here, as such sites are very likely to have evolved under diversifying selection pressures.

3.3 Results and Discussion

In this study, we have employed a homology-based strategy to reconstruct the TLR pathway in both chicken and zebra finch - the two avian species for which complete genome sequences are currently available. In addition 66,719 zebra finch ESTs were clustered into 9681 contigs consisting of 2 or more clustered ESTs, and 15,402 singleton sequences. These sequences were searched to identify any possible homologs not found in the sequenced zebra finch genome as well as to confirm the expression of the identified TLR pathway orthologs in this species.

63 mammalian proteins sequences with known involvement in this innate immune cascade have been described and these proteins were used to identify corresponding one-to-one orthologous genes in the two bird species. Overall,

the TLR pathways in the two avian species studied are remarkably similar. With the exception of a duplication of TLR7 in zebra finch, both bird genomes code for an identical subset of the mammalian TLR pathway genes examined (Table 3.1). The speciation event that ultimately led to the extant lineages of both chicken and zebra finch represents one of the basal divergences within the avian phylogeny (HACKETT *et al.* 2008). This was one of the reasons why the zebra finch was chosen as the second avian species to have its genome sequenced. Our results suggest that the minimally required TLR pathway for avian species was determined prior to this divergence, with no subsequent loss of established pathway components occurring in the subsequent 90 million years.

Table 3-1. TLR pathway genes and Antimicrobial Peptide genes identified in chicken and zebra finch genomes. Chicken genes are represented by corresponding Ensembl prediction identifier, derived from Ensembl 51: Nov 2008. Finch Location refers to chromosome and starting base pair for the orthologous gene in the current release of the zebra finch genome.

Gene Name	Chicken Ensembl Gene	Finch location	Average % id between avian and human genes
Toll Like Receptors and associated external proteins			
TLR1-1	ENSGALG00000017485	4:48369726	52
TLR1-2	ENSGALG00000022995	4:48357924	
TLR2-1	ENSGALG00000009237	4:27726290	51
TLR2-2	ENSGALG00000009239	4:27734012	
TLR3	ENSGALG00000013468	4:39286156	61
TLR4	ENSGALG00000007001	17:3832626	47
TLR5	ENSGALG00000009392	3:8554975	52
TLR6	Mammalian		
TLR7	ENSGALG00000016590	1:18349048	64
TLR7-2		1:18334615	
TLR8	Lost		
TLR9	Lost		
TLR10	Mammalian		
ggTLR15	ENSGALG00000008166	3:28850976	
ggTLR21	ENSGALG00000000774	Un:124095996	
MD-2	ENSGALG00000015648	2:124381556	32
CD-14	ENSGALG00000021559	13:15,730,730	34
SIGIRR	ENSGALG00000004267	5:7139564	58
Adaptor Molecules			
MyD88	ENSGALG00000005947	2:5029035	72
MAL	ENSGALG00000001077	24:7597116	55
TICAM-1	ENSGALG00000024109	28:3824923	35
TICAM-2	Mammalian		
SARM	ENSGALG00000003595	19:7001321	72
Non-Kinase Pathway components			
TRAF6	ENSGALG00000007932	5:7001321	76
TRAF3	ENSGALG00000011389	5:51243956	89
ECSIT	Lost		
UBC13	ENSGALG00000011292	1A:44129387	95
UeV1A	ENSGALG00000008015	20:14968123	96
TRADD	ENSGALG00000003195	11:6503633	50
FADD	ENSGALG00000007625		41
TOLLIP	ENSGALG00000006697	5:14626848	89
NAP1	Lost		
TAB1	ENSGALG00000012150	1A:50302173	89
TAB2	ENSGALG00000012356	3:47349347	88
TAB3	ENSGALG00000016284	1:10283061	88
DDX3X	ENSGALG00000016231	1:6619028	92
PELLINO1	ENSGALG00000008837	3:1634246	99
PELLINO2	ENSGALG00000012124	5:58508038	96
PELLINO3	Lost		

TLR Pathway Kinases			
IKK-A	ENSGALG00000003289	6:3349399	81
IKK-B	ENSGALG00000003740	22:3346354	83
IKK-E	ENSGALG00000013356	1A:32965546	69
IKK-G	Lost		
IRAK1	Lost		
IRAK2	ENSGALG00000008407	12:21143145	46
IRAK3	Lost		
IRAK4	ENSGALG00000009586	1A:29368030	64
MEKK1	ENSGALG00000014718	Z:47924788	87
MKK3-6	ENSGALG00000004735	14:10991844	95
MKK7	Lost		
TAK1	ENSGALG00000015596	3:77416948	93
TBK1	ENSGALG00000009840	1A:32965546	86
TANK1	ENSGALG00000011131	7:11474255	52
NIK1	ENSGALG00000000685	27:1236599	62
JNK	ENSGALG00000006109	6:17318879	99
BTK	ENSGALG00000004958	4A:19635363	86
MAPK1	ENSGALG00000001501	15:8343306	100
RIP1	ENSGALG00000012827	2:43103429	50
RIP2	ENSGALG00000015899	2:130575938	67
RIP3	Lost		
Transcription factors and associated proteins			
IRF1	ENSGALG00000006785	13:11407450	63
IRF3	Lost		
IRF5	ENSGALG00000001405	26:3269351	59
IRF7	ENSGALG00000014297	5:15855987	43
NFkB1	ENSGALG00000012304	4:21423294	73
NFKB2	ENSGALG00000005653	6:15893040	76
IKBa	ENSGALG00000010063	5:36152194	71
IKBb	Mammalian		
Known Chicken Antimicrobial Peptides			
AvBD1	ENSGALG00000022815		
AvBD2	ENSGALG00000016669	3:110796397	80
AvBD3	ENSGALG00000016670		
AvBD4	ENSGALG00000019843	3:110735427	66
AvBD5	ENSGALG00000016671	3:110739722	68
AvBD6	ENSGALG00000016668		
AvBD7	ENSGALG00000022817	3:110802774	61
AvBD8	ENSGALG00000019844	3:110810151	61
AvBD9	ENSGALG00000019845	3:110813816	90
AvBD10	ENSGALG00000016667	3:110822873	78
AvBD11	ENSGALG00000019846	3:110848147	71
AvBD12	ENSGALG00000019847	3:110851155	60
AvBD13	ENSGALG00000019848	3:110858416	80
AvBD14	CAL47019		
LEAP2	ENSGALG00000007099	13:1155326	65
Cathelecidin	ENSGALG00000019696		

TLRs

The TLR repertoire coded for in the chicken genome has been well characterized (HIGGS *et al.* 2006; IQBAL *et al.* 2005; TEMPERLEY *et al.* 2008). We identified orthologous sequences for all ten of the previously identified chicken TLRs in the zebra finch genome as well as a novel TLR generated through the tandem duplication of the TLR7 gene (Table 2.1). No evidence of the zebra finch TLRs was found in the clustered EST dataset.

The two predicted zebra finch TLR7-like genes code for proteins that differ in composition at 21 amino acid sites, which is comparable to the sequence divergence observed between the TLR7 genes of cow and sheep. Analysis of the TLR7 gene family using the maximum likelihood methods which take into account the evolutionary relationships between the sequences indicates a statistically significant excess of amino acid changing substitutions on the zebra finch TLR7-2 branch ($p = 0.0226$) though the low level of sequence diversity between the two finch TLR7-like genes precludes the confident identification of individual amino acid sites subject to positive selection subsequent to the gene duplication event (Table 3.2).

Previous studies reported the tandem duplication of the TLR2 gene in the chicken lineage (TEMPERLEY *et al.* 2008). Our analysis of the zebra finch genome also identified two syntenic TLR2 genes, implying that the duplication of this gene in birds occurred over 100 million years ago and suggesting that an expanded TLR2 subfamily is likely to be a feature of most bird species. A phylogenetic tree constructed using publicly available full-length vertebrate TLR2 sequences fails to reconstruct the expected phylogeny for the TLR2 family in birds, as the gene duplicates within each species cluster together rather than with their respective orthologous sequence from the other avian species (Figure 3.1A). When the alignment of the avian TLR2 sequences are scrutinized, a 190 amino acid sequence corresponding to LRR8 to LRR15 (Figure 2A and 2B) is observed in the extracellular domain that generates a phylogenetic signal in agreement with the divergence of TLR2-1 and TLR2-2

prior to the last common ancestor of chicken and zebra finch (Figure 3.1B). In contrast, two separate regions corresponding to the N-terminal 250 amino acids and the C-terminal 342 amino acids display a higher degree of similarity in TLR2-1 and TLR2-2 comparisons within each avian species, than in between species ortholog comparisons (Figure 3.2A and 3.2B). The information derived from the phylogenetic trees using different regions of the TLR2 protein sequence is suggestive of gene conversion events occurring between the TLR2 duplicate genes in these avian species. The absence of gene conversion in the extracellular 190 amino acid region coupled with the homogenization of both terminal regions of the TLR2 proteins in birds, suggests that the non-converted region is the key domain driving the maintenance of both TLR2 gene duplicates in avian lineages.

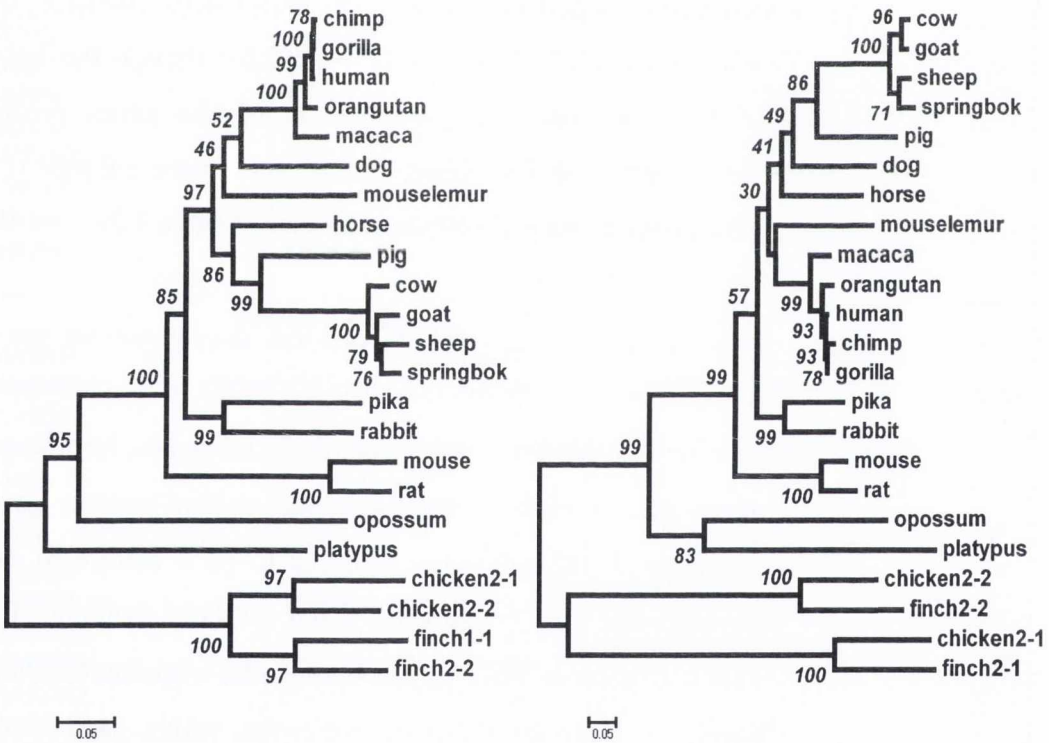


Figure 3-1. A. Neighbour joining tree of vertebrate TLR2 sequences constructed using full length protein sequences constructed using MEGA version 4.0 (Poisson Correction model, 1,000 bootstrap replicates). B. Neighbour-joining tree of vertebrate TLR2 sequences constructed using the 190 amino acid sequence corresponding to the non-gene converted region of avian TLR2s.

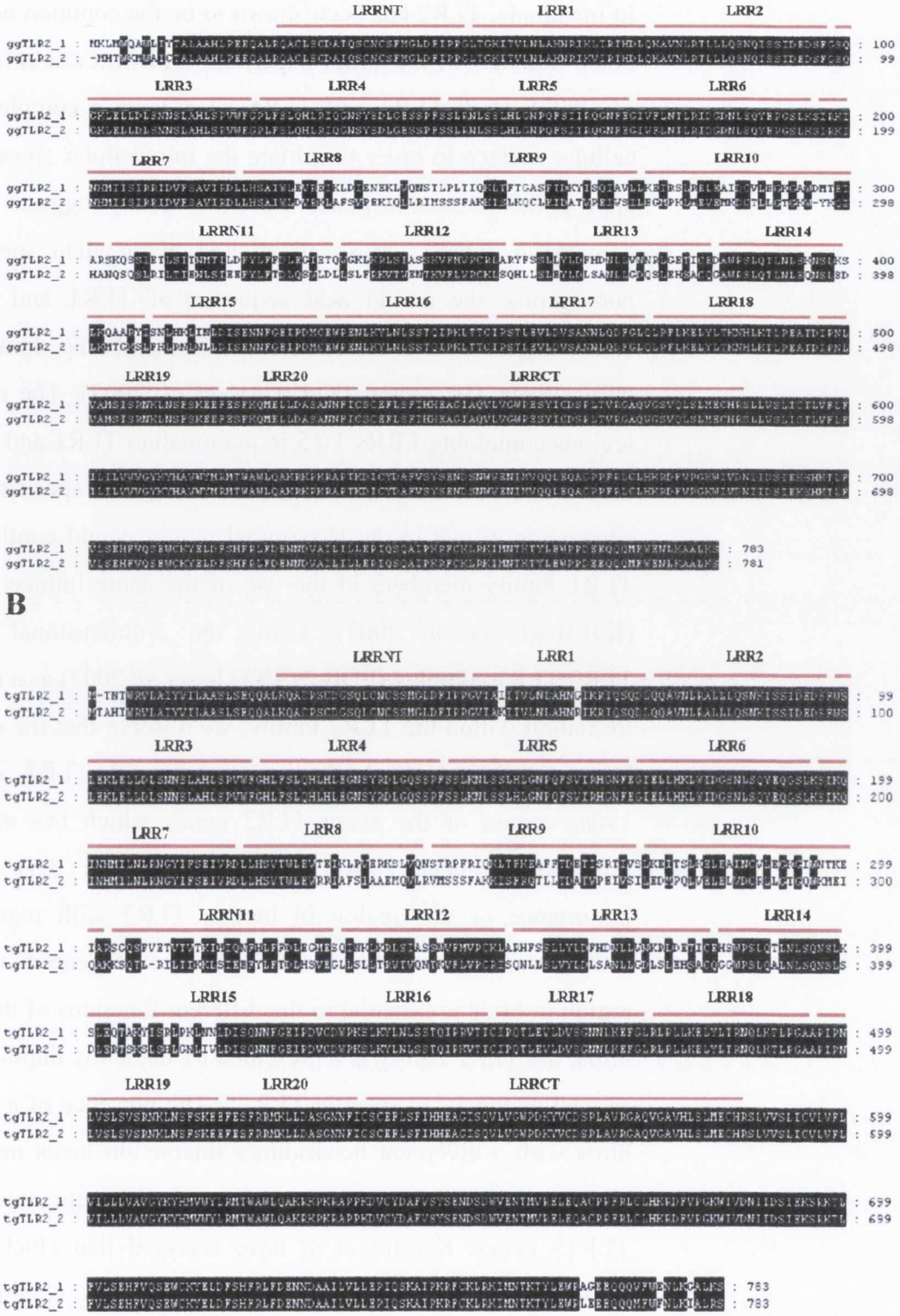


Figure 3-2. Multiple sequence alignment of chicken TLR2-1 and TLR2-2 with predicted locations of Leucine Rich Repeats indicated. B. Multiple sequence alignment of chicken TLR2-1 and TLR2-2 with predicted locations of Leucine Rich Repeats indicated.

In mammals, TLR2 has been shown to be the common heterodimer partner for other members of the TLR1 superfamily (TAKEUCHI *et al.* 2001; TAKEUCHI *et al.* 2002). Both TLR1 and TLR6 must form a complex with TLR2 on the cellular surface in order to initiate the intracellular signalling cascade and the appropriate immune response. Recent evidence suggests that independent gene conversion events, across all studied mammalian species, have served to homogenize the amino acid sequence of TLR1 and TLR6 in the region consisting of LRR16-LRR19, the transmembrane domain and most of the intracellular TIR region (KRUITHOF *et al.* 2007). The non-conversion of the regions containing LRRs 1-15 in mammalian TLR1 and 6 could have allowed the two proteins to develop diverse ligand specificity, while the gene conversion events in the N-terminal region would continue to facilitate these TLR1 family members in the use of the same intracellular adaptor proteins (KRUITHOF *et al.* 2007). Using the 3-dimensional structure of human TLR2/TLR1 complex (PDB 2Z7X) (JIN *et al.* 2007) as a model for heterodimer formation within the TLR2 family, we observe that the extracellular region of human TLR2 which physically interacts with TLR1, is homologous to the 190aa region of the avian TLR2 genes which has not evolved under the influence of gene conversion events (Figure 3.3). Given the apparent importance of this region in human TLR2 with regard to protein-protein interaction, it is possible that the absence of gene conversion in this 190aa region in birds is essential to the different functions of the duplicated genes. In birds, the TLR1 family is represented by tandemly duplicated TLR1-like genes as well as the avian specific TLR15. The presence of a second TLR2 gene in birds with a divergent heterodimer interaction locus may satisfy the need in birds for a novel dimerisation partner for either the avian specific TLR1-2 or TLR15 genes. Keestra *et al* have reported that chicken TLR1-1 can only activate NF-κB transcription factors when in combination with TLR2-2 (KEESTRA *et al.* 2007) and we have previously shown highly correlated upregulation of TLR15 and TLR2 following in vivo *Salmonella enterica* serovar *Typhimurium* infection in chickens (HIGGS *et al.* 2006). As chicken TLR2-2 appears to be the functional homolog of the single mammalian TLR2

gene (KEESTRA *et al.* 2007), we examined the avian branch leading to TLR2-1 for evidence of positive Darwinian selection, which might have contributed to the evolution of this gene segment following the TLR2 gene duplication event in birds. A multiple sequence alignment and phylogenetic tree was constructed using the corresponding 190 amino acid region from other vertebrates (Figure 3.1B) and codeml was used to evaluate the selective constraints acting on this gene in avian lineages. Branch specific models indicate a significant difference in the rate of amino acid substitutions on the branch leading to TLR2-1 in birds compared to all other branches in the phylogeny (Table 3.2). In addition, branch-site specific models identified 6 sites that have been subject to positive selection in this region of the avian TLR2-1 gene following its divergence from TLR2-2 (Figure 3.4 A and B and Table 3.2). The presence of positively selected sites in this 190aa region of TLR2-1 on the branch immediately following the duplication of the ancestral avian TLR2 gene could indicate strong selective pressure on this gene to adopt a novel function – mostly likely driven by ligand or heterodimer partner specificity. Overall, the evolution of the two TLR2 loci in birds provides a striking example of gene conversion of peptide segments that potentially has enabled conservation of functional features of the proteins while simultaneously allowing non-converted regions to evolve novel roles.

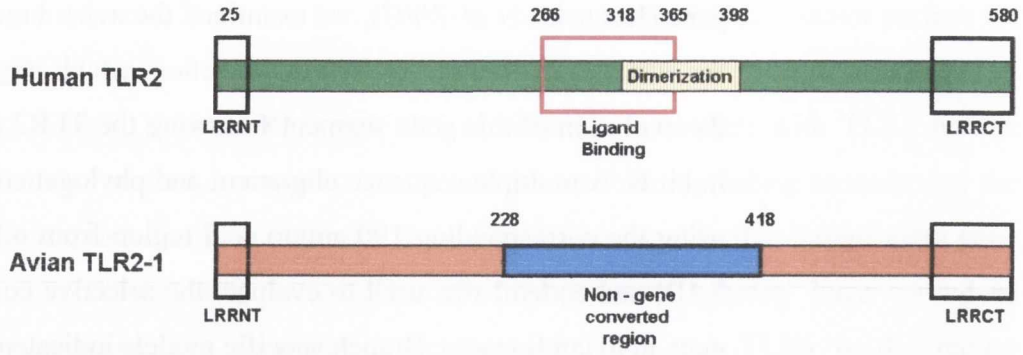


Figure 3-3. Schematic representations of the TLR2 gene from human (top) and the avian TLR2-1 gene (bottom). In human TLR2 the red box indicates the region responsible for binding of the tri-acylated lipopeptide ligand, while the yellow lined box indicates the region of extracellular contact between TLR2 and TLR1 in dimer formation. In avian TLR2-1 the blue box indicates the portion of the gene in which gene conversion has not taken place. The N and C terminal LRRs are represented a black box on both TLR genes. Amino acid numbers corresponding to the start of each domain are indicated above the genes. The ligand-binding and dimerization domains for human TLR2 were identified from the crystal structure described in (Jin et al. 2007).

Table 3-2. Results of CODEML analyses of the avian TLR7 and TLR2 under different branch and branch-site models. Sites identified as positively selected on the avian TLR2-1 branch are numbered based on the chicken TLR2-1 sequence. [A] Positively selected sites identified under Model A with posterior probabilities ≥ 95% on TLR2-1 branch post gene duplication. (Sites with posterior probabilities ≥ 99% are shown in bold).

