Using genomic and proteomic information to characterize the evolution of genes involved in development and adaptation in vertebrates under differential conditions of selective pressure

Imran Khan

(SFRH/BD/48518/2008)

PhD Thesis

CIIMAR,

Centro Interdisciplinar de Investigação Marinha e Ambiental,

Rua dos Bragas, n.289,

4050-123, Porto,

Portugal.

July, 2014.

"Nothing in biology makes sense except in the light of evolution"

Theodosius Dobzhansky (1964)

Acknowledgements

First of all, I offer my profoundest gratitude to my supervisor, Prof. Agostinho Antunes, for his continuous support and guidance during all these years, thank you Prof. Agostinho for your encouragement, support, motivation, easily approachable nature and freedom and all the excellent comments for the all the results and scintillating discussions. I will always be grateful for your mentoring and supporting me and having faith in me. My co-supervisor Prof. Vitor Vasconcelos, was always an encouraging presence and support during this whole time.

I would also like to extend special thanks to Dr. Warren E. Johnson and Dr. Stephen J. O'Brien for their valuable suggestions and comments which helped me in improving my manuscripts.

I will like to thank my previous research supervisors and colleagues Dr. K. Thangaraj, Dr. L. Singh, Dr. G. Chaubey, Dr. S. P. Goyal, Dr. J. Aspi, Dr. I. Kojola, Late Dr. M. Ruokonen, Dr. B. Habib, Dr S. Gupta and Dr. Q. Qureshi for instilling the scientific aptitude in me.

A special mention goes to Emanuel, Siby, Anoop, Tiby, R. Borges, Daniela, J. Paulo Guillermin, Barbara, Dany, Christiana, Liliana my colleagues from LEGE and special thanks to Anoop, Siby and Emanuel for making my stay memorable and helping me adjust to Porto life and lab.

I would also like to acknowledge my friends in India Farhan, Govind and Ranjana for their valuable friendship, they surely have special place in my life.

I would also like to thank the officials at FCUP and CIIMAR specially Rosario, M. Arouca and S. Santos and Micaela Vale for facilitating the official work.

I don't have any words to thank the love of my life Nahid and parents for their constant support and motivation.

.

The Portuguese foundation for science and technology (FCT), is sincerely acknowledged for the PhD grant (SFRH/BD/48518/2008), making my stay and study possible. My research was also funded in part by the FCT project PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490) and partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Program and national funds through FCT under the project PEst-C/MAR/LA0015/2013.

Thank you all once again!

Imran Khan

Contents

List of Tables

List of Figures

Abstract

In this thesis it is presented in depth comparative evolutionary genomics and adaptive evolutionary analyses of gene and gene families involved in developmental, detoxification, immune defense, gamete interaction and sensory perception.

Adaptation of mammals to terrestrial life was facilitated by the unique vertebrate trait of body hair, which occurs in a range of morphological patterns. Keratin associated proteins (KRTAPs), the major structural hair shaft proteins, are largely responsible for hair variation. The study of the KRTAP gene family revealed genetic variations related with ecological adaptation of species. The gene loss in the KRTAP gene family in dolphin was related with hair less, a feature of relevance for fast swimming, whereas expanded KRTAP repertoire in sloth favored the hosting of hair symbionts. Gene expression variation probably also influenced hair diversification patterns. For example, humans have identical KRTAP repertoire relatively to apes, but much less hair.

Olfaction, the ability to smell, is one of the most important sensory functions, which is governed by olfactory receptors (OR) expressed predominantly at the cell-surface of olfactory sensory neurons (OSNs) that are located in the main olfactory epithelium of the nasal cavity. Although the olfactory gene repertoire of mammals has been linked to ecological specialization, patterns of adaptation have not been explicitly addressed in other vertebrates. Here, the OR diversity was explored in a phylogenetic and ecologically diverse group of sauropsida, including 48 birds and two reptiles (*Alligator mississippiensis and Chelonia mydas*) to assess how ecological patterns may have influenced their OR gene family repertoire and olfactory abilities. Ecological adaption was found to determine the olfactory ability in birds and reptiles shaped by different proportion of functional OR gene families contributing to the diversification of OR ability in birds and reptiles, with different gene families under rapid expansion and positive selection. The presence of positive selected sites in ORs also suggests the ongoing adaptive diversification of olfaction.

The cytosolic glutathione transferases (cGSTs) are known for their dynamic and interactive defense mechanism providing protection against cytotoxic electrophilic substrates and adaptation to exposure to cellular stress. Six out of the seven major cytosolic GST classes were found in birds. The sequence comparison and phylogeny of avian cGSTs revealed that birds GST have similar active binding site as mammals. The gene duplication and positive selection played an important role in diversification of avian GSTs. The positive selection in duplicated GSTA and GSTT genes suggest

likely their important role in protection from reactive species and xenobiotic compounds. We also found positive selection in GSTO and GSTZ, the evidence of positive selection was also supported by radical changes in amino acid properties.

Toll-like receptors (TLRs) are multigene family members involved in immune response and are a fascinating example of the evolutionary arms race between host and pathogen. Most of the TLR gene family members predate the origin of basal vertebrate lineages. The TLR gene family evolution was shaped by events of gene gain and loss varying among species, with episodes of gene duplication occurring mostly in the fish lineage. The coelacanth TLRs showed shared features with both tetrapods and fishes. TLR genes were earlier thought to be restricted to certain vertebrate lineages but this study revealed wider distribution with 26 TLR subfamilies in diverse vertebrate species/lineages. The retrieved results point towards the wider distribution of the TLR gene family in vertebrates. Different rates of gene gain and loss originated variable numbers of TLRs across different lineages of vertebrates due to specific dynamics required to recognize a large variety of pathogens. The shared synteny among coelacanth, fishes and tetrapods suggest an evolutionary transition and proximity with these lineages and also with its unique immune system which lacks Immunoglobulin M, required in the first line of immune defense. Avian TLR genes, including both viral and non-viral TLRs, evolved under positive selection. The strong selective pressure found in viral TLR immune genes is likely explained by long term co-evolutionary dynamics between birds and viruses. Overall those findings support the important role of the TLR gene family in host-pathogen arms race.

The ZP gene family is involved in egg envelope formation throughout the vertebrate lineage. The differences in type, number of genes and rapid evolution of the ZP genes determine the species-specific gamete interactions. These changes altogether may create reproductive barriers and lead to speciation, representing a crucial step in species evolution. Contrary to mammals that are monospermic, birds and reptiles undergo physiological polyspermy, without detrimental consequence on later development and thus studying the comparative genomics and adaptive evolution of the egg envelope ZP subgenome within and between these diverse lineages of vertebrates can be insightful to understand the evolutionary role of ZP genes in reproduction and speciation. The vertebrates had lineage and species-specific changes of gene gain, loss and functional diversification of the ZP gene family. The rapid evolution of ZP egg envelope proteins is possibly caused by diverse but entwined evolutionary forces, which includes cryptic female choice, sperm competition, hybridization avoidance and sexual conflict, ultimately resulting in speciation. The rapid evolution of ZPs in monospermic mammals and physiological

polyspermic birds and reptiles suggest that polyspermy avoidance is not the only driving force for the rapid evolution of ZPs and diverse but entwined evolutionary forces, like cryptic female choice, sperm competition, hybridization avoidance and sexual conflict are possibly involved and ultimately result in speciation.

The extensive comparative evolutionary genomics, proteomics and adaptive evolution of genes and gene family involved in development (KRTAP gene family), sensory perception (Olfactory receptors gene family), detoxification (cGSTs gene family), immune system (TLR gene family) and reproduction (ZP gene family), revealed important evolutionary and adaptive features acting in vertebrates. The loss of hairs, related with the role of KRTAPs in the water to land transition, the role of ecological adaptation in determining the OR subgenome and its relation with olfactory bulb ratio (OBR), the gene gain lead to copy number variations and rapid adaptive evolution of cGST gene family with possible relation with rapid increase in toxin diversity. The differential gene gain, gene loss and strong positive selection in both viral and non-viral TLR gene family suggested the host-pathogen arms race, and the relevance of viruses in the evolution of mammals and birds. The ZP gene family evolution revealead rapid evolution in monospermic mammals and polyspermic birds and reptiles which strongly support that other factors then polyspermic avoidance are likely responsible for their rapid evolution, e.g. cryptic female choice, sperm competition, hybridization avoidance and sexual conflict are possibly involved and ultimately result in speciation.

Resumo

Nesta tese apresentamos um estudo comparativo de genómica e adaptação evolutiva de genes e famílias de genes envolvidas no desenvolvimento, desintoxicação, defesa imunológica, interação de gâmetas e percepção sensorial.

A adaptação de mamíferos à vida terrestre foi facilitada pelo desenvolvimento exclusivo em vertebrados de pêlos corporais, que ocorrem numa gama variada de padrões morfológicos. As proteínas associadas à queratina (KRTAPs), as proteínas mais comuns nos pêlos, são responsáveis em grande medida pela sua variação. Encontramos variações na família de genes KRTAP relacionadas com a adaptação ecológica das espécies. A perda de genes da família KRTAP em golfinhos desprovidos de pêlo facilita uma natação rápida enquanto que a expansão do repertório KRTAP em preguiças facilitou a associação de simbiontes. Variações na expressão genética também influenciam os padrões de diversificação de pêlos. Por exemplo, os humanos, têm reportórios genéticos idênticos aos dos macacos, mas exibem muito menos pêlo.

A capacidade olfativa, que permite cheirar, é uma das mais importantes funções sensoriais, e é devida a receptores olfativos (OR) expressos predominantemente na superfície celular de neurónios sensoriais olfativos (OSNs) localizados no epitélio olfativo da cavidade nasal. Embora o reportório de genes olfativos em mamíferos tenha sido ligado a uma especialização ecológica, os padrões de adaptação não têm sido claramente detetados em outros vertebrados. Neste trabalho, exploramos a diversidade de ORs em grupos filogeneticamente e ecologicamente diversos de sauropsideos, incluindo 48 aves e duas espécies de répteis, *Alligator mississippiensis* e *Chelonia mydas*, de forma a avaliar como os padrões ecológicos podem influenciar o repertório de genes OR e as capacidades olfativas. Descobrimos que a adaptação ecológica determina as capacidades olfativas em aves e repteis, sendo moldada por proporções diferentes de famílias de genes OR. Diferentes famílias de genes contribuem para a diversificação da capacidade olfativa em aves e répteis; diferentes famílias de genes estão sob rápida expansão e seleção positiva nestas linhagens. A presença de locais selecionados positivamente em ORs também sugere a diversificação adaptativa em curso do olfato.

As glutationa-s-tranferases citosólicas (cGSTs) são conhecidas pela sua dinâmica e pelo mecanismo de ação defensiva, proporcionando proteção contra substratos citotóxicos electrofílicos e adptação à exposição a stress celular.

Encontramos seis das sete mais importantes classes de GST em aves. A comparação sequencial e filogenética de cGST de aves mostrou que têm locais de ligação similares às dos mamíferos. A duplicação de genes e a seleção positiva tem uma papel importante na diversificação do papel das cGST de aves. A seleção positiva nos genes duplicados GSTA e GSTT suporta o seu importante papel na proteção contra espécies reativos e compostos xenobioticos. Também encontramos seleção positiva em GSTO e GSTZ que possivelmente aponta para o seu importante papel secundário. Os resultados da duplicação de genes junto com os da seleção positiva são também suportados pelas alterações radicais das propriedades dos aminoácidos.

Os receptores *Toll-like* (TLRs) são uma família multigénica envolvida na resposta imune e são exemplos fascinantes da relação entre hospedeiro e patogenes. A maior parte das famílias de genes TLR foram originadas cedo na evolução com a maior parte dos TLRs encontrados na linhagem basal dos vertebrados. A evolução da família de genes TLR foi moldada por eventos de perdas e ganhos de genes, tendo sido encontrados diferente taxas de perdas e ganhos de genes em diferentes espécies, com a maior parte das duplicações génicas a ocorrer cedo na linhagem dos peixes. Os TLRs do celacanto mostraram algumas características comuns com tetrápodes e peixes. Os genes TLR que inicialmente foram restritos a certas linhagens foram encontrados em novos genomas estudados recentemente. Os nossos resultados apontam para uma distribuição mais alargada da família de genes TLR nos vertebrados que foi moldada por perdas e ganhos de genes, levando a um grande e variável número de TLRs em vertebrados, de forma a reconhecerem uma extensa variedade de patogenes. A sintenia partilhada do celacanto com os peixes e os tetrapodes é possivelmente devida a uma proximidade evolutiva com o ancestral que primeiro caminhou sobre a terra e é primordial para se entender a evolução de tetrapodes e o seu sistema imunológico único sem IgM. Os genes TLR de aves evoluíram sob seleção positiva, incluindo os genes TLR virais e não virais. No seu conjunto, os nosso resultados suportam o importante papel da família de gene TLR na relação hospedeiro/patogene. A forte pressão seletiva encontrada nos genes imunológicos virais TLR sugere a hipótese da coexistência prolongada de aves e vírus.

A familia de genes ZP está envolvida na formação da cobertura do ovo ao longo da linhagem de vertebrados. As diferenças no tipo, número de genes ZP e a rápida evolução destes genes contribuem para a interação dos gâmetas sendo espécie-especifica. Estas alterações são muito importantes para a evolução das espécies levando à especiação pela criação de barreiras reprodutivas. Ao contrário dos

mamíferos que são monospermicos, as aves e os repteis são polispermicos, sem ter uma consequência negativa no desenvolvimento e por isso o estudo do subgenoma ZP entre diferentes linhagens pode ajudar ao melhor entendimento do papel evolutivo dos genes ZP na reprodução e na especiação. A família de genes ZP sofreu modificações especificas nas várias espécies de vertebrados como pode ser demonstrado pelo ganho e perda de genes e pela diversificação funcional nas várias linhagens de vertebrados. A rápida evolução das proteínas ZP é possivelmente causada por diversas e complexas pressões evolutivas, que incluem escolha críptica de fêmeas, competição por esperma, hibridação incompatível e conflito sexual, que em último caso pode levar à especiação. O evitar da polispermia não parece ser a única força motriz para a evolução rápida das ZP em mamíferos monospermicos, e nas aves e repteis que são polispermicas.

1 Introduction

1.0.1 Background

The excess fecundity and consequent competition to survive in every species, provides the preconditions for the process Darwin called natural selection (Darwin et al. 1859). Natural selection is arguably the most important idea in biology. Darwin theory of natural selection specifies the preservation of favorable variations and the rejection of injurious variations by nature, whereas neutral variations without harmful or useful effect are not affected by natural selection and help in maintaining polymorphism (Darwin et al. 1859). Natural selection not only can produce evolutionary change and increases the fitness of an organism making them better adapted to the changing environment. Adaptive evolution can increase the frequency of one allele favoring the beneficial phenotypic variation or trait by increasing the fitness of the organism. The gradual accumulation of these favorable changes creates diversification, which ultimately leads to the origin of new species. Tough Darwin explained that beneficial traits are favored by nature, Darwin used a provisional hypothetical mechanism to explain inheritance and termed it Pangenesis, but he was still not clear about how these changes happen and were inherited in nature.

Gregor Mendel was the first to explain the concept of heredity which was more or less right and was published in 1866 but went unnoticed till 1900 when it was separately rediscovered by Hugo de Vries, Carl Correns and Erich von Tschermak, which set the pace for modern day genetics (http://www.genome.gov/25520238). The chromosome theory proposed chromosome as heritable unit and later genes were found to reside in chromosomes with genetic maps showing their linear arrangement on chromosomes. In 1944 the DNA was isolated as the genetic material (Avery et al. 1944) and in 1953 the double helix structure was resolved and the genetic code was cracked by the end of 1967 and DNA was sequenced for the first time (Sanger, Air, et al. 1977). The invention of polymerase chain reaction (PCR) in 1983 lead to the rapid amplification of DNA sequences in the lab setup, ultimately resulting in large scale DNA sequence data revolutionized the field of genetics and evolutionary biology and leading to sequencing the first genome of a free living organism (Fleischmann et al. 1995). Modern days advanced genomic sequencing and computational analyses are able to provide worth full of DNA sequence information enabling large scale sequence comparisons at various level, e.g. gene, gene families

and genome level for underlying and understanding the secrets of evolution and its effect on life, changing drastically our perception of the molecular basis of adaptation.

The large number of published examples in the recent past precludes referencing all because of space limitations. This is due in part to the large influx of genomic sequence data resulting from genome-sequencing projects and increased sequencing efficiency combined with the development of new statistical analyses utilizing comparative sequence data and polymorphism data to uncover selective events. The identification of genes and gene regions subjected to positive selection can lead to predictions regarding the putative functionally important regions of genes. Exciting new areas of investigation include genomic approaches to identify the frequency of selective events from large sequencing surveys.

Molecular evolution of genes, gene families and genomes

The genes are defined as inheritable unit of genomic sequence (DNA or RNA) associated with regulatory regions, transcribed regions, and or other functional sequence regions (Pennisi 2007)(Pearson 2006). The advancement in genome sequencing technology has broadened the scope of comparative genomics from single gene to gene families and whole genomes. DNA is the heritable material and phenotypic variations are one of the outcomes of changes in that heritable material. These changes are produced by various mechanisms that could be external, such as environment like radiation or chemical, or could be internal like replication slippage, insertion or deletion (indels) of segments of DNA, gene, chromosome or whole genome duplication. Thus, the extent of mutation can range from single nucleotide to duplication of genes to whole genome duplication and can have several outcomes, which could be beneficial, harmful or without any change. The single gene can also be involved in multiple, apparently unrelated, phenotypes and are thus called pleiotropic genes (Hodgkin 1998)(Hartl 2005) or multiple genes (polygenic) can converge to result in a single phenotype. The changes within a gene could be caused by nucleotide substitution (synonymous or non-synonymous) or indel, these in turn can create beneficial (adaptive), harmful (negative) or neutral changes in gene. Broadly speaking, a gene could be responsible for a given trait (phenotype; e.g. coat color) and different forms of a gene called allele leads to difference in traits, e.g. melanism in the Cat family is caused by variants of coat color gene like MC1R and ASIP (Eizirik et al. 2003). Thus, advantageous changes can gradually become fixed in the population improving the fitness of organisms.

The other major benefit of whole genome sequencing projects is the study of multigene families. The multigene family is a group of genes descended from a

common ancestral gene and therefore resemble in functions and have DNA sequence similarity. The gene duplication at gene, chromosome and genome level are important source for providing raw material (paralogs) for genetic innovation (Ohno 1970). The duplicated gene may be found arranged in tandem or at distinct location in genomes. Following the birth-and-death model, new genes are created by gene duplication and some duplicate genes stay in the genome for a long time, while others are inactivated or deleted from the genome (Nei and Rooney 2005). The duplicated genes can perform a new function (neofunctionalization), or share the ancestral function (subfunctionalization), or become inactivated by pseudogenization.

Two different model have been proposed to explain the subfunctionization (Force et al. 1999), (1) the duplication-degeneration-complementation model (DDC) and (2) the Escape from Adaptive Conflict (EAC), both models having a similar outcome. The DDC model assumes that neutral drift results in complementary retention of sub-functions shared between the two gene copies, whereas EAC assumes that the ancestral gene evolves to intermediate multifunctional gene and descendant of duplication carry on the shared function (Hittinger and Carroll 2007). The concerted evolution is another evolutionary mechanism for tandemly arranged duplicated genes, where paralogs within a species are more closely related to each other than their orthologous repeats in a related species due to homogenization caused by gene conversion and homologous recombination (Ganley and Kobayashi 2007). Thus, gene gain and gene loss are main driving forces for gene family evolution creating diverse number of paralogs, which may undergo functional diversification and innovation and increasing the adaptability of an organism.