Model	NP	Parameters Estimates	Likelihood	P-Value	Positive Selection
Finch TLR7-2 - Branch Specific					
One-ratio	37	$\omega = 0.2151$	-21272.9		Not allowed
Two ratios	38	$\boldsymbol{\omega_0 = 1.2133}$ (foreground) $\omega_1 = 0.2151$ (background)	-21270.3	< 0.05	Positive Selection
Avian TLR2-1 - Branch Specific (Nonconverted 190aa region)					
One-ratio	44	$\omega = 0.3420$	-6772.82		Not allowed
Two ratios	45	$\boldsymbol{\omega_0 = 2.2200}$ (foreground) $\omega_1 = 0.3295$ (background)	-6768.71	< 0.01	Positive Selection
Avian TLR2-1 - Branch-Site (Nonconverted 190aa region)					
Model A	45		-6639.57		Not allowed
Model A ($\omega_2 \neq 1$)	46	($\omega_0 = 0.16$), $f_0 = 0.41$ ($\omega_1 = 1$), $f_1 = 0.25$ $\boldsymbol{(\omega_2 = 5.07)}$, $\boldsymbol{f_{2+3} = 0.34}$	-6636.32	< 0.05	[A] 281,A294,**G337,** L370, H375, **I388**

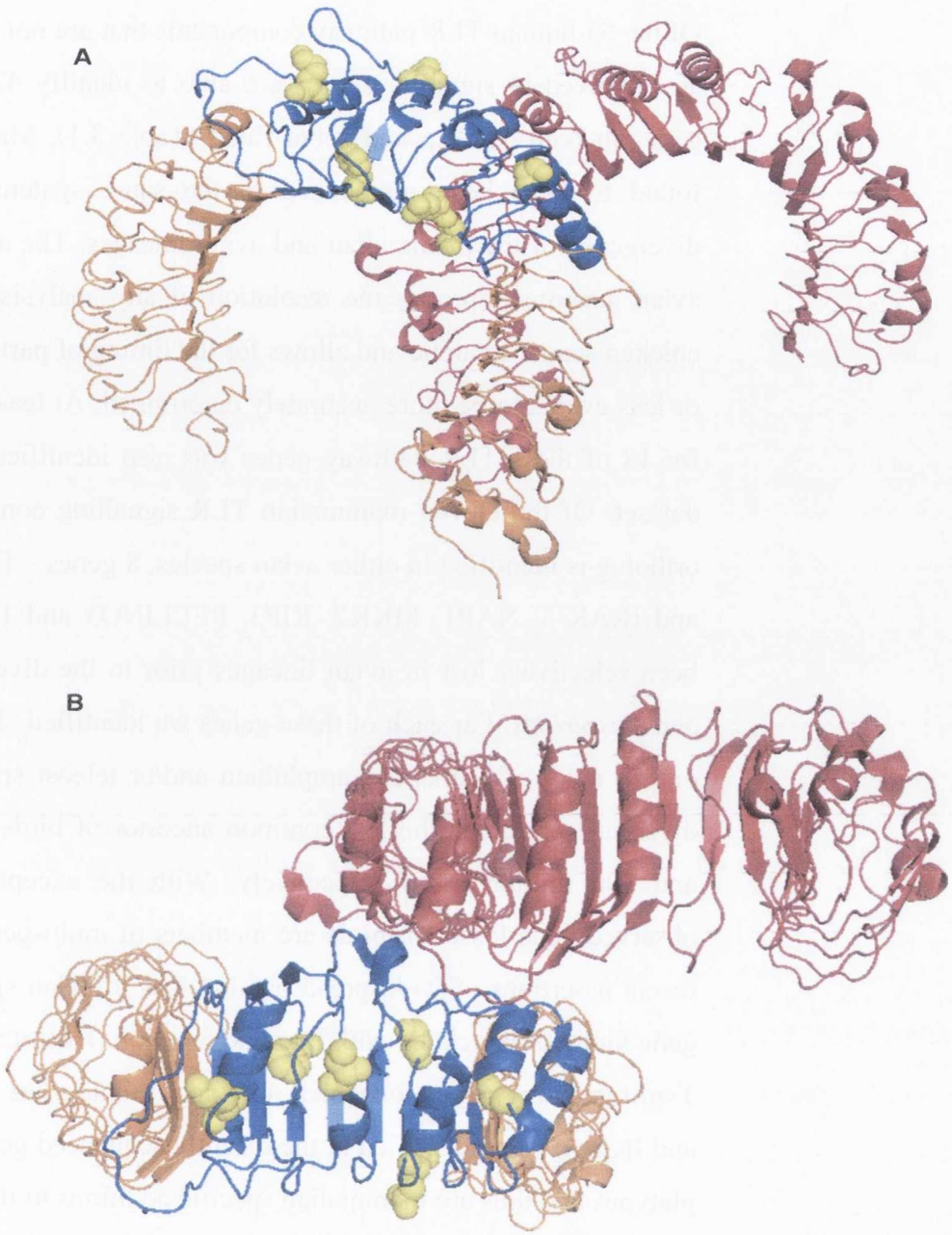


Figure 3-4. Overall structure of human TLR1-TLR2 heterodimer complex. TLR1 is shown in red. TLR2 regions homologous to the tracts subject to gene-conversion in avian TLR2s are shown in blue. The central non-converted region of TLR2 is shown in orange with six sites subject to positive selection in avian TLR2-1 shown in yellow. B. Top down view.

Signalling Components

Of the 53 human TLR pathway components that are not Toll-like-receptors but are involved in signalling, we were able to identify 42 putative orthologous genes in both chicken and zebra finch (Table 3.1). Many of these genes are found to have been maintained in the same syntenic context during the divergence of the mammalian and avian lineages. The availability of a second avian genome improves the resolution of any analysis carried out using the chicken sequence alone and allows for the timing of particular gene duplication or loss events to be more accurately determined. At least partial sequence data for 18 of these TLR pathway genes was also identified in the clustered EST dataset. Of the eleven mammalian TLR signalling components for which no ortholog is identified in either avian species, 8 genes – ECSIT, IKK-γ, IRAK-1 and IRAK-3, NAP1, MKK7, RIP3, PELLINO3 and IRF3 – appear to have been selectively lost in avian lineages prior to the divergence of *Galliformes* and *Passerines*. For each of these genes we identified 1:1 orthologus genes in one or more of sequenced amphibian and/or teleost species genomes, whose divergence predates the last common ancestor of birds and mammals by 150 and 300 million years respectively. With the exception of NAP1, all the observed gene losses in birds are members of multi-gene families, supporting recent assertions of widespread paralog loss in avian species – particularly in gene families associated with immune function (HUGHES and FRIEDMAN 2008). Two remaining genes for which no avian ortholog has been found (TICAM-2 and IKB-β) were identified in the recently sequenced genomes of opossum and platypus and thus are mammalian specific additions to the TLR pathway.

AMPs

In contrast, striking differences are observed in the range of AMP genes coded for by the two studied avian species. Separate studies have identified a single cluster of 14 beta-defensin like genes on chromosome 3 in chicken as well as a single cathelicidin gene on chromosome 12 (LYNN *et al.* 2004; XIAO *et al.* 2004). We identified a homologous but significantly expanded cluster consisting of 22 beta-defensin genes on chromosome 3 in the zebra finch

genome assembly (Figure 3.5) but no evidence was found for any complete cathelicidin gene despite close scrutiny of the predicted syntenic region. Of the previously described chicken defensins, orthologs of AvBD6 and AvBD14 have not been found in zebra finch, indicating that these have originated in the chicken through gene duplication events. Although no clear zebra finch genes orthologous to chicken AvBD1 or AvBD3 are identified here, the latter appears to have served as the progenitor gene for a series of gene duplication events, ultimately resulting in the addition of 9 novel defensin genes in the zebra finch lineage (tgAvBD15 – tgAvBD23), while the three zebra finch defensins tgAvBD24 – tgAvBD26 are most likely to have been derived from the gene coding for AvBD1 (Figure 3.5B). Evidence supporting expression of the novel zebra finch defensins - tgAvBD17 (DV954612) and tgAvBD23 (FE728335) was found by searching the publicly available EST sequences from Genbank suggesting that these novel zebra finch defensins represent transcribed, functional genes. To investigate whether the novel zebra finch defensins contain variable sites which might have evolved under diversifying selective pressures, we examined these genes along with the two corresponding chicken paralogs using site-specific codeml models. Both models M2a and M8, which allow for variable selective pressures among amino acid sites, were found to fit the alignment significantly better than their respective neutral models - M1a and M7 - with both selection models indicating a significant proportion of sites evolving under positive selection (Table 3.2). BEB for both selection models subsequently identified 12 sites that are likely to be subject to positive selection. All 12 sites were located in the mature peptide region; by contrast, the signal peptide appears to be highly conserved with few amino acid changes occurring in this region (Figure 3.5C). As they evolve to exploit different ecological niches, disparate bird species like chicken and zebra finch face a shifting range of microbial challenges, and the duplication of AMP genes and their subsequent diversification would allow species to compete in the "arms race" with the pathogenic microbes specific to their local environmental. Concurrent with this direct antimicrobial role, species-specific repertoires of AMPs could impact on the activity of the adaptive immune response in these

species through influence on chemotaxis and cellular activation. Positive selection of advantageous amino acid substitutions is one means by which AMPs with novel functions can develop. It is possible that the selected variation of the amino acid sites in these closely related finch genes by altering alter important physicochemical properties such as hydrophobicity and charge may play a role in the targeting of these novel defensins to particular microbes.

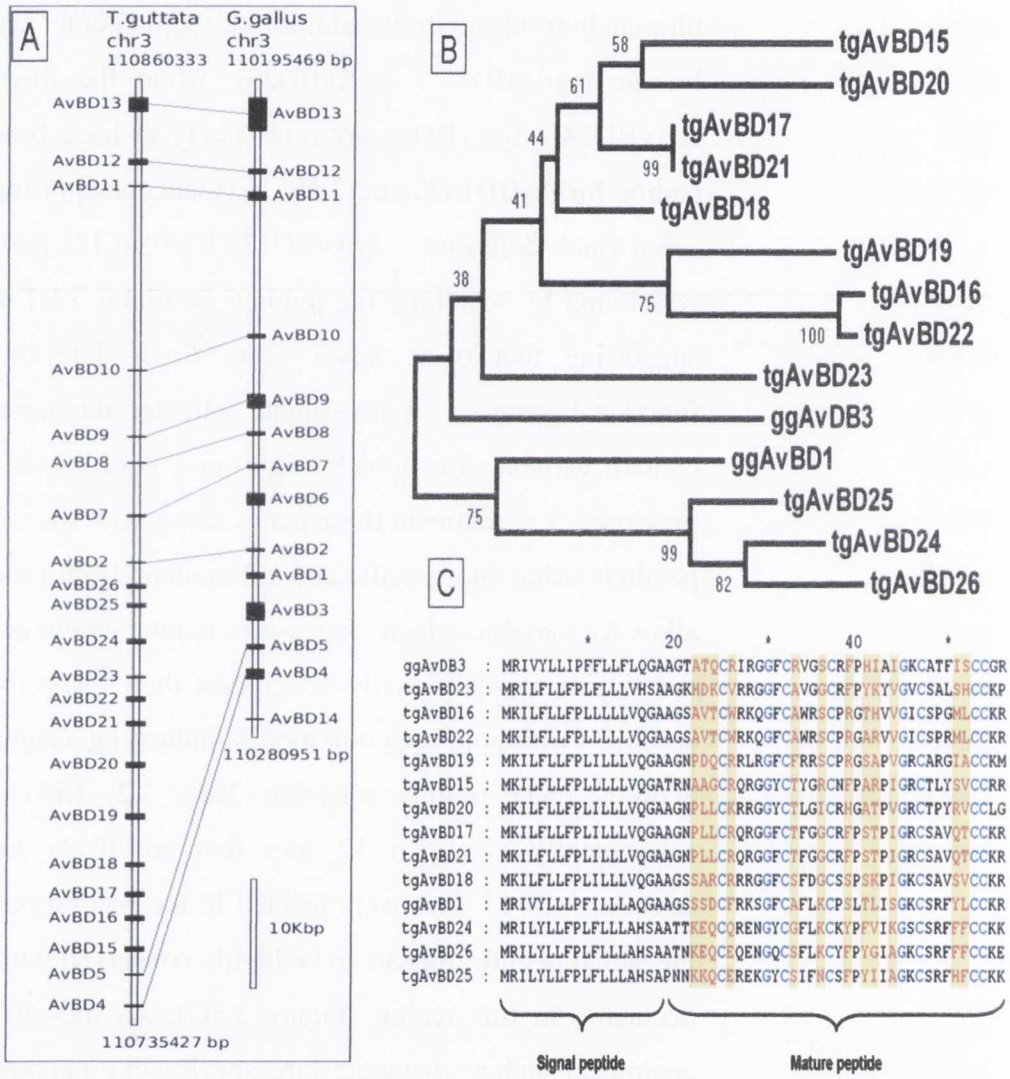


Figure 3-5. Analysis of avian specific beta-defensin cluster. A. Syntenic map of avian beta-defensin cluster on chromosome 3 showing between species orthologs and species specific genes. B. Neighbour joining tree of novel zebra finch beta-defensins (AvBD15-AvBD24) and chicken AvBD1 and AvBD3. C. Multiple sequence alignment of novel zebra finch beta-defensins and chicken AvBD1 and AvBD3. Conserved residues common to all peptides are shown in blue. Positively selected residues are shown in red.

Table 3-3. Results of CODEML analyses of the zebra finch specific beta-defensin genes under different models of variable ω ratios among sites. Positively selected sites identified under models M2A and M8 with posterior probabilities ≥ 95% are listed. (Sites with posterior probabilities ≥ 99% are shown in bold and underlined). Positively selected sites are numbered based on the chicken ggAvBD3 sequence.

Model	P	Parameters	Likelihood	ω	Positively selected sites
M0: one ratio	1		-1664.77	0.7396	
M1:neutral	1	P_0 = 0.43, (ω_0 = 0.06) P_1= 0.57, (ω_1 = 1)	-1582.36	0.5969	
M2:selection	3	P_0 = 0.37, (ω_0 = 0.06) P_1 = 0.27,(ω_1 = 1) **P_2 = 0.35, (ω_2 = 3.1)**	-1566.51	1.3571	22S,**23S**,24D,**26F,33A**, 36K,38P,41T,**42L**,44S, 51Y,52L
M7: beta	2	p = 0.23, q = 0.2	-1580.62	0.6199	
M8: beta and ω	4	P_0 = 0.60, p = 0.22 , q = 0.46 **P_1 = 0.40, ω = 2.62**	-1564.07	1.2507	**22S,23S,24D,26F,33A**, 34F,**36K,38P,41T,42L, 44S**,49R,**51Y,52L**

3.4 Conclusions

The TLR pathway (or a functionally homologous counterpart) is a key mediator of immune responses in a wide array of diverse organisms, ranging from insects to vertebrates. Unfortunately, as a scientific discipline immunology has long displayed a strong anthropocentric bias, and most of the current knowledge regarding the molecular composition of this pathway has been derived from studies using classical model organisms, such as mice, which are closely related to humans. As a result of this, the structure and organisation of this pathway in non-mammalian vertebrates is poorly understood and an evolutionary interpretation on the origin, conservation, duplication and loss of pathway components in vertebrate species is limited. Here we use bioinformatics techniques to reconstruct the evolution of the TLR

pathway for two disparate avian species and show that birds possess a streamlined version of the mammalian TLR pathway when compared to mammals, as well as showing that little variation exists in the components of this pathway among avian species. We provide tentative evidence that there may be less functional redundancy amongst components of this pathway in birds, as many multi-gene families associated with this signalling cascade display an avian-specific pattern of paralog loss.

The sequencing of numerous different genomes has shown that notable diversity exists in the repertoire of immune genes possessed by different species. In the midst of this variation, TLRs have been conserved as primary sentinels of the innate immune system, charged with initial detection and response to invading pathogens. Using all the currently available vertebrate genome sequences it is now possible to reconstruct the evolutionary history of this family of receptor across 600 million years (ROACH *et al.* 2005) and observe that although the overall number of TLR genes per species is relatively stable across most of the phylogeny, each of the four major taxonomic groupings – fish, amphibians, birds and mammals - has seen independent expansions and contractions of individual TLR subfamilies (Figure 3.6). As a result birds carry a repertoire of TLRs which is clearly distinct from all other vertebrate lineages including that of mammals – the only other well studied group of tetrapods.

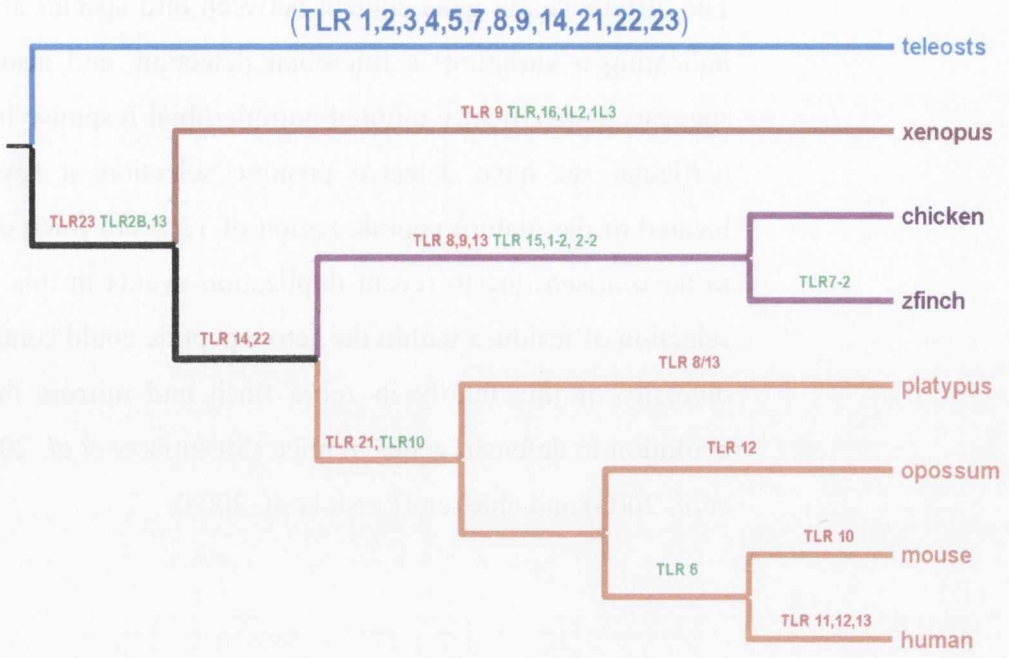


Figure 3-6. Vertebrate phylogenetic tree indicating the likely points of TLR gene-gain and gene-loss events. TLRs marked in green signify gene-gain events the branch indicated. TLRs marked in red indicate gene-loss events on the branch indicated.

In birds we have shown that the TLR2 family has been significantly expanded to include 2 TLR2-like genes, 2 TLR1-like genes and TLR15. The evolution of the TLR2 genes in birds subsequent to duplication shows a similar pattern to that suggested for the mammalian TLR1/TLR6 genes (KRUITHOF *et al.* 2007). Gene conversion between avian TLR2-1 and TLR2-2 has served to homogenize the regions coding for LRRNT-LRR7 and LRR16-TIR of both these genes and suggests strong selective pressures acting to limit the diversification of these regions. In the case of the LRRNT–LRR7 region, limiting divergence may allow for both avian TLR2 genes to interact with an identical but as yet undefined extracellular protein similar to the interaction between TLR4 and CD14. For the LRR16-TIR domain restrained divergence may ensure that the same intracellular adaptor protein is utilized by both paralogs. At the same time independent evolution of the 190aa region corresponding to LRR8-LRR15 which encompasses both the ligand binding and dimerization domains would allow the TLR2 duplicates to discriminate with regard to potential ligands and dimerization partners.

The differences in gene content between bird species are limited to the TLRs, indicating a variation in microbial detection, and also to the AMPs, which suggests a specifically tailored antimicrobial response in each bird species. In particular we have detected positive selection at several amino acid sites located in the mature peptide region of 12 zebra finch defensins which appear to have arisen due to recent duplication events in this lineage. Such positive selection of residues within the active peptide could contribute to the functional diversity of this family in zebra finch and mirrors the finding of adaptive evolution in defensin genes in mice (MORRISON *et al.* 2003), primates (SEMPLE *et al.* 2003) and chicken (LYNN *et al.* 2004).

4. Characterisation of the Bovine Beta-Defensin Repertoire

Abstract

The evolution of an advanced adaptive immune system occurred, not to supersede the existing innate immune system in complex organisms, but to complement it. Therefore most modern higher mammals and vertebrates retain many of the components of this innate immune system. Antimicrobial Peptides represent one of the most important effector arms of this innate response. β-defensins constitute a widely distributed and highly diverse sub family of AMPs. These proteins appear to function both as antibacterial agents themselves as well as attractants for other components of the immune system to sites of infection. In addition, recent evidence suggests a role for these peptides in functions as diverse as coat colour differentiation in dogs as well as sperm maturation and motility in various mammals. A consequence of these discoveries of the pleiotropic functions of β-defensins has been increasing interest into the evolutionary history of this gene family, and in particular into species-specific family expansions or contractions occurring in various mammals. This chapter describes a systematic search of the sequenced bovine genome to characterise this extensive gene family in *Bos taurus*, providing new insight into the pattern of evolution of β-defensin genes among species.

4.1 Introduction

The evolution of more complex multicellular organisms has necessitated the development of intricate, multilateral defence systems to counteract the diverse range of pathogenic insults that these organisms can encounter. The same basic principles of pathogen recognition and generated response have been described across a wide range of biological taxa. Of the generated effector responses of an immune response, one of the most ancient and evolutionarily well conserved is the production of small protein molecules known as Antimicrobial Peptides (AMPs).

Defensin genes code for an extensive family of small cationic anti-microbial peptides (AMPs) that constitute an important effector arm of the innate immune system across a wide range of species (GANZ 1999) as well as representing a functional link between the innate and adaptive immune responses in higher organisms (YANG *et al.* 1999). The principal subdivision of the family into α-defensins and β-defensins is characterised by distinctive spacing of cysteine residues in their active peptide regions as well as their patterns of tissue expression (LEHRER and GANZ 2002). A third defensin sub-type, θ-defensins, appear to be a recently acquired, primate-specific class of peptides, which are generated by the merging of two α-defensin-like precursors (TANG *et al.* 1999). α-Defensins have only been described in mammals, but they are widely distributed in that order (LYNN and BRADLEY 2007). β-Defensins are the more comprehensively studied sub-class and possess the widest taxonomic distribution, being found in vertebrates and invertebrates as well as plants (BOMAN 2003), thus indicating an ancient point of origin. Multiple gene duplication and loss events and subsequent sequence diversification in the mammalian lineage has resulted in a large family of proteins with diverse amino acid sequence but virtually identical tertiary structure based on the characteristic disulphide bridging between cysteine residues (BAUER *et al.* 2001), (See chapter 1 for full review).

Prior to the availability of completely sequenced genomes the identification of new AMPs was hindered both by the type of sequence resources available and the sequence characteristics of this class of effector molecules. In the absence of a genomic sequence for a species, ESTs provide a potential but under-powered resource for bioinformatic identification of AMPs. Many immune genes are expressed at low levels in most cell types and therefore are likely to be under-represented in the cDNA pool used to generate an EST library. Associated with this are the often restricted patterns of tissue expression observed for many immune related genes whereby some genes may only be expressed in a restricted subset of cells or tissues. Because of these limitations, EST datasets sourced from non-immune related tissue types are likely to be deficient in immune related genes. In addition many immune response genes are inducible, and do not display constitutive expression in the absence of pathogenic stimuli and EST libraries from healthy, non-infected individuals and tissues would be unlikely to provide evidence of expression for such inducible genes. Further difficulties in novel AMP identification arise from the fact that AMPs themselves tend to be small and variable in size, usually in the range of 6-100 amino acids. They tend to be rapidly evolving, and as a consequence display a low level of sequence conservation, even amongst gene-duplicates. These characteristic along with a highly variable secondary structure and architecture mean that it is often difficult to identify novel AMPs using traditional bioinformatic techniques.

Recent sequencing of high quality drafts of many vertebrate genomes has allowed for a comparative genomic approach to be used in characterizing the β-defensin repertoire in various species (PATIL *et al.* 2005; RADHAKRISHNAN *et al.* 2005; SCHUTTE *et al.* 2002). In species with recent common ancestors, comparable numbers and types of β-defensin have been identified. However, even in closely related species or clades, gain or loss of individual β-defensin genes has been demonstrated – a possible consequence of differing pathogenic insults each individual species faces during the course of its evolution within its own ecological niche (RADHAKRISHNAN *et al.* 2005).

The bovine lineage was one of the first in which β-defensin-like molecules were discovered (DIAMOND *et al.* 1991; SELSTED *et al.* 1993; TARVER *et al.* 1998). These previous studies resulted in the identification of 18 complete and partial bovine β-defensin (BBD) sequences. This number was noticeably smaller than the 35 human, 33 chimp, 38 dog and 45 mouse purportedly functional β-defensin genes subsequently identified by both *in silico* and laboratory methods (PATIL *et al.* 2005). In this study, we describe a bioinformatics strategy using complementary search and characterisation techniques for the identification and analysis of the entire β-defensin repertoire encoded in the bovine genome: the homology-based search methods of the BLAST family of programs (ALTSCHUL *et al.* 1997) and the more sensitive Hidden Markov Models (HMM) (EDDY 1998). The use of HMM profiling searches allows us to avail of position-specific information unique to the β-defensin gene family and improve the sensitivity of gene search efforts compared to the use of BLAST alone which assumes all positions in a protein to be equally important. The presence of conserved cysteine containing domains in these peptides made them good candidates for the construction of HMMs. Specifically the spacing of cysteines within the mature peptide region allows for HMMs to identify β-defensins genes but also distinguish them from α-defensin genes whose six-cysteine motif has a different spacing pattern (LEHRER and GANZ 2002).

4.2 Materials and Methods

All publicly available protein sequences corresponding to the known human, mouse and dog β-defensin sequences (PATIL *et al.* 2005) were retrieved from GenPept (http://www.ncbi.nlm.nih.gov/GenPept) (See Appendix). The draft 4.0 version of the sequenced and assembled bovine genome was downloaded from Ensembl (http://www.ensembl.org/). In order to carry out Hidden Markov Model (HMM) (Eddy 1998) searches of the bovine genome for β-defensin motifs, the entire genome was translated in all six reading frames using a purpose written Perl script. All homologous sequences from human, mouse and

dog containing the signature six cysteine conserved motif were extracted and aligned using the T-Coffee multiple sequence alignment program (NOTREDAME *et al.* 2000) and subsequently used in the construction of the HMM by the hmmbuild program in HMMER 2.1.1 (http://hmer.wustl.edu) (Eddy 1998). The generated HMM profile was then searched (hmmsearch) against the translated genome to identify putative β-defensin like regions. In addition, the Ensembl predicted bovine gene set (N=20,118) was also searched for sequences displaying a high degree of similarity to mammalian β-defensins. Because mammalian defensins tend to form clusters any genomic region in which putative bovine defensins were identified was more extensively examined using additional iterative BLAST and HMM searches based on the potential bovine defensin sequences identified in our initial search, as well as their direct 1:1 orthologous sequences from each of the other three mammalian species until no more novel sequences were revealed. All genes were named based on their orthologous in other species as defined by the phylogenetic and syntenic analysis. If two identical bovine genes were identified they were both named based on their most likely ortholog with the second gene differentiated with the suffix A (for example BBD109 and BBD109A). If two or more non-identical genes were identified in the bovine genome which had a higher similarity to each other than to any β-defensin gene in other mammals, they were designated with the suffix "Like" - representing genes which have duplicated and subsequently diverged in the bovine lineage.

Chromosomal location and strand orientation of the identified β-defensins were determined using the BLAST Like Alignment Tool (BLAT) at the UCSC genome browser (http://genome.ucsc.edu) (KENT 2002). Genomic DNA corresponding to putative defensins was retrieved using BLAT and a 15kb 5' flanking region was used for prediction of corresponding first exons for each predicted mature peptide region as well as definition of accurate intron/exon boundaries using both GenScan (http://genes.mit.edu/GENSCAN.html) (BURGE and KARLIN 1997) and Genewise2 (BIRNEY *et al.* 2004). The complete repertoire of bovine β-defensin genes was further analysed by alignment with

homologous sequences from other vertebrate species using the T-Coffee multiple sequence alignment program (NOTREDAME *et al.* 2000), while neighbour-joining phylogenetic analysis of the proteins was carried out using MEGA v.4.0 (KUMAR *et al.* 2001).

4.3 Results

Prior to this study, 18 putative bovine β-defensins had been identified through a combination of genomic sequence analysis (ROOSEN *et al.* 2004) and direct sequencing of isolated purified proteins from blood neutrophils (SELSTED *et al.* 1993). The completed genomic sequence for Bos taurus provided us the opportunity to reconstruct the full defensin family for this species. An HMM profile was constructed based on the alignment of the characteristic six-cysteine containing motif of known mammalian β-defensins. This motif was used to search the sequenced bovine genome which had been translated in all six reading frames. This search identified four regions in the bovine genome located on chromosomes 8, 13, 23 and 27 which each code for multiple putative β-defensin-like motifs. Genome-wide homology searches of several mammalian genomes had previously identified four syntenic β-defensins clusters in dog, mouse, rat, opossum and platypus with the human and chimpanzee genomes also containing a fifth chromosomal region with β-defensins like genes (PATIL *et al.* 2005; SCHUTTE *et al.* 2002). The identification of four potential chromosomal regions in the bovine genome indicated that the overall genomic organization of β-defensin clusters observed in other mammals is most likely also maintained in *Bos taurus*.

The potential β-defensin sequences identified by our initial HMM profile were BLAST searched against known β-defensins in Genbank and this showed that the bovine chromosome 27 cluster is orthologous to the established mammalian syntenic cluster A, while the regions identified on bovine chromosomes 8, 23, and 13 correspond to the conserved mammalian clusters B, C and D respectively. Sequence analysis of β-defensins discovered in other species has revealed that genes within a particular β-defensin cluster tend to be more

similar to each other than to those in other clusters within the same species (SEMPLE *et al.* 2006). In order to identify every potential β-defensin gene within each of these regions in the bovine genome, cluster-specific HMMs were constructed for each one based on the amino acid sequences of the orthologous mouse, rat, human and dog genes previously identified as belonging to that cluster (PATIL *et al.* 2005). Searches of the appropriate genomic location using these HMMs as well as BLAST searches using the relevant β-defensins protein sequences from dog, human and mouse led to the discovery of additional cysteine containing motifs within each of the predicted bovine clusters, which had not been identified by the initial, genome wide searches.

In total, 78 open reading frames (ORFs) bearing a similarity to the characteristic six-cysteine-containing domain of the β-defensin family of genes were identified in our comprehensive bioinformatic search of the sequenced bovine genome. Of these 6 were located on ChrUN – a pseudo chromosome created for sequences which have not yet been accurately mapped to their correct genomic location while 43, 5, 4 and 20 were mapped to clusters A, B, C and D respectively. To identify first exons corresponding to each of these second exon sequences, the 5′ upstream genomic region (10,000bp) was analysed using the gene prediction software GenScan, as well as homology-based searches using full-length orthologous β-defensin sequences from other species. These methods established possible corresponding first exons for 64 of the 78 β-defensins second exons predicted in the bovine genome. For several of the remaining ORFs an incomplete or prematurely terminated first exon was identified, though it was not possible to determine whether these sequences represent true pseudogenised β-defensin genes or are simply the result of inaccurate base calling during the genome sequencing project. For a number of other ORFs corresponding to predicted β-defensin six-cysteine containing motifs, gaps in the sequence of the upstream region or the short nature of the assembled contig are most likely reasons for lack of an associated first exon.

The general genomic structure of mammalian β-defensin genes consists of two exons and separated by a single intron of variable length. A single exception, β-defensin 105 possesses a short protein-coding third exon (PAZGIER *et al.* 2006). In all other mammalian species investigated the first exon codes for a hydrophilic signal peptide whilst the second exon codes for the mature peptide with a short preceding anionic pro-peptide (LEHRER and GANZ 2002). Likewise all of the predicted full-length bovine β-defensin genes identified contain a precursor peptide coded for by the first exon. These signal peptides range in length from 18-24 amino acids and have an abundance of hydrophobic residues, particularly leucine. In the predicted bovine gene set, the mature peptide consisting of a pro-sequence and the six-cysteine defensin motif ranges in length from 35-42 amino acids in length. For all the full length β-defensin genes identified in this analysis, the intron-exon boundaries and the length of the intervening intron between exon one and exon two was consistent with those previously estimated for orthologous genes in dog and human (PATIL *et al.* 2005).

The short sequence length and the enormous level of sequence divergence observed amongst the β-defensin genes in mammalian species means that an accurate and comprehensive phylogeny of the entire family is difficult to reconstruct although in general, genes within syntenic clusters tend to form separate clades in phylogenetic trees constructed for each mammalian species (PATIL *et al.* 2005). As a result of this, the relationships among β-defensin genes can only be reliably investigated within each of the individual syntenic clusters described above, whereby lineage specific duplications and deletions of cluster members can be detected.

4.3.1 Syntenic Cluster A

At least one β-Defensin gene has been identified in all the vertebrate genomes so far sequenced indicating that the evolution of the accepted β-defensin gene family predates the divergence of fish from tetrapods (ZOU *et al.* 2007). Subsequent species-and lineage-specific tandem duplication of genes has led to

the expanded clusters common to all extant mammalian species. Syntenic cluster A represents the most ancient cluster which has been conserved throughout the evolution of all mammals and is orthologous to the single defensin cluster observed in both chicken and zebra finch bird species (CORMICAN *et al.* 2009; WHITTINGTON *et al.* 2008). Previous bioinformatics studies have extensively mapped this cluster in human, mouse, rat and dog (PATIL *et al.* 2005; SCHUTTE *et al.* 2002) and have shown that while many orthologous genes within this cluster are maintained in the same order and syntenic position in all species, there is considerable variation as to the overall makeup of this β-defensin locus in cross species comparisons.

In the bovine genome, we mapped 30 full-length β-defensin sequences to this cluster, spanning 1.9mb on chromosome 27 (Figure 4.2B). In addition 7 further genes were located on chrUN, and most likely belong to this cluster in bovine due to the high degree of sequence similarity between them and the 30 genes already mapped to chromosome 27. By comparison, this syntenic cluster in dog (the species most closely related to bovine for which a complete genome draft is available) contains only 15 genes and is considerably more compact, spanning 356kb on chromosome 16 (PATIL *et al.* 2005). Humans carry 11 β-defensin genes in this cluster, which also includes a single α-defensin cluster on chromosome 8 (span 639kb). Like bovine, both mouse and rat possess an expanded cluster in comparison to the dog and human genomes. The mouse cluster, located on chromosome 8 spans 1118kb and codes for 29 full length β-defensin sequences as well as an α-defensin cluster while rat codes for 24 β-defensin sequences and an α-defensin cluster in an 812kb span of chromosome 16 (PATIL *et al.* 2005).

Comparisons between these four previously characterised species and the results of our search of the bovine genome indicate that 8 genes (β-defensin 1, 103, 104, 105, 106, 107, SPAG11c and SPAG11e (nomenclature based on human sequences) are conserved across all species (Figure 1 and Figure 2). To determine the likely origin of the other paralogous genes in this cluster we

constructed a neighbor-joining tree of all cluster members from human, dog, mouse and cow (Figure 4.1A). As predicted from our syntenic mapping of this cluster, the 8 orthologs conserved across all species form independent strongly supported clades in our tree indicating that these β-defensin lineages arose prior to the last common ancestor of all eutherian mammals. In all cases the branching pattern in these subtrees agrees with the established phylogenetic relationships of artiodactyls, carnivores rodents and primates (MURPHY *et al.* 2001). Aside from of these conserved cluster members, there exist several species- or clade-specific subsets of genes. In the bovine genome we identified 16 genes which form a well supported bovine specific clade in our tree (Figure 4.1). The high degree of sequence similarity among these 16 genes and the short branch lengths linking them in our phylogenetic tree would indicate a relatively recent point of origin for this sub-group of β-defensins. The absence of orthologs for any of these genes in the canine genome would indicate that these genes are most likely artiodactyl specific and arose through a series of gene duplication events subsequent to the divergence point separating carnivore and artiodactyls. Most of the 18 bovine β-defensin genes identified prior to this study were members of this separate bovine subgroup. This cluster had only been extensively identified in the bovine lineage, though individual members of this clade have been reported in other artiodactyls including sheep (Huttner et al. 1998), reindeer (UniProt Q0MR48), water buffalo (UniProt A3RJ36) and goat (UniProt Q0PGY0).

The bovine genome also codes for 7 genes which form a strongly supported subclade with the genes coding for β-defensin 1 (Figure 4.1). None of the 7 β-defensin 1 like genes are located adjacent to each other on the bovine chromosome 27 indicating that they are unlikely to have arisen though tandem duplication of an existing gene (Figure 4.2B). At this stage, however the almost identical amino acid composition of these β-defensin 1 like genes makes it impossible to determine whether these predicted genes represent 7 different β-defensin 1 derivatives within the bovine genome or represent errors resulting from the assembly process. This question will perhaps be answered though

high throughput sequencing of the entire 1.9mb region in a large number of bovine breeds and related species. In addition the bovine genome also codes for two β-defensin 109 genes (BBD109 and BBD109A) (Figure 4.1). The duplication and subsequent diversification of these genes in the bovine lineage could be indicative of some species-specific or clade-specific pathogenic challenge against the cow/artiodactyls, which is not a factor for the human lineage. Identification of two distinct, non-identical, β-defensin 109-like genes in clustered porcine EST sets (data not shown), displaying a high degree of similarity to the bovine predictions, indicates that this particular duplication is a feature of all artiodactyls rather than being bovine specific. By comparison, DEFB109 in humans is represented by two identical copies on chromosome 8 as well as three pseudogenised loci (SEMPLE *et al.* 2006). Two genes corresponding to canine β-defensin cbd103 were also identified in this study and denoted BBD103 and BBD103like (Figure 4.1). The cbd103 gene appears to have independently duplicated in the canine lineage giving rise to canine specific cbd102. The level of sequence divergence between the two canine paralogs is far greater than between the two bovine DEFB103 like genes indicating that the bovine duplication event is more recent that the one in the canine lineage.

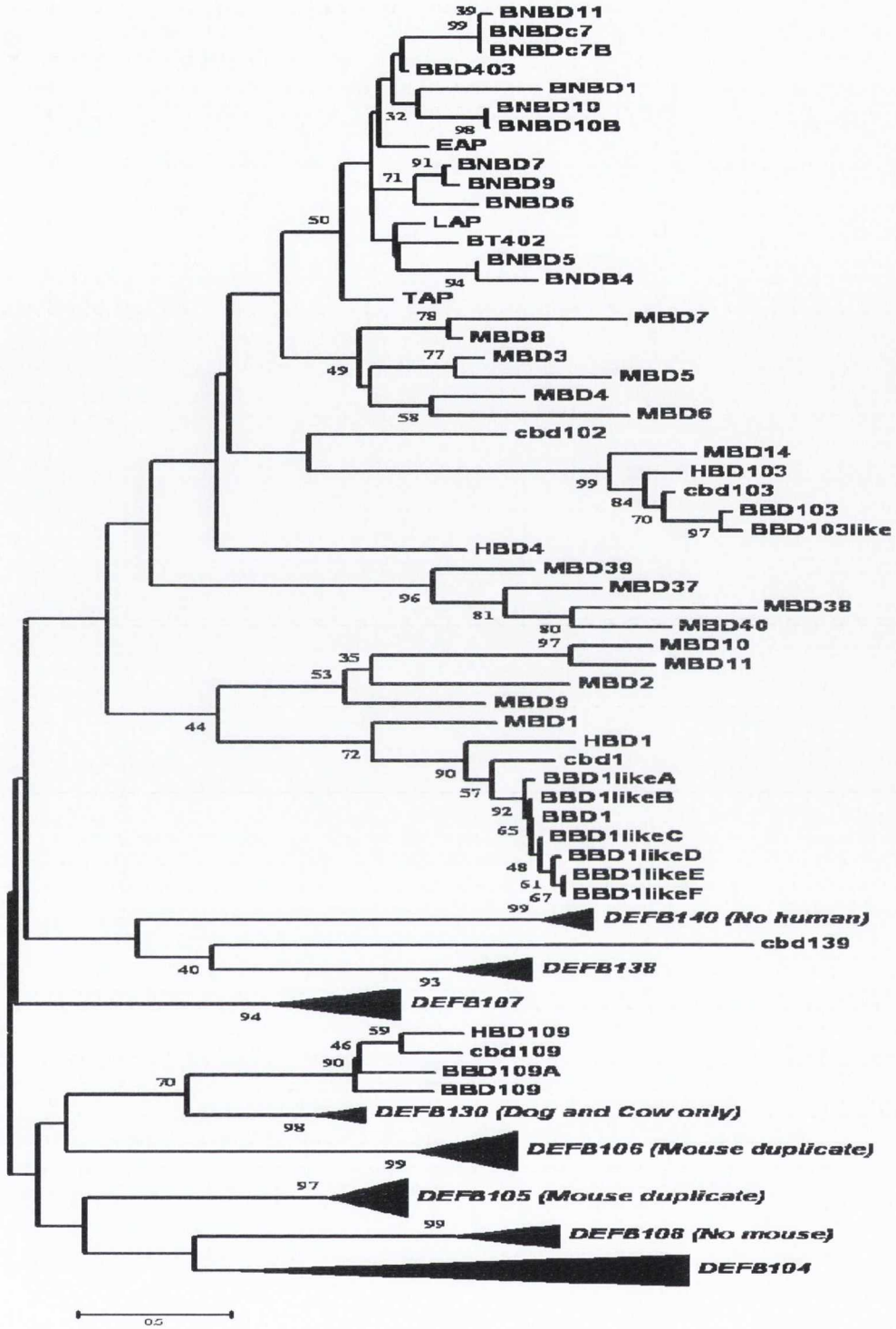


Figure 4-1. Neighbour-joining phylogenetic tree constructed using full-length sequences from the bovine , canine, mouse and human β-defensin repertoires. Branches labelled with *italicised numbers* have been collapsed and represent β-defensin families where a human, mouse, canine and bovine orthologs have been identified and clustered together in the tree with strong bootstrap support

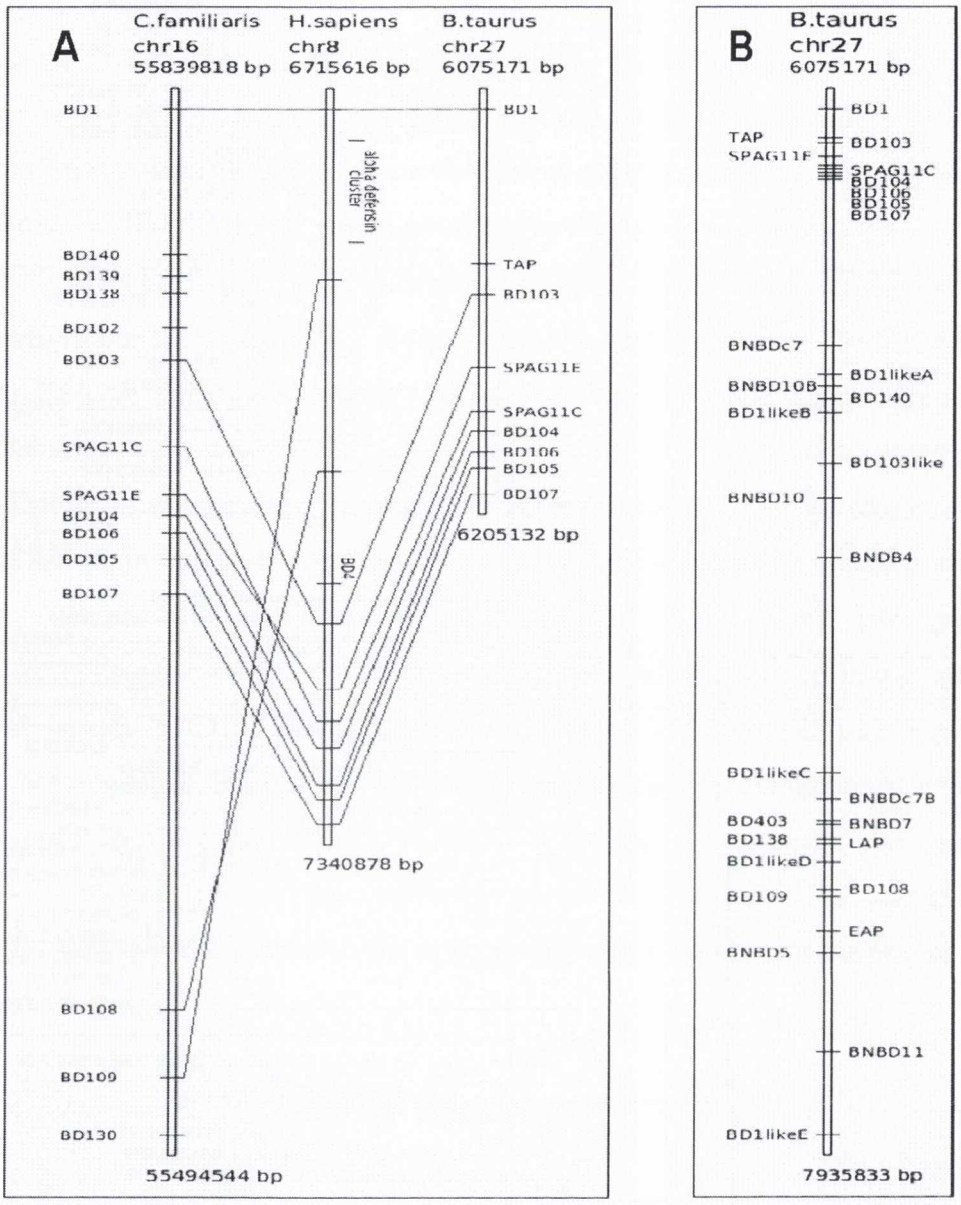


Figure 4-2. A. **Syntenic map to scale of bovine, human and canine β-defensin clusters mapping to bovine chromosome 27 showing the portion of the bovine cluster containing genes in a direct 1:1 relationship with those in other species. The extended bovine cluster has been cropped for ease of viewing.** B. **Map of entire 1.9mb cluster locus on chromosome 27 in cow showing the order and clustering of the novel bovine β-defensin genes.**

4.3.2 Syntenic Clusters B and C

The conserved mammalian syntenic clusters B and C arose as a single genomic locus early in mammalian evolution. While not observed in platypus, the single extant monotreme species for which a genome sequence is available (WHITTINGTON *et al.* 2008), the metatherian opossum genome carries a single cluster, with genes apparently orthologous to members of both mammalian cluster B and C (BELOV *et al.* 2007). This cluster later split into two, subsequent to the divergence of metatherians and eutherians, approximately 180 million years ago (BELOV *et al.* 2007). In contrast to the other two mammalian clusters, the regions encompassing mammalian clusters B and C show very little evidence of either gene gain or gene loss events in any of the eutherian mammals for which the defensin repertoire has been characterised (PATIL *et al.* 2005) (Figure 4.3). In bovine, cluster B consists of BBD131 and BBD134-136, orthologous genes to the observed members of this cluster in human and dog (PATIL *et al.* 2005) and spans 93kb of chromosome 8. Syntenic cluster C in bovine spans 51kb and consists of BBD110-BBD114 with the apparent loss in rodents of the gene orthologous to DEFB112 the only observed change in gene content of this genomic region in the species investigated prior to this study (PATIL *et al.* 2005). In bovine no evidence of a gene was detected orthologous to human HBD133, which is located downstream of DEFB114 in human and MBD49 in mouse (Figure 4.4). This gene appears to have been pseudogenised in the last common ancestor of both the canine and bovine lineages.