The changes in genomic sequences could range from point mutation to whole genome duplications. The large scale genomic data from diverse forms of life have enabled to extend the large scale comparative genomics from genes to gene families and genomes. As not only the changes in gene sequences can result in diversity but also the changes in gene family composition together with differences in genomic organization can have remarkable evolutionary outcome.

Genome sequencing and assembly play an important role in genomics and determines the quality and quantity of information that can be retrieved from the genomes. The genomes are strings of DNA, i.e. nucleotide base pairs A, T, G and C. The sequencing methods provide the way to read these base pairs, for example chain termination method of DNA sequencing (Sanger, Nicklen, et al. 1977) (Grada and Weinbrecht 2013). The initiation of sequencing projects using the shotgun

sequencing of randomly fragment genomic DNA resulted in the ground breaking mile stone of the human genome sequencing in the early 21th century (Lander et al. 2001)(Venter et al. 2001) and this set the face for advancement in sequencing technologies changing the pace of sequencing in present days to the highthroughput sequencing of numerous genomes Genome 10K Project (Genome 2009) **https://genome10k.soe.ucsc.edu/** leading to development of comparative genomics and interdisciplinary sciences which ultimately help in better understanding evolution of life and life forms.

The major objectives of the present study were to perform comparative evolutionary genomics of genes and gene families involved in various biological functions, like developmental (hair phenotype variations in mammals), sensory perception (olfactory receptors super gene family), detoxification (GSTs; Glutathione-S-transferase), immune defense (TLR; Toll like receptors gene family) and, reproduction and speciation (ZP; Zona pellucida egg envelope subgenome).

1.0.2 Brief introduction to the methods used to study the comparative genomics and adaptive evolution of genes and gene families

The comparative genomics and adaptive evolutionary analyses of gene and gene families are insightul in elucidating many evolutionary puzzles. The complete exploration of super gene or gene family members provide important information on the type of changes that have taken place in gene family members, from gene gain (expansion) to gene loss (deletion or pseudogeniztion), to genomic arrangements of a gene family. Thus, getting the complete gene family repertoire from different genomes is the first crucial step that sets the base of comparative genomics of gene family members. The correct estimation of super gene or gene families from a genome is important for the correct characterization of multiple members into families, subfamilies, classes and subclasses. In adition, the proper synteny analysis elucidates the correct assignment of orthologous and paralogous relationships. The precise grouping/clustering of gene family members together with the accurate orthologous and paralogous relationship is useful for all downstream analysis leading to confirmation of functional divergence and adaptive evolution of genes and gene families.

Thus, our comparative genomics studies broadly involve: (1) identification of a valid relationship between the biological phenomenon or phenotype and genes/gene–families; (2) exploration of genomes using the proper search strategy to infer the complete repertoire of genes and gene families; (3) the characterization of the

complete gene repertoire found within a genome into functional and nonfunctional genes; (4) the characterization of a complete gene repertoire found within a genome into proper groups, subfamilies, classes, etc; (5) assignment of orthologous and paralogous relationships; (6) checking evolutionary rates of gene and gene families within and between lineages to understand the role of functional divergence and adaptive evolution; (7) evaluate the location and importance of sites under different evolutionary rates (positive and/or negative selection) and assess the importance of these sites in function and structural integrity and consequences (harmful and beneficial) of the deduced changes; (8) relate all the findings and determine the factors influencing the evolution of genes and gene family repertoire, and functional divergence shaping a particular phenotype.

Genome curation

The advent of next generation high-throughput sequencing technologies resulted in large scale sequencing providing numerous whole genomes sequences (Grada and Weinbrecht 2013). The sequencing depth/coverage of the genomes determines the assembly quality and thus brings many challenges. The most important is gene annotation and this become complicate for low coverage genomes and ultimately affects the level of data curation. Getting almost complete high coverage finished assembly, as available for the human genome, requires a lot more efforts and expenses due to manual correction given repetitive regions, e.g. several million bases of repeat-rich heterochromatin, which are nearly impossible to achieve for other genomes. Many genome assemblies have only been assembled to the scaffold level. The genome assemblies are hierarchical in which the shortest sequence unit is known as contigs. The contigs are assembled to form scaffolds, and scaffolds are assembled into chromosomes resulting into a finished complete genome assembly. The genome assemblies are used for annotation of protein-coding genes, pseudogenes and non-coding RNAs along with homologous relationships and can be available freely from online resources like http://www.ensembl.org/ (Flicek et al. 2011), NCBI genbank (Benson et al. 2013). Most of the conserved genes are found annotated in these databases but care should be taken for cross checking and verification before moving on with the data analysis, e.g. the homology (orthologs and paralogs) relationship can be verified by reciprocal blast hits or more accurately by verifying chromosome synteny (Altschul et al. 1990a) (Muffato et al. 2010a) (A. Louis et al. 2013). The annotation of gene and gene families in these databases is not always complete due to species-specific variations and gradual increase in new genomic information (Young et al. 2010). The proper annotation and characterization

of gene families requires the complete prior knowledge of the all available gene family members. This information is used as query to search the genomes applying blast methods (Altschul et al. 1990a). All the reliable hits obtained are then fetched and used for annotation allowing the complete characterization of genes and gene families. These results in turn are dependent of the genome coverage and the type of gene and gene families in question, e.g. with high coverage genomes the proper search strategy will result in compete and correct information on any gene and gene family, whereas with low coverage genomes this will mostly depend of the size of genes, as small size genes like OR and VR that are intronless and ~1000bp can even be fetched from low coverage genomes (Young et al. 2010) (Hayden et al. 2010).

Multi sequence alignment

The accurate multiple sequence alignment is the basic requirement for all downstream analysis. This is very critical for adaptive evolutionary analyses and alignment errors can lead to false positives (Gharib and Robinson-Rechavi 2013) (Markova-Raina and Petrov 2011). Therefore different methods are available for achieving errors free alignments (Edgar 2004) (Thompson et al. 1994) (Löytynoja and Goldman 2005) (Notredame et al. 2000), which can be further improved by visual inspection and manual correction. Software like Gblocks (Talavera and Castresana 2007) or GUIDANCE (Penn et al. 2010) can be used to remove the gapped region from the alignments.

Phylogenetics for genes and gene families

Phylogenetics is the science of estimating the evolutionary past. The molecular phylogeny is based on the comparison of DNA or protein sequences. In the age of rapid and rampant gene sequencing, molecular phylogeny has truly come into its own, emerging as a major tool for making sense of a sometimes overwhelming amount of information (Baldauf 2003). As evolution is related with homology, the homologous sequences are used for reconstructing a phylogeny, which could include orthologs (diversified from a common ancestor by a speciation event) or paralogs (due to species specific duplication). The phylogenetic tree based on orthologs retrace the species tree, helping understanding the evolutionary relatedness between set of species, whereas phylogeny based on paralogs can resolve the evolutionary relatedness of gene duplicates of a gene family within genomes (Baldauf 2003). The phylogeny inferring methods can be broadly divided into two categories: (1) using

distance data and (2) using discrete data. The distance methods are based on the genetic distance between two sequences and these include neighbor joining, UPGMA, minimum evolution. The discrete methods include the parsimony, maximum likelihood (ML), and MCMC-based Bayesian inference. The parsimony requires minimum evolutionary change to explain the observed data. In maximum likelihood the estimates of the probabilities of DNA base substitutions are modeled by continuous-time Markov chains and used for construct the extant state of observed sequence and tree topology (Guindon and Gascuel 2003) (Stamatakis 2014). Bayesian methods basically use a posterior distribution for a parameter, composed of a phylogenetic tree and a model of evolution (Huelsenbeck and Ronquist 2001). The rates at which one nucleotide replaces another during evolution can be modeled by various substitution models and these models are mostly used in maximum likelihood and Bayesian tree reconstruction (Darriba et al. 2012)

Adaptive Evolution in Protein-Coding Genes:

Various statistical modeling techniques have been developed to study almost every aspect of molecular evolution and genomics. The models for codon evolution allow to determine the level of natural selection during gene sequence evolution. The codon models distinguish between the synonymous rate (dS) and the nonsynonymous rate (dN) of evolution within a gene. The ratio of these rates (dN/dS), is referred as omega (w), which provide the measure of the direction and intensity of natural selection pressure acting on a protein (Yang and Bielawski 2000) (Kosiol and Anisimova 2012) (Anisimova and Liberles 2007) (Anisimova and Kosiol 2009). If the absence of selection the rate of nonsynonymous evolution would be the same as the synonymous rate, with dN/dS= 1 (i.e., neutral evolution). However, most proteins are dominated by purifying selection (i.e., the removal of functionally deleterious mutations), thus their nonsynonymous rate will be less than the synonymous rate, with dN/dS < 1. In case of positive or Darwinian selection the nonsynonymous rate can exceed the synonymous rate, with dN/dS > 1 resulting in increased fitness (Goldman and Yang 1994). Independently proposed similar codon models (Muse and Gaut 1994) serve as the foundation for the large collection of codon substitution models currently available. The codon models are extensively used to investigate the process of molecular innovation and divergence, and are the subject of substantial research efforts. In-depth description of models and methodological developments can be found elsewhere (Anisimova and Kosiol 2009).

Several modifications of the codon models are implemented in PAML (Yang 1997) (Yang 2007a), including branch, branch site, clade and site, allowing the detection of variable selection pressure over time, over sites and both. These models are fitted by specifying models and NSsites in the control file codeml.ctl. Likelihood Ratio Test (LRT) is used for testing assumptions (model parameters) through comparison of two competing hypotheses. The above models consider comparisons of nested models, where the null hypothesis (H0) is a restricted version (special case) of the alternative hypothesis (H1). Twice the difference in log-likelihood ($2\Delta l = l1-l0$) is compared with the chi square (χ^2) distribution with the degree of freedom equal to the difference in the number of parameters between the two models (Yang 1998a). The site models allow the dN/dS ratio to vary among sites in the alignment (Nielsen and Yang 1998a) (Yang 2000). These are specified by (model= 0) and variable NSsites = 0, 1, 2, 7, and 8 that will fit five models to the same data in one go. Two pairs of models appear to be particularly useful, forming two likelihood ratio tests of positive selection. The first compares M1a (Nearly Neutral) and M2a (Positive Selection), while the second, more powerfully, compares M7 (beta) and M8 (beta& w) (Wong et al. 2004). The Bayes empirical Bayes (BEB) (Yang et al. 2005) calculates the posterior probabilities for site classes, and is used to identify sites under positive selection if the likelihood ratio test is significant. The free-ratio model (model = 1) assumes an independent dN/dS ratio for each branch. This model is very parameter-rich and its use is discouraged. The model = 2 allows the user to have several dN/dS ratios for different branches (n-ratio) of interest specified by branch level, e.g. in the branch models allow the dN/dS ratio to vary among branches in the phylogeny and are useful for detecting positive selection acting on particular lineages (Yang 1998a)(Yang and Nielsen 1998). For example, the two ratio model estimates dN/dS ratio for the branch of interest (specified in tree by branch label) and it can be compared with the null model one ratio model, which can also be used for neutrality test by comparing with a model with fix omega =1. The branch-site models (Zhang et al. 2005) allow to vary both among sites and across branches to detect positive selection affecting a few sites along particular lineages (called foreground branches). The alternate model (model= 2 NSsites = 2) is compared with the corresponding null model with (fix_omega = 1 and omega = 1). The posterior probabilities of positive selected sites are inferred by BEB. Clade model C is specified by model = 3 and Nssites = 2, while clade model D is specified by model = 3 and NSsites = 3, using ncatG to specify the number of site classes (Bielawski and Yang 2004). The model C can be compared with the null models M1a (Nearly Neutral). M1a test for functional divergence among clades is more prone to false positives under simple evolutionary conditions. The

new null model M2a_rel (NSsites = 22 now specifies the site model M2a_rel) is proposed to better account for among-site variation in selective constraint (Weadick and Chang 2012). The clade model can be sued for testing multiple clades and is useful in infering positive selection (Weadick and Chang 2012).

The other maximum likelihood based methods for detection of positive selection are implemented in Datamonkey (Pond and Frost 2005a) and in HyPhy (hypothesis testing using phylogenies platform; Pond et al. 2005). Single likelihood ancestor counting (SLAC) is a heavily modified and improved derivative of the Suzuki–Gojobori counting approach. Fixed effects likelihood (FEL) is a likelihood-based and statistically rigorous method to fit an independent dN and dS to every site in the context of codon substitution models and test whether dN = dS. Random effects likelihood (REL) allows both dS and dN to vary across sites (Pond and Frost 2005a) Fast Unconstrained Bayesian AppRoximation (FUBAR) (Murrell et al. 2013) is a method much faster and statistically more robust than REL (which can produce misleading results). The imprint of natural selection on protein coding genes is often difficult to identify because selection is frequently transient or episodic, i.e. it affects only a subset of lineages. The mixed effects model of evolution (MEME) takes this into consideration and is a recommended method to find signatures of episodic selection even when the majority of lineages are subject to purifying selection (Murrell et al. 2012). All these methods can be used to complement the positive selection results providing stronger confidence. The protein level approach implemented in TreeSAAP (Woolley et al. 2003) can provide significant information about the positive radical changes in amino acid properties, which can lead to functional and structural variation of a protein. Finally, homology modelling can reveal value information to understand the molecular changes of proteins under positive selection, which can be quite helpful in devising newer applications such as for drug design and medication.

The functional divergence between duplicate genes can also be estimated based on amino acid sites variation present in duplicated genes, which likely responsible for any functional change, e.g. the DIVERGE (Gu et al. 2013) software test and predict type I and type II amino acid patterns in duplicated genes. Type I represents amino acid patterns that are highly conserved in one duplicate cluster but highly variable in the other; these sites may have experienced shifted functional constraints. Type II represents amino acid patterns that are highly conserved within both duplicate clusters but are conserved in a biochemically different state. For example, negatively

charged amino acids may be conserved in one gene and positively charged in the other.

1.0.3 Structure of the thesis

The results obtained during my PhD program titled “Using genomic and proteomic information to characterize the evolution of genes involved in development and adaptation in vertebrates under differential conditions of selective pressure” are presented in form of six chapters. Each chapter has its own introduction, methods, results and discussion sections. The subsequent chapters 2 to 6 provided an in depth exploration of five different gene families across varied species and lineages of vertebrates. The studied gene families were involved in developmetal (KRTAP), sensory perception (OR), detoxification (cGSTs), immune response (TLRs) and reproduction (ZPs) across diverse species and lineages to shed light on how natural selection pressure favours genetic variations at the gene and genome level to cope with morphological/phenotypic innovations and adaptive radiations. Various approaches at the genome, gene and protein level were used for the comparative evolutionary genomics studies. The use of complementary methods provided strong support of the findings obtained.

The first chapter of this thesis focus on the general introduction, giving an overview of the background, materials and methods with relevance to the developed studies. The importance of comparative evolutionary genomics studies is discussed to elucidate the role of evolution in shaping genes and genomes with respect to natural selection and future implication in developing strategties for biodiversity and conservation. In chapter 2-6 we provide in-depth exploration of five diverse gene families.

In chapter 2 we studied KRTAP gene family involved in mammalian hair development and hair phenotypic variations using the model species with characteristic hair phenotypes, e.g. hairless dolphin and hairy sloth. In chapter 3 we explored olfactory receptor gene family in 48 avian and 2 reptilian genomes eclucidating role of ecological adaptation in determining OR subgenomes. In chapter 4 we addressed detoxification enzymne cytosolic Glutathione S-transferase (cGSTs) gene family variation across diverse vertebrate lineages to see how species and lineage specific variation in gene numbers together with positive selection help in protection against the reactive species and xenobiotics. In chapter 5 we studied Toll like receptors (TLRs) gene family involved in vertebrate immune defense system to see how diverse TLRs (viral and non-viral TLRs) are evolving in different lineages. The role of

host-pathogen arms race in shaping the TLRs gene family repertoire and rapid adaptive evolution was also assessed. The chapter 6 deals with ZP (zona pellucida) gene family involved in egg envelop formation and gamete interaction. The objective of this study was to see how ZP gene family is shaped in different lineages of monospermic mammals and polyspermic birds and reptiles .

The KRTAPs play an important role in mammalian hair formation and are responsible for characteristic variations in hair phenotypes. The study found loss of KRTAP gene families in hairless dolpin and expansion in sloth, supporting the important evolutionary role of KRTAP gene family in shaping the hair phenotypes in diverse ecological adaptations. The sensory perception in one of the important communication sector responsible for various signal perception, e.g. vemeronasal receptor =pheromones detection; taste receptors = diet and protection from ingestion of poisonous or harmful substances; olfactory receptors (ORs) = olfaction. The ORs form the largest multigene family in vertebrates with around 1000 genes in mammals. The exploration of OR gene family repertoire in 48 birds and 2 reptilian genomes revealed reduced OR subgenomes in birds as compared to reptiles and suggest important role of ecological adaption in determining OR subgenome. The strong association was found between the olfactory bulb ratio (OBR) and total number of olfactory receptors present in genomes The detoxification system is important for the protection and cleansing of toxins ingested or produced inside the body during metabolic processes. The cytosolic Glutathione S-transferase (cGSTs) gene family involved in the removal and protection against xenobiotic substrates in vertebrates was studied using in depth adaptive analyses in vertebrates, including 48 avian genomes. We found that the gene gain and loss in vertebrates together with positive selection in avian GSTs member is related with adaptive requirement against rapidly increasing diverse variety of xenobiotic compounds. The TLR gene superfamily forms the first line of protection against the invading pathogens and this relationship leads to host pathogen arms race and the coevolution of host and pathogens. Thus, the Toll like receptors (TLR) gene family was studied in vertebrates, with special focus of birds. The exploration of different TLRs (viral and non viral) using in-depth adaptive analysis at codon and proteins level in mammalian and bird lineages revealed rapid evolution of both viral and non viral TLRs suggesting host pathogen arms race in mammals and birds. The adaptive evolution can also influence variation leading to reproductive barriers and ultimately speciation, which is well supported by the rapid evolution of genes involved in reproductive system (e.g. gamete interaction). Thus, the Zona pellucida (ZP) egg envelope subgenome was studied in

vertebrates, particularly in birds that showed the largest ZP subgenome, with most of the gene members influenced by positive selection. Comparison of the ZPs rates of evolution across various lineages revealed that the omega estimates is also higher in mammals and reptiles, suggesting that ZP genes evolve at similar rates in both physiological poly spermic birds and reptiles and mono spermic mammals.

The thesis titled “Using genomic and proteomic information to characterize the evolution of genes involved in development and adaptation in vertebrates under differential conditions of selective pressure” are presented herewith.

I. Comparative genomics and adaptive evolution of genes and gene families involved in phenotypic variation and adaptation

2

Mammalian keratin associated proteins (KRTAPs) subgenomes: disentangling hair diversity and adaptation to terrestrial and aquatic environments

2.1 Abstract

Background

Adaptation of mammals to terrestrial life was facilitated by the unique vertebrate trait of body hair, which occurs in a range of morphological patterns. Keratin associated proteins (KRTAPs), the major structural hair shaft proteins, are largely responsible for hair variation.

Results

We exhaustively characterized the KRTAP gene family in 22 mammalian genomes, confirming the existence of 30 KRTAP subfamilies evolving at different rates with varying degrees of diversification and homogenization. Within the two major classes of KRTAPs, the high cysteine (HS) subfamily experienced strong concerted evolution, high rates of gene conversion/recombination and high GC content. In contrast, high glycine-tyrosine (HGT) KRTAPs showed evidence of positive selection and low rates of gene conversion/recombination. Species with more hair and of higher complexity tended to have more KRATP genes (gene expansion). The sloth, with long and coarse hair, had the most KRTAP genes, of which 141 of 175 were intact. By contrast, the “hairless” dolphin had 35 KRTAPs and the highest pseudogenization rate (74% relative to the 19% mammalian average). Unique hair-related phenotypes, such as scales (armadillo) and spines (hedgehog), were correlated with changes in KRTAPs. Gene expression variation probably also influences hair diversification patterns. For example humans have the identical KRTAP repertoire as apes, but much less hair.