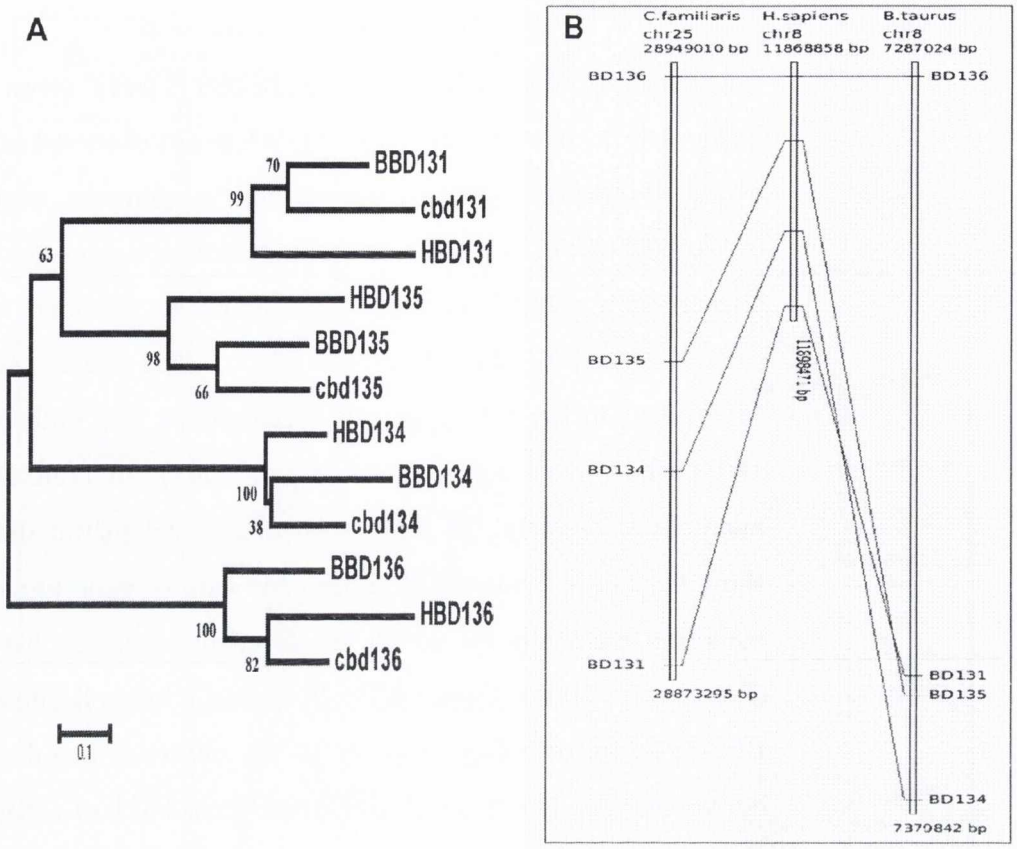


Figure 4-3. A Neighbour joining phylogenetic tree constructed using full length sequences from the bovine, canine and human β-defensin repertoires. All human and canine defensin sequences are available at http://www.ncbi.nlm.nih.gov/GenPept. B Syntenic map of bovine, human and canine β-defensin clusters mapping to bovine chromosome 8.

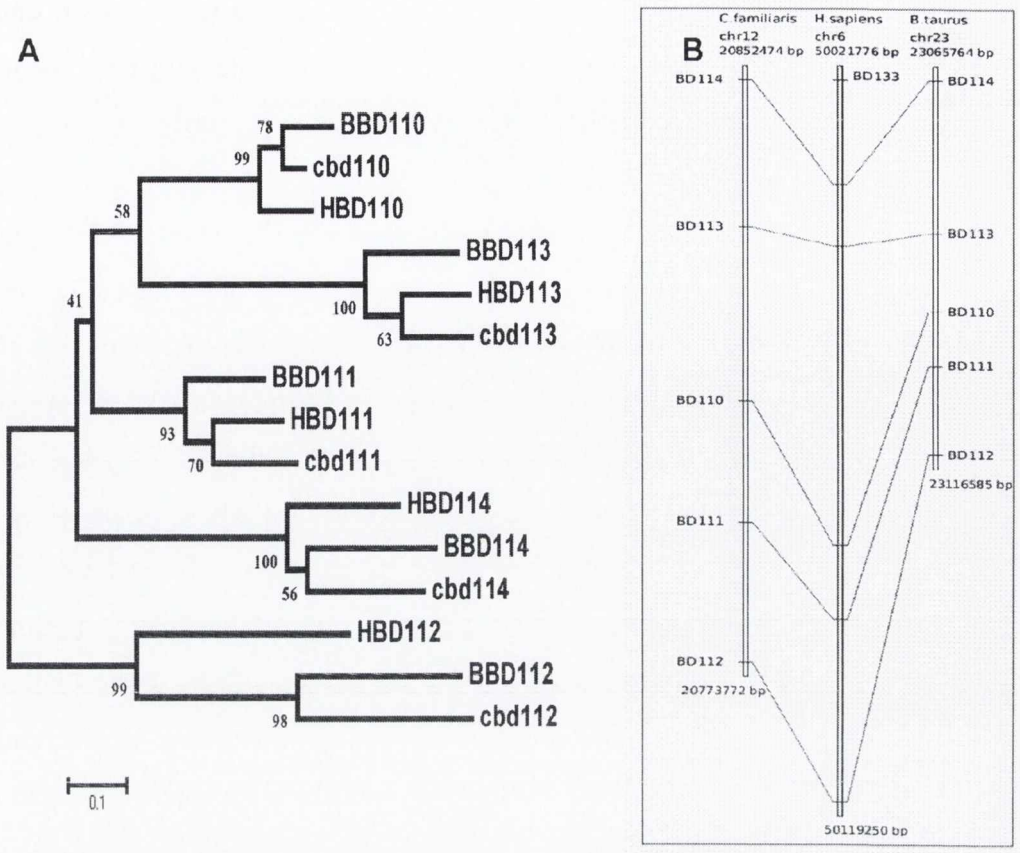


Figure 4-4. A Neighbour joining phylogenetic tree constructed using full length sequences from the bovine, canine and human β-defensin repertoires. B Syntenic map of bovine, human and canine β-defensin clusters mapping to bovine chromosome 23.

4.3.3 Syntenic cluster D

Analysis of the platypus and opossum genomes suggests that syntenic cluster D was the second mammalian β-defensin cluster to be established. A single gene – denoted DEFB6 - was identified in the platypus genome with clear homology to mammalian DEFB116 and DEFB119. In the opossum genome, four genes were identified in a cluster on chromosome 1 which phylogenetic analysis established as orthologs of members of syntenic cluster D from eutherian mammals. Similar analysis carried out in mouse, human and dog has shown that in higher order mammals this cluster is relatively stable with few gene duplication or gene loss events disrupting the syntenic order and gene orientation established early eutherian mammal evolution (RADHAKRISHNAN *et al.* 2007). The only major change observed in this cluster is limited to primate

lineages in which the continuous syntenic cluster observed in all other mammals has split into two separate clusters located in humans on chromosome 20q and 20p (SEMPLE *et al.* 2003).

Our analysis of the bovine genome identified 18 full length and 1 partial β-defensin genes forming bovine syntenic cluster D on a 320kb span of chromosome 13. By comparison the orthologous canine cluster consists of 17 full length and 1 partial genes spanning 393kb of chromosome 24 and the human cluster contains 14 predicted functional genes as well as one partial gene. The partial gene denoted DEFB117, is observed in all mammalian species and consists solely of an exon 2 like sequence coding for a mature cysteine containing region. To date, no credible first exon has been identified in any species though the remarkable level of sequence conservation of this potential beta-defensin across all mammals would suggest that selective pressures have limited its sequence divergence and is thus suggestive of functionality.

Of the remaining 18 β-defensins forming the bovine syntenic cluster D, 12 are conserved in a 1:1:1 orthologous relationship between human, dog and bovine (Figure 4.5B). Interestingly, the bovine genome appears to encode a number of relatively recent gene duplicates within this cluster when compared to the repertoire found in other mammals. BBD122 and 125 are each represented by two non-identical genes in the tree, which bear a closer similarity to each other than to any other defensin in the bovine genome indicting that the gene duplication events leading to these two gene expansions occurred subsequent to the divergence of the carnivore and artiodactyl lineages. Of note is the recent finding that DEFB122 has been pseudogenised in the human lineage through a premature stop codon insertion due to a nonsense mutation (RADHAKRISHNAN *et al.* 2007). In addition the bovine genome codes for an ortholog of DEFB142, a defensin previously only identified in the dog genome.

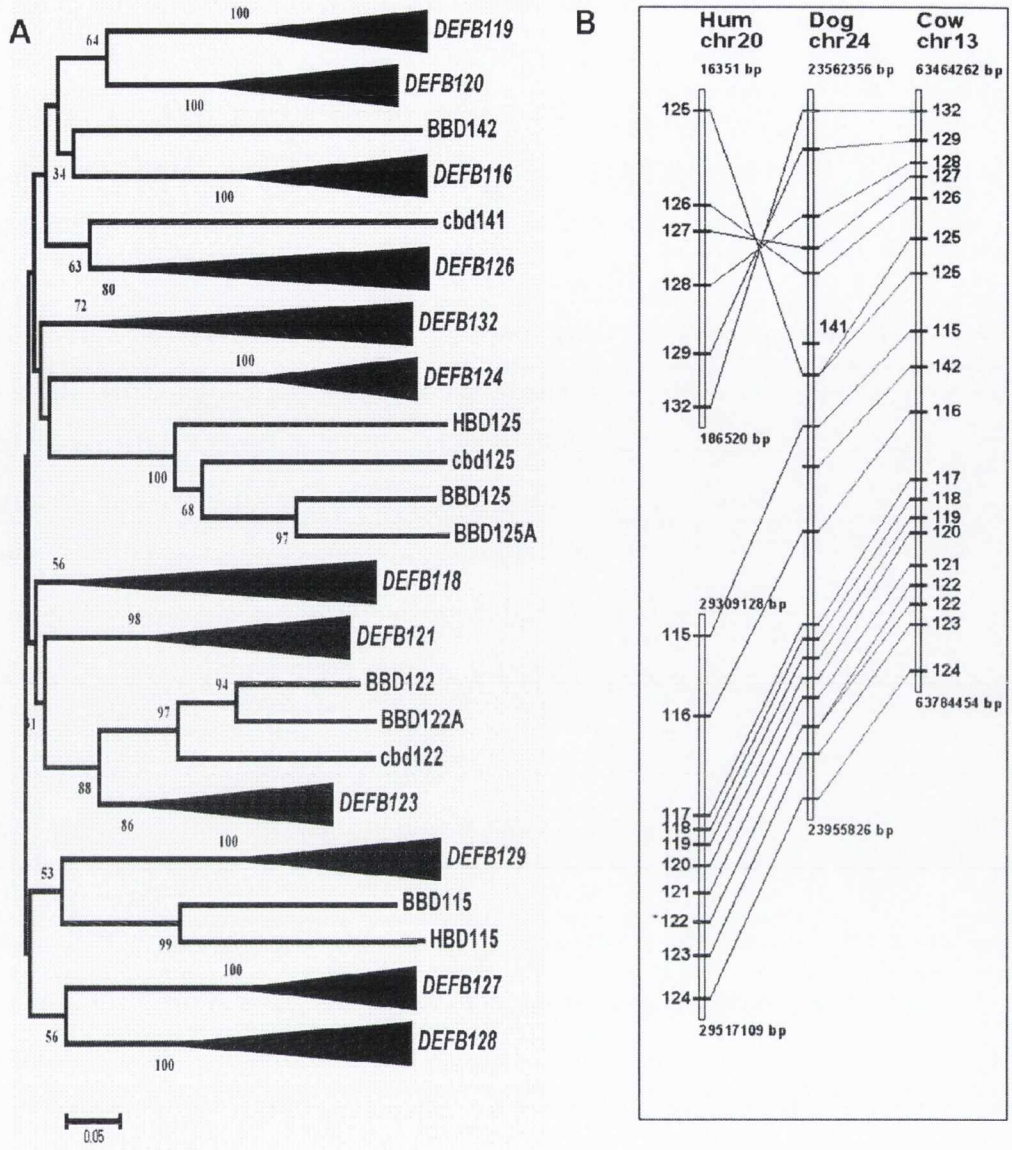


Figure 4-5: **A Neighbour joining phylogenetic tree constructed using full length sequences from the bovine, canine and human β-defensin repertoires. Only bootstrap support values of > 30% are highlighted. Collapsed branches labelled with italicised numbers represent β-defensin families where a human, canine and bovine ortholog have been identified and clustered together in the tree with strong bootstrap support. These clades have been collapsed for ease of viewing. B Synteny map of bovine, human and canine β-defensin clusters mapping to bovine chromosome 13.**

As part of our initial work on characterization of bovine β-defensins it was decided to carry out functional testing of one of the newly discovered bovine genes. BBD123 from syntenic cluster D was selected and synthesized and its functional activity against a pathogen panel that included both gram-positive and gram-negative bacteria was examined. BBD123 was chosen because

similar work had been previously been carried out using the human ortholog (DEFB123) (MOTZKUS *et al.* 2006), which allowed cross species comparisons of effectiveness. Searches of the bovine EST database returned 23 separate sequences corresponding to this gene providing further support of our in-house PCR based experiments confirming our BBD123 bioinformatic prediction as a true, expressed bovine gene. This work was carried out by Dr. Kieran Meade and Sarah Cahalane in St. Vincent's Hospital, Dublin and established that BBD123 is primarily expressed in the bovine testis, but also at lower levels in the uterus, mammary cells, spleen and small intestine. This wide expression in tissues exposed to potential pathogens suggested an important defence role for BBD123 (CORMICAN *et al.* 2008). The pathogen panel tested included *Escherichia coli* and *Staphylococcus aureus,* two common causes of mastitis in cattle, and *Salmonella typhimurium* and *Listeria monocytogenes* which can cause gastro-intestinal problems in both animals and man.

As a follow-on study, it was decided to analyse all the members of the bovine chr13 syntenic cluster D for evidence of sites that have evolved under positive selection. As mammals evolved to occupy new niches, they were faced with a new range of microbial pathogens. Evolution of antimicrobial peptides, with new sensitivities, capable of targeting novel infectious agents would have conferred a selective advantage. Positive selection is necessary for the development of such novel functions. As β-defensins are a critical component of the innate immune response in an 'arms race' against fast evolving microbes it is possible that these molecules are subject to positive selection.

4.3.4 Positive Selection Analysis

Of the 48 amino acids that make up the ungapped peptide alignment of the syntenic cluster D family members in cow, 14 sites were predicted to be subject to positive selection (PS) across the gene phylogeny in our analysis (MEADE *et al.* 2008). Nine of these PS sites are located in the peptide region containing the β-sheet loop region around the third and fourth cysteine, characteristic of the β-defensin structure. Another three amino acids located in

the tail region (C terminus) of each peptide were also found to be PS sites. No PS sites were detected in the signal peptide, indicating that positive selection only occurs in the functionally active antimicrobial region of the peptide (Figure 4.6). A similar analysis carried out using the human syntenic cluster D members returned almost identical results (data not shown).

The results of this positive selection analysis were used to make informed modifications of BBD123 at sites that were most relevant during the functional evolution of the syntenic cluster D members. The sequence modifications made were in relation to charge and hydrophobicity and hydrophillicity in the tail region – 3 properties associated with direct antimicrobial activity which can significantly affect the specificity and antimicrobial activity of a peptide. The newly modified BBD123 sequence was then tested to determine the effect that these functional parameters on antimicrobial activity. We found that targeted changes informed by positive selection appear more likely to influence pathogen specificity rather than direct antimicrobial activity against any particular bacteria (MEADE *et al.* 2008). In addition we found that all four peptide modifications made at positively selected sites with relation to charge, hydrophobicity mad hydrophilicity increased antimicrobial activity against methicillin-resistant *Staphylococcus aureus* (MRSA) compared with the native form of bovine β-defensin 123. This work harnesses the combined power of new technologies in computational and molecular biology to design improved modified antimicrobial agents from already effective native peptides and provides direction for follow-on studies to identify peptide modifications likely to improve efficacy against targeted pathogens.

Figure 4-6. Multiple sequence alignment of peptide sequence for all bovine β-defensin members located in cluster D on chromosome 13. Signature cysteine residues are shaded red, with positively selected sites shaded yellow. Sites with > 50% gaps were removed from the alignment.

4.4 Conclusions

The discovery of novel antimicrobials is a welcome development in light of the increasing prevalence of antibiotic resistance. AMPs are among the most ancient components of the immune system (SELSTED and OUELLETTE 2005) but their extensive role in mammalian defense has only recently become apparent. We have used the recently released bovine genome sequence to identify novel β-defensin family members in cattle, increasing the count to 65 putative β-defensin genes encoded for in *Bos taurus*. This is the largest expansion of this gene family identified in any mammalian species and supports the use of complementary search strategies in complete quantification of this gene family in a species. Our phylogenetic and syntenic positioning comparisons of the novel bovine defensins with the repertoire encoded in other eutherian mammals indicates the discontinuous nature of the β-defensin catalogue many species, where individual defensin genes can be either gained or lost in a particular species, possibly due to selective pressures exerted on that lineage. The diversity observed in mammals as the make up of the β-defensin repertoires mirrors the variable nature of this gene set we previously

observed for avian species (See chapter 3). In the process of domestication, cattle were subjected to much higher population densities than wild bovids. This would have exposed them to a greater abundance and diversity of microbes. This may explain why the bovine genome appears to host the highest number of β-defensins of any mammal and is also a plausible explanation for the positive selection signatures reported in the bovine specific β-defensins (LUENSER and LUDWIG 2005).

In this study we also detected positive selection at several amino acid sites located in the mature antimicrobial region of the genes comprising the youngest mammalian syntenic cluster - denoted cluster D. The expression of β-defensins of mammalian syntenic cluster D have previously been shown to be predominantly localized to the male reproductive tract (JELINSKY *et al.* 2007) and display a striking differential expression patterns when different regions of the testis and epididymis are examined. Such differential patterns of expression are unlikely to be pathogen driven and are perhaps indicative of a secondary non-antimicrobial function for these peptides and evidence is accumulating as to the importance of several of these proteins in the reproductive process (YUDIN *et al.* 2005; ZHOU *et al.* 2004).

5. Positive Darwinian selection influences the evolution of mammalian CD3 subunits

Abstract

The CD3 family of peptides are fundamental signalling constituents of lymphocyte α/β and γ/δ T-cell receptors involved in recognition of MHC on the surface of APCs. The formation of heterodimers of CD3δ/ε and CD3γ/ε as well as homodimeric CD3ζ contributes to both the expression and activity of functional TCRs. We have used all available mammalian orthologous sequences to perform phylogeny-based tests for each of the CD3 subunits in mammals. By comparing the rate of amino-acid changing mutations with those that affect only synonymous sites, we identify positions on the active peptides that are subject to positive selection: changing more rapidly in evolutionary time than expected under a neutral model. Mapping these sites subject to adaptive evolution, we find that they are exclusively located in the extracellular domain of CD3 γ, δ and ε and not in other parts of those peptides. We also show that these positively selected sites are significantly associated in 3-D space with exposed sites, especially potential glycosylation sites. The alteration of the glycosylation pattern of these peptides by positive selection could influence the specificity of each of the CD3 subunits within the TCR complex as well as providing a means of by which CD3 peptides could evade unwanted interaction with pathogenic bacteria.

5.1 Introduction

Many proteins implicated in the generation of an effective immune response have been shown to have unusually divergent amino acid sequences when compared across mammalian species (HUGHES 2002). One commonly proposed theory is that variability in the primary amino acid sequence of these genes allows the immune system to compete in an "arms race" with often rapidly evolving pathogenic organisms. Long and occasionally acrimonious discussions since the late 1960s have centred on the significance of the roles played by neutral genetic drift and positive Darwinian selection in the evolution of such genes. There is now a tentative consensus that most genetic mutations are deleterious to the organism in which they occur and are culled from the population through negative selection, and that observable genetic variants (substitutions) are predominantly neutral having no effect on the fitness of the organism. However an increasing number of examples particularly in the fields of immunology and reproduction have identified genes where substitutions which could confer a fitness benefit to the organism in which they occur are positively selected and tend to increase in frequency within a population.

The signals of differing selective pressures, both positive and negative have traditionally been studied through comparison of synonymous (silent) nucleotide changes with non synonymous (amino acid replacing) substitutions (YANG and BIELAWSKI 2000). Although it is sometimes desirable to compare non-synonymous changes with rates of change in adjacent intronic sites (2005), a comparison of the rates at which synonymous changes occur at synonymous sites (d_S) and nonsynonymous changes occur at nonsynonymous sites (d_N) within the coding region, gives an indication of the type of selective pressures influencing the evolution of a protein. A d_N/d_S ratio, (herein denoted as ω) > 1 is indicative of an excess of amino acid changing substitutions when compared the rate of silent (neutral) nucleotide changes and is taken as evidence of positive selection. An $\omega < 1$ indicates purifying selection where amino acid changes are not tolerated and are removed from a population, while $\omega = 1$

indicates a protein evolving neutrally. Although the ω value is calculated at the DNA level, the actual selective pressures acting on a protein are exerted when the translated peptide has been folded into its correct 3-dimensional structure. Amino acid residues in a peptide which are either functionally or structurally critical to a protein are often constrained in terms of allowable substitutions by the necessity to maintain enzyme-active sites or a location of protein-protein interactions (CAFFREY *et al.* 2004). A clearer picture as to the biological significance of predicted positively selected sites can be determined for proteins for which the 3D structure has been determined and whose protein-protein interactions have been described.

One such set of genes for which extensive structural and protein interaction data has been determined are the mammalian CD3 family (ARNETT *et al.* 2004; KJER-NIELSEN *et al.* 2004). In vertebrates the α/β and γ/δ T cell receptors (TCRs) mediate interaction of T lymphocytes with Antigen Presenting Cells (APCs) through recognition of antigen loaded peptide MHC (pMHC) expressed on the APC surface (ALARCON *et al.* 2003). The various vertebrate TCR chains are composed of an antigen-binding variable region coded for by clonotypically rearranged gene segments and a membrane proximal constant domain but not a cytoplasmic signalling domain by which to activate intracellular signalling pathways (DAVIS *et al.* 1998). To accomplish this signalling , the TCR dimers form an eight unit multimeric complex with CD3 δ/ε, and CD3 γ/ε heterodimers and the structurally unrelated CD3ζζ homodimers (CALL *et al.* 2002). The three CD3 subunits involved in heterodimer formation – δ, ε and γ consist of a substantial extracellular Ig domain, a cysteine rich stalk region, a transmembrane region and a highly conserved intracellular domain (IRVING and WEISS 1991; RETH 1989). The intracellular domains contain multiple immunoreceptor tyrosine based activation motifs (ITAMs) which are phosphorylated upon activation and subsequently interact with downstream signalling mediators including Zap70 – a key component in the T-Cell pathway leading to the activation of the transcription factor NFκB (LIN and WEISS 2001)

As well as their role in signalling, CD3 subunits have also been implicated in the efficient cell surface expression of the completely assembled TCR complex. Assembly of the TCR complex in the endoplasmic reticulum occurs in a carefully regulated order. Initial events include the assembly of the CD3 γ/ε and δ/ε heterodimers. In the absence of both the CD3γ and CD3δ subunits, misfolding of CD3ε occurs, leading to the formation of inactive homodimers, indicating that the co-translation of the other two CD3 subunits are required for the correct folding of the CD3ε subunit (HUPPA and PLOEGH 1997). Following CD3 dimer formation the TCR-α chain is bound by CD3 δ/ε while CD3 γ/ε interacts with the β chain of the TCR (HALL *et al.* 1991). These interactions allow for formation of intrachain disulphide bonds between the TCR chains and the insertion of the CD3ζ homodimer into the complex (HUPPA and PLOEGH 1997). Once assembled the entire protein complex is trafficked to the T Cell surface.

The three CD3 subunits involved in heterodimer formation – δ, ε and γ are coded for by genes clustered together in a 50-kb part of chromosome 11 in humans (Figure 5.1). The CD3γ and CD3δ genes are separated by less than 2kb but with opposing transcriptional orientation while CD3ε is located a further 22KB closer to the centromere. Non mammalian vertebrates encode for a homolog of CD3ε as well as a single gene – designated CD3γ/δ - which displays similarity to both the CD3γ and CD3δ mammalian genes. The duplication and subsequent divergence of this CD3γ/δ gene is thought to have occurred 250-300mya, subsequent to the last common ancestor of birds and mammals (GOUAILLARD *et al.* 2001). Maximum likelihood (ML) based models which allow for heterogeneity in the ω ratios among sites and lineages within a phylogeny have been successfully applied to identify a signal of positive selection in a variety of diverse proteins (SWANSON *et al.* 2001; YANG *et al.* 2003). For the first time, we present such an analysis of mammalian CD3 subunits for a signature of selection and reveal that all three genes CD3ε, CD3γ and CD3δ contain multiple positively selected sites in the Ig-fold of their

extracellular domains. The location of sites where residue altering mutations far exceed synonymous substitutions in these genes when compared across the mammalian phylogeny suggests that variability in these domains is advantageous in mammals as well as contributing further information regarding the overall 3D structure of the entire TCR-CD3 complex.

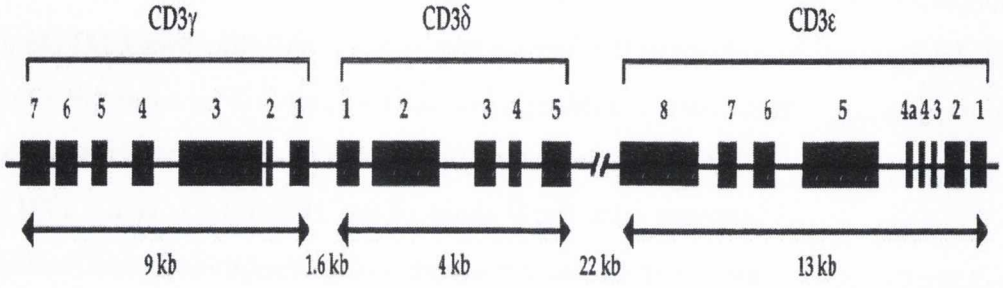


Figure 5-1. Genomic organisation of the human CD3 gene cluster. Figure adapted from Göbel et al. 2000, The Journal of Immunology.

5.2 Materials and methods

All mammalian protein and mRNA full length sequences corresponding to the three CD3 subunits (CD3ε. CD3δ and CD3γ) were retrieved from the Genbank database. Duplicate sequences were removed and the remaining sequences were used in BLAST searches (ALTSCHUL *et al.* 1997) of the all the mammalian sequenced genomes at ENSEMBL (http://www.ensembl.org/), in order to identify further putative orthologs for each CD3 gene (See Appendix). Once identified, all CD3 subunit orthologs were aligned using TCOFFEE (NOTREDAME *et al.* 2000), with the aligned protein sequences serving as a template for alignment of the corresponding DNA sequences. Phylogenetic trees were constructed using MEGA 4.0 (TAMURA *et al.* 2007) and PHYLIP. Alignments were manually adjusted to correct regions of obvious misalignment.

The CODEML program of the PAML suite of software (http://abacus.gene.ucl.ac.uk/software/paml.html) was use to test for variable

selective pressures acting on these genes in mammals. To determine lineage specific variation, a one ratio model, where a single ω ratio was estimated for all branches in the phylogeny was compared to a free ratios model in which the ω ratio for each branch was estimated from the data. Statistical significance of the difference in likelihood between the two models was performed using a Likelihood Ratio test and compared to a $\chi 2$ distribution with degrees of freedoms (df) equal to the number of parameter differences between the two models.

In order to estimate the selective constraints acting on individual codons in the CD3 subunit genes, the CODEML program was again used to estimate the site to site variation in ω. Three comparisons of nested evolutionary models were used to test for selection signatures. M0, a model where one ratio is assumed for all sites is compared with M3, which has discrete classes of sites where the ω and proportions of each class is estimated from the data. . M1 (neutral) assumes two site classes with one value fixed at $\omega = 1$ and is compared with M2a (positive selection), which adds a third site class with ω estimated from the data as a free parameter. M7 (beta) is compared with M8 (beta & ω), with ω in M7 is restricted to values between zero and one by the assumption of a beta distribution B (p, q). The alternative model, M8 has an additional class of sites with proportions and ω estimated from the data. The significance of the differences between the nested models was estimated by likelihood ratio tests which compare twice the difference in the likelihoods of the models to a chi2 distribution with 4 (M0/M3), two (M1/M2a) and two (M7/M8) degrees of freedom (df). When the LRT suggests a proportion of positively selected sites in an alignment, the Bayes empirical Bayes (BEB) approach is used to calculate the probability that each sites belongs in the positively selected site class in the appropriate models – M2a and M8. Sites estimated to have a high probability of belonging to the class of sites with $\omega > 1$ are likely to have evolved under strong diversifying selection pressures.

Three dimensional structures for the human homologs of the CD3 subunits were retrieved from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb/). The PDB entry 1XIW consists of the 3D structure of the extracellular domains of human CD3ε and CD3δ while PDB entry 1SY6 comprises the corresponding region for human CD3ε bound to CD3γ. Human structures were taken as typical of CD3 proteins in mammals. This assumption is validated by comparison with the crystallised structures of these peptides in other species where root mean square deviations (rmsd) for residues in super imposed peptides were < 1.5Å for both CD3ε and CD3γ (KJER-NIELSEN *et al.* 2004). Sites predicted to be positively selected were mapped onto the 3-D structure for each protein using the PYMOL viewer to display PDB structures files. A measure of the degree of clustering (M_s) of positively selected residues on the 3-D structure was estimated by calculating the average inverse distance between all possible pairs of positions for the positively selected amino acid subset (SCHUELER-FURMAN and BAKER 2003). 1000 random datasets of similar size were constructed from the same PDB co-ordinate file. The value of M_s for the positively selected sites compared to that of the 1000 random datasets to indicate whether the positively selected residues are located close to each other on the 3-D structure of each of the three CD3 subunits when compared to random samplings of residues. In order to determine the degree of burial of positively selected sites within the 3-D structure of a protein the number of alpha-carbons within 6.0 Å of each positively selected site was calculated and ranked with a similar analysis of every amino acid in the protein sequence. Residues with higher numbers of alpha-carbons located within 6.0 Å were deemed most likely to be buried in the core of the protein structure.

5.3 Results

Neighbour joining and maximum likelihood trees were constructed for all three CD3 subunits which contain an extracellular Ig-like domain. In the cases of CD3ε, CD3γ and CD3δ, the reconstructed trees were not fully consistent with the previously determined molecular phylogeny for mammals (MURPHY *et al.* 2001). The highly conserved nature of a significant portion of sites in the alignments for all three genes limits the number of informative sites available for accurate reconstruction of the correct phylogeny. All subsequent analysis was carried out using both the previously described species trees and the gene trees generated during our own analysis with no substantive difference in the generated results.

In order to detect evidence of adaptive evolution in the different CD3 subunits we used the maximum likelihood methods implemented in PAML to detect variability in selective pressures acting on amino acid sites in each protein. The likelihood of several null hypothesis models are tested against "nested" models that allow for selection to have occurred in the dataset. The calculated likelihood for each null model was compared by LRT to its appropriate alternative model in order or accept/reject the null hypothesis that there is no positive selection acting on these genes. These tests allow us to determine if there are amino acids sites which are subject to positive selection and/or if there are phylogenetic lineages or branches which are so selected.

To test for variable ω ratios among lineages the one-ratio model (GOLDMAN and YANG 1994), which assumes the same ω ratio for all lineages, was compared using the LRT to the free-ratio model (YANG *et al.* 1994), which estimates an independent ω ratio for each branch in a phylogeny. For all three CD3 subunits the average ω, estimated across all sites and lineages was found to be significantly smaller than one. This would indicate that the dominant selective pressure acting on all three genes in mammalian species is purifying selection, operating to constrain amino acids sites which are perhaps structurally or functionally fundamental to the peptides. The free-ratio to one-

ratio likelihood comparison which averages ω over all sites in a branch from a phylogeny, is sensitive to the presence of a significant portion of highly conserved sites. Similar problems can be encountered when using the more conventional pairwise species ω analysis. Despite this, for both CD3δ and CD3γ as well as CD3ε ,the free ratio model was significantly favoured over the one ratio model. These results are indicative of variable selective pressures between lineages in all three CD3 subunits. For CD3ε, ($2\delta l = 59.62$, $P > 0.05$ df=26) the branches leading to cow and rhesus monkey were estimated to have $\omega > 1$ as well as the branches leading to the split of perisodactyls (horses etc) and carnivores and the Great Ape branch preceding the divergence of the chimp and human lineages (Figure 5.2C). This is tentative evidence that CD3ε in these branches of the mammalian tree has evolved under positive Darwinian selection. For CD3γ, ($2\delta l = 60.62$, $P > 0.05$ df=40), several branches across the cetartiodactyl, perisodactyl, carnivore and primate lineages as well as a single branch in the rodent phylogeny were predicted to have an average $\omega > 1$ and thus display a strong signal of diversifying selection (Figure 5.2A). For CD3δ, ($2\delta l = 56.82$, $P > 0.05$ df=38) a similar pattern of pattern of positively selected branches is observed (Figure 5.2B) . It would indicate that for all three genes, there is a significant difference in the rate in which amino acid substitutions are accumulated in any branch in this mammalian phylogeny.

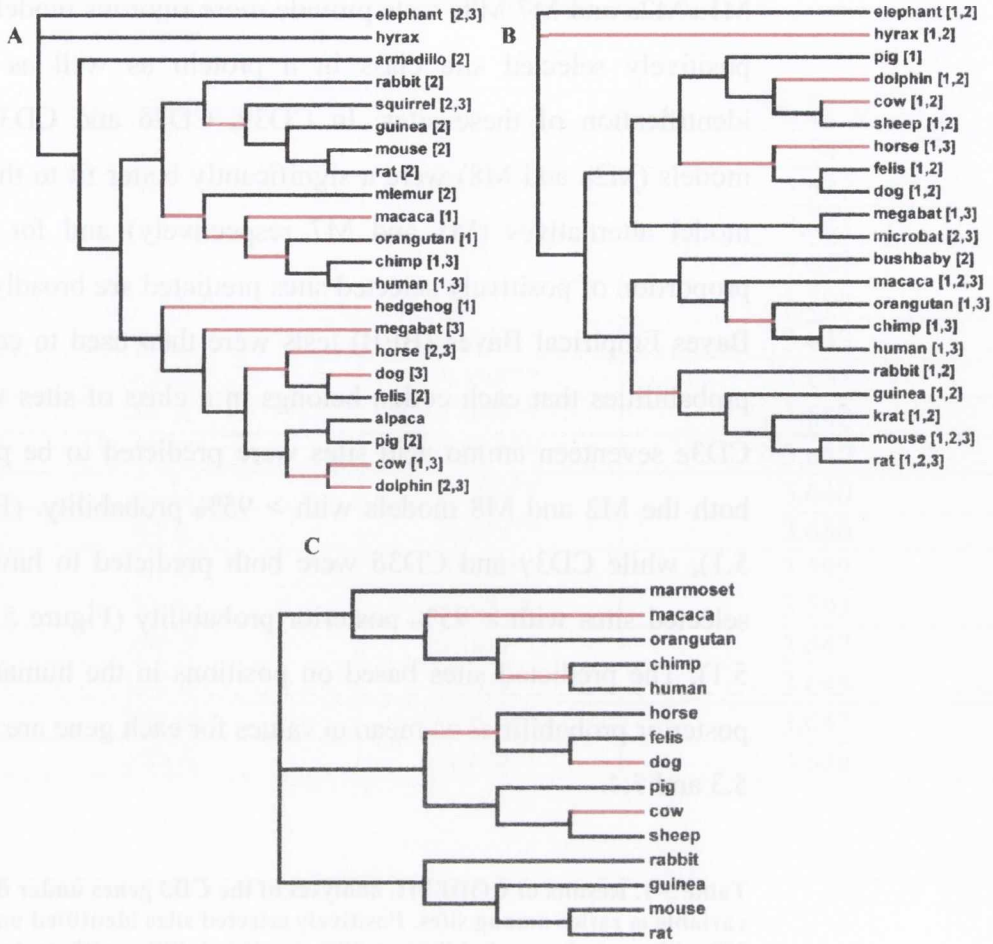


Figure 5-2. A. Trifurcated phylogenetic tree of mammalian species CD3γ sequences used in this analysis, with branches predicted to have an omega > 1 shown in red. The position of the potential glycosylation sites coded for by each species are displayed in square brackets. B. CD3δ. C. CD3ε

Next the models M0 and M3 were compared for all three CD3 subunits. The LRT for all three comparisons is significant (Table 5.1) suggesting that there are non-uniform selective pressures influencing the evolution of some codons in these proteins. The selection model M3, estimates a significant proportion of positively selected sites in each protein. However, while the M0/M3 comparison has been shown to be a reliable test of variability of selective pressures in codons it also appears to be nonconservative with regard to estimation of the proportion of positively selected sites in an alignment (ANISIMOVA *et al.* 2002)

M1a/M2a and M7/M8a tests provide more rigorous models for estimation of a positively selected site class in a protein as well as for the subsequent identification of these sites. In CD3γ, CD3δ and CD3ε the two selection models (M2a and M8) were a significantly better fit to the data that their null model alternatives (M1 and M7 respectively) and for all three genes the proportion of positively selected sites predicted are broadly similar (Table 5.1). Bayes Empirical Bayes (BEB) tests were then used to calculate the posterior probabilities that each codon belongs in a class of sites with omega > 1. For CD3ε seventeen amino acid sites were predicted to be positively selected in both the M2 and M8 models with > 95% probability. (Figure 5.2 and Table 5.1), while CD3γ and CD3δ were both predicted to have thirteen positively selected sites with > 95% posterior probability (Figure 5.2, Figure 5.3, Table 5.1). The predicted sites based on positions in the human protein, as well as posterior probabilities as mean ω values for each gene are shown in Tables 5.2, 5.3 and 5.4.

Table 5-1. Results of CODEML analyses of the CD3 genes under different models of variable ω ratios among sites. Positively selected sites identified under models M2A and M8 with posterior probabilities ≥ 95% are listed. (Sites with posterior probabilities ≥ 99% are shown in bold and underlined). Positively selected sites are numbered based on the human sequences.

	N	Codons	dN/dS	M0vsM3 P-value	M1vsM2 P-value	M7vsM8 P-value	Positively Selected Sites
CD3ε	15	182	0.691	<0.001	<0.001	<0.001	34,44,**53,55**,64,**66,67,68,69,81,83**,86,97,**100,101,102,103**
CD3γ	22	163	0.639	<0.001	<0.001	<0.001	**24,47,48,49**,60,61,**63,68,86**,88,**90,92,93**
CD3δ	21	158	0.596	<0.001	<0.001	<0.001	**25**,29,**42,43,50,51,52,53,**57,**70,72**,75,77

Table 5-2. Positively selected sites in CD3ε with a posterior probability greater than 0.95. (*=>0.95, **=>0.99)

Site (Human)	$P_{(\omega>1)}$	Mean ω
34 T	0.965*	3.545
44 T	0.958*	3.527
53 P	1.000**	3.643
55 S	1.000**	3.643
64 K	0.984*	3.598
66 I	1.000**	3.643
67 G	0.998**	3.637
68 S	1.000**	3.642
69 D	0.997**	3.640
81 H	0.999**	3.640
83 S	0.995**	3.630
86 E	0.982*	3.592
97 V	0.982*	3.593
100 P	1.000**	3.642
101 R	1.000**	3.643
102 G	0.999**	3.641
103 S	0.999**	3.639

Table 5-3 Positively selected sites in CD3γ with a posterior probability greater than 0.95. (*=>0.95, **=>0.99)

Site (Human)	$P_{(\omega>1)}$	Mean ω
24 S	1.000**	2.595
47 D	0.999**	2.589
48 A	1.000**	2.594
49 E	1.000**	2.595
60 K	0.982*	2.558
61 M	0.985*	2.564
63 G	1.000**	2.595
68 D	0.999**	2.593
86 Q	1.000**	2.595
88 K	0.985*	2.594
90 S	1.000**	2.595
92 N	1.000**	2.595
93 K	1.000**	2.595

Table 5-4. Positively selected sites in CD3δ with a posterior probability greater than 0.95. (*=>0.95, **=>0.99)

Site (Human)	$P_{(\omega>1)}$	Mean ω
25 P	1.000**	2.559
29 L	0.967*	2.502
42 T	0.997**	2.554
43 W	0.997**	2.555
50 T	0.994**	2.550
51 L	0.993**	2.548
52 L	1.000**	2.559
53 S	1.000**	2.559
57 R	0.985*	2.533
70 I	0.995**	2.551
72 R	0.999**	2.558
75 G	0.969*	2.503
77 T	0.988*	2.537

Overall the results for all three CD3 subunits provide strong statistical support that positive selection of amino acid substitutions contributes to the evolution of these genes in mammals. By contrast CD3ζ, a TCR component with no substantial extracellular domain shows no evidence of positive selection. In this analysis, sites must have sustained repeated advantageous substitutions throughout the phylogeny in order to be predicted as positively selected and this study is not designed to detect adaptive evolution that acts only along a few lineages or sites (YANG and NIELSEN 2002). It is generally accepted that amino acid sites which are highly conserved between species are either structurally or functionally important (SCHUELER-FURMAN and BAKER 2003). Conserved structural residues tend to cluster together in protein hydrophobic cores (SCHUELER-FURMAN and BAKER 2003), whereas highly conserved residues on an exposed region of a peptide are most likely indicative of enzyme-active sites or a location of protein-protein interactions (CAFFREY *et al.* 2004). In the case of enzymatic peptides or those involved in protein-protein interaction, positive selection promoting the fixation of advantageous mutations in or near the active site is proposed as a means by which substrate specificity or binding affinity can be altered in different species (BARKMAN *et al.* 2007). In order to hypothesize what selective pressures could be influencing

repeated substitutions at the sites predicted to be positively selected we mapped the sites back to the primary peptide sequence and 3-dimensional structure for the three subunits.

In all three CD3 genes only the extracellular domain was found to contain sites which have evolved under positive selection (Figures 5.3, 5.4, 5.5). The stalk, transmembrane and intracellular ITAM domains are highly conserved across the mammalian phylogeny by the action of strong purifying selection. Amino acid substitutions in these regions – particularly the ITAM domains, must severely diminish the ability of the CD3 subunits to function properly. NMR and X-ray crystallography have been used to generate 3-dimensional models of the extracellular regions of dimers of CD3ε in complex with CD3δ as well as CD3γ (Figures 5.6 and 5.7). As can be seen from these figures, CD3ε displays a very different spatial arrangement of positively selected sites compared to CD3δ and CD3γ. For CD3ε, all bar one of positively selected sites are surface exposed residues. None of the positively selected residues occur within the highly conserved region which has previously been described as crucial for the binding to the CD3γ and CD3δ subunits (ARNETT *et al.* 2004; KJER-NIELSEN *et al.* 2004), although one selected site – THR 34 is located in a region where it could directly interact with the dimer partners. A single selected site on the CD3δ peptide (ILE 70) is located in close proximity to this proline residue upon dimer formation (Figure 5.6). The remaining positively selected sites on CD3ε are found on both "faces" of the peptide with several sites predicted on the face proposed to be adjacent to the TCR chains in the completely formed complex. (Figure 5.6).

For both CD3δ and CD3γ the positively selected sites are localised distal to the transmembrane sequence of the molecule in the region proposed to point away from the TCR when the TCR-CD3 complex is formed (SUN *et al.* 2004). All positively selected sites on both CD3δ and CD3γ are surface exposed. Most of the positively selected sites in the two peptides are located in or adjacent to the 3 loops regions – connecting (1) the B-C strand, (2) C-E strands (C-D strands

for CD3δ), and (3) the F-G strands while none of the loop regions located at the "bottom" of the molecule display an excess of amino acid substitutions.

Pairwise cluster analysis of all three CD3 subunits indicate that for both CD3δ and CD3γ positively selected sites are significantly clustered in the proteins tertiary structure with respect with the most membrane proximal amino acids present in the available structures, when compared to random subsets of amino acids of similar size drawn from the same peptide. In contrast, positively selected sites in CD3ε are not significantly clustered in 3-dimensional space on the human CD3ε molecule. The observed significant clustering of selected sites in CD3δ and CD3γ could be indicative of positive selection acting at a particular site of specific activity or peptide binding.

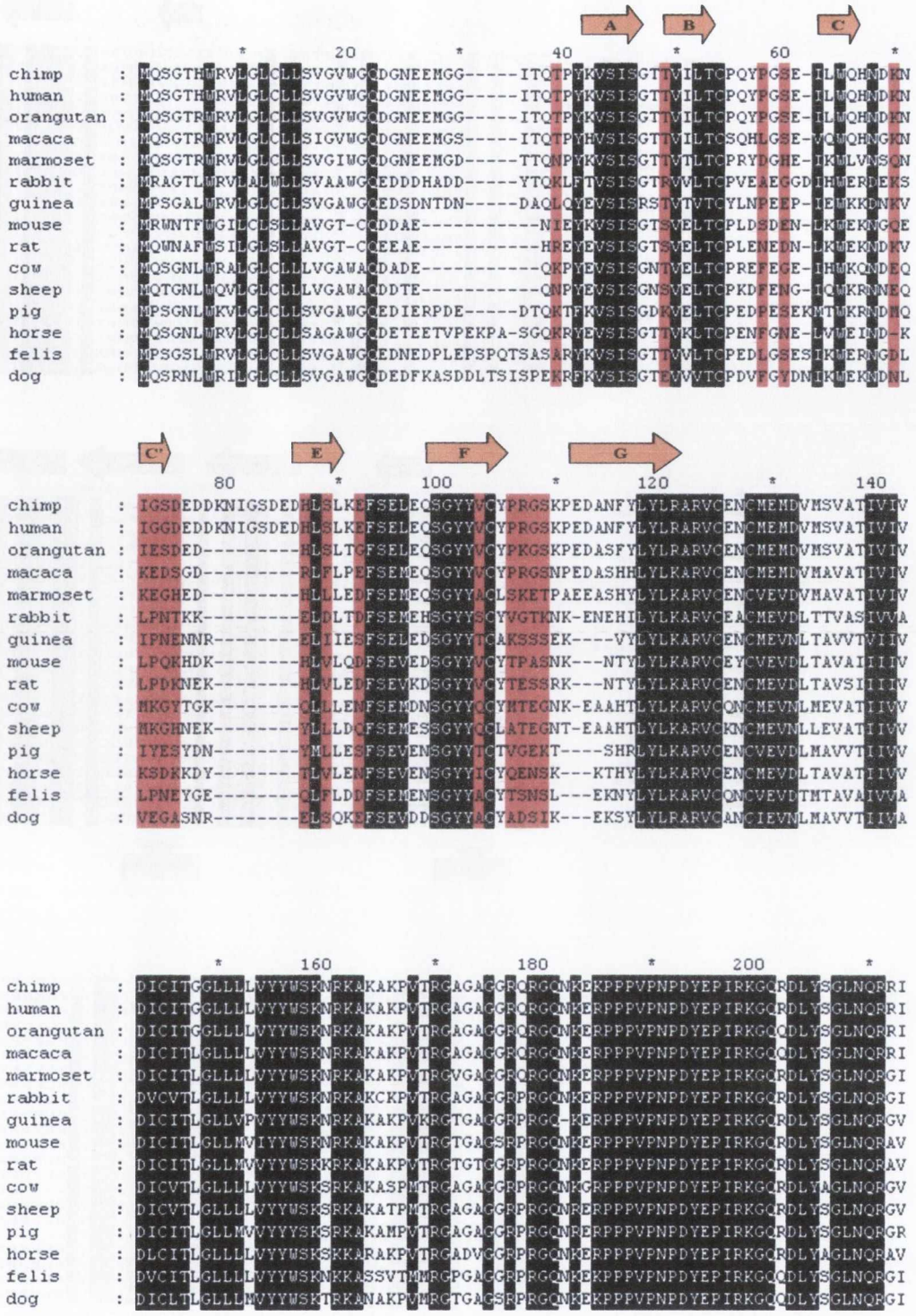


Figure 5-3. Alignment of mammalian full length protein sequences for CD3ε. Positions of Beta-strands deduced from human peptide are shown. Positively selected sites highlighted in red.