Conclusions

We hypothesize that differences in KRTAP gene repertoire and gene expression, together with distinct rates of gene conversion/recombination, pseudogenization and positive selection, are likely responsible for micro and macro-phenotypic hair diversification among mammals in response to adaptations to ecological pressures.

Keywords

Concerted evolution, gene family, Keratin Associated Proteins, Keratin, Hair, Gene conversion, Recombination, Positive selection

2.2 Background

Terrestrial life in extant vertebrates was accompanied by the formation of diverse and rigid body coverings (scales, feathers and hairs), along with other cornified appendages (e.g. horns, hoofs, claws and nails) that evolved in response to strong selective pressures. These coverings protected vertebrates and allowed them to adapt to environmental pressures, including heat, ultra violet radiation, water loss, and mechanical forces (Alibardi 2003) (Chuong and Homberger 2003). These adaptations involved genes responsible for skin appendage development, such as ectodysplasin-signaling pathway genes (eda, edar, edaradd, xedar and troy) and keratinization genes (Pantalacci et al. 2008) (Alibardi et al. 2009) (Alibardi 2009) as hard keratin appendages were essential for land colonization through the formation of efficient protective barriers. Various small glycine-rich proteins, which likely evolved from progenitor proteins present in basic (reptilian) amniotes (Alibardi 2006), gave rise to the glycine-rich proteins of scales and claws in reptiles, the beak and feathers in birds and the keratin-associated proteins present in mammalian corneous derivatives. It has been suggested that glycine-rich proteins, with similar chemical composition, immunological characteristics, and molecular weight as beta keratins, may represent the reptilian counterpart of keratin-associated proteins present in the hair, nails, hooves, and horns of mammals (Alibardi 2006) (Alibardi et al. 2006). The glycine rich proteins such as HGT in mammals and HGP (high glycine proline) in reptiles and birds had a primary role in the formation of hard protective keratin appendages, contributing to the successful radiation of mammals, reptiles and birds (Alibardi et al. 2009) (Alibardi 2009) (Alibardi 2006) (Alibardi et al. 2006) (Vandebergh and Bossuyt 2011).

Since keratinization protects the body by forming a barrier between the body and the outside world, genes involved in keratinization evolve rapidly in response to changing environments (e.g. evidence of positive selection in the chimpanzee and hominids KRTAP4-5)(George et al. 2011). Changes in gene family composition (gene gain, and gene loss/pseudogenization) have often been linked with adaptive evolution and changes in the number of related genes could affect expression levels (Sun et al. 2012). In terrestrial vertebrates, the formation of hard cornified skin appendages involves interactions between fibrous (keratin) and matrix proteins (KRTAPs) (Alibardi et al. 2009) (Alibardi 2009) (Rogers et al. 2007). The fibrous alpha-keratins, type I and II, appear to have evolved in stem vertebrates (Zimek and Weber 2005) (Alibardi 2006) and recent studies suggest that there are hair-specific alpha keratins

orthologs in amphibians, reptiles, and birds (Alibardi et al. 2011)(Vandebergh and Bossuyt 2011)(Eckhart et al. 2008). Importantly, the structural and functional conservation of keratin intermediate filaments (KIFs) within mammals contrasts with the large diversity of mammalian hair phenotypes (Hesse et al. 2004) (Wu et al. 2008) (Alibardi 2004) and highlights the importance of understanding the molecular diversification of the keratin associated protein (KRATP) multigene family.

Hair is a dynamic mini-organ formed by ectodermal-mesodermal interactions (Botchkarev and Paus 2003) (Millar 2002) (Schneider et al. 2009) (Hardy 1992) and is broadly divided into the root sheath (outer and inner), hair shaft, and matrix zone. Hair has microscopic differences (e.g. cuticular, medullar and cross section), which have long been used as forensic markers for identifying human ethnicity and classifying mammalian species (Franbourg et al. 2003) (Sahajpal et al. 2009) (Bahuguna and Mukherjee 2000) (Jenkins and Powell 1994). Hair-fiber formation is a cyclical process, which involves growth (anagen), regression (catagen), and resting phases (telogen), followed by the shedding of the hair shaft. The process involves the expression of both hair-keratin intermediate filament proteins and their keratin-associated proteins (Shimomura et al. 2002) (Rogers et al. 2008) (Rogers et al. 2002) (Powell et al. 1995)(Pruett et al. 2004). This cycle is of particular importance in diverse processes such as determining hair size, shedding fur for body surface cleansing, and changing the body cover to adapt to changing environments, such as from hot summers to cold winters (Stenn and Paus 2001).

The existing diversity of hair in extant mammals has evolved through innovations and changes in numerous genes and their corresponding proteins. Humans have 54 functional alpha-keratin genes comprising 28 type I and 26 type II keratins (Rogers 2004) (Schweizer et al. 2006) (Hesse et al. 2004) arranged in two clusters on chromosomes 17q21.2 and 12q13.13 (M. Rogers et al. 2004) (Rogers et al. 2005), which include 11 type I and 6 type II hair keratins (Rogers et al.) (Rogers et al. 2000). Hair keratin types I and II undergo higher-ordered copolymerization-forming keratin intermediate filaments (KIFs) (Steinert et al. 1994) (Powell et al. 1991) (Powell and Rogers 1997) (Fujikawa et al. 2012), which are embedded into a matrix formed by keratin associated proteins (KRTAPs) involved in the formation of hard cornified resilient hair shafts (Shimomura and Ito 2005) (Lee et al. 2006) (Koehn et al. 2010). The KRTAP multigene family is divided into two broad groups, high cysteine and high glycine-tyrosine, which together comprise 30 subfamilies based on amino acid composition and phylogenetic relationships (Wu et al. 2008). In humans, KRTAPs include approximately 100 gene members that are arranged in tandem and are

clustered on chromosomes 11p15.5, 11q13.4, 17q21.2, 21q22.1, and 21q22.3 (Rogers et al. 2001) (Rogers et al. 2002) (M.A. Rogers et al. 2004) (Shibuya, Obayashi, et al. 2004) (Yahagi et al. 2004) (Rogers et al. 2007) (Rogers et al. 2008). Given the role of the KRTAP multigene family in the formation of hair morphology, we have characterized them in the genomes of 22 diverse mammalian species to provide insights on KRTAP evolution and diversification. We found contrasting KRTAP gene family repertoires among mammals, as well as differences in rates of gene expansion, contraction and pseudogenization. The two major groups of KRTAPs showed distinct evolutionary patterns with high concerted evolution influencing species-specific copy number variation and gene homogenization in high cysteine KRTAPs. In contrast, high glycine-tyrosine genes had more dynamic evolutionary patterns with less gene conversion and recombination, lower GC content, and evidence of positive selection (e.g. subfamily 20), which may also have been an important force of the evolution in subfamilies of high glycine-tyrosine.

2.3 Results

2.3.1 Genome scans

Advances in genome sequencing have made it easier to explore multigene families across different genomes. Expansion, contraction and pseudogenization, along with genomic/chromosomal organization (gene clusters) of gene families, are important mechanisms driving genome evolution and influencing fitness within lineages or species (Fumasoni et al. 2007), as suggested by lineage- or species-specific variations in genes involved in pathogen recognition, stress response and structural proteins (Lespinet et al. 2002) (Leister 2004) (Zhang 2003a) (Lynch and Force 2000). Here, we explored the KRTAP multigene family in the genome assemblies of 22 mammalian species (Figure 2.1 and Additional File 2.1) including: (1) alpaca (*Vicugna pacos*) low-coverage 2.51X, assembly, vicPac1, Jul 2008, (2) armadillo (*Dasypus novemcinctus*) low-coverage 2X, assembly, dasNov2, Jul 2008, (3) bushbaby (*Otolemur garnettii*) low-coverage 1.5X, assembly, otoGar1, May 2006, (4) cow (*Bos taurus*) coverage 7X, assembly Btau_4.0, Oct 2007, (5) dolphin (*Tursiops truncatus*) low-coverage 2.59X, assembly, turTru1, Jul 2008, (6) elephant (*Loxodonta africana*) coverage 7X, assembly, Loxafr3.0, Jul 2009, (7) gibbon (*Nomascus leucogenys*) whole genome coverage 5.6x, assembly, Nleu1.0, Jan 2010, (8) gorilla (*Gorilla gorilla*) gorGor3, Dec 2009, (9) guinea pig (*Cavia porcellus*) high-coverage 6.79X, assembly, cavPor3, Mar 2008, (10) hedgehog (*Erinaceus europaeus*) low-coverage 1.86X, assembly, eriEur1, Jun 2006, (11) horse (*Equus caballus*) coverage

6.79X, assembly, Equ Cab 2, Sep 2007, (12) marmoset (*Callithrix jacchus*), NCBI build 1.1, (13) megabat (*Pteropus vampyrus*) low-coverage 2.63X assembly, pteVam1, Jul 2008, (14) mouse lemur (*Microcebus murinus*) low-coverage 1.93X, assembly, micMur1, Jun 2007, (15) orangutan (*Pongo abelii*) NCBI build 1.2, (16) panda (*Ailuropoda melanoleuca*) high-coverage, assembly, ailMel1, Jul 2009, (17) pig (*Sus scrofa*) from NCBI build 3.1, high-coverage, assembly, Sscrofa10, Jun 27, 2011, (18) rabbit (*Oryctolagus cuniculus*) high-coverage, assembly, oryCun2, Nov

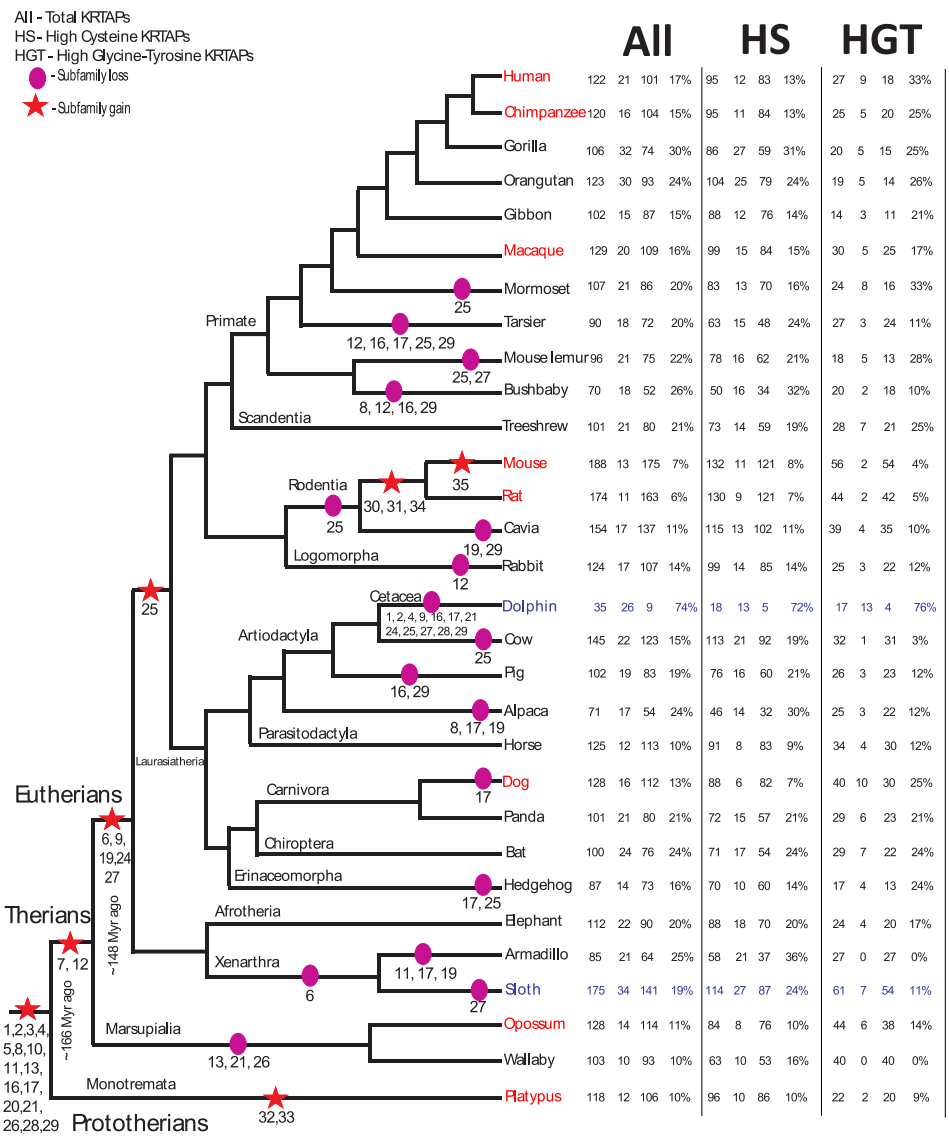


Figure 2.1: The topological tree representing evolution of KRTAP gene family repertoires in 30 mammalian species. Twenty two from the present study and eight from Wu et al, 2008 (Reference 16) marked in red). Stars and circles respectively show the gain and loss of subfamilies, represented by numbers below. The number mentioned under each headings ALL, HS and HGT are number of total, pseudogene, intact genes and percentage of pseudogene respectively.

2.3.2 Characterization of KRTAP gene family

The KRTAP multigene family consists of ~100-180 gene members divided into two major classes, High Cystine (HS) and High Glycine/Tyrosine (HGT), which in turn are divided into many subfamilies with unique motifs and sequence repeats. We assigned all KRTAP multigene family members to their respective subfamilies following previously published guidelines (Wu et al. 2008). We built species-specific

phylogenetic trees to classify the gene subfamilies for each genome (Additional file 2.2, Figures 1-21), as well as a phylogenetic tree incorporating all members belonging to the high glycine-tyrosine KRTAP multigene family from 22 genomes (Figure 2.2 and Additional file 2.2, Figure 24). We also observed that one-to-one orthologous relationships diminished as species diverged over time (Additional file 2.2, Figures 22 and 23). We used the amino acid composition, unique motifs and sequence repeats, as well as blast results, to classify intact genes, partial genes and pseudogenes. The most-closely-related subfamilies are generally located in close proximity and in tandem arrangements in the genome, as are the members of the same subfamily. The low coverage of these genomes has only a limited effect on the overall KRTAP findings because these are intronless genes of ~1000bp and even with a small number of overlapping reads it is possible to cover the entire gene. Thus the methods we used provide highly reliable approximation of KRTAP gene family evolution. The representation of pseudogenes found in low and high coverage genomes do not seem to be biased, as the average percentage of pseudogenes in both types of covered genomes was almost similar, 19% and 20% respectively (excluding dolphin and varying from 9.6% to 30% across all mammalian species). This suggests that the pseudogene percentage found in the dolphin genome (74%) is likely a real evolutionary feature. We found high number of genes in low coverage genomes, such as in sloth (2.05X) and wallaby (2X) with 175 and 103 genes, respectively, similar to the high number of genes found in mammalian species with high coverage genomes (e.g. pig and panda with 102 and 101 genes, respectively). Both dolphin (2.59x) and sloth (2.05x) genomes were of low coverage, but showed contrasting differences in the KRTAP repertoire that were correlated with differences in the species-specific hair phenotypes rather than the degree of genome coverage. Together, these results support the premise that low coverage genomes are suitable for the study of such gene families as has been previously suggested (Young et al. 2010), unlike genes with multiple introns that are difficult to study in low-coverage genomes. However, these results are approximate, rather than fully accurate, as most of these genes are tandem duplicates, which could only be characterized in detail with well-finished genomes, such as human and mouse, which would be unfeasible for most of the other genomes in the near future (Young et al. 2010).

To prove the absence of KRTAP genes in high coverage genomes with intact gene clusters, we performed synteny analysis and searched for human orthologs that should be flanking the missing KRTAP genes. For example, in pig the 5' and 3 ' human orthologs flanking the KRTAP cluster 5 was missing, indicating that this

region has most-likely not been sequenced and that further research and/or higher genomic coverage is needed for confirmation. We verified the synteny of conserved orthologs flanking the missing genes for the subfamily KRTAP25 in the callitrix, cow and elephant, along with KRTAP25, KRTAP19 and KRTAP29 in cavia, and KRTAP12 in rabbit.

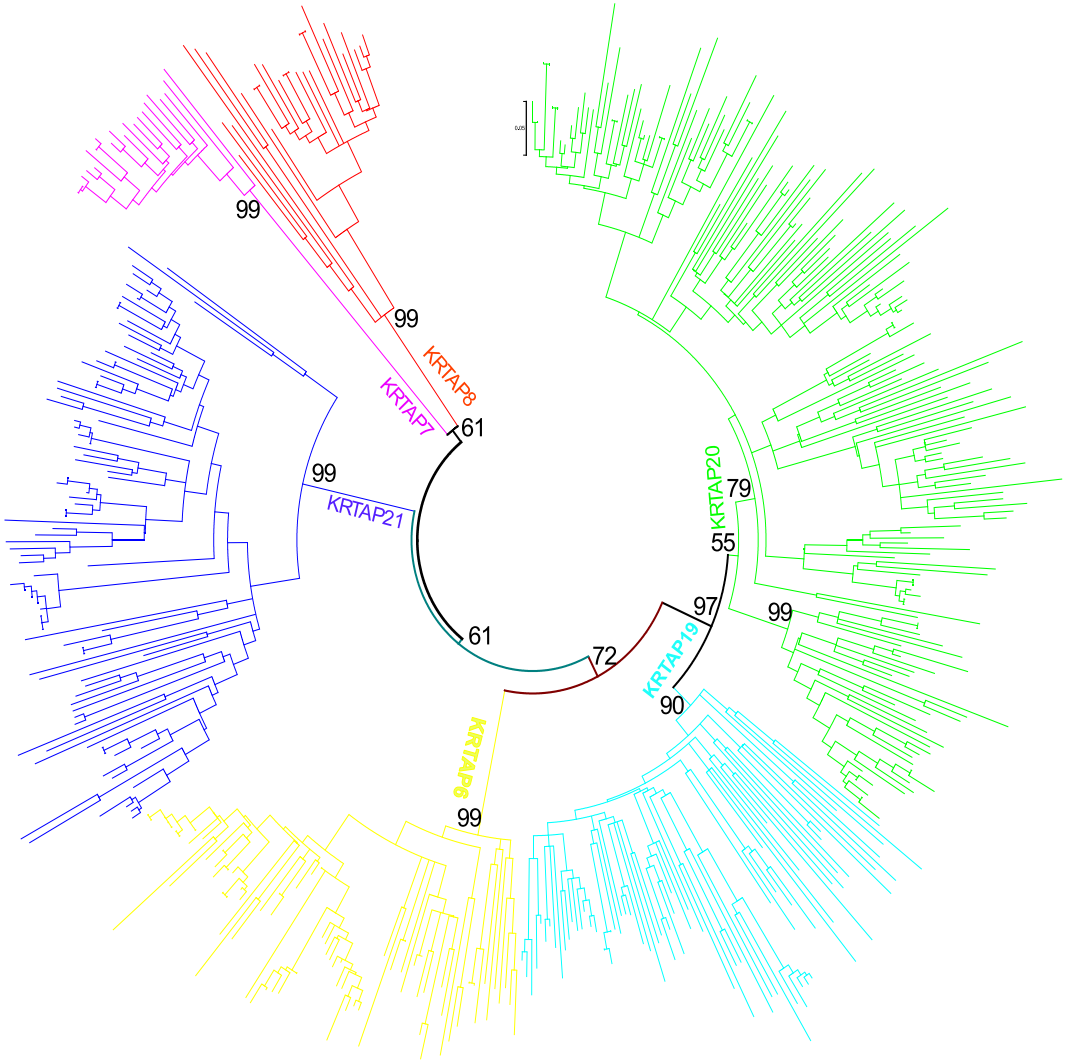


Figure 2.2: The phylogeny of all high glycine-tyrosine gene family members of 22 mammalian genomes. Neighbor-joining method used with P-distance and interiors branch test with 1,000 replications. The different color represents different subfamilies of high glycine-tyrosine KRTAP.