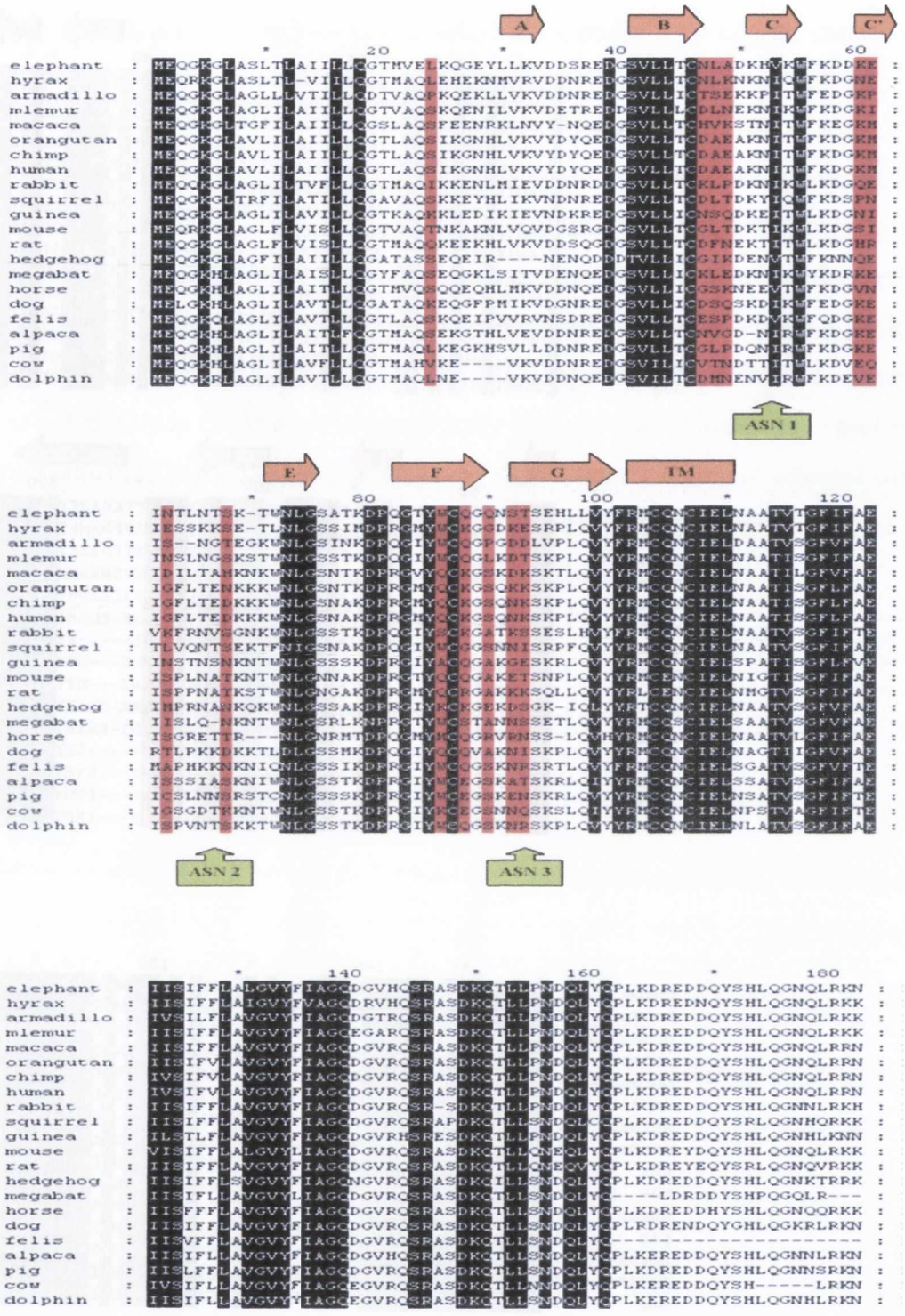


Figure 5-4 .Alignment of mammalian full length protein sequences for CD3γ. Positions of Beta-strands deduced from human peptide are shown. Positively selected sites highlighted in red. The approximate positions of potential glycosylation sites in mammals are highlighted with a green box.

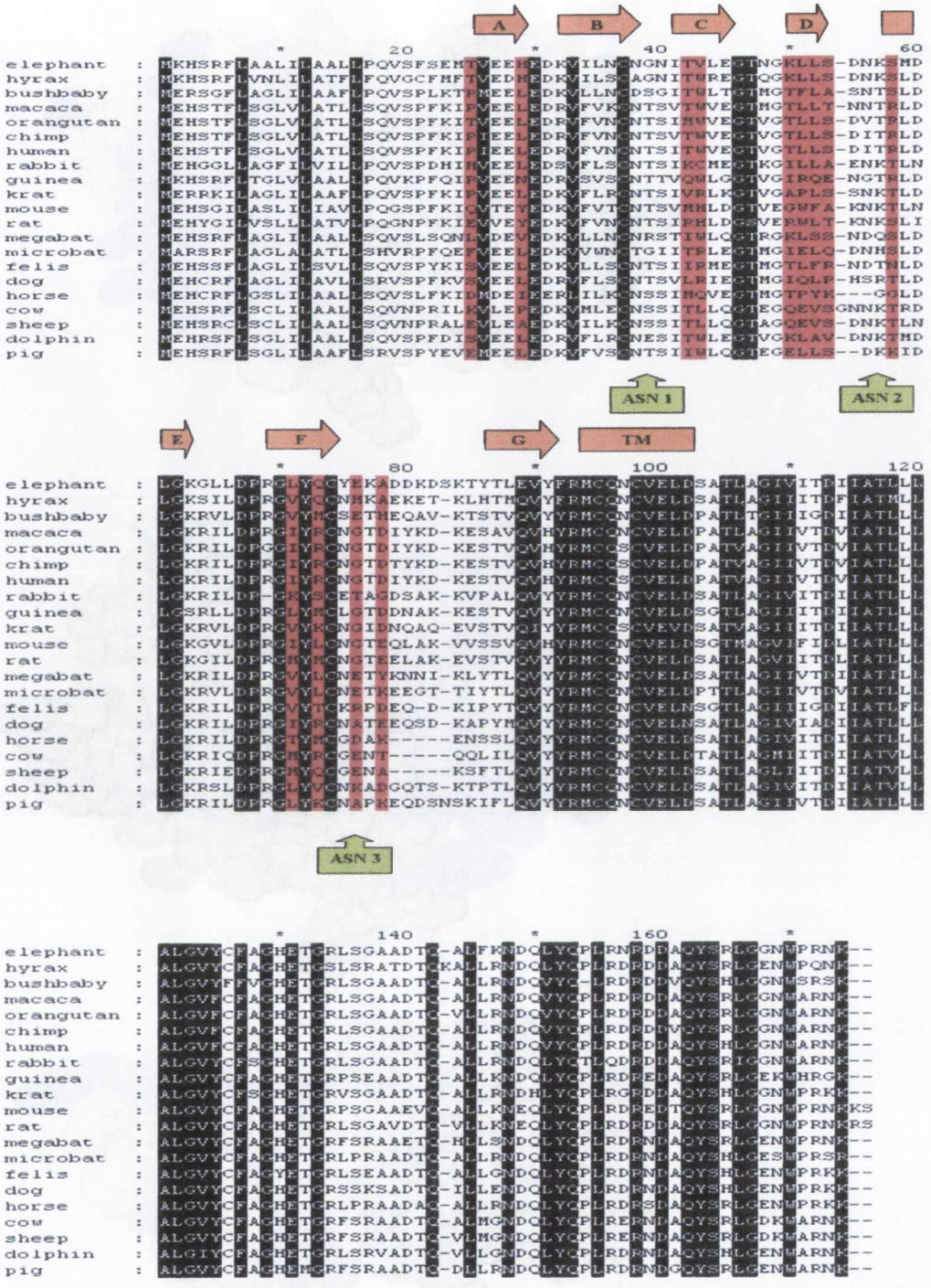


Figure 5-5. Alignment of mammalian full length protein sequences for CD3δ. Positions of Beta-strands deduced from human peptide are shown. Positively selected sites highlighted in red. The approximate positions of potential glycosylation sites in mammals are highlighted with a green box.

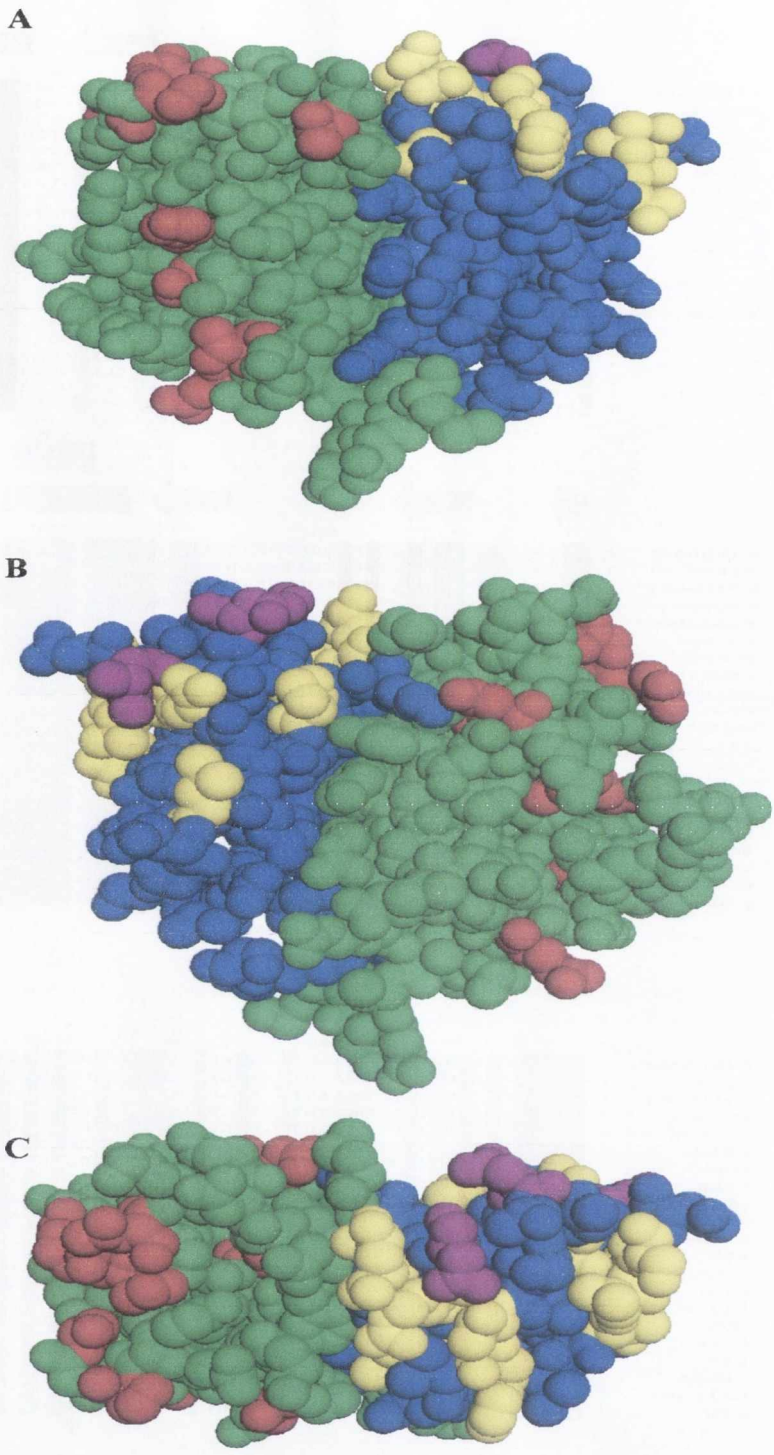


Figure 5-6 A. Human CD3ε and CD3δ dimer complex derived from PDB 1X1W. A – Front view in space-fill format (CD3ε (peptide shown in green) on the left. Positively selected sites in CD3ε are shown as red spheres. CD3δ (peptide shown in blue) on right. Positively selected sites in CD3δ are shown as yellow spheres. The amino acid in each of the three loop regions most likely to encode of a potential glycosylated Asparginine residue are highlighted in purple. B. Reverse view of CD3ε and CD3δ dimer. B. Top down view of CD3ε and CD3δ dimer

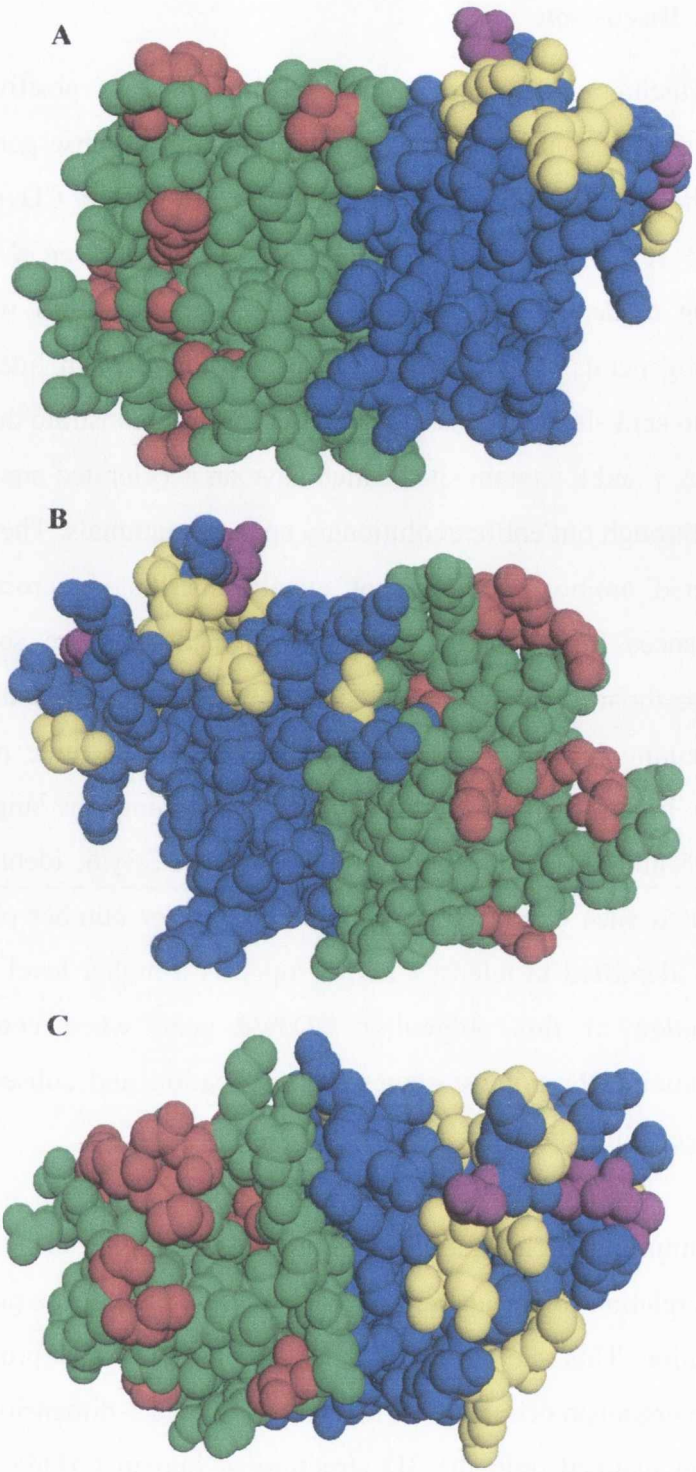


Figure 5-7. A. Human CD3ε and CD3γ dimer complex derived from PDB 1SY6. A – Front view in space-fill format (CD3ε (peptide shown in green) on the left. Positively selected sites in CD3ε are shown as red spheres. CD3γ (peptide shown in blue) on right. Positively selected sites in CD3γ are shown as yellow spheres. The amino acid in each of the three loop regions most likely to encode of a potential glycosylated Asparginine residue are highlighted in purple. B. Reverse view of CD3ε and CD3γ dimer. B. Top down view of CD3ε and CD3γ dimer

5.4 Discussion

Convincing evidence regarding the influence of positive selection on the evolution of many mammalian host immune response genes is accumulating. Of these previously described genes, a number of CD molecules including CD45 (FILIP and MUNDY 2004) and CD2 (LYNN *et al.* 2005) have shown strong evidence of selection. In this analysis, LRTs were used to model heterogeneous selective pressures across mammalian lineages and individual amino acid sites. The results presented here demonstrate that all three subunits, CD3ε, γ and δ contain sites which have an accelerated amino acid replacement rate through out entire evolutionary span of mammals. The predicted positively selected amino acids are not evenly distributed across the CD3 subunit sequences. In all three cases the affected sites are solely located in the extracellular Ig domain whereas the transmembrane and intracellular ITAM containing regions are strongly conserved in all three proteins. A previous study in non-mammalian vertebrates, examining the single gene homolog of CD3δ and CD3γ in teleosts – designated CD3γ/δ, identified five positively selected sites (CRUZ *et al.* 2007). The smaller number of positively selected sites identified in teleost CD3γ/δ suggests a higher level of constraint on the evolution of this progenitor CD3γ/δ gene when compared to the two mammalian homologs suggesting duplication and subsequent divergence of this gene in mammals.

An amino acid change which improves the activity or specificity of a protein with relation to its function is proposed to be fixed in a population by positive selection. This selection for amino acid changes which provide a fitness benefit to an organism occurs at the level of the folded 3-dimensional protein structure. When mapped onto the 3D structure of human CD3δ/ε and CD3γ/ε dimers (Figures 5.6 and 5.7), the locations of the sites which are positively selected and thus have a high rate of amino acid turnover across the mammalian phylogenetic tree provides new insight into the form and function of the CD3 heterodimers in the TCR complex. The exact orientation of the CD3 dimers within the TCR complex has not been completely defined, although the

extracellular domains of CD3 units appear more crucial for CD3 dimer formation than the tetracysteine stalk domain which makes a greater contribution to the interaction of the preformed heterodimers with the TCR chains (CALL *et al.* 2002; CALL and WUCHERPFENNIG 2007). However, the extracellular domains of the CD3 dimers are believed to make contact with the constant domains of both the TCR alpha and TCR beta chains. Structural studies have shown that CD3ε forms an almost identical folded structure when bound to both CD3γ and CD3δ (XU *et al.* 2006) and this would appear to indicate that CD3ε is not the effective binding unit as CD3δ/ε and CD3γ/ε dimers have significantly different affinities for the TCR alpha and TCR beta chain respectively. The more scattered spatial arrangement of the predicted positively selected sites on the surface of CD3ε compared to the other two subunits could indicate that CD3ε is the more solvent exposed part of the dimer structure when bound to the TCR and that the lower regions of both CD3γ and CD3δ which are more highly conserved and contain no predicted positively selected sites are constrained for interaction with the TCR chains. The conserved nature of this region in both CD3γ and CD3δ as well as previous evidence demonstrating physical proximity of CD3γ and the TCR beta chain (BRENNER *et al.* 1985) and the exposure of multiple monoclonal antibody targeted epitopes on CD3ε in the fully formed TCR complex indicates that CD3ε functions more as a scaffold rather than an interface peptide for any extracellular interactions in the TCR-CD3 complex.

5.4.1 Localisation of sites near glycosylation sites

The localisation of the majority of the predicted positively selected amino acids in CD3γ and CD3δ to the three loops segments linking the B-C, C-E and F-G beta-strands or the immediately adjacent strand regions poses interesting questions regarding the nature of the selective pressures acting in the region particularly in the context of the proposed orientation of this region in the fully formed TCR-CD3 complex (SUN *et al.* 2004). One feature of this region in both CD3γ and CD3δ is the presence of a variable number of potential N-linked glycosylation sites in different species. Sugar moieties attached to

glycosylation sites in immune genes have been implicated in such diverse activities as influencing correct folding of the 3D protein structure, control of cell-surface expression of proteins, maintenance of peptides in correct orientation for ligand binding as well as prevention of non specific binding of glycoproteins with similar molecules or with pathogenic peptides (RUDD *et al.* 1999). For CD3γ and CD3δ the contribution of the linked carbohydrates to the activity of the peptides has not been fully elucidated though both peptides have been shown to be co-translationally glycosylated at two separate Asn-X-Ser/Thr triplets (herein referred to as a glucon) in humans (HUPPA and PLOEGH 1997), although it appears the role played by glycosylation on each subunit may be different. Mutation studies have shown that glycosylation of potential sites on CD3γ, is not necessary for correct assembly and surface expression of TCR and that the correctly folded 3D structure of CD3γ is observed even in the absence of carbohydrate addition (DIETRICH *et al.* 1996). In contrast the glycosylation of CD3δ has been shown to be necessary for binding of this subunit to CD3ε and subsequent assimilation of the dimer into the TCR complex (DIETRICH *et al.* 1996).

In mammalian species, all three of the loops linking the B-C, C-E and F-G beta-strands or the strand regions immediately adjacent to them, are capable of containing potential N-glycosylation sites but the sites present across the mammalian phylogeny are not conserved with respect to their position. Similarly there is considerable variation in numbers of potential N-linked sites that are present in between species (Figures 5.2). Even closely related species show remarkable differences with regard to their potential glycosylation sites (Figures 5.2). For example, in CD3γ the bovine gene contains two potential glycosylation sites – indicated as ASN 1 and ASN 3 from Figure 5.4. By contrast pig, the most closely related species to bovine in our analysis codes for a single potential glucon (ASN 2 from Figure 5.4), which is different to either of the two putative sites in cow. In CD3γ, positively selected sites are predicted within two of the potential sites where a glucon can occur in mammals (ASN 2 and ASN 3 – Figure 5.4). ASN 1 and 2 are separated by

fourteen amino acids in the primary sequence but are situated adjacent to each other on the 3D structure (Figure 5.7). The surrounding region contains six positively selected sites where an excess of amino acids substitutions has taken place in mammals. The third region in which an N-linked glycosylation site can occur in mammals (ASN 2, Figure 5.4) also shows an excess of amino acid substitutions including two of the amino acid sites in the potential N-linked triplet. In CD3δ the first potential glycosylation site is almost universally conserved in the mammals sampled in this analysis. Unlike the other potential regions, none of the amino acids in or surrounding this glycosylation triplet are positively selected. In contrast to CD3γ the necessity of glycosylation for CD3δ function most likely explains the higher degree of conservation observed at the potential N-linked sites in CD3δ compared to CD3γ, particularly at this first loop region (Figures 5.4 and 5.5). As with CD3γ the other two loop regions contain several rapidly evolving sites.