Species-specific subfamily differences, changes in the total number of genes, functional genes, pseudogenes, amino acid content (changes in sulfur content are responsible for disulphide bonds, which provide rigidity, strength and flexibility to hair) and size polymorphism in genes within subfamilies may be responsible for the species-specific hair characteristics and the marked variability found in hair patterns among mammalian species.

2.3.3 Genomic organization of the KRTAP gene family

The KRTAP gene family consists of 30 subfamilies, 24 of which are high cysteine and six are high glycine-tyrosine. The complete KRTAP gene family is arranged into five clusters at five different genomic locations (Figure 2.3). Each cluster contains members of one or more subfamilies arranged in a tandem array. The genomic organization of the KRTAP gene family is similar in all species studied, with only slight variations. Subfamilies KRTAP 1, 2, 3, 4, 9, 16, 17 and 29 are present in cluster one. All high glycine-tyrosine (HGT) KRTAP subfamilies, together with KRTAP 11, 13, 24-27 subfamilies, form cluster two. Subfamilies KRTAP10 and KRTAP12 form cluster three, whereas cluster four consists of subfamily KRTAP28 and cluster five of subfamily KRTAP5.

Cluster 5 shows some variation. For example in primates, KRTAP cluster 5 is divided into two paralogous gene clusters, most likely through segmental duplication, with both clusters having members of the KRTAP5 subfamily (Figure 2.3.) In all of the other mammals studied, genes of the KRTAP5 subfamily form a single cluster. The KRTAP subfamilies that are clustered together in the genome (Figure 2.3) are phylogenetically closely related (e.g. all subfamilies of high glycine-tyrosine KRTAPs are located in close proximity in cluster 2 represented by HGT in Figure 2.3, which supports their functional relatedness and common ancestry arising from duplications and divergence. The conserved genomic organization of the KRTAP gene clusters over more than 166 Myr (i.e. divergence of therian from the monothermes mammals) (Bininda-Emonds et al. 2007) confirms the strong evolutionary constrain acting on their genomic arrangement (Walsh 2001). The conserved clustering of KRTAPs seems to be related with its ordered expression in follicle (McLaren et al. 1997).

2.3.4 KRTAP Gene family dynamics and hair characteristics

Previously, the KRTAP gene repertoire had been assessed in eight mammalian species (Wu et al. 2008), all terrestrial species with few characteristic differences in hair phenotypes. Here we expanded on previous results by analyzing 22 additional mammalian species consisting of a much more diverse group of mammals including species from different mammalian orders with diverse hair characteristics, such as the armadillo (modified scales), hedgehog (spines), alpaca (fiber), sloth (hosting symbionts in hair crusts) and dolphin (mostly hairless and aquatic) (Chen et al. 2011)(Vincent 2002) (Lichtenstein and Vilá 2003)(Suutari et al. 2010)

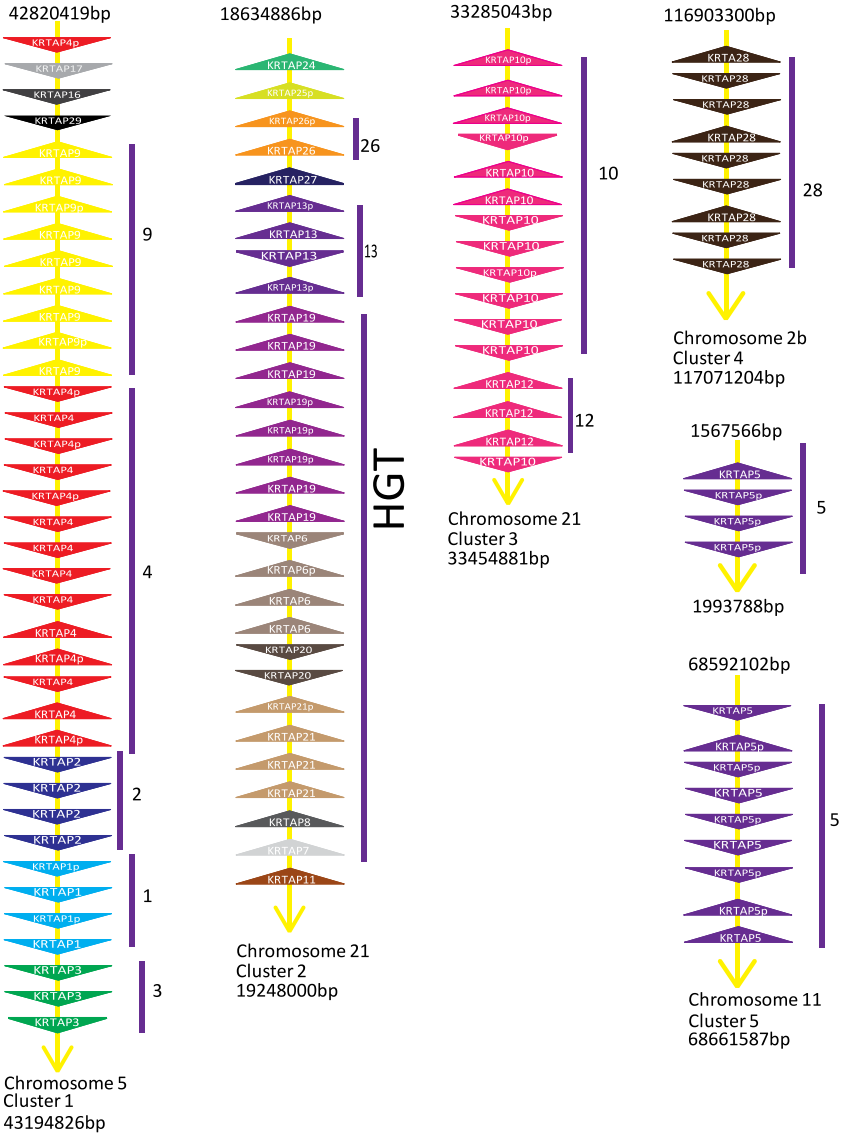


Figure 2.3: Genomic organization of KRTAP gene family in the gorilla genome. The KRTAP gene family is arranged in five different clusters, nucleotide numbers show the chromosomal position of each cluster, with name of cluster and chromosome in which they are present. Each triangle represents a gene member; where p means a pseudogene, same subfamily members are shown with same colors. The triangle points the direction of transcription. The distance between the genes is not to scale.

(Higginbotham et al. 2014) (Thewissen et al. 2009), along with several more-closely related species (e.g. members of hominidae family in primates).

We identified near complete KRTAP gene repertoires in 22 mammalian genomes, including 11 high-coverage genomes (Table 2.1, Figure 2.1, and Additional File 2.1). Our findings suggest that the most recent common ancestor of mammals is supposed to have had 53% (16 of 30) of the known KRTAP subfamilies (1-5, 8, 10,

11, 13, 16, 17, 20, 21, 26, 28, and 29) (Figure 2.1). Extant monotremes (platypus) and marsupials (opossum and wallaby) have slightly different subfamilies representation (60%, 18 of 30 subfamilies, and 50%, 15 of 30, respectively), while eutherians have up to 93% (28 of 30) of the KRTAP subfamilies. This shows that the diversification of the KRTAP gene family occurred early in mammalian evolution, likely starting after the split of sauropsids (leading to birds and reptiles) and synapsids (leading to mammals-like reptiles) around 350 Myr ago (Warren et al. 2008) (Bininda-Emonds et al. 2007). Sauropsids developed beta keratins present in hard appendages like feathers, beaks, scales and claw, etc., and synapsids developed mKRTAPs present in hair, nails, hoofs, claws, etc. It has been suggested that glycine-rich proteins with the chemical composition, immunological characteristics, and molecular weight of beta keratins may represent the reptilian counterpart of keratin-associated proteins present in hairs, nails, hooves, and horns of mammals (Alibardi et al. 2006) (Alibardi 2006) (Alibardi et al. 2009). The evolution of glycine rich proteins such as HGT in mammals and HGP (high glycine proline) in reptiles and birds is related with the formation of hard keratin appendages and contributed to the successful radiation of reptiles, birds and mammals. Further expansion and diversification of the KRTAP gene family, favored by high rates of concerted evolution in HS-KRTAPs and positive selection in HGT-KRTAPs, led to the species-specific hair characteristics observed in extant mammals. Additional analyses of sauropsida and the mammalian keratins and associated proteins are likely to reveal insights into the patterns of adaptive radiation present in extant reptilian, birds and mammalian keratin derivatives. Subfamilies 7 and 12 first appear in therian mammals after their divergence from monothremes around 166 Myr ago (Bininda-Emonds et al. 2007). Subfamilies 6, 9, 19, 24, and 27 are specific to placental mammals (eutherians), and thus appeared after their divergence from marsupials around 148 Myr ago (Bininda-Emonds et al. 2007). Subfamily 25 is absent in Afrotheria and Xenarthra, which suggests an origin within placental mammals only after the divergence from the atlantogenata clade (Figure 2.1 and Table 2.1). Monotremes and marsupials lack subfamily 9, which we observed to have expanded dramatically in the basal placental mammal xenartha (sloth) to 50 members. We noted that the KRTAP gene family shows species-specific variation as expected due to concerted evolution, and some of the subfamilies are restricted to particular species, namely subfamilies 30, 31, and 34 are present only in mouse and rat, subfamily 35 in mouse, and subfamilies 32 and 33 in platypus (Wu et al. 2008) (Figure 2.1 and Table 2.1). We also observed remarkable differences among these KRTAP genes (Table 2.1, Figure 2.1 and Additional File 2.1), including a dramatic

gene expansion with 175 members (50 in subfamily 9 and 37 genes in subfamily 20, respectively) in sloth (*Choloepus hoffmanni)*, a nocturnal hairy mammal with long, coarse and shaggy fur that serves as a host for different microorganism (Suutari et al. 2010) (Higginbotham et al. 2014) (Table 2.1, Figure 2.1, and Figure 2.4). Similarly, we found gene expansion in subfamily 20 (27 genes copies) in the rodent guinea pig (*Cavia porcellus*), and 38 genes copies in the marsupial wallaby (*Macropus eugenii*). Subfamily 28 has expanded in rabbit (*Oryctolagus cuniculus*) (23 genes copies), which belongs to order lagomorpha. We typically observed functional genes in (HS-KRTAPs) subfamilies 11, 16, 17, 24-27 and 29 varying from a minimum of one to a maximum of three. The subfamilies 11, 16, 17 and 25 have a maximum of one functional gene member, subfamilies 24 and 29 have a maximum of two functional genes (present in orangutan and cow, respectively), and subfamily 26 has a maximum of three members (present in sloth and elephant). Subfamily 7, belonging to the high glycine-tyrosine group, has a maximum of one functional gene member (Table 2.1). We found that closely related species, e.g. among Hominidae family (human, chimpanzee, gorilla and orangutan), have very similar gene repertoires with only slight differences (e.g. humans have the highest number of HGT pseudogenes; Figure 2.1). The results obtained from the additional primates species studied (including hominids) support earlier finding that expression level differences could be a cause of hair phenotype variation among primates and of the relative lack of hair in humans (Wu et al. 2008). We also observed the apparent reduction in the KRTAP gene repertoire in alpaca (fiber), armadillo (modified scales), hedgehog (spines), and dolphin (mostly hairless and aquatic) (Figure 2.1, Figure 2.5), probably due to the replacement or modification of hair function with extensive specialization. For example, we observed high rates of pseudogenization (Figure 2.1, Figure 2.4, Figure 2.5 and Table 2.1) (74% compared to the mammalian average of 19%) and only nine intact genes in the dolphin (*Tursiops truncatus*). This aquatic mammal is almost hairless, with only a few hairs (bristles) on the upper lip of the rostrum, which are shed soon after birth, leaving hairless pits on the rostrum of adults that have specialized sensory function (Palmer and Weddell 1964) (Meyer et al. 2012) (Czech-Damal et al. 2011) (Mauck et al. 2000) (Jenkins, J.2009) (Thewissen et al. 2009) (Figure 2.4). The epidermal surface also undergoes high proliferation and sloughing of epidermis cells in order to maintain a smooth skin, a major advantage for swimming (Fish and Hui 1991)(Hicks et al. 1985).

Table 2.1: Number of KRTAP gene present in each subfamily in twenty-two mammalian species

	KRTAP	Gorilla	Pongo	Gibbon	Calllitrix	Tarsius	Mouse lemur	Otolemur	Treeshrew	Cavia	Rabbit	Dolphin	Cow	Pig	Alpaca	Horse	Panda	Bat	Hedge hog	Elephant	Armadillo	Sloth	Wallaby
HS-KRTAP	KRTAP1	4(2)	4(0)	3(0)	4(1)	4(0)	4(0)	4(0)	2(1)	4(0)	4(0)	0(0)	4(0)	4(1)	5(1)	4(0)	3(0)	4(1)	4(1)	4(0)	2(0)	4(1)	4(1)
	KRTAP2	4(0)	3(0)	4(0)	4(0)	3(0)	4(0)	2(0)	5(1)	4(0)	4(0)	0(0)	4(0)	4(0)	3(0)	4(0)	5(1)	3(0)	4(1)	4(0)	4(0)	w3(0)	5(0)
	KRTAP3	3(0)	6(0)	4(1)	4(0)	3(0)	2(0)	4(1)	2(1)	4(0)	4(0)	2(2)	4(0)	4(0)	5(2)	4(0)	4(1)	4(0)	4(0)	4(1)	3(1)	4(0)	4(0)
	KRTAP4	15(6)	15(4)	10(2)	8(1)	8(1)	9(2)	14(3)	12(1)	16(1)	14(1)	0(0)	16(5)	7(0)	7(3)	12(0)	4(1)	6(0)	8(1)	13(4)	9(3)	8(0)	17(4)
	KRTAP5	13(8)	14(6)	15(3)	19(5)	16(8)	8(4)	13(2)	11(4)	18(5)	18(1)	1(1)	18(5)	0(0)	2(2)	11(2)	13(3)	6(4)	5(0)	12(4)	8(7)	8(1)	7(1)
	KRTAP9	9(2)	12(2)	9(1)	5(0)	6(1)	4(1)	8(0)	5(0)	12(2)	6(1)	0(0)	19(3)	5(0)	4(1)	8(2)	2(0)	9(3)	9(1)	4(0)	7(2)	50(5)	0(0)
	KRTAP10	13(5)	17(2)	13(0)	13(2)	8(2)	5(1)	7(1)	10(2)	12(1)	10(2)	2(2)	15(3)	14(1)	1(1)	11(0)	8(1)	14(5)	1(1)	8(1)	3(1)	1(0)	5(3)
	KRTAP11	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	0(0)	1(0)	1(0)
	KRTAP12	3(0)	4(2)	2(0)	3(2)	0(0)	0(0)	3(1)	1(0)	12(1)	0(0)	1(0)	6(0)	4(0)	1(0)	10(1)	5(1)	1(0)	10(1)	11(2)	1(0)	5(2)	7(0)
	KRTAP13	4(2)	8(4)	7(3)	6(1)	4(2)	4(2)	8(0)	6(1)	8(2)	8(3)	10(6)	8(3)	14(10)	8(0)	10(2)	8(1)	12(3)	7(1)	11(5)	9(5)	16(13)	0(0)
	KRTAP16	1(0)	2(1)	1(0)	1(0)	0(0)	0(0)	1(0)	1(0)	1(0)	1(0)	0(0)	1(0)	0(0)	1(0)	1(0)	1(0)	1(0)	2(1)	1(0)	1(0)	2(1)	1(0)
	KRTAP17	1(0)	1(0)	1(0)	1(0)	0(0)	1(0)	1(1)	1(0)	1(0)	1(0)	0(0)	1(0)	1(0)	0(0)	1(0)	1(0)	1(0)	0(0)	1(0)	0(0)	1(0)	1(0)
	KRTAP24	1(0)	2(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	1(0)	0(0)	1(0)	1(1)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	0(0)
	KRTAP25	1(1)	1(1)	1(0)	0(0)	0(0)	1(1)	0(0)	1(1)	0(0)	1(1)	0(0)	0(0)	1(1)	1(1)	1(0)	1(0)	1(1)	0(0)	0(0)	0(0)	0(0)	0(0)
	KRTAP26	2(1)	1(0)	2(0)	1(0)	1(0)	1(1)	1(1)	1(0)	1(0)	1(0)	1(1)	2(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	3(0)	1(0)	4(1)	0(0)
	KRTAP27	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	0(0)	1(0)	1(0)	1(0)	0(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	0(0)	0(0)
	KRTAP28	9(0)	11(1)	12(2)	10(1)	7(1)	4(3)	9(4)	11(3)	19(1)	23(5)	0(0)	10(2)	14(2)	2(2)	9(1)	12(6)	4(0)	11(0)	8(1)	7(2)	5(2)	10(1)
	KRTAP29	1(0)	1(1)	1(0)	1(0)	0(0)	0(0)	1(1)	1(0)	0(0)	1(0)	0(0)	2(0)	0(0)	2(2)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)
	KRTAP30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	KRTAP31	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	KRTAP32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	KRTAP33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	KRTAP34	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	KRTAP35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	TOTAL-HS	86(27)	104(25)	88(12)	83(13)	63(15)	78(16)	50(16)	73(14)	115(13)	99(14)	18(13)	113(21)	76(16)	46(14)	91(8)	72(15)	71(17)	70(10)	88(18)	58(21)	114(27)	63(10)
HGT-KRTAP	KRTAP6	4(1)	3(1)	3(0)	3(0)	8(0)	4(0)	4(1)	5(0)	7(1)	9(1)	1(1)	5(0)	4(0)	9(0)	4(0)	7(2)	6(0)	4(1)	5(1)	0(0)	0(0)	0(0)
	KRTAP7	1(0)	1(0)	1(0)	1(0)	1(0)	2(0)	2(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(0)	1(1)	1(0)	1(0)	1(0)
	KRTAP8	1(0)	1(0)	1(0)	1(0)	2(0)	0(0)	0(0)	1(1)	1(0)	1(0)	3(0)	2(0)	4(1)	0(0)	3(1)	1(0)	1(0)	1(0)	1(0)	1(0)	4(2)	1(0)
	KRTAP19	8(3)	9(4)	5(3)	9(5)	10(3)	5(0)	6(0)	9(2)	0(0)	2(0)	6(5)	8(1)	3(2)	0(0)	8(2)	8(3)	7(2)	4(0)	6(1)	0(0)	7(1)	0(0)
	KRTAP20	2(0)	2(0)	1(0)	2(1)	4(0)	6(1)	5(1)	10(3)	27(3)	8(1)	6(6)	5(0)	5(0)	12(2)	12(0)	4(0)	6(3)	2(2)	6(1)	12(0)	37(3)	38(0)
	KRTAP21	4(1)	3(0)	3(0)	8(2)	2(0)	3(1)	3(1)	2(1)	3(0)	4(1)	0(0)	11(0)	9(0)	3(1)	6(1)	8(1)	8(2)	5(1)	5(0)	13(0)	12(1)	0(0)
	TOTAL-HGT	20(5)	19(5)	14(3)	24(8)	27(3)	18(5)	20(2)	28()7	39(4)	25(3)	17(13)	32(1)	26(3)	25(3)	34(4)	29(6)	29(7)	17(4)	24(4)	27(0)	61(7)	40(0)
	ALL-KRTAP	106(32)	123(30)	102(15)	107(21)	90(18)	96(21)	70(18)	101(21)	154(17)	124(17)	35(26)	145(22)	102(19)	71(17)	125(12)	101(21)	100(24)	87(14)	112(22)	85(21)	175(34)	103(10)

Number of pesudogene is represented in parenthesis.

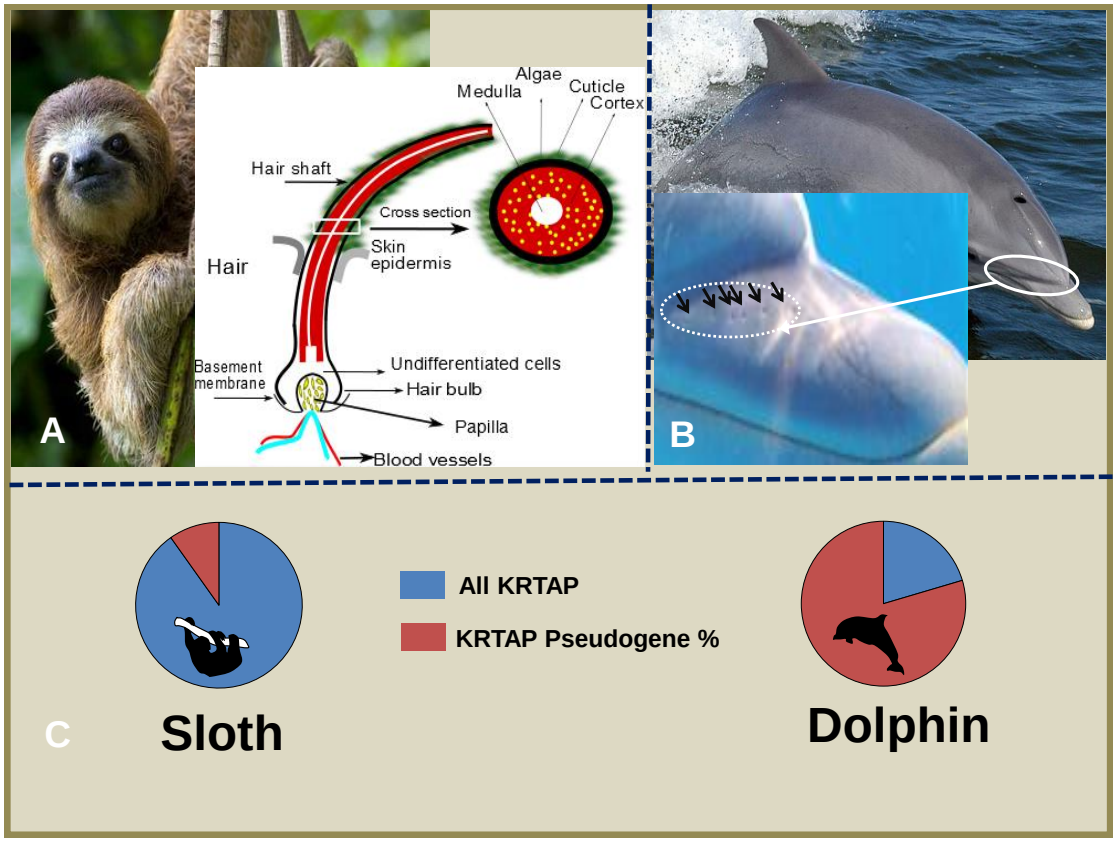


Figure 2.4: Hair characteristic adaption in terrestrial and aquatic mammals. Sloth an arboreal mammal with high density of hair harboring algae. Representation of sloth hair with algal growth and cross section of hair showing the major layers of hair shaft (A). Bottlenose dolphin with rostrum selected in circle and detailed in image with arrows point the hairless vibrissae crypts of dolphin (B). Overall number of KRTAP genes and percentage of pseudogene present in sloth and dolphin (C).

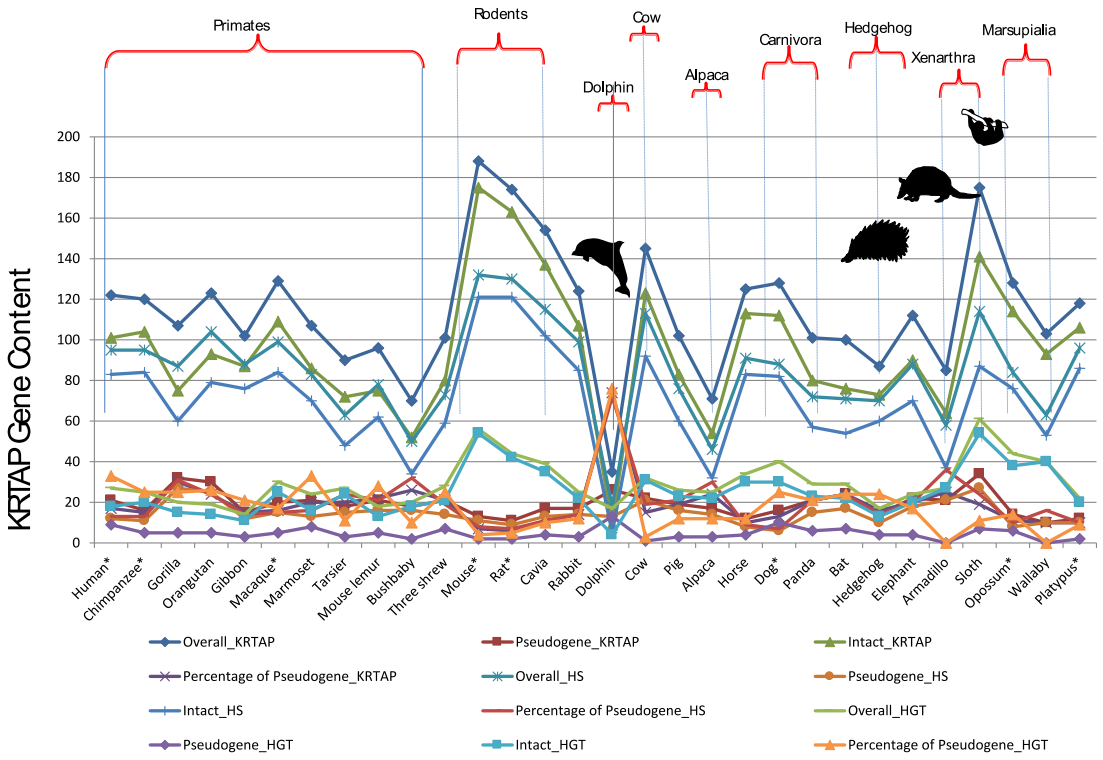


Figure 2.5: Variation in KRTAP gene family in mammals and relation with hair characteristic features.

2.3.5 Concerted evolution, GC bias and sequence divergence

Tandemly arranged gene members of multigene families often show more similarity among each other than with their counterpart orthologs in other species, which suggest that they evolved in similar or concerted fashion. This would further lead to species-specific variation, as observed in KRTAP gene family. Two mechanisms play an important role in concerted evolution. Recombination increases the copy number of gene by providing raw material for further functional innovations and diversification, and gene conversion, which principally homogenizes genes, can help insure the rapid synthesis of a gene product (protein) that may be required during a precise stage of cell cycle. Gene conversion also decreases the evolutionary distance among paralogous members and shifts the substitutions from weak (A or T) to strong (G or C) by increasing GC content through biased gene conversion (gBGC) (Kostka et al. 2011) (Duret and Arndt 2008)(Escobar et al. 2011) (Galtier and Duret 2007). The negative correlation between evolutionary distance calculated by synonymous substitution rates and GC content provides the level of divergence between the members of a subfamily (Noonan et al. 2004).

Table 2.2: Amino acid composition of KRTAPs subfamily genes in mammals

KRTAP	Subfamily	Cysteine	Glycine	Leucine	Proline	Glutamine	Serine	Threonine	Tyrosine
HS-KRTAP	KRTAP1	25.98	9.68	1.83	9.26	7.16	13.89	9.41	1.70
	KRTAP2	28.83	4.35	1.59	14.50	5.29	9.17	9.28	0.52
	KRTAP3	19.65	3.92	8.05	15.43	3.12	8.51	10.97	1.29
	KRTAP4	34.84	3.29	1.15	9.87	6.14	17.60	7.19	0.72
	KRTAP5	33.67	23.00	0.35	5.02	3.74	20.58	0.28	0.36
	KRTAP9	35.30	3.57	1.17	11.64	7.41	12.57	12.79	1.81
	KRTAP10	25.92	1.95	3.66	13.79	5.85	18.86	4.85	0.73
	KRTAP11	12.84	6.86	3.79	8.25	7.16	16.22	10.81	2.64
	KRTAP12	22.54	4.26	4.15	13.85	6.24	17.94	4.27	1.11
	KRTAP13	11.11	10.67	6.62	7.25	3.83	21.93	5.99	7.71
	KRTAP16	19.24	1.89	2.54	14.12	5.33	17.40	5.94	2.24
	KRTAP17	36.09	28.93	0.00	5.29	3.73	10.37	2.83	0.15
	KRTAP24	9.73	5.89	7.90	9.49	4.16	17.18	6.87	6.55
	KRTAP25	8.17	3.92	6.86	8.82	8.17	16.99	1.96	7.52
	KRTAP26	10.38	6.58	9.55	11.13	4.08	18.98	6.24	4.36
	KRTAP27	8.87	4.33	6.69	7.97	8.32	17.15	7.90	2.04
	KRTAP28	38.43	30.92	0.04	1.78	4.19	7.71	2.68	1.20
	KRTAP29	16.25	6.13	3.74	11.10	7.96	16.66	6.92	2.29
HGT-KRTAP	KRTAP6	14.37	39.89	5.40	0.11	0.02	6.96	0.16	22.07
	KRTAP7	7.82	20.26	5.99	6.85	0.29	11.93	5.65	11.76
	KRTAP8	6.15	22.47	4.52	6.64	0.14	8.83	2.69	19.15
	KRTAP19	8.07	35.88	4.51	1.53	0.10	10.39	0.33	18.18
	KRTAP20	10.62	32.08	4.97	2.22	0.13	7.01	0.54	25.05
	KRTAP21	17.88	34.32	0.76	1.76	0.16	11.50	1.77	21.10
Average percentage of eight amino acids									

Using Geneconv (Sawyer 1989) and RDP3 (Martin et al. 2010) we found higher rates of gene conversion and recombination events in the high cysteine KRTAPs compared with the high glycine-tyrosine KRTAP genes (Additional file 2.3 and 2.4). KRTAP subfamilies also displayed different rates of gene conversion in different species. For example, in gorilla we found 44 gene pairs of the KRTAP10 under gene conversion, compared to only 17 gene pairs found for this subfamily in gibbon (Additional file 2.3). The high level of gene conversion also reduces orthologous relationship between genes of two different species.Sequences with higher synonymous substitution rates (dS) had higher overall GC content (GC %) and third-codon GC content (GC3%), and lower synonymous substitution rates in the high-cysteine genes than the high glycine-tyrosine genes. The negative correlation between GC content and synonymous

substitution rate (dS) is consistent with the higher rates of concerted evolution observed in high cysteines (Figure 2.6 A and B). The high GC content in the HS – KRTAP gene family compared with the HGT-KRTAP could be a consequence of the high number of gene conversion events.

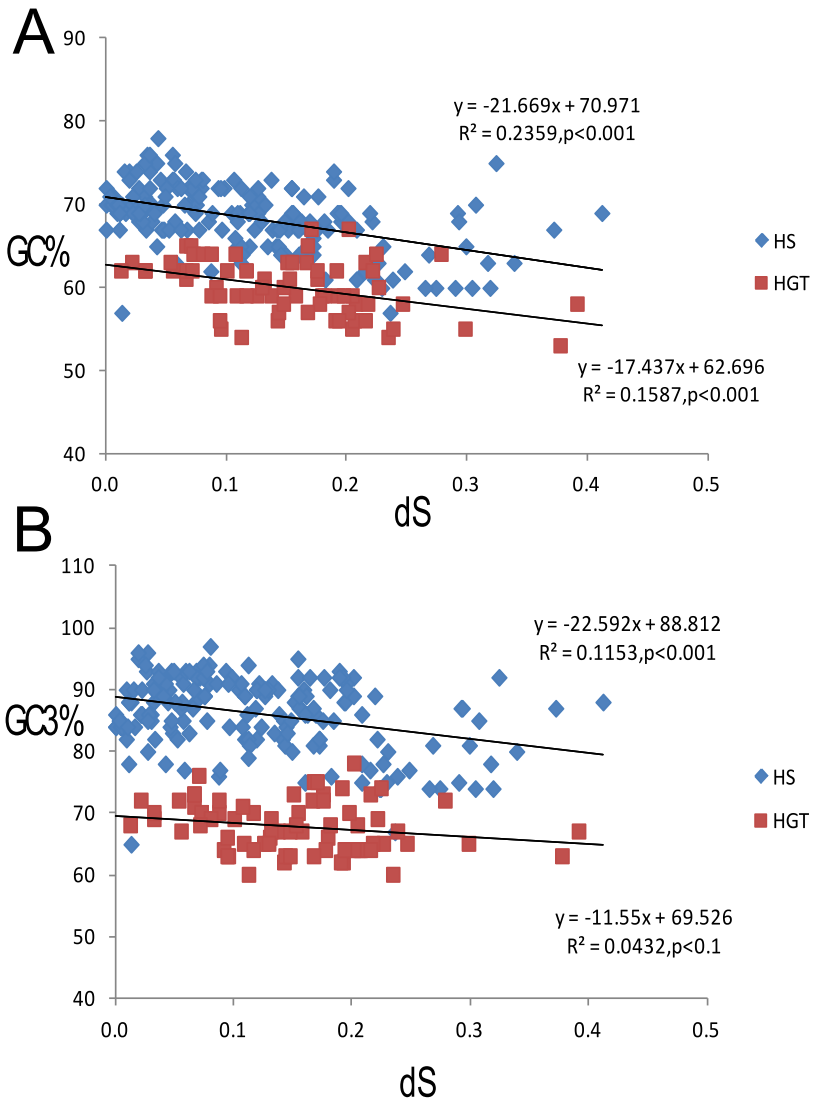


Figure 2.6: GC-content dynamics. GC-biased gene conversion (gBGC) and evolutionary distance between the KRTAP genes, shown by the correlation between the synonymous substitution rates (dS) and GC content (GC%) among paralogous members of each subfamily (A) and third codon GC content (GC3%) (B). Negative correlation points towards the gene conversion. High cyteine KRTAP (HS) and high glycine-tyrosine KRTAP (HGT) are represented by blue and red squares respectively. The linear regression is shown.

2.3.6 Adaptive evolution

Gene expansion provides the essential raw material for positive selection to act (Han et al. 2009), which in turn accelerates the diversification of duplicated copies by increasing the number of nonsynonymous substitutions (dN) relative to the synonymous substitutions (dS) through positive selection (dN/dS>1). The PAML package (Yang 2007b) was used to identify signatures of positive selection.

Specifically, we used likelihood ratio test for positive selection (Yang 1998b) (Nielsen and Yang 1998b) to test site-specific models comparing twice the difference in log-likelihood between two models to chi square distribution with two degrees of freedom. For expanded subfamilies, such as in the case of the KRTAP20 in wallaby with 38 members, we tested if this species-specific expansion has been influenced by adaptive evolution. We tested two nested pairs of site-specific models (M1a vs. M2a and M7 vs. M8), where M1a and M7 state no positive selection ($\omega \leq 1$) and M2a and M8 indicate positive selection ($\omega \geq 1$). In both cases the likelihood of positive selection was significantly higher ($p<0.0001$), retrieving similar sites under positive selection. The likelihood ratio test is a conservative approach, which can be biased by false positives in the presence of high recombination rates (Anisimova et al. 2003). Thus, we evaluated the possibility of gene conversion/recombination in the KRTAP20 subfamily that has expanded dramatically in the wallaby genome, but did not detect such evidence. The results of positive selection tests are shown in Table 2.3. The positive selection acting on KRTAP probably favored the diversification and adaptation to different environments. We also performed positive-selection analyses for other KRTAP genes (genes free from gene conversion and recombination), but found no significant results (possibly due to less number of sequences available). Future studies with increased number of taxa would help unravel the patterns of positive selection signatures in other KRTAP genes.

2.3.7 Differential evolution of the HS and HGT KRTAPs

The KRTAP multigene family has experienced dynamic evolution and diversification within and among genomes as observed in the 30 diverse subfamilies of high cysteine and high glycine-tyrosine subfamilies. These two groups have evolved differently, with the high cysteine group showing high rates of gene conversion within subfamilies, with some exhibiting characteristic differences in copy number, while others have been more conserved. This may be an adaptive mechanism promoting a high order of amplification of similar copies to meet the high demand for the structural proteins required to adapt to changing environmental conditions (e.g. sloth have extensive hairs that can harbor microbial communities including symbionts, while the dolphin is “hairless” in response to a more-predictable and constant environment and to create less resistance when swimming). We also compared the differential evolutionary patterns between high cysteine and high glycine–tyrosine genes using the Pearson correlation coefficient for the number of genes in each subfamily between species. The coefficient value for high cysteine is significantly higher than for high glycine-tyrosine (Figure 2.7A) and the coefficient values for the two are positively correlated ($p<0.001$)

Table 2.3: Likelihood ratio test for PAML site models within Wallaby

Model	Parameters	lnL	2ΔlnL (LRT)
M0	ω: = 0.65464	-1673.250363	NA
M1a	p_0: 0.57003 p_1: 0.42997 ω0: 0.09616 ω1: 1.00000	-1576.668729	
M2a	p_0: 0.49130 p_1: 0.31322 p_2: 0.19547 ω_0: 0.10488 ω_1: 1.00000 ω_2: 3.48389	-1556.441999	M1a vs. M2a 40.45346 (p = 1.643E-09)
M7	p= 0.33061 q= 0.37447	-1581.277897	
M8	p_0= 0.74935 p= 0.55372 q= 1.09725 (p_1= 0.25065) ω_1= 2.75722	-1560.301816	M7 vs. M8 41.952162 (p = 7.765E-10)

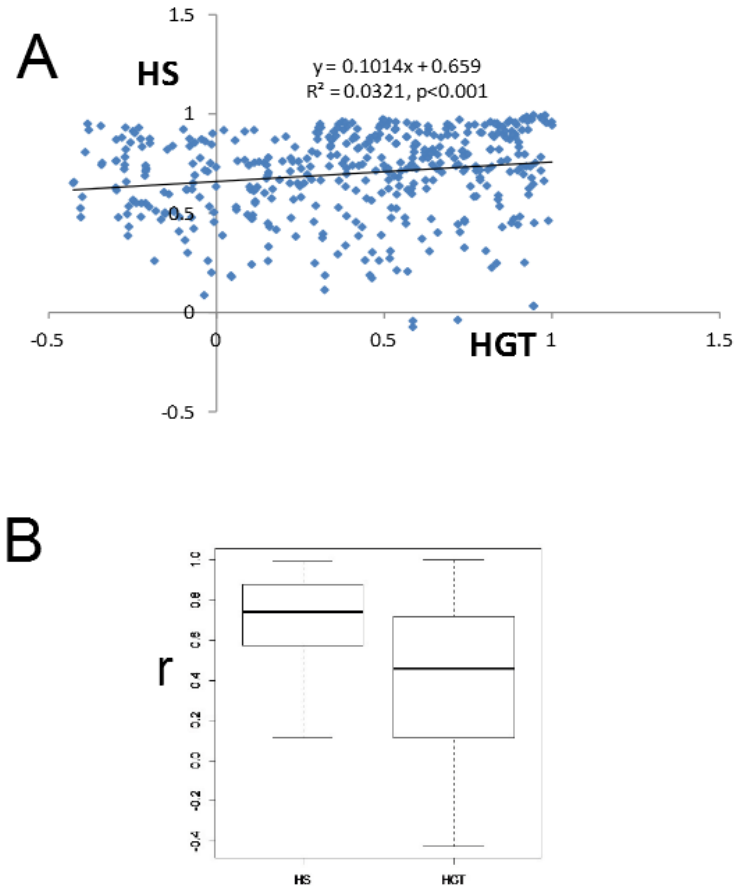


Figure 2.7: Pearson correlation coefficients (r) show the evolutionary differentiation of KRTAP genes. Pearson correlation coefficients (r) values of the high cysteine and high glycine-tyrosine KRTAP are positively correlated. The linear regression is shown. (A) The boxplot for Pearson correlation coefficients (r) of gene numbers of each subfamily between species shows, high cysteine KRTAP genes have higher correlation coefficient than high glycine-tyrosine KRTAP genes (B).