It has been suggested that the lack of conservation of potential glycosylation sites between species implies that the attached glycans do not play a pivotal role in interaction of CD3 subunits to the TCR chains and that the presence of correctly folded CD3 subunits in the absence of glycosylation suggests a minimal role in accurate folding of these peptides (DIETRICH *et al.* 1996). One possible function of the attached glycans on CD3 subunits is mediation of interaction with other peptides on the lymphocyte surface. Functional links between CD3δ and the co-receptors CD4 and CD8 have been reported (DOUCEY *et al.* 2003; VIGNALI *et al.* 1996). Both of these peptides are extensively glycosylated and interaction between sugar moieties on these peptides and CD3δ could be involved in directly linking the peptides or in orientation of the peptides allowing direct interaction between other regions of the proteins.

As one of the key signalling mediators responsible for activation of T lymphocytes, CD3 proteins represent attractive targets for pathogens as inhibition of these peptides could prevent the generation of an immune

response. The lack of monoclonal antibodies targeting the CD3δ and CD3γ when compared to the extensive array of antibodies to CD3ε suggests that very little of the native peptide of the two proteins is exposed when in the TCR complex. The attached glycans may provide a protective shield across the top of the molecule preventing non-specific and host-pathogen protein-protein interactions. High turnover of amino acids in these regions could allow the "movement" of potential glycosylation sites across the top of the molecules, providing a more effective species specific shield to compensate for any minor differences in the 3D structure between mammals. The addition of an N-linked site to a region of a peptide requires substitution of not just amino acids to code for the Asn-X-Ser/Thr triplet but also modification of the amino acids surrounding these sites in order accommodate the enzymes responsible for the linking of carbohydrate to the peptide (SWANSON *et al.* 2001). The localisation of positively selected "hotspots" in the regions in and around the three loop segments at the top of CD3δ and CD3γ could indicate that the requirement for glycosylation site fluidity is driving the high rate of amino acid substitutions occurring in these regions in mammals.

6. Final Discussion and Future Directions.

The dawning of the genomic era has provided immunologists with unique opportunities to answer previously intractable questions relating to the origin, evolution and diversification of vertebrate immune system. From such studies has emerged a picture of the immune proteome in which the majority of functional genes can be classified as concatenations of limited number of motifs and domains with the architecture of such immune gene products representing the linear arrangement of these domains within a polypeptide chain. Genome wide, cross-species comparisons have shown that the development of the complex immune system associated with vertebrate species did not depend on the evolution of novel domains, with contemporary vertebrate immune genes created by assembly of existing domains in new ways or gene-wide duplication and subsequent diversification of already established genes. These two mechanisms have contributed to the incredible diversity observed in the make up of the immune systems in the many vertebrates so far studied, and presumably have allowed for a species-specific sharpening of the defence responses unique to each organism. By cataloging and understanding both the similarities and differences we can begin to unravel the functional complexity that exists in vertebrate immunology. The objective of this thesis was to examine and characterise some of these complexities.

6.1 TLR15

Prior to this thesis, comparative studies had shown that the TLR repertoire carried by each vertebrate species is formed from different combinations of genes derived from the six major TLR gene family sub-divisions. Most species studied possess at least one receptor from each of these sub-divisions and presumably this is sufficient to provide a wide-ranging detection system for most of the pathogens faced by each species. One implication of this is that the common ancestor of all vertebrates possessed an innate immune system with the ability to detect lipoprotein, peptidoglycan, LPS, flagellin, double- or single strand RNA, and CpG DNA as PAMPs prior to the expansion and

establishment of all the present day lineages. Occasionally however, gene-duplication events can give rise to a potentially novel receptor with a ligand specificity different to its ancestral gene. Such a gene could provide a selective advantage to its possessor as it may allow for the identification of a novel PAMP, different from those recognised by other related TLRs in a species or could perhaps assume the role of a receptor that had previously been lost in that lineage. In this thesis we have identified a novel TLR - denoted TLR15 - which appears to be unique to the diapsid lineage. This gene appears to be a new member of the TLR2 family of receptors. In mammals this family has undergone at least two gene-duplication events along the TLR1 lineage. Whilst the clade specific TLR15 can clearly be catagorised as belonging to this TLR2 family, it has sufficiently diverged to make a precise phylogenetic relationship to other diapsid TLRs difficult to resolve. A resolution to this issue may be possible with the sequencing of further closely related genomes. At this point, little is known about this novel TLR in relation to gene expression and transcriptional regulation though ongoing molecular studies within our group will hopefully reveal the mechanism of TLR15 gene regulation. In addition efforts are ongoing to determine the specific ligand recognised by this TLR in order to provide insight into its functional role. As member of the TLR2 family, the most likely ligand for TLR15 is a derivative of the diacyl or triacyl lipopeptides recognised by the other members of the family. These PAMPS range from the yeast derived zymosan to the peptidoglycan associated with the exterior membrane surface of Gram-negative bacteria. The members of the TLR2 family in vertebrates thus provide a extensive protective screen against microbes. The addition of the TLR15 gene to the repertoire of receptors carried by diapsid species may reflect the need for a more specific recognition mechanism to counteract the spectrum of pathogenic challenge faced by this vertebrate subgroup.

6.2 Avian TLR Pathway

The application of a bioinformatics protocol to the sequenced genomes of chicken and zebra finch species has led to the identification of avian orthologs for most of the components involved in TLR pathway signalling in mammals. Overall, we have shown that as a biological system the TLR pathway shows remarkable conservation at both the gene and network level when compared across the 300 million years since the last common ancestor of birds and mammals. In addition we have noted that within the pathway, paralog loss in several multigene families occurred very early in evolution of the avian lineage and very little gene turnover has subsequently occurred in the streamlined TLR signalling pathway even amongst highly divergent bird species. One of the critical findings of this work was the discovery that gene-conversion and positive selection had played a role in the evolution of the TLR2 paralogs in birds. In mammals, TLR2 appears to function as a universal component of functional heterodimers formed with other members of the TLR2 subfamily - TLRs 1 and 6. The formation of such heterodimers serves to increase the ligand-recognition spectrum of these TLRs. The presence of two functional TLR2-like genes in birds is a unique feature when compared to other vertebrates. We have shown that gene conversion events between the two functional TLR2 paralogs in both chicken and zebra finch has served to homogenize the intracellular TIR signalling domain between both genes presumably to allow for the use of the same intracellular TIR domain containing adaptor protein. However, the evolution of the extracellular portion of the both TLR genes is characterised by the presence of a 190 amino acid region encompassing both the expected ligand binding and dimerisation domains. The differences detected in this region could allow for recognition of an alternative pathogen associated ligand or the use of differential dimerisation partner by the two avian TLRs. Overall these results would indicate that in avian species, TLR2 genes are under strong selection pressures for both maintenance and adaptation of function.

6.3 Bovine β -defensins

Through comprehensive, genome-wide screening we have identified a total of 67 β-defensin genes in the sequenced bovine genome, representing by far the largest expansion of this gene family in any mammalian species yet studied. We have shown that for the most part, this large number of bovine β-defensin genes is a result of the presence of 18 related, recent gene duplicates within the established mammalian syntenic cluster A. Preliminary analysis of the draft sequences of both the sheep and pig genomes would indicate the majority of the gene duplication events that resulted in this expanded cluster in cows occurred subsequent to the divergence of each of these species from their last common ancestor. The presence of these apparently bovine or artiodactyl specific genes mirrors the findings of previous studies which showed that unique subgroups of β-defensin genes exist in almost all mammalian lineages examined including rodents, primates and carnivores. Overall our findings in conjunction with these previous studies show that in mammals the evolution of the β-defensin repertoire is both a dynamic and ongoing process. The challenge remains to determine the functional significance of these species-specific differences across a range of mammals including cow. Recent years has seen a broadening of the view as to the physiological functions of mammalian β-defensins. Once perceived exclusively as direct antimicrobial peptides they have now been implicated in a variety of other immune and non-immune functions. For each newly discovered bovine β-defensin genes further molecular studies are required to ascertain both their pattern of expression and functional relevance to the evolving nature of the bovine immunome.

6.4 Mammalian CD3

The adaption of immune defences both within and between species undoubtedly occurs as organisms compete in an "arms race" with pathogens in their local environment. Surface expressed molecules involved in the generation of an immune response represent attractive propositions for undergoing adaptive changes as they must evolve to both recognise microbes which are often highly mutable themselves, and also to evade the actions of

micro-organism which try to block or mimic their actions. We have investigated the CD3 gene family in mammals and the results presented in this thesis demonstrate that all three mammalian CD3 genes have been under strong and sustained positive selection throughout the evolution of modern mammals, representing a time span of approximately 100 million years. In particular, we have noted that a highly localised pattern of positively selected residues in CD3δ and CD3γ, the two mammalian specific CD3 subunits. The physical location of these positively selected amino acids along with the predicted orientation of the CD3 dimer within the TCR complex raises a number of questions as to the selective pressures driving the rapid rate of amino acid turnover in these genes.

In this thesis I have attempted to show that the appropriate employment of suitable bioinformatic techniques can enable us to take an evolutionists view of the inherent diversity in the immunomes of vertebrate species. Approaches such as this have enormous power to help us to characterise the changes accumulated by species in order to adapt to the different immunological challenges that are faced in its ecological niche. In addition these bioinformatic techniques could have practical applications, in particular in guiding the design and production of novel therapeutic agents based on informed modification of existing naturally occurring peptide sequences.

7. Bibliography

2005 Initial sequence of the chimpanzee genome and comparison with the human genome. Nature **437:** 69-87.

ALTSCHUL, S. F., T. L. MADDEN, A. A. SCHAFFER, J. ZHANG, Z. ZHANG *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res **25:** 3389-3402.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2002 Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol **19:** 950-958.

ANISIMOVA, M., and D. A. LIBERLES, 2007 The quest for natural selection in the age of comparative genomics. Heredity **99:** 567-579.

ARNETT, K. L., S. C. HARRISON and D. C. WILEY, 2004 Crystal structure of a human CD3-epsilon/delta dimer in complex with a UCHT1 single-chain antibody fragment. Proc Natl Acad Sci U S A **101:** 16268-16273.

BARKMAN, T. J., T. R. MARTINS, E. SUTTON and J. T. STOUT, 2007 Positive selection for single amino acid change promotes substrate discrimination of a plant volatile-producing enzyme. Mol Biol Evol **24:** 1320-1329.

BARTON, G. M., J. C. KAGAN and R. MEDZHITOV, 2006 Intracellular localization of Toll-like receptor 9 prevents recognition of self DNA but facilitates access to viral DNA. Nat Immunol **7:** 49-56.

BELL, J. K., G. E. MULLEN, C. A. LEIFER, A. MAZZONI, D. R. DAVIES *et al.*, 2003 Leucine-rich repeats and pathogen recognition in Toll-like receptors. Trends Immunol **24:** 528-533.

BELLA, J., K. L. HINDLE, P. A. MCEWAN and S. C. LOVELL, 2008 The leucine-rich repeat structure. Cell Mol Life Sci **65:** 2307-2333.

BELOV, K., C. E. SANDERSON, J. E. DEAKIN, E. S. WONG, D. ASSANGE *et al.*, 2007 Characterization of the opossum immune genome provides insights into the evolution of the mammalian immune system. Genome Res **17:** 982-991.

BELVIN, M. P., and K. V. ANDERSON, 1996 A conserved signaling pathway: the Drosophila toll-dorsal pathway. Annu Rev Cell Dev Biol **12:** 393-416.

BIRNEY, E., M. CLAMP and R. DURBIN, 2004 GeneWise and Genomewise. Genome Res **14:** 988-995.

BOGUSKI, M. S., T. M. LOWE and C. M. TOLSTOSHEV, 1993 dbEST--database for "expressed sequence tags". Nat Genet **4:** 332-333.

BOMAN, H. G., 2003 Antibacterial peptides: basic facts and emerging concepts. J Intern Med **254:** 197-215.

BRENNER, M. B., I. S. TROWBRIDGE and J. L. STROMINGER, 1985 Cross-linking of human T cell receptor proteins: association between the T cell idiotype beta subunit and the T3 glycoprotein heavy subunit. Cell **40:** 183-190.

BROGDEN, K. A., 2005 Antimicrobial peptides: pore formers or metabolic inhibitors in bacteria? Nat Rev Microbiol **3:** 238-250.

BURGE, C., and S. KARLIN, 1997 Prediction of complete gene structures in human genomic DNA. J Mol Biol **268:** 78-94.

CAFFREY, D. R., S. SOMAROO, J. D. HUGHES, J. MINTSERIS and E. S. HUANG, 2004 Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? Protein Sci **13:** 190-202.

CALL, M. E., J. PYRDOL, M. WIEDMANN and K. W. WUCHERPFENNIG, 2002 The organizing principle in the formation of the T cell receptor-CD3 complex. Cell **111:** 967-979.

CALL, M. E., and K. W. WUCHERPFENNIG, 2007 Common themes in the assembly and architecture of activating immune receptors. Nat Rev Immunol **7:** 841-850.

CHOU, A., and J. BURKE, 1999 CRAWview: for viewing splicing variation, gene families, and polymorphism in clusters of ESTs and full-length sequences. Bioinformatics **15:** 376-381.

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. J Biol Chem **274:** 10689-10692.

CHUANG, T. H., J. LEE, L. KLINE, J. C. MATHISON and R. J. ULEVITCH, 2002 Toll-like receptor 9 mediates CpG-DNA signaling. J Leukoc Biol **71:** 538-544.

CORMICAN, P., K. G. MEADE, S. CAHALANE, F. NARCIANDI, A. CHAPWANYA *et al.*, 2008 Evolution, expression and effectiveness in a cluster of novel bovine beta-defensins. Immunogenetics **60:** 147-156.

CRUZ, F., D. G. BRADLEY and D. J. LYNN, 2007 Evidence of positive selection on the Atlantic salmon CD3gammadelta gene. Immunogenetics **59:** 225-232.

DACHEUX, J. L., J. L. GATTI and F. DACHEUX, 2003 Contribution of epididymal secretory proteins for spermatozoa maturation. Microsc Res Tech **61:** 7-17.

DIETRICH, J., A. NEISIG, X. HOU, A. M. WEGENER, M. GAJHEDE *et al.*, 1996 Role of CD3 gamma in T cell receptor assembly. J Cell Biol **132:** 299-310.

DOUCEY, M. A., L. GOFFIN, D. NAEHER, O. MICHIELIN, P. BAUMGARTNER *et al.*, 2003 CD3 delta establishes a functional link between the T cell receptor and CD8. J Biol Chem **278:** 3257-3264.

DOYLE, S. L., and L. A. O'NEILL, 2006 Toll-like receptors: from the discovery of NFkappaB to new insights into transcriptional regulations in innate immunity. Biochem Pharmacol **72:** 1102-1113.

EDDY, S. R., 1998 Profile hidden Markov models. Bioinformatics **14:** 755-763.

FILIP, L. C., and N. I. MUNDY, 2004 Rapid evolution by positive Darwinian selection in the extracellular domain of the abundant lymphocyte protein CD45 in primates. Mol Biol Evol **21:** 1504-1511.

FITCH, W. M., 2000 Homology a personal view on some of the problems. Trends Genet **16:** 227-231.

FITZGERALD, K. A., and L. A. O'NEILL, 2000 The role of the interleukin-1/Toll-like receptor superfamily in inflammation and host defence. Microbes Infect **2:** 933-943.

FRAZER, K. A., L. ELNITSKI, D. M. CHURCH, I. DUBCHAK and R. C. HARDISON, 2003 Cross-species sequence comparisons: a review of methods and available resources. Genome Res **13:** 1-12.

GANZ, T., 1999 Defensins and host defense. Science **286:** 420-421.

GANZ, T., 2003 Defensins: antimicrobial peptides of innate immunity. Nat Rev Immunol **3:** 710-720.

GIBB, G. C., O. KARDAILSKY, R. T. KIMBALL, E. L. BRAUN and D. PENNY, 2007 Mitochondrial genomes and avian phylogeny: complex characters and resolvability without explosive radiations. Mol Biol Evol **24:** 269-280.

GIBBARD, R. J., P. J. MORLEY and N. J. GAY, 2006 Conserved features in the extracellular domain of human toll-like receptor 8 are essential for pH-dependent signaling. J Biol Chem **281:** 27503-27511.

GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736.

GORDEN, K. K., X. X. QIU, C. C. BINSFELD, J. P. VASILAKOS and S. S. ALKAN, 2006 Cutting edge: activation of murine TLR8 by a combination of imidazoquinoline immune response modifiers and polyT oligodeoxynucleotides. J Immunol **177:** 6584-6587.

HANCOCK, R. E., and R. LEHRER, 1998 Cationic peptides: a new source of antibiotics. Trends Biotechnol **16:** 82-88.

HANCOCK, R. E., and H. G. SAHL, 2006 Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol **24:** 1551-1557.

HAZIOT, A., E. FERRERO, F. KONTGEN, N. HIJIYA, S. YAMAMOTO *et al.*, 1996 Resistance to endotoxin shock and reduced dissemination of gram-negative bacteria in CD14-deficient mice. Immunity **4:** 407-414.

HEMMRICH, G., D. J. MILLER and T. C. BOSCH, 2007 The evolution of immunity: a low-life perspective. Trends Immunol **28:** 449-454.

HIGGS, R., P. CORMICAN, S. CAHALANE, B. ALLAN, A. T. LLOYD *et al.*, 2006 Induction of a novel chicken Toll-like receptor following Salmonella enterica serovar Typhimurium infection. Infect Immun **74:** 1692-1698.

HOLLOX, E. J., U. HUFFMEIER, P. L. ZEEUWEN, R. PALLA, J. LASCORZ *et al.*, 2008 Psoriasis is associated with increased beta-defensin genomic copy number. Nat Genet **40:** 23-25.

HOSHINO, K., O. TAKEUCHI, T. KAWAI, H. SANJO, T. OGAWA *et al.*, 1999 Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. J Immunol **162:** 3749-3752.

HUANG, S., S. YUAN, L. GUO, Y. YU, J. LI *et al.*, 2008 Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. Genome Res **18:** 1112-1126.

HUPPA, J. B., and H. L. PLOEGH, 1997 In vitro translation and assembly of a complete T cell receptor-CD3 complex. J Exp Med **186:** 393-403.

IMLER, J. L., and J. A. HOFFMANN, 2001 Toll receptors in innate immunity. Trends Cell Biol **11:** 304-311.

IMLER, J. L., and L. ZHENG, 2004 Biology of Toll receptors: lessons from insects and mammals. J Leukoc Biol **75:** 18-26.

ISHII, A., M. KAWASAKI, M. MATSUMOTO, S. TOCHINAI and T. SEYA, 2007 Phylogenetic and expression analysis of amphibian Xenopus Toll-like receptors. Immunogenetics **59:** 281-293.

JENSEN, R. A., 1976 Enzyme recruitment in evolution of new function. Annu Rev Microbiol **30:** 409-425.

JENSEN, R. A., 2001 Orthologs and paralogs - we need to get it right. Genome Biol **2:** INTERACTIONS1002.

JIN, M. S., S. E. KIM, J. Y. HEO, M. E. LEE, H. M. KIM *et al.*, 2007 Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell **130:** 1071-1082.

JONES, D. T., 1999 Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol **292:** 195-202.

KANEHISA, M., and S. GOTO, 2000 KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res **28:** 27-30.

KELLEY, J., B. DE BONO and J. TROWSDALE, 2005 IRIS: a database surveying known human immune system genes. Genomics **85:** 503-511.

KENT, W. J., C. W. SUGNET, T. S. FUREY, K. M. ROSKIN, T. H. PRINGLE *et al.*, 2002 The human genome browser at UCSC. Genome Res **12:** 996-1006.

KJER-NIELSEN, L., M. A. DUNSTONE, L. KOSTENKO, L. K. ELY, T. BEDDOE *et al.*, 2004 Crystal structure of the human T cell receptor CD3 epsilon gamma heterodimer complexed to the therapeutic mAb OKT3. Proc Natl Acad Sci U S A **101:** 7675-7680.

KRUITHOF, E. K., N. SATTA, J. W. LIU, S. DUNOYER-GEINDRE and R. J. FISH, 2007 Gene conversion limits divergence of mammalian TLR1 and TLR6. BMC Evol Biol **7:** 148.

KUMAR, H., T. KAWAI and S. AKIRA, 2009 Toll-like receptors and innate immunity. Biochem Biophys Res Commun.

KUMAR, S., K. TAMURA, I. B. JAKOBSEN and M. NEI, 2001 MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17:** 1244-1245.

LEHRER, R. I., and T. GANZ, 2002 Defensins of vertebrate animals. Curr Opin Immunol **14:** 96-102.

LEMAITRE, B., and F. M. AUSUBEL, 2008 Animal models for host-pathogen interactions. Curr Opin Microbiol **11:** 249-250.

LEULIER, F., and B. LEMAITRE, 2008 Toll-like receptors--taking an evolutionary approach. Nat Rev Genet **9:** 165-178.

LEULIER, F., A. RODRIGUEZ, R. S. KHUSH, J. M. ABRAMS and B. LEMAITRE, 2000 The Drosophila caspase Dredd is required to resist gram-negative bacterial infection. EMBO Rep **1:** 353-358.

LUENSER, K., and A. LUDWIG, 2005 Variability and evolution of bovine beta-defensin genes. Genes Immun **6:** 115-122.

LYNN, D. J., and D. G. BRADLEY, 2007 Discovery of alpha-defensins in basal mammals. Dev Comp Immunol **31:** 963-967.

LYNN, D. J., A. R. FREEMAN, C. MURRAY and D. G. BRADLEY, 2005 A genomics approach to the detection of positive selection in cattle: adaptive evolution of the T-cell and natural killer cell-surface protein CD2. Genetics **170:** 1189-1196.

LYNN, D. J., A. T. LLOYD and C. O'FARRELLY, 2003 In silico identification of components of the Toll-like receptor (TLR) signaling pathway in clustered chicken expressed sequence tags (ESTs). Vet Immunol Immunopathol **93:** 177-184.

MADERA, M., and J. GOUGH, 2002 A comparison of profile hidden Markov model procedures for remote homology detection. Nucleic Acids Res **30:** 4321-4328.

MAIZELS, N., 2005 Immunoglobulin gene diversification. Annu Rev Genet **39:** 23-46.

MANNING, G., G. D. PLOWMAN, T. HUNTER and S. SUDARSANAM, 2002 Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci **27:** 514-520.