(Figure 2.7B). The high GC content and negative correlation between GC content and synonymous substitution rates also support the higher rates of gene conversion observed in high cysteine genes relative to high glycine-tyrosine KRTAPs, suggesting that high cysteine are under high rates of concerted evolution promoted by gene conversion and recombination events (see Additional files 2.3 and 2.4). By contrast, HGT-KRTAP had a more-dynamic evolutionary pattern, with less evidence of gene

2.3.8 Size Polymorphism and amino acid composition affects KRTAP matrix formation and interactions with hair KIFs

The KRTAP family is widely grouped into three major categories based on amino acid composition: (i) high sulfur (<30% cysteine content), including subfamilies 1, 2, 3, 10-13, 16, 24-27, 29, 31, 34 and 35; (ii) ultrahigh sulfur (>30% cysteine content), including

subfamilies 4, 5, 9, 17, 28, 30, 32 and 33; and (iii) high glycine/tyrosine, including subfamilies 6, 7, 8, 19, 20 and 21. The amino acid composition is shown in Table 2.2. Subfamily gene members also showed size polymorphism (Kariya et al. 2005) (Parry et al. 2006) (Rogers and Schweizer 2005) mostly due to cysteine-rich repeats, which create difference in cysteine content. Cysteine is important for the formation of strong disulphide bonds. Thus, changes in cysteine composition can result in differential interaction among KRTAPs and between KIFs and KRATPs leading to combinatorial complexity and thereby creating morphological differences in hair fiber strength, rigidity and flexibility (Shimomura and Ito 2005).

2.4 Discussion

Gene families are formed by gene duplication, a process that provides important raw material for functional innovation and adaptive selection. Gene families vary in size from a few to thousands of gene members, which makes it difficult to identify and characterize them without sufficient genome sequences. The genome sequencing projects have made it possible to explore complex gene families involved in different phenotypes. Here we explored the mammal-specific KRTAP gene family, which is the major constituent of the hair proteome and plays a primary role in hair formation and thus long been associated with phenotypic differences in hair and wool. This study assessed patterns of variation using comparative genomic approaches in the KRTAP gene family. Our study used 22 diverse mammalian genomes that encompassed closely related species, such as the family Hominidae of primates, comparing apes with dense hair cover with human with much less hair cover, along with species with diverse hair related characteristic, such as alpaca (fibre), armadillo (modified scales), hedgehog (spines), sloth (hosting hair symbionts) and dolphin (mostly hairless and aquatic), to obtain greater insights into the KRTAP gene family evolution relative to mammalian hair and phenotypic variations.

We found high molecular diversity within the KRTAP gene family, with 30 subfamilies (24 belonging to high cysteine and six belonging to high glycine-tyrosine KRTAP) (Additional file 2.1 and Table 2.1) and approximately 100-180 KRTAP gene members, which are arranged in five clusters at five different chromosome locations in a genome (Figure 2.3). Most KRTAP subfamilies are found in all mammalian orders, with variations in expansion, contraction, presence/absence, different rates of pseudogenization and sequence variation (length polymorphisms and amino acid changes). For example, we found species-specific differences in the size and compositions of some subfamilies (e.g. subfamilies 4, 5 and 9) probably caused by

unequal crossing over accompanied with high GC content. Moreover, we also found lineage-specific trends, such as in marsupials, where both wallaby and opossum lacked subfamilies 13, 21 and 26 and showed expansion of the KRTAP20 subfamily, which is under positive selection (Table 2.3). However, highly conserved sequences and the maintenance of the same number of members in subfamilies 1, 2, and 3 suggests that high rates of gene conversion maintain homogeneity and with evolution occurring through a process of punctuated equilibrium (Wu et al. 2008) (Gould and Eldredge 1993) (Mattila and Bokma 2008). Similarly, the conserved synteny of KRTAP gene clusters shows that there are also strong constraints acting on this gene family and supports the important role of KRTAP gene family in shaping hair characteristics.

Together with the high molecular diversity of the KTRAP gene family observed in our study, considerable intraspecies diversification has been reported with copy number variations in ethnic human populations and allelic variations in sheep (Marotta et al. 2012) (Gautam et al. 2012) (Gong, Zhou, Yu, et al. 2011) (Gong et al. 2010) (Gong, Zhou, and Hickford 2011) (Zhou et al. 2012). The high polymorphism of KRTAPs (e.g. length polymorphism and amino acid changes) may influence its expression, protein structure, and/or post-translation modifications. This will subsequently effect the binding patterns of KRTAP and KIF, influencing wool/hair fiber structure and wool/hair quality traits (Yu et al. 2009) (Jenkins and Powell 1994) (Gong, Zhou, Dyer, et al. 2011) (Liu et al. 2011). Indeed, changes in cysteine composition can result not only in differential interactions among KRTAPs, but also between KRATPs and KIFs, leading to combinatorial complexity and thereby creating morphological differences in hair fiber strength, rigidity and flexibility. Evidence of linkage reported between KRTAP6-8 and wool fiber diameter (quantitative trait) in sheep (Parsons et al. 1994) may be related with similar characteristics in alpaca (fiber), which has one of the largest number of KRTAP6 genes (n = 9) in mammals. Further exploration of KRATP gene family in sheep could help shed light on the improvement of hair/wool traits (McLaren et al. 1997) (McKenzie et al. 2010)(Parsons et al. 1994)(Purvis and Jeffery 2007).

Interestingly, we found differences in KRTAP gene repertoire related with hair features. A very expanded KRTAP gene family repertoire (175 total genes and 141 intact genes) was found in sloth, an arboreal mammal with long, dense and coarse body hair cover, providing increased surface area for hosting symbiotic microorganisms (green algae) in hair crusts (Suutari et al. 2010) (Higginbotham et al. 2014). By contrast, we have detected a reduced number of functional KRTAP genes and high percentage of KRTAP pseudogenes (74%) in dolphin (aquatic mammal), highlighting the much lower KRTAP gene requirement in this smooth-skinned species that only has a few hairs (bristle) at

the rostrum (Thewissen et al. 2009) (Palmer and Weddell 1964) (Meyer et al. 2012). These are lost soon after birth and in adults the hairless pits are adapted for sensory functions (Figure 2.4), illustrating the adaptive potential of hair follicles to diversify into more specialized sensory organs (Czech-Damal et al. 2011) (Mauck et al. 2000).

We also observed that several unique hair-related phenotypes in some of the species, such as scales in armadillo, fiber in alpaca and spines in hedgehog, are linked with an inverse correlation between the number of intact KRTAP genes and the number of pseudogenes. The mammalian species with unique and modified protective hair phenotypes like scales (armadillo) and spines (hedgehog), with less hair requirements, had less KRTAPs (Chen et al. 2011)(Vincent 2002). Alpaca have long been selected for economic importance of its fiber, raising the possibility that the reduced number of genes in alpaca could be due to inbreeding and genome homogenization during domestication (Lichtenstein and Vilá 2003)(O'Brien et al. 2010) . The "hairless" dolphin had a large number of KRTAP pseudogenes relative to intact genes (Figure 2.5). In contrast, the sloth showed a high positive correlation with intact KRTAP genes, suggesting that changes in KRTAP can be related to morphological diversity of hair phenotypes (Figure 2.5). Although the number of KRTAP genes in sloth is similar to those found in rat and mouse, they differ in the composition of the KRTAP subfamilies. The KRTAP subfamilies 30, 31, 34 and 35 are exclusively present in the mouse and rat genomes (absent in all the other mammalian genomes) (Wu et al. 2008). The increased number of KRTAP genes in mouse and rat may have favored heat insulation adaptation by using hair cover to maintain constant body temperature in these nocturnal mammals (Wu et al. 2008), especially given their small body sizes, which causes more-rapid heat dissipation given their large surface area to volume ratio (Ruben and Jones 2000). The observed differences suggest that different KRTAP genes have unique specialized roles.

In contrast, we did not find any correlation between the comparatively hairless human and other primates, as has been recognized previously (Wu et al. 2008). This favors the hypothesis that diversification of keratinization structures in mammals is not only explained by the variation of KRTAP gene numbers, but also by other biological mechanisms generating diversity, such as gene expression differences (which can be further influenced by the polymorphism of KRTAP genes).

We suggest that the diverse repertoire and variability in KRTAPs (at gene, family and genome level) provides extraordinary combinatorial complexity (Henikoff 1997) for interaction between KRTAPs and Keratin intermediate filaments, resulting in a rich

diversity of pathways for evolutionary change, which together with differences in higher order expression of KRTAP genes results in the diverse hair morphological characteristic visible in extant mammals. Overall, we conclude that KRTAPs play an important role in evolution and diversification of hair character across mammals and are responsible for unique features of hair.

2.5 Conclusions

The present study explored KRTAP gene family evolution in various mammalian species inhabiting diverse terrestrial and aquatic environments. The two groups of the KRTAP gene family, high cysteine and high glycine-tyrosine KRTAP genes, have evolved differently, resulting in species-specific diversification of this multi-gene family and leading to wide morphological diversity in hair characteristics in extant mammals. We conclude that differences in KRTAP gene family repertoires, together with changes in expression patterns, are responsible for shaping unique hair characteristics in diverse mammalian species. These differences are more pronounced between aquatic and terrestrial species and demonstrate the important adaptive role of hair in terrestrial colonization and the radiation of mammals from water to land. Future studies comparing the KRTAP repertoire in key model organisms, such as alpaca and sheep, may provide insights to understanding the role of KRTAP gene variations in hair fibre traits and its use in textile industry.

2.6 Methods

2.6.1 Gene Identification

All KRTAP genes are relatively small (ca. 1 kb) and generally have single exon (Rogers and Schweizer 2005). Some KRTAP genes appear to possess small introns. However these are similar to repeat regions present in the gene (Shibuya, Kudoh, et al. 2004) and can be translated in-frame with the coding exon, leading to the conclusion that all KRTAP are intron-less (Wu et al. 2008) (Wu et al. 2009). The presence of KRTAP gene clusters in mammalian genomes makes it easy to identify and fully characterize the gene family in genomes with high coverage, but in low coverage genomes it requires much more manual inspection and in-depth screening to insure an almost complete or maximum possible repertoire of non-redundant KRTAP genes (Additional file 2.1). In order to identify the complete gene repertoire in the KRTAP gene family, all previously annotated gene sequences were taken and used as query in blast searches against

the genomes from ensemble http://www.ensembl.org/Multi/blastview and NCBI genome data base http://www.ncbi.nlm.nih.gov/blast/Blast.cgi using BLASTN algorithm (Altschul et al.) and E-value cut-off of 10 . We retrieved multiple hits for each query and selected all the non-redundant hits by extending 500 bp at both 5' and 3' ends. Non-redundant hits, which were seen to be clustered in the same region (chromosome, contig, genescaffold, scaffold, supercontig) were merged together to form a single extended common DNA fragment, bearing all these hits and the ends of this fragment were further extended to maximum 0.3 Mbp were ever possible. Finally all the hits were used to identify and annotate KRTAP gene using program BLAST 2 Sequences (Tatusova and Madden 1999) TFASTX and TFASTY incorporated in Fasta programs (Pearson et al. 1997) and ORF finder from NCBI http://www.ncbi.nlm.nih.gov/gorf/gorf.html and Mobyle (Néron et al. 2009). The identified genes were blast searched against non redundant NBCI blast data base, all best hits which resulted in KRTAP or KRTAP like sequences were finally taken as KRTAP genes. The KRTAP genes were further classified into intact/complete genes, partial genes and pseudogene with interrupting frame-shift mutations and/or stop codons.

2.6.2 Phylogenetic analysis

We employed phylogenic tree building method to further classify the identified KRTAP gene repertoire to their respective subfamilies. For each species the intact genes were used for building phylogentic tree. All intact KRTAP genes were translated to amino-acid and aligned using ClustalW incorporated in MEGA5 (Tamura et al. 2011). The alignments were visually inspected and manually corrected. This final protein sequence alignment was used to build the KRTAP gene tree with the Neighbour-Joining method with P-distance and the interior branch test evaluated with 1,000 replications (Saitou and Nei 1987) (Additional file 2.1 figures 1-24). We make use of unique motifs and repeat sequence structure present in KRTAP subfamilies along with phylogeny and blast results to further help identify and classify partial and pseudogenes to the respective subfamilies.

2.6.3 Positive selection

The ratio of nonsynonymous/synonymous substitution rates (dn/ds) is defined as omega (ω). The value of omega ratio is used as a measure of natural selection acting on protein. The values $\omega < 1$, $\omega = 1$, and $\omega > 1$ represent negative purifying selection, neutral evolution, and positive selection, respectively. PAML package (Yang 2007b) implements various models for the detection of variable ω ratio among lineages and among sites. Likelihood ratio test is used to test hypotheses through comparison of

implemented models (Yang 1998b) (Nielsen and Yang 1998b). The site models allow ω to vary among sites and we used site models to detect signals of positive selection in KRTAP genes. The sequences were translated to amino-acid and aligned using ClustalW incorporated in MEGA5 (Tamura et al. 2011). The multiple sequence alignments were back translated, and all alignment were visually inspected and manually corrected. The genes influenced by gene conversion and recombination are prone to false positive results (Anisimova et al. 2003) and thus were not used in our selection study. The site models are specified with (model = 0) variable Nssites = 0 1 2 7 8. Likelihood ratio test was used to compare two pairs of models M1a (NearlyNeutral) and M2a (PositiveSelection), M7 (beta) and M8 (beta & ω) (Yang 2007b).

2.6.4 Gene conversion and Recombination study

We used the program Geneconv http://www.math.wustl.edu/~sawyer/geneconv/ (Sawyer 1989) to detect statistically significant events of sequence homogenization on paralogs using Global Bonferroni corrected P values. The lower P values indicate greater support for gene conversion. The sequences were translated to amino-acids and aligned. The multiple sequence alignments were back translated. All alignments were visually inspected and manually corrected and used as input. Geneconv gives both global and pairwise fragments involved in gene conversion. We also used the RDP3 software (Martin et al. 2010) to detect recombination events using RDP, Bootscan, MaxChi and Chimaera with 1,000 permutations and cutoff p value of 0.01 employing Bonferroni correction.

The evolutionary distance between genes can be calculated with synonymous substitution, which are immune to selection and are not decreased by negative selection (Zhang 2003a). The sequence divergence was estimated using approximate synonymous substitution rates (dS) using modified Nei-Gojobori (P-distance) method with transition/transversion ratio of 2. GC content was estimated using MEGA5 (Tamura et al. 2011). More than two sequences are needed to detect the signals of recombination therefore subfamilies having more than three genes were used for studies of gene conversion (Additional files 2.3 and 2.4).

2.6.5 Statistical analysis

In order to study the differential evolutionary pattern of high cysteine and high glycine-tyrosine KRTAP genes, we compared the pairwise-pearson correction coefficient (Figure 2.7) of the number of genes present in each subfamily (Table 2.1). We also compared the correlation between GC content (GC% and GC3%) and synonymous

substitution rates (Figure 2.6) using the Nei-Gojobori (P-distance) method with transition/transversion ratio of 2 in MEGA4 (Tamura et al. 2007)

Authors' contributions

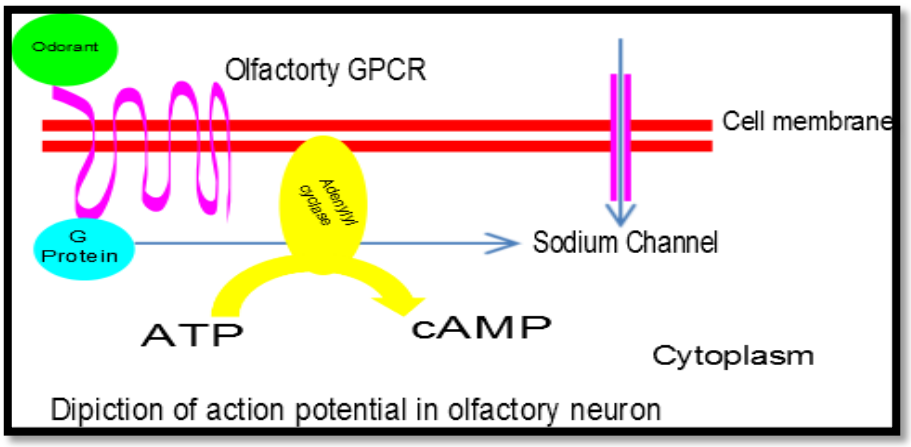
IK performed all the genomics, phylogenetics, and evolutionary analyses and drafted the manuscript. EM participated in the genome mining analysis and drafting of the study. VV participated in the drafting and coordination of the study. SJOB participated in the drafting and coordination of the study. WEJ participated in the drafting and coordination of the study. AA participated in the design, genetic analyses, drafting and coordination of the study.All authors read and approved the final manuscript.

Acknowledgements

IK was funded by a PhD grant (SFRH/BD/48518/2008) from Fundação para a Ciência e a Tecnologia (FCT). SJO was supported in part by Russian Ministry of Science Mega-grant no.11.G34.31.0068. AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490) .We would also like to thank Siby Philip and João Paulo Machado for useful discussions during this work.

3

Olfactory receptor (OR) gene families determines the ecological adaptations in Sauropsida



Dipiction of action potential in olfactory neuron

3.1 Abstract

Olfaction, one of the most important sensory functions, is governed by olfactory receptors (OR) expressed predominantly at the cell-surface of olfactory sensory neurons (OSNs) that are located in the main olfactory epithelium of the nasal cavity. Although the olfactory gene repertoire of mammals has been linked to ecological specialization, patterns of adaptation have not been explicitly addressed in other vertebrates. Here we explored OR diversity in a phylogenetic and ecologically diverse group of sauropsida, including 48 bird and two reptile species *Alligator mississippiensis and Chelonia mydas* to assess how ecological patterns may have influenced their OR gene family repertoire and olfactory abilities. We found that reptiles have a larger OR gene repertoire (~1000 genes) than birds (~200 to ~700 genes). All functional OR genes were members of one of 11 recognized families divided into either class I (OR 1-14) and class II (OR 51 and 52) genes, OR families 55 and 56 were missing in all 50 sauropsida genomes. Gene families 1, 3, 7 and 12 were less common in birds than reptiles, possibly accounting for observed differences in olfactory sensitivities. Families 51 and 52, both specifically adapted for sensory acuity in aquatic environments, were dramatically expanded in the two non-avian reptilian lineages (alligator and sea turtle). In contrast, family 14 (γ-c clade), which is responsible for identifying hydrophobic compounds, has expanded notably in birds. Phylogenomic Bayesian-assignment and principal-component analyses showed that OR families in birds were closely correlated with ecological adaptations (i.e. birds of prey, water birds, land birds and vocal learners). OR families 5/ 8/ 9 were more numerous in predatory bird species and the alligator, suggesting a link between OR repertoire and adaptive specializations linked with carnivory. OR families 2/13, 51 and 52 was correlated with aquatic adaptations (waterbirds), OR families 6 and 10 were more pronounced in vocal-learning birds, while most of the specialized landbirds had much expanded OR family 14 (γ-c clade). Passerines had reduced OR families diversity relative to other bird species. We also found that the evolutionary expansion of OR family 14 (y-c clade) in birds and 51 and 52 in turtle were shaped by positive selection, possibly leading to functional innovation and diversification as most of the birds within the expanded y-c clade had reduced numbers of other OR families. Together, these patterns of gene presence, absence, expansion, contraction, positive selection and concerted evolution suggest that gene loss/gene gain and positive selection played an important role in avian OR gene-family evolution, leading to ecological diversification.

3.2 Introduction

Olfactory receptors are largely responsible for odor perception and the detection of chemical cues, facilitating the differentiation of tens of thousands of unique odorants. This makes olfaction an important physiological function that is crucial to the survival of animals because of its role in recognizing suitable food, mates, offspring, territories and the presence of predators or prey (Niimura and Nei 2006) (Nei et al. 2008) (Adipietro et al. 2012).

Olfactory receptors (ORs) are intron less small sized (1000bp) seven trans-membrane (TM) G-protein coupled receptors (GPCRs)(Buck and Axel 1991) many characteristic conserved motifs. The ligand binding sites responsible for detection of specific odor molecules are conserved between orthologs (between species same function) and variable among paralogs (neo or sub functionalization). Most of these sites are located in the third and seventh trans membrane domains (Man et al. 2004). OR expression occurs primarily in the main olfactory epithelium and to a lesser extent in the vomeronasal organ (Lévai et al. 2006), suggesting that they might have some overlapping functions (Baxi et al. 2006). ORs are also expressed in the testis, where they have a role in sperm chemotaxis (Spehr et al. 2003). The ectopic expression of OR genes in non-olfactory tissues also implies that OR genes likely have additional functionalities (Cruz et al. 2009). Though the relationship between odors and ORs is not clear, it has been hypothesized that a combinatorial coding scheme might allow a single OR to identify multiple odors and also permit different ORs to identify similar odors (Malnic et al. 1999).

The vertebrate and invertebrate ORs followed distinct evolutionary origins (Niimura 2009a). The ORs evolved independently multiple times during animal evolution (Niimura 2012), creating considerable differences in OR gene family repertoire(Bargmann 2006a) (Sato et al. 2008) (Wicher et al. 2008) (Bargmann 2006b) (Benton et al. 2006). In vertebrates, ORs are considered to be the largest multigene family (Niimura 2009b), with characteristic and dramatic variation in OR gene family repertoire among diverse species and lineages, ranging from a single intact gene in elephant sharks to more than 1000 genes in mammals (~1200 genes in rat and opossum and ~1900 intact genes in elephant) (Niimura 2009b) (Niimura and Nei 2005) (Zhang and Firestein 2002). The classification of these genes is complex. The vertebrate OR gene family is divided into two types, Type I with α, β, γ, δ, ε, and ζ groups and Type II ORs form a single group η. The type II OR have been lost in amniotes as it is only found in fish and amphibian. The Class I have diversified in fishes

and amphibians with presence of (δ, ε, and ζ) restricted to these two lineages (Niimura 2009b). The α and γ ORs are tetrapod specific (except, only one **γ** gene in Zebra fish). The β group is reported in both tetrapod and fishes. Based on genetic similarity, the OR genes are divided into 18 families in mammals. The class I with four families (51, 52, 55 and 56)**,** postulated to bind to water-borne molecules, and 14 families (1-14) belonging to Class II, hypothesized to bind mainly to airborne molecules (Hayden et al. 2010)(Olender et al. 2004)(Nguyen et al. 2012)(Quignon et al. 2005) (Glusman et al. 2000). It has been suggested that ecological adaptation has been instrumental in structuring mammalian olfactory subgenomes (Hayden et al. 2010) across species by modifying the number and diversity of OR genes and families (Steiger et al. 2010). Additional modification would have occurred through gene duplication, positive selection and gene conversion, leading to the formation of new gene families and ultimately increasing the adaptive capacity (Steiger et al. 2010). Through time, this OR diversity would have facilitated the adaptation of vertebrates to varied ecological niches at both the broad evolutionary scale (e.g. among fishes, amphibians, reptiles, birds and mammals), as well as among more-recently diverged species (Niimura 2009b).

In this study we characterized the OR gene family repertoire in 48 avian and two reptilian genomes to assess how ecological adaptation may have shaped the OR gene family diversification and olfactory abilities.

3.3 Results and discussion

3.3.1 The genome coverage and olfactory receptor repertoire

The ORs are intron less ~ 1000 bp genes and the whole genome comparative genomic as so far the best approach for in depth elucidation of evolutionary dynamics such a large multi gene families. We performed extensive Blast searches for OR genes in 48 avian and two reptilian genomes. The two reptiles had a larger OR repertoire (*Alligator mississippiensis*, 989 genes and *Clenonia mydas*, 964 genes) than any of the avian species (which had ~200 to ~700 genes). Some bird species exhibited evidence of OR gene expansion, such as the little egret (490 genes), parrot (484), chicken (675), hoatzin (467) and zebra finch (688), while others had reduced numbers of ORs, as for example the medium ground finch (182), rifleman (222) and manakin (227) (see Supplementary Table 3.1). The number of OR genes identified in the chicken and zebra finch were significantly larger than those in other birds. This may be because these genomes were assembled using traditional long-read sequencing instead of the short-read sequencing employed in the other genomes. However, a scatter plot

between the numbers of identified OR genes and sequencing depths (Supplementary Figure 3.1 and Supplementary Table 3.1), did not show a strong positive correlation, which suggests that the differences are due to real biological features that arise from avian OR evolution, like extensive expansion of OR14 like γ-c clade in birds e.g. chicken with 428 γ-c ORs and zebrafinch 552 γ-c ORs.

3.3.2 OR repertoire and the enhanced role of gene loss (pseudogenization) in avian lineage

The gene gain and gene loss provide the essential raw material for evolution to act upon. These changes results in numerous changes at various level from gene (e.g. subfunctionalization, neofunctionlization and pseudogenization), gene groups (e.g. families, subfamilies and classes), and changes in genomic landscape due to genomic arrangements (e.g. tandem arrangements of gene member often results in concerted evolution). The role of these forces is much more pronounced in large families like ORs due to the wide range of evolutionary forces. As previously suggested the olfactory capacity of an organism can be determined by the number of functional and/or nonfunctional ORs.

The complete OR repertoire was divided into three major categories, partial genes less than 650 bp in size and/or without start and /or stop codon, pseudo genes with less than 650bp in size and/or with stop codon or frame shift mutation, and functional/intact genes ORs with 650bp or more size with proper reading frame and without stop codon. In total 16,503 OR genes were found in 48 bird genomes with 6704 partial and 7855 pseudo ORs and 1944 functional ORs. If partial genes were considered to be nonfunctional, then the total number of nonfunctional OR increased to 14,559; whereas the sum total of partial and functional genes was just 8648 this overall supports that gene loss have major contribution in evolution bird ORs repertoire. The comparison ORs for individual bird species also supported the above results.

The comparison between the two reptilian species lineage alligator and turtle reveals shows interesting variation in functional and nonfunctional ORs tough both have almost same number of total ORs genes 989 and 964 respectively. The alligator have 405 functional ORs compared to turtle which only have 205 functional genes. Similarly inclusion of partial gene as pseudo resulted in 584 and 714 ORs in alligator and turtle respectively whereas if all the partial genes are assumed to be functional, the total becomes 638 and 459 for alligator and turtle respectively which ultimately supports that gene loss is much pronounced in turtle than alligator. The above results suggest that ORs evolution is shaped by individual olfaction requirements possibly due to different

modalities, e.g. ecological, behavior and physiological requirements (hayden streiger, dehara, hayden new sound ecocolation etc).Thus characterization of the complete repertoire of ORs into functional and nonfunctional helped in understanding the role of gene gain and gene loss in shaping the OR repertoire and thus the olfactory ability in different species and lineages.

3.3.3 OR gene family assignment and dynamics

The OR gene family is divided into 18 families in tetrapod. Each OR gene found was assigned to one of the 18 known OR families as per HORDE database (the Human Olfactory Data Explorer) #43 http://genome.weizmann.ac.il/horde/. The assignment of gene families

3.3.4 Phylogenetic grouping

The phylogenetic analyses was performed using all functional ORs from the avian and reptilian genomes from present study, representative ORs covering α, β, γ, δ, ε, ζ and η groups (Niimura 2009)? and 1-18 class II and 51-56 class I families from Horde database http://genome.weizmann.ac.il/horde/. The phylogeny supported that birds and reptilian ORs form nine groups which is in accordance with the groping of mammalian OR families into 11 OR families **(Hayden et al**). These groups were supported with high bootstrap support for more than 90% of these,: OR 1/3/7, 2/13, 4, 5/8/9, 6, 10-12, 14 (γ-c clade), 51 and 52 (Figure 3.1). The most-traditional OR families formed a monophyletic clade (with families 4, 6, 10, 11, 12, 14, 51 and 52), whereas clades 1/3/7, 5/8/9 and 2/13 were monophyletic and formed 2-3 different families intermixed into fairly well-defined clades, perhaps because of their functional redundancy and/or combinatorial coding (Malnic et al. 1999)(Cruz et al. 2009).

3.3.5 OR gene family diversity and olfactory ability

The elucidation of complete OR repertoire from To explore the evolutionary dynamics of OR genes in avian and reptilian genomes, we identified the complete repertoire of ORs gene families that are known to recognize odors in vertebrates. In each species, the presence of an OR family and the number of genes in each family was compared with their perceived use of olfactory signals. We found that the percentage of functional OR gene families varied across species and phylogenetic groups (Figure 3.2). The ratio of functional OR gene families ratio of what?? ranged from 0% to 95% within avian species. Some families of OR genes varied more than others, e.g. families 1/3/7, 11

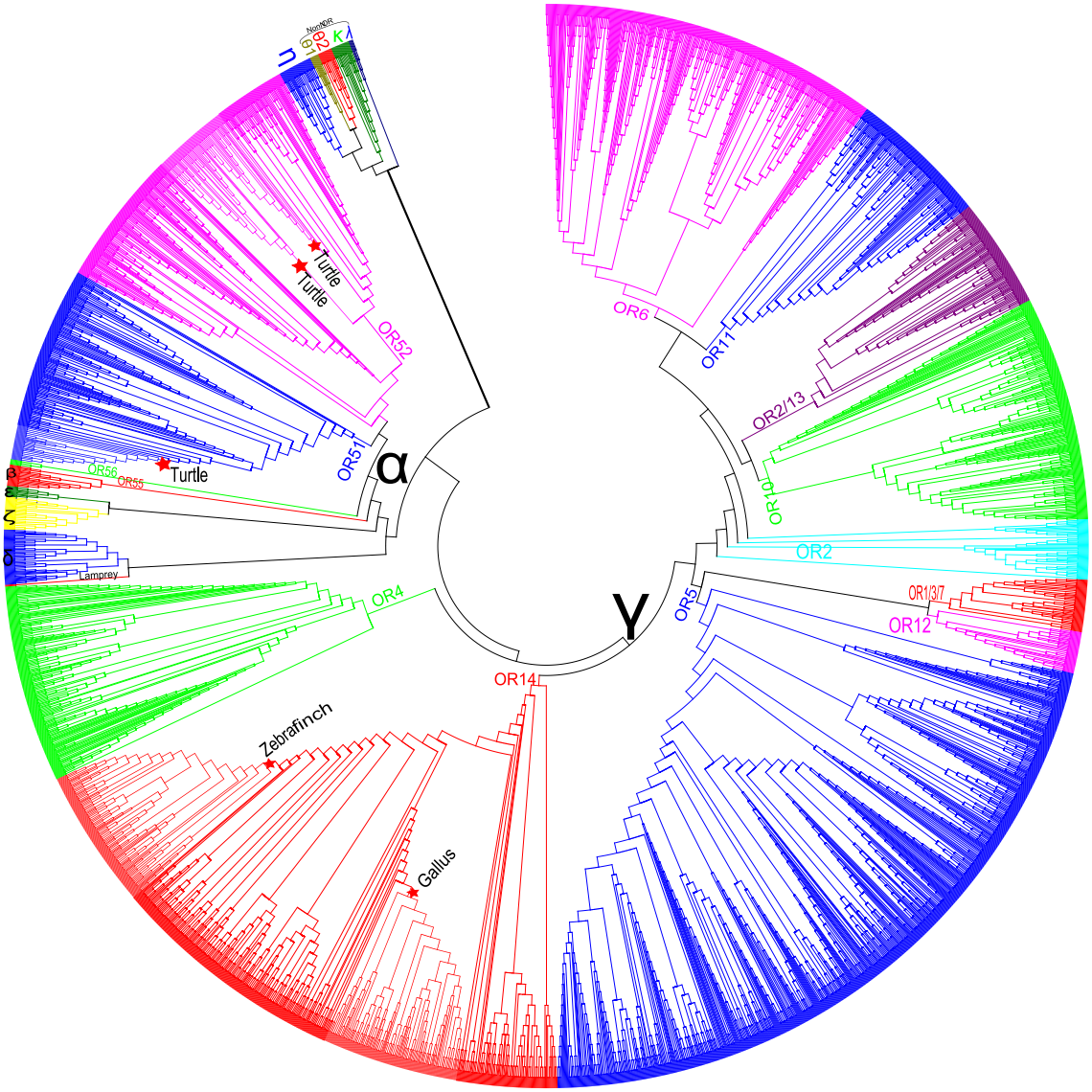


Figure 3.1: Neighbor-Joining phylogeny of the OR gene family considering all functional genes found (n=2599) in the 48 avian and two reptilian genomes studied here, together with other known representative OR gene families (1-14 and 51-56), OR groups α-η and non-OR GPCR θ, κ, λ from human and zebrafish retrieved from Niimura 2009 and Niimura 2012.

OR gene families varied across species and phylogenetic groups (Figure 3.2). The ratio of functional OR gene families ratio of what?? ranged from 0% to 95% within avian species. Some families of OR genes varied more than others, e.g. families 1/3/7, 11 and 12 were relatively rare and thus may contribute less to olfactory sensibility. In contrast, families 2/13, 5/8/9, 4, 6, 10, 14, 51 and 52 appear to have contributed significantly to the diversification of olfactory receptors among birds. In particular, class 1 families 51 and 52 expanded in number and diversity dramatically in the two reptiles studied and class II family 14 (γ-c clade) was most abundant in birds. Families 51 and 52 are predicted to be sensitive to hydrophilic compounds present in aquatic environments. Along with the expanded Class I OR51 and 52 genes in sea turtle and

alligator, they also have a comparatively large number of genes from class II families 1-14, which could be related with their relative dependence on terrestrial habitats for breeding and other functions (Kishida et al. 2007). We found that among the reptiles studied to date, sea turtle and alligator have a more well-developed OR family repertoire relative to other previously characterized squamata (lizard and snake; (Kishida and Hikida 2010) (Dehara et al. 2012).

In contrast, Passerine birds had an overall reduced OR gene family diversity Do you mean reduced number of genes or reduced divergence or reduced SNP variation. This suggests a possible loss of function and/or that other OR families have been recruited to mitigate the loss of functions that would have been handled by missing OR families (e.g. the family 14 in zebrafinch has expanded and has been under positive selection). Overall, families 2, 13, 51 and 52 contributed to aquatic lineages and families 6, 10 were more determinant in vocal learners. However, birds of prey, had a comparatively high percentage of OR families 5/8/9. These were also the largest OR families observed in alligator, which like birds that hunt, depend heavily on hunting or scavenging for prey. Further studies are needed to understand the particular role of these OR families in predatory groups and to determine if these findings are generalizable across a larger set of species with diverse ecological adaptations and life styles.

3.3.6 OR gene family and ecological adaptation

In order to assess the role of ecological adaptation in the determination of the functional OR avian subgenomes, the 48 bird species were grouped into four major ecological groups: land birds, water birds, vocal learners and birds of prey. Figure 3.3 represents the heat map of the OR familial percentage based on the four avian ecological groups defined.

Using principal component analysis (PCA) and Bayesian assignment tests to assess the degree of correlation of specific OR gene families with groups of land birds, water birds, vocal learners and birds of prey, we were able to place most of the species into groups that reflected their ecological niches (Figure 3.4 and Figure 5.5). The naïve Bayes assignment algorithm and PCA assigned most species into their respective ecogroups, with only a few exceptions (Figure 3.4 and Figure 3.5).

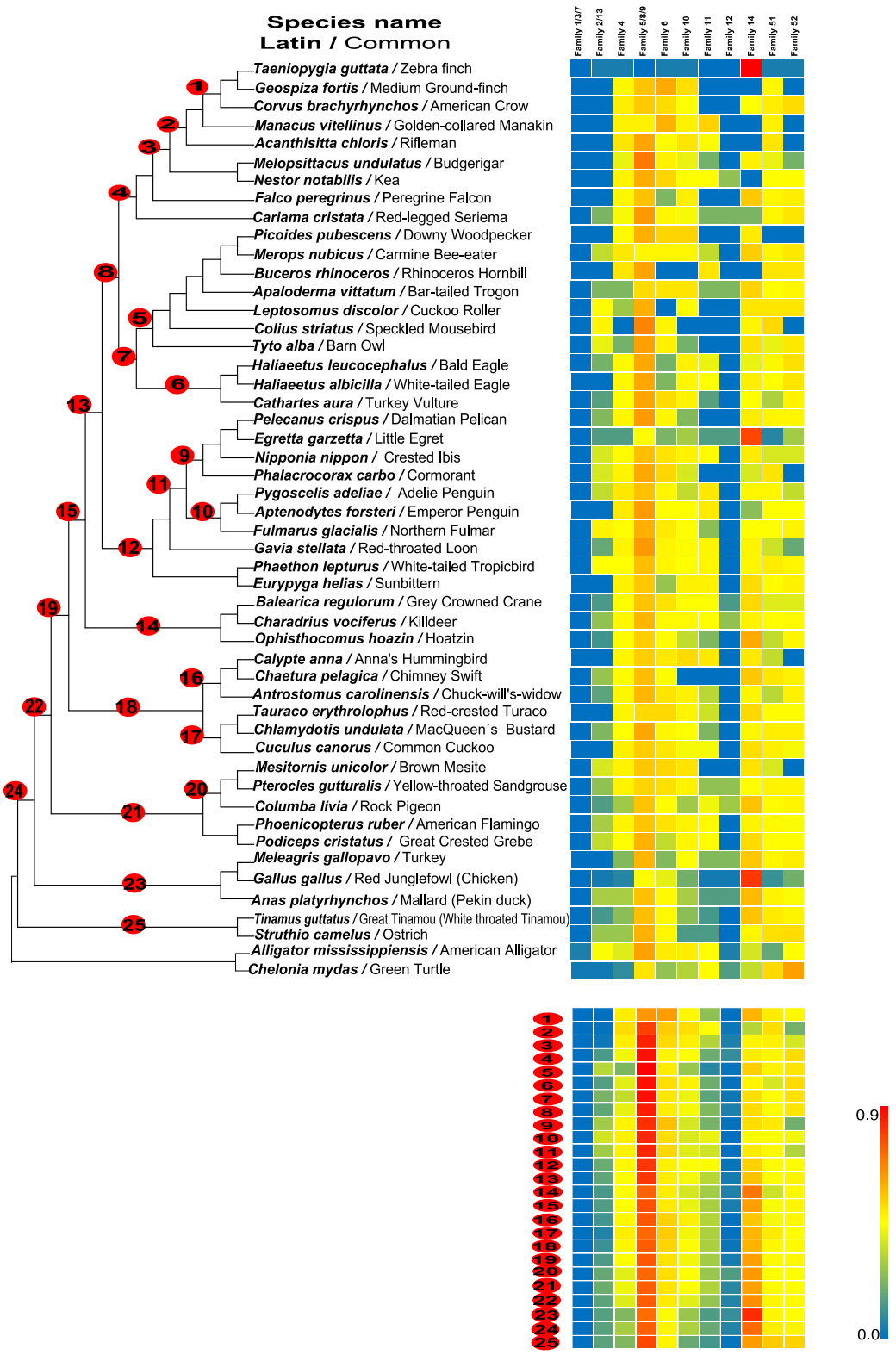


Figure 3.2: A consensus phylogeny of the avian genomes with alligator and turtle as outgroups showing the heat map of relative percentage (0 to 100%) of functional OR gene families in each species. B. The corresponding ancestral states nodes in the tree in A, reconstructed (labeled A1-A25).