MATSUSHIMA, N., T. TANAKA, P. ENKHBAYAR, T. MIKAMI, M. TAGA *et al.*, 2007 Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. BMC Genomics **8:** 124.

MCGETTRICK, A. F., and L. A. O'NEILL, 2004 The expanding family of MyD88-like adaptors in Toll-like receptor signal transduction. Mol Immunol **41:** 577-582.

MEADE, K. G., S. CAHALANE, F. NARCIANDI, P. CORMICAN, A. T. LLOYD *et al.*, 2008 Directed alteration of a novel bovine beta-defensin to improve antimicrobial efficacy against methicillin-resistant Staphylococcus aureus (MRSA). Int J Antimicrob Agents **32:** 392-397.

MEDZHITOV, R., 2007 Recognition of microorganisms and activation of the immune response. Nature **449:** 819-826.

MILLER, D. J., G. HEMMRICH, E. E. BALL, D. C. HAYWARD, K. KHALTURIN *et al.*, 2007 The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss. Genome Biol **8:** R59.

MINAKHINA, S., and R. STEWARD, 2006 Nuclear factor-kappa B pathways in Drosophila. Oncogene **25:** 6749-6757.

MIZEL, S. B., A. P. WEST and R. R. HANTGAN, 2003 Identification of a sequence in human toll-like receptor 5 required for the binding of Gram-negative flagellin. J Biol Chem **278:** 23624-23629.

MORRISON, G. M., C. A. SEMPLE, F. M. KILANOWSKI, R. E. HILL and J. R. DORIN, 2003 Signal sequence conservation and mature peptide divergence within subgroups of the murine beta-defensin gene family. Mol Biol Evol **20:** 460-470.

MOTZKUS, D., S. SCHULZ-MARONDE, A. HEITLAND, A. SCHULZ, W. G. FORSSMANN *et al.*, 2006 The novel beta-defensin DEFB123 prevents lipopolysaccharide-mediated effects in vitro and in vivo. FASEB J **20:** 1701-1702.

MURPHY, W. J., E. EIZIRIK, S. J. O'BRIEN, O. MADSEN, M. SCALLY *et al.*, 2001 Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294:** 2348-2351.

NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148:** 929-936.

NOTREDAME, C., D. G. HIGGINS and J. HERINGA, 2000 T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol **302:** 205-217.

O'NEILL, L. A., 2008 The interleukin-1 receptor/Toll-like receptor superfamily: 10 years of progress. Immunol Rev **226:** 10-18.

O'NEILL, L. A., and A. G. BOWIE, 2007 The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nat Rev Immunol **7:** 353-364.

O'NEILL, L. A., K. A. FITZGERALD and A. G. BOWIE, 2003 The Toll-IL-1 receptor adaptor family grows to five members. Trends Immunol **24:** 286-290.

OHASHI, K., V. BURKART, S. FLOHE and H. KOLB, 2000 Cutting edge: heat shock protein 60 is a putative endogenous ligand of the toll-like receptor-4 complex. J Immunol **164:** 558-561.

ORTUTAY, C., M. SIERMALA and M. VIHINEN, 2007 Molecular characterization of the immune system: emergence of proteins, processes, and domains. Immunogenetics **59:** 333-348.

ORTUTAY, C., and M. VIHINEN, 2009 Immunome knowledge base (IKB): an integrated service for immunome research. BMC Immunol **10:** 3.

PATIL, A. A., Y. CAI, Y. SANG, F. BLECHA and G. ZHANG, 2005 Cross-species analysis of the mammalian beta-defensin gene family: presence of syntenic gene clusters and preferential expression in the male reproductive tract. Physiol Genomics **23:** 5-17.

PAZGIER, M., D. M. HOOVER, D. YANG, W. LU and J. LUBKOWSKI, 2006 Human beta-defensins. Cell Mol Life Sci **63:** 1294-1313.

PERTEA, G., X. HUANG, F. LIANG, V. ANTONESCU, R. SULTANA *et al.*, 2003 TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics **19:** 651-652.

POLLASTRI, G., and A. MCLYSAGHT, 2005 Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics **21:** 1719-1720.

POWERS, J. P., and R. E. HANCOCK, 2003 The relationship between peptide structure and antibacterial activity. Peptides **24:** 1681-1691.

PURCELL, M. K., K. D. SMITH, L. HOOD, J. R. WINTON and J. C. ROACH, 2006 Conservation of Toll-Like Receptor Signaling Pathways in Teleost Fish. Comp Biochem Physiol Part D Genomics Proteomics **1:** 77-88.

RADHAKRISHNAN, Y., M. A. FARES, F. S. FRENCH and S. H. HALL, 2007 Comparative Genomic Analysis of a Mammalian {beta} - Defensin Gene Cluster. Physiol Genomics.

RADHAKRISHNAN, Y., K. G. HAMIL, S. YENUGU, S. L. YOUNG, F. S. FRENCH *et al.*, 2005 Identification, characterization, and evolution of a primate beta-defensin gene cluster. Genes Immun **6:** 203-210.

REHAUME, L. M., and R. E. HANCOCK, 2008 Neutrophil-derived defensins as modulators of innate immune function. Crit Rev Immunol **28:** 185-200.

ROACH, J. C., G. GLUSMAN, L. ROWEN, A. KAUR, M. K. PURCELL *et al.*, 2005 The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci U S A **102:** 9577-9582.

ROOSEN, S., K. EXNER, S. PAUL, J. M. SCHRODER, E. KALM *et al.*, 2004 Bovine beta-defensins: identification and characterization of novel bovine beta-defensin genes and their expression in mammary gland tissue. Mamm Genome **15:** 834-842.

RUDD, P. M., M. R. WORMALD, R. L. STANFIELD, M. HUANG, N. MATTSSON *et al.*, 1999 Roles for glycosylation of cell surface receptors involved in cellular immune recognition. J Mol Biol **293:** 351-366.

SACKTON, T. B., B. P. LAZZARO, T. A. SCHLENKE, J. D. EVANS, D. HULTMARK *et al.*, 2007 Dynamic evolution of the innate immune system in Drosophila. Nat Genet **39:** 1461-1468.

SAWYER, S., 1989 Statistical tests for detecting gene conversion. Mol Biol Evol **6:** 526-538.

SCHEEFF, E. D., and P. E. BOURNE, 2005 Structural evolution of the protein kinase-like superfamily. PLoS Comput Biol **1:** e49.

SCHUELER-FURMAN, O., and D. BAKER, 2003 Conserved residue clustering and protein structure prediction. Proteins **52:** 225-235.

SCHULTZ, J., R. R. COPLEY, T. DOERKS, C. P. PONTING and P. BORK, 2000 SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res **28:** 231-234.

SELSTED, M. E., M. J. NOVOTNY, W. L. MORRIS, Y. Q. TANG, W. SMITH *et al.*, 1992 Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils. J Biol Chem **267:** 4292-4295.

SELSTED, M. E., and A. J. OUELLETTE, 2005 Mammalian defensins in the antimicrobial immune response. Nat Immunol **6:** 551-557.

SEMPLE, C. A., P. GAUTIER, K. TAYLOR and J. R. DORIN, 2006 The changing of the guard: Molecular diversity and rapid evolution of beta-defensins. Mol Divers **10:** 575-584.

SEMPLE, C. A., M. ROLFE and J. R. DORIN, 2003 Duplication and selection in the evolution of primate beta-defensin genes. Genome Biol **4:** R31.

SLACK, J. L., K. SCHOOLEY, T. P. BONNERT, J. L. MITCHAM, E. E. QWARNSTROM *et al.*, 2000 Identification of two major sites in the type I interleukin-1 receptor cytoplasmic region responsible for coupling to pro-inflammatory signaling pathways. J Biol Chem **275:** 4670-4678.

SONNHAMMER, E. L., and E. V. KOONIN, 2002 Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet **18:** 619-620.

STANLEY, M. A., 2002 Imiquimod and the imidazoquinolones: mechanism of action and therapeutic potential. Clin Exp Dermatol **27:** 571-577.

STOVEN, S., I. ANDO, L. KADALAYIL, Y. ENGSTROM and D. HULTMARK, 2000 Activation of the Drosophila NF-kappaB factor Relish by rapid endoproteolytic cleavage. EMBO Rep **1:** 347-352.

SUN, Z. Y., S. T. KIM, I. C. KIM, A. FAHMY, E. L. REINHERZ *et al.*, 2004 Solution structure of the CD3epsilondelta ectodomain and comparison with CD3epsilongamma as a basis for modeling T cell receptor topology and signaling. Proc Natl Acad Sci U S A **101:** 16867-16872.

SWANSON, K., S. GORODETSKY, L. GOOD, S. DAVIS, D. MUSGRAVE *et al.*, 2004 Expression of a beta-defensin mRNA, lingual antimicrobial peptide, in bovine mammary epithelial tissue is induced by mastitis. Infect Immun **72:** 7311-7314.

SWANSON, W. J., Z. YANG, M. F. WOLFNER and C. F. AQUADRO, 2001 Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A **98:** 2509-2514.

TAKEDA, K., and S. AKIRA, 2007 Toll-like receptors. Curr Protoc Immunol **Chapter 14:** Unit 14 12.

TAMURA, K., J. DUDLEY, M. NEI and S. KUMAR, 2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol **24:** 1596-1599.

TEMPERLEY, N. D., S. BERLIN, I. R. PATON, D. K. GRIFFIN and D. W. BURT, 2008 Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss. BMC Genomics **9:** 62.

TENOR, J. L., and A. ABALLAY, 2008 A conserved Toll-like receptor is required for Caenorhabditis elegans innate immunity. EMBO Rep **9:** 103-109.

UEMATSU, S., and S. AKIRA, 2008 Toll-Like receptors (TLRs) and their ligands. Handb Exp Pharmacol**:** 1-20.

VIGNALI, D. A., R. T. CARSON, B. CHANG, R. S. MITTLER and J. L. STROMINGER, 1996 The two membrane proximal domains of CD4 interact with the T cell receptor. J Exp Med **183:** 2097-2107.

WEBER, A. N., M. A. MORSE and N. J. GAY, 2004 Four N-linked glycosylation sites in human toll-like receptor 2 cooperate to direct efficient biosynthesis and secretion. J Biol Chem **279:** 34589-34594.

WERLING, D., and T. W. JUNGI, 2003 TOLL-like receptors linking innate and adaptive immune response. Vet Immunol Immunopathol **91:** 1-12.

WRIGHT, S. D., R. A. RAMOS, P. S. TOBIAS, R. J. ULEVITCH and J. C. MATHISON, 1990 CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein. Science **249:** 1431-1433.

XIA, X., and Z. XIE, 2001 DAMBE: software package for data analysis in molecular biology and evolution. J Hered **92:** 371-373.

XU, C., M. E. CALL and K. W. WUCHERPFENNIG, 2006 A membrane-proximal tetracysteine motif contributes to assembly of CD3deltaepsilon and CD3gammaepsilon dimers with the T cell receptor. J Biol Chem **281:** 36977-36984.

YAMAMOTO, M., S. SATO, H. HEMMI, S. UEMATSU, K. HOSHINO *et al.*, 2003 TRAM is specifically involved in the Toll-like receptor 4-mediated MyD88-independent signaling pathway. Nat Immunol **4:** 1144-1150.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci **13:** 555-556.

YANG, Z., N. GOLDMAN and A. FRIDAY, 1994 Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol **11:** 316-324.

YOUNT, N. Y., and M. R. YEAMAN, 2004 Multidimensional signatures in antimicrobial peptides. Proc Natl Acad Sci U S A **101:** 7363-7368.

ZHANG, J., R. NIELSEN and Z. YANG, 2005 Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol **22:** 2472-2479.

ZOU, J., C. MERCIER, A. KOUSSOUNADIS and C. SECOMBES, 2007 Discovery of multiple beta-defensin like homologues in teleost fish. Mol Immunol **44:** 638-647.

8. Appendix

Chapter 2: Chicken Toll Like Receptors Uniprot Accession numbers:

ggTLR1	A2A251	Toll-like receptor 1 type 2
ggTLR16	A5YBP4	Toll-like receptor 16
ggTLR2.1	Q9DD78	Toll-like receptor 2 type-1
ggTLR2.2	Q9DGB6	Toll-like receptor 2 type-2
ggTLR3	Q0PQ88	Toll-like receptor 3
ggTLR4	Q7ZTG5	Toll-like receptor 4
ggTLR5	Q4ZJ82	Toll-like receptor 5
ggTLR7	Q5GQ97	Toll-like receptor 7

Chapter 2: Mammalian Toll Like Receptors Uniprot Accession numbers (hs = human, mm= mouse)

hsTLR1	Q15399	Toll-like receptor 1
hsTLR2	O60603	Toll-like receptor 2
hsTLR3	O15455	Toll-like receptor 3
hsTLR4	O00206	Toll-like receptor 4
hsTLR5	O60602	Toll-like receptor 5
hsTLR6	Q9Y2C9	Toll-like receptor 6
hsTLR7	Q9NYK1	Toll-like receptor 7
hsTLR8	Q9NR97	Toll-like receptor 8
hsTLR9	Q9NR96	Toll-like receptor 9
hsTLR10	Q9BXR5	Toll-like receptor 10
mmTLR11	Q6R5P0	Toll-like receptor 11
mmTLR12	Q6QNU9	Toll-like receptor 12
mmTLR13	Q6R5N8	Toll-like receptor 13

Chapter 3: TLR Pathway Components Uniprot Accession Numbers

SARM1	Q6SZW1	Sterile alpha and TIR motif-containing protein 1
TRIF	Q8IUC6	TIR domain-containing adapter molecule 1
TRAM	Q86XR7	TIR domain-containing adapter molecule 2
TRAF6	Q9Y4K3	TNF receptor-associated factor 6
TIRAP-MAL	Q56UI0	MyD88 adapter-like protein
TAB1	Q15750	Mitogen-activated protein kkk 7-interacting protein 1
TAB2	Q9NYJ8	Mitogen-activated protein kkk 7-interacting protein 2
TAB3	Q8N5C8	Mitogen-activated protein kkk 7-interacting protein 3
TRAF3	Q13114	TNF receptor-associated factor 3
BTK	Q06187	Tyrosine-protein kinase BTK
SOCS1	O15524	Suppressor of cytokine signaling 1
TAK1	O43318	Mitogen-activated protein kkk 7 TAK1
MD-2	Q9Y6Y9	Lymphocyte antigen 96 MD-2
TOLLIP	Q6FIE9	Toll interacting protein
TBK1	A8K4S4	TANK-binding kinase 1
CD14	P08571	Monocyte differentiation antigen CD14
RIP1	Q13546	RIP1 Receptor-interacting serine/threonine kinase
IKKa	O15111	Inhibitor of nuclear factor kappa-B kinase subunit alpha
IKKb	O14920	Inhibitor of nuclear factor kappa-B kinase subunit beta
IKKg	Q7LBY6	NFkappaB essential modulator (IKBKG protein) IKKG
IKKe	Q14164	Inhibitor of nuclear factor kappa-B kinase subunit epsilon

MKK3-6	P46734	Mitogen-activated protein kinase kinase 3 MKK3/6
MKK7	O14733	Mitogen-activated protein kinase kinase 7 MKK7
NIK1	Q99558	Mitogen-activated protein kkk 14
NFKB1	P19838	Nuclear factor NF-kappa-B p105 subunit
NFKB2	Q00653	Nuclear factor NF-kappa-B p100 subunit
IKBa	P25963	NF-kappa-B inhibitor alpha NFKBIA
IKBb	Q15653	NF-kappa-B inhibitor beta NFKBIB
UBC13	P61088	UBC13 Ubiquitin-conjugating enzyme E2
UeV1A	Q13404	Ubiquitin-conjugating enzyme E2 variant 1 UEV1A
ECSIT	Q9BQ95	Evolutionarily conserved signaling intermediate in Toll
MEKK1	Q13233	Mitogen-activated protein kinase kinase kinase 1
p38-MK14	Q16539	Mitogen-activated protein kinase 14
JNK-MK08	P45983	Mitogen-activated protein kinase 8
ERK2-MK01	P28482	Mitogen-activated protein kinase 1
TANK	Q92844	TRAF family member-associated NF-kappa-B activator
NAP1	Q9BU70	Nef-associated protein 1
SINTBAD	A2A9T0	TANK-binding kinase 1-binding protein
IRAK1	P51617	Interleukin-1 receptor-associated kinase 1
IRAK2	O43187	Interleukin-1 receptor-associated kinase-like 2
IRAKM	Q9Y616	Interleukin-1 receptor-associated kinase 3
IRAK4	Q9NWZ3	Interleukin-1 receptor-associated kinase 4
IRF1	P10914	Interferon regulatory factor 1
IRF3	Q14653	Interferon regulatory factor 3
IRF5	Q13568	Interferon regulatory factor 5
IRF7	Q92985	Interferon regulatory factor 7
SIGIRR	Q6IA17	Single Ig IL-1-related receptor
FADD	Q13158	FAS-associated death domain protein

Chapter 4: Canine Defensins Genbank Accession numbers

cbd1	AAY59708.1
cbd102	AAY59709.1
cbd103	AAY59710.1
cbd104	AAY59711.1
cbd105	AAY59712.1
cbd106	AAY59713.1
cbd107	AAY59714.1
cbd108	AAY59715.1
cbd109	AAY59716.1
cbd110	AAY59717.1
cbd111	AAY59718.1
cbd112	AAY59719.1
cbd113	AAY59720.1
cbd114	AAY59721.1
cbd116	AAY59722.1
cbd117	AAY59723.1
cbd118	AAY59724.1
cbd119	AAY59725.1
cbd120	AAY59726.1
cbd121	AAY59727.1
cbd122	AAY59728.1
cbd123	AAY59729.1
cbd124	AAY59730.1
cbd125	AAY59731.1
cbd126	AAY59732.1
cbd127	AAY59733.1

cbd128	AAY59734.1
cbd129	AAY59735.1
cbd130	AAY59736.1
cbd131	AAY59737.1
cbd132	AAY59738.1
cbd134	AAY59739.1
cbd135	AAY59740.1
cbd136	AAY59741.1
cbd138	AAY59742.1
cbd139	AAY59743.1
cbd140	AAY59744.1
cbd141	AAY59745.1
cbd142	AAY59746.1
cfspag11c	AAY59747.1
cfspag11e	AAY59748.1

Chapter 4: Human Defensins Genbank Accession numbers

HBD1	NP_005209.1
HBD4	NP_004933.1
HBD103	NP_061131.1
HBD104	NP_525128.2
HBD105	NP_689463.1
HBD106	NP_689464.1
HBD107	AAZ81951.1
HBD108	NP_001002035.1
HBD109	AAY59749.1
HBD110	AAY59750.1
HBD111	AAY59751.1
HBD112	AAY59752.1
HBD113	AAY59753.1
HBD114	AAY59754.1
HBD115	AAY59755.1
HBD116	AAY59756.1
HBD117	AAY59757.1
HBD118	NP_473453.1
HBD119	NP_695021.2
HBD120	NP_697018.1
HBD121	NP_001011878.1
HBD123	NP_697019.1
HBD124	AAZ81952.1
HBD125	NP_697020.2
HBD126	NP_112193.1
HBD127	NP_620713.1
HBD128	AAP47223.1
HBD129	NP_543021.1
HBD130	AAY59758.1
HBD131	AAQ09523.1
HBD132	NP_997352.1
HBD133	AAY59759.1
HBD134	AAY59760.1
HBD135	AAY59761.1
HBD136	AAY59762.1
HSSPAG11C	NP_478110.1
HSSPAG11E	NP_478114.1

Chapter 4: Mouse Defensins Genbank Accession numbers

MBD1	NP_031869.1
MBD2	NP_034160.1
MBD3	NP_038784.1
MBD4	NP_062702.1
MBD5	NP_109659.2
MBD6	NP_473415.1
MBD7	NP_631966.1
MBD8	NP_694748.2
MBD9	NP_631965.1
MBD10	NP_631971.1
MBD11	NP_631967.1
MBD12	NP_690015.1
MBD13	NP_631969.1
MBD14	NP_898847.1
MBD15	NP_631968.1
MBD16	AAY59763.1
MBD17	AAY59764.1
MBD18	AAY59765.1
MBD19	AAY59766.1
MBD20	AAY59767.1
MBD21	AAY59768.1
MBD22	AAY59769.1
MBD23	AAY59770.1
MBD24	AAY59771.1
MBD25	AAY59772.1
MBD26	AAY59773.1
MBD27	AAY59774.1
MBD28	AAY59775.1
MBD29	AAN77095.1
MBD30	AAY59776.1
MBD33	AAY59777.1
MBD34	NP_898856.1
MBD35	NP_631970.1
MBD36	AAY59778.1
MBD37	NP_859011.1
MBD38	NP_898857.1
MBD39	NP_898859.2
MBD40	NP_898860.3
MBD41	AAY59779.1
MBD42	AAY59780.1
MBD43	AAY59781.1
MBD44	AAY59782.1
MBD50	AAY59783.1
MBD51	AAY59784.1
MBD52	AAY59785.1
MBD53	AAY59786.1
MMSPAG11C	AAY59787.1
MMSPAG11	NP_694755.1

Chapter 5: CD3δ Accession numbers from Genbank and Ensembl databases

Human	NP_000723
Chimp	XP_508789
Orangutan	ENSPPYP00000004501
Macaca	XP_001097302
Bushbaby	ENSOGAP00000008358
Elephant	ENSLAFP00000009442
Dolphin	ENSTTRP00000004007
Hyrax	ENSPCAP00000000083
Megabat	ENSPVAP00000005893
Kangaroo rat	ENSDORP00000003213
Rabbit	ENSOCUP00000000382
Guinea Pig	ENSCPOP00000011850
Microbat	ENSMLUP00000012796
Cat	ENSFCAP00000010929
Dog	XP_536556
Cow	NP_001029205
Rat	NP_037301
Mouse	NP_038515
Sheep	NP_001009382
Pig	NP_998940
Horse	XP_001502881

Chapter 5: CD3γ Accession numbers from Genbank and Ensembl databases

Human	NP_000064
Chimp	ENSPTRP0000000744
Orangutan	ENSPTRP00000007441
Macaca	XP_001093643
Mouse Lemur	ENSMICP00000008453
Dolphin	ENSTTRP00000003995
Pig	NP_001008686
Rat	NP_001071114
Mouse	NP_033980
Dog	XP_546501
Horse	XP_001502888
Cow	NP_001035562
Hedgehog	ENSEEUP00000009283
Alpaca	ENSVPAP00000005180
Elephant	ENSLAFP00000009439
Rabbit	ENSOCUP00000005934
Squirrel	ENSSTOP00000012873
Guinea Pig	ENSCPOP00000001293
Cat	ENSFCAP00000010926
Hyrax	ENSPCAP00000011440
Megabat	ENSPVAP00000005892

Chapter 5: CD3ε Accession numbers from Genbank and Ensembl databases

Human	NP_000724
Chimp	XP_001160698
Orangutan	ENSPPYP00000004500
Macaca	XP_001097204
Marmoset	ABA29629
Dog	NP_001003379
Rabbit	NP_001075470
Cow	NP_776436
Pig	NP_999392
Cat	NP_001009862
Mouse	NP_031674
Sheep	NP_001009418
Rat	NP_001101610
Horse	ENSECAP00000012382
Guinea Pig	ENSCPOP00000001293