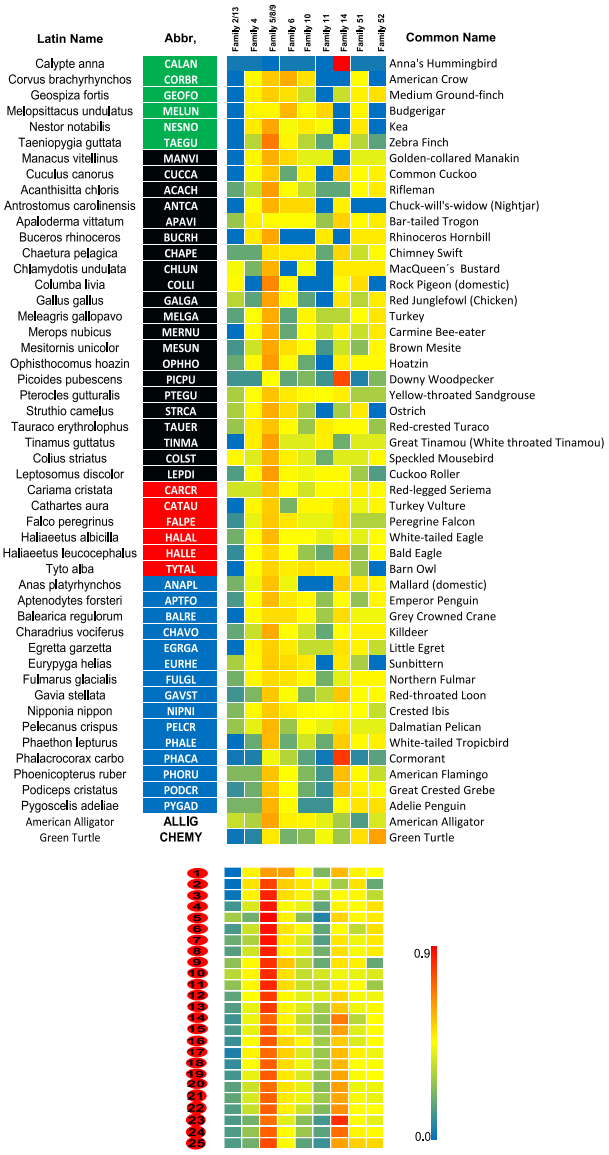


Figure 3.3: Heat map partition of informative OR gene families considering the broad ecological traits groups in birds (Land birds, Water birds, Vocal learners and Birds of prey).

These results support that OR families 51, 52 and 2/13 were more closely associated with the aquatic group. OR families 6 and 10 contributed the most to defining the vocal learner group of species and family 5/8/9 the birds of prey, while most of the specialized land birds had an expanded number of genes from the OR family 14 (γ-c clade). These patterns demonstrate that the OR repertoires are correlated with the ecological adaptations of each species and are related less with shared ancestry (phylogenetic relationships). Similar ecological partitioning of gene characteristics was also apparent in reptiles (lizard, sea turtle and alligator), as lizard grouped with land birds, sea turtle with aquatic birds and alligator (a semi-aquatic species), was more closely aligned with aquatic and land birds. We also found that desert birds clustered together in PCA (data not shown) and that some species were outliers, with patterns

that were distinct from the other species in their ecological group, especially those with an expanded OR14 repertoire (e.g. zebrafinch). Species that were not clearly associated with any group, such as fulmar and ostrich, have adaptive characteristics that fit the life-history patterns of one or more ecological partitions (e.g. fulmar is a seabird and also a bird of prey; ostrich is grouped with prey birds like Turkey vulture and Houbara Bustard, and all three species are found in desert environment). Shared traits could also explain some of the deviations observed in egret and duck, which grouped into land birds instead of aquatic birds (Figure 3.4).

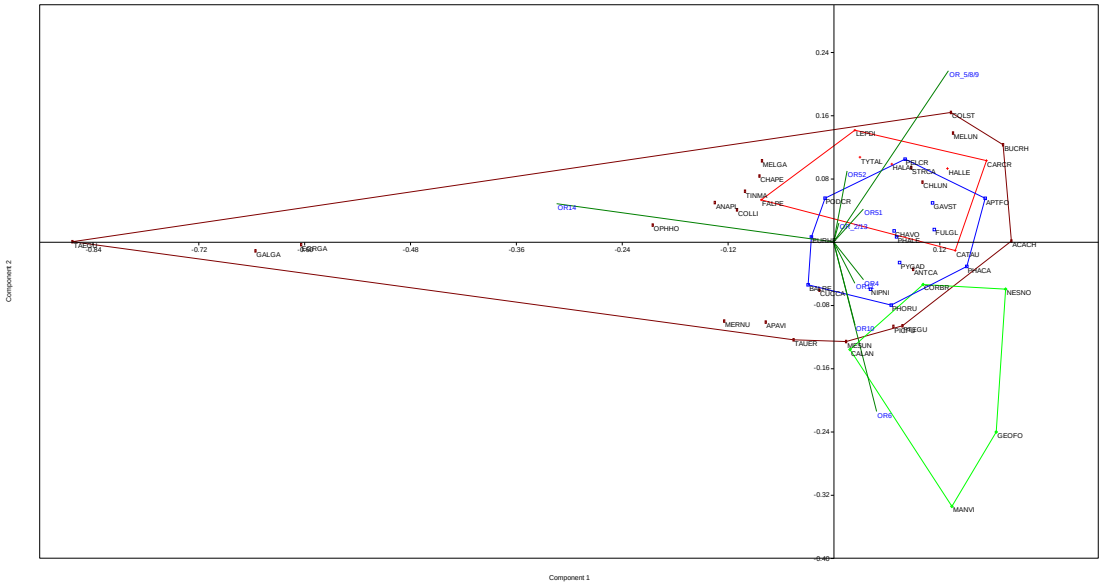


Figure 3.4: PCA scatterplots showing the partitioning of ecological traits groups (Land birds, Water birds, Vocal learners and Birds of prey) and the OR families contribution in each group. The two components explained more than 68% of the data variance (ANOSIM r=0.58).

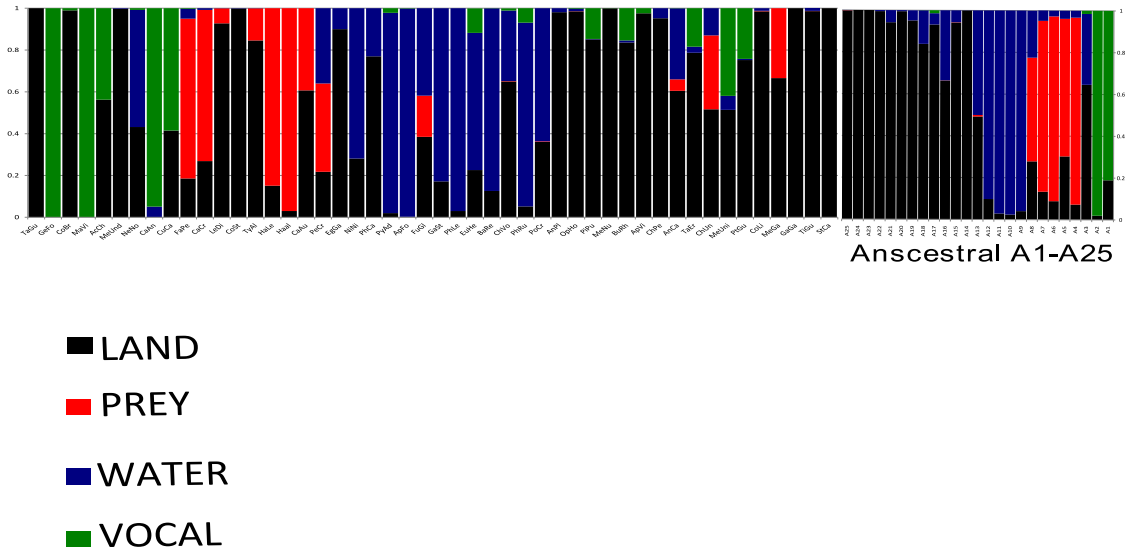


Figure 3.5: Naïve Bayesian assignment of the avian species into different ecological groups (Land birds, Water birds, Vocal learners and Birds of prey) based on the OR gene family contribution and of the OR ancestral states (A1-A25) considering the different ecological groups (Land birds, Water birds, Vocal learners and Birds of prey).

We selected 25 ancestral nodes as shown in Figure 3.2 for ancestral states reconstruction to assess whether those states follow the ecological grouping of birds. We found that the reconstructed ancestral states also followed the ecological avian groups using PCA and naïve Bayes assignment tests (Supplementary Figure 3.2 and Figure 3.5; e.g. nodes A1 and A2 grouped in vocal learners).

3.3.7 OR gene families and adaptive evolution

To detect evidence of positive selection, we used SLAC, FEL, REL, FUBAR and MEME, as well as an integrative approach that considered multiple phylogenies based on the inferred potential breakpoints. This approach is generally more reliable compared with PAML, which depends on a single phylogeny and may lead to an increased number of false positives, especially when recombination and gene conversion rates are high (Steiger et al. 2010). Using both individual and integrative approaches, we found signals of positive selection in the expanded OR family 14 in birds (eight bird species) and in the OR family 51 and 52 in *Chelonia mydas* (Supplementary Table 3.2), suggestive that positive selection is playing a role in the functional diversification and ecological adaptation of the OR genes.

The alignment-wide test of positive selection using the Parris method was significant with a p-valve <0.001 for OR14 family in *Gallus gallus*, *Taeniopygia guttata* and *Egretta garzetta* and for OR52 in *Chelonia mydas* (Supplementary Table 3.3). Sites with positive selection patterns identified with two or more methods and in two or more bird species were plotted on the *Gallus gallus* OR14 protein sequence (Supplementary Figure 3.3) and clearly demonstrated that the majority of these positive-selected sites in these birds are restricted to the protein trans-membrane (TM) domains. Most of these positive-selected sites were located in TM5, which is believed to be involved in ligand binding (Steiger et al. 2010). Most of the sites found in the turtle OR51 (Supplementary Table 3.4a) and OR52 (Supplementary Table 3.4b) were also located in TM domains. These positive-selected sites provide additional evidence of the important role that ecological adaptation has in the evolution of olfactory capabilities.

3.4.Materials and Methods

3.4.1 Annotation of olfactory receptor (OR) genes in bird genomes

To identify the OR genes in bird genomes, we downloaded the known amino acid sequences of OR genes of *Anolis carolinensis* (green anole), *Gallus gallus* (chicken) and *Taeniopygia guttata* (zebra finch) from the paper of (Steiger et al. 2009). We used similar procedure to that of (Steiger et al. 2009) to identify the putative OR genes in 48 birds, as well as in alligator (*Alligator mississippiensis*) and turtle (*Chelonia mydas*).

Firstly, TBLASTN searches with an E-value cut-off of 10 were conducted to identify candidate OR loci. Then the results of TBLASTN were clustered together according to the locations of BLAST hits in the genome. For a given locus, the best hit with smallest E-value and with length of >=150bp was retained for subsequent analysis. And for the candidates lacking start/stop codons, we searched 90bp of the upstream to find start codons and 90bp of the downstream to find stop codons.

Secondly, RepeatProteinMask was adopted to distinguish OR genes from non-OR GPCRs. The above known full-length OR sequences from (Steiger et al. 2009) and 328 non-OR GPCR sequences from (Lagerstrom et al. 2006) were merged together as the library to run RepeatProteinMask for each genome. Based on the results of RepeatProteinMask, the candidate loci from the TBLASTN step that matched non-OR GPCR regions (overlapping length/candidate length > 50%) were filtered out.

The remaining OR candidates can be classified into three categories: intact genes with normal start codons and stop codons and more than 650bp in size thus can code for seven TM domains, partial genes without start codon or stop codon or both, and pseudogenes with frame shift mutations and/or premature stop codons.

3.4.2 OR assignments of group, families and subfamilies

In order to assign all functional genes to their respective OR families we performed HMMER searches against a local database consisting of protein profiles of all known OR families present in HORDE database (OR1-14 and OR51-56) and other known OR groups from river lamprey, zebrafish and frog (Freitag et al. 1999) (Niimura 2009b) thereby covering all known ORs (α- η) from all major vertebrate groups. The sensitive search against the database allowed us to assign each OR gene based on best similarity to the closest known OR gene profiles with high confidence. The accuracy of assignment was tested, by assigning known human and lizard ORs against the database each known OR were very correctly assigned to their respective family.

3.4.3 Phylogeny of avian and bird ORs

The amino acid sequences of all intact functional OR genes ≥ 650bp found in this study were aligned using MUSCLE (Edgar 2004) and the alignment was manually corrected and used to construct a Neighbor-Joining tree in MEGA5 (Tamura et al. 2011) with Poisson correction method and 1000 replicates (Felsenstein 1985). We used all available previously described representative ORs families (OR1-14 and OR51-56) and groups (α-η) from zebrafish, river lamprey, frog and human (Niimura 2009b), which improved the resolution of the phylogenetic tree.

3.4.4 Positive selection

The ratio of nonsynonymous and synonymous mutations ($\omega=d_n/d_s$) provides an estimate of changes that are advantageous, reflecting positive selection ($\omega>1$), neutral ($\omega=1$), or disadvantageous, reflecting negative selection ($\omega<1$)(Yang 1997). Because of gene conversion and recombination, no single tree can represent a correct phylogeny, and methods like PAML, which are based on single phylogeny, can give false positives. Therefore we used five different individual methods along with an integrated approach. These methods allow the use of multiple phylogenies based on inferred potential breakpoints and thus are more accurate to detect signals of positive selection under these conditions compared with PAML, which depends on a single phylogeny and can possibility lead to more false positives due to recombination and gene conversion. All these methods are implemented in Datamonkey web server http://www.datamonkey.org (Pond and Frost 2005a) and also in the Hyphy package

(Pond et al. 2005). These includes Single Likelihood Ancestral Counting (SLAC), Fixed Effects Likelihood (FEL), Random Effects Likelihood (REL), (Pond and Frost 2005b) Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012), FastUnconstrained Bayesian AppRoximation (FUBAR)(Murrell et al. 2013) and integrative approach. SLAC model uses ancestral sequences reconstruction. FEL calculates site-by-site dn/ds without assuming a prior distribution. REL assume a prior distribution across site. FUBAR ensures robustness against model misspecification. MEME is the most appropriate to detect episodic diversifying selection affecting indvidiual codon sites. The integrative approach incorporates all sites detected by SLAC, FEL, REL, FUBAR and MEME. The sites detected by two different methods can be supportive of positive selection. Combined with the PARRIS method, our approach provides a robust inference of positive selection in recombining coding sequences by allowing for variable tree topologies and branch lengths across detected recombination breakpoints and variable synonymous substitution rates across sites. These methods make use of multiple phylogenies resulting from each recombinant fragment and thus are less prone to false positives. All these methods were used with default settings.

3.4.5 Principal component analysis (PCA) and analysis of similarities (ANOSIM)

PCA analysis of all functional genes was done using PAST v1.89 (Øyvind Hammer et al. 2001) The covariance matrix was used to assess patterns of variation in OR family distribution in different bird groups based on their shared traits (namely land birds, water birds, vocal learners and birds of prey). The significance of these groupings was tested using a non parametric test for analysis of similarities (ANOSIM) (Clarke 1993) between groups using Euclidean distances and derivations of R-statistics. The observed values were compared to 95% confidence interval of a simulated distribution.

3.4.6 Ancestral state reconstruction

The ancestral state construction of OR gene repertoire for nodes 1-16 Fig 2 was done using Mesquite v2.75 (Maddison, W. P and D.R. Maddison 2011) using the consensus avian phylogeny from the (Jarvis et al. 2014)(Zhang et al. 2014). The parsimony method using continuous character was used to estimate the ancestral OR familial distribution at each node (Fig. S3). The OR family distribution at each ancestral node was determined based on the assignment test.

3.4.7 Bayesian assignments

Naïve bayes assignment is a machine learning algorithm implemented in the WEKA package (Whitten IH and Frank E 2005). It uses independent assumptions to determine how best to categorize a data set based on the expressed variation (here based on OR

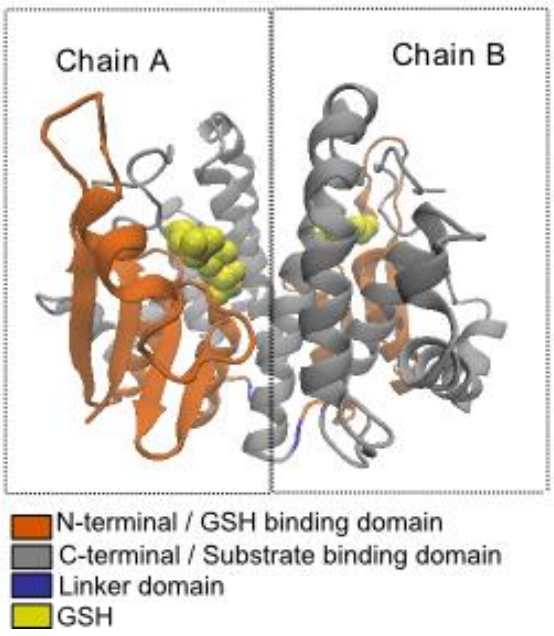
familial distribution and ecological trait categories including land birds, water birds, vocal learners and birds of prey). This trained data set is then used to assign each species to a respective ecological group based on OR family distribution. The species to be assigned (the target species) is removed from the training set and subsequently assigned.

Acknowledgements

IK was funded by a PhD grant (SFRH/BD/48518/2008) from Fundação para a Ciência e a Tecnologia (FCT). AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490). SJO was supported as Principal Investigator by Russian Ministry of Science Mega-grant no.11.G34.31.0068.

4

Avian cytosolic glutathione transferases: Gene expansion and adaptive evolution suggest protective role against diverse xenobiotics and cellular stress



4.1 Abstract

Background

The cytosolic glutathione transferases (cGSTs) are known for their dynamic and interactive defense mechanism providing protection against cytotoxic electrophilic substrates and adaptation to cellular stress exposure.

Results

The genomic scan of 48 avian genomes revealed that birds lack cGSTP and possess only six out of seven major cytosolic GST classes found in most non-avian vertebrates. Sequence comparison and the phylogeny of avian cGSTs revealed that cGST in birds have similar active binding site as mammals. We found duplication of cGSTA and cGSTT. The positive selection played an important role in diversification of avian cGSTs and the duplicated cGSTA and cGSTT genes evolve under strong positive selection, which supports their important role in protection from reactive species and diverse xenobiotic compounds. We also found positive selection in cGSTO and cGSTZ, which likely suggests their important secondary role in detoxification. The positive selection results were also supported by the positive radical changes in amino acid properties leading to structural and functional diversification of avian cGSTs.

Conclusion

Our study shows that gene duplication and positive selection played an important role in the molecular evolution and functional diversification of avian cGSTs. The duplicated cGST genes evolved under positive selection, suggesting their relevant role in the protection against diversified endogenous and exogenous substrates resulting from varied stress conditions.

4.2 Introduction

The evolution of gene families plays an important role in species adaptation and diversification. Gene gain and gene loss shape the evolution of gene families. The new members often gain new function (neofunctionalization) or new paralogs share one of the original functions of the ancestral gene (subfunctionalization) (Zhang 2003) (OHNO 1970). The glutathione transferases (GSTs) (EC 2.5.1.18), historically known as glutathione S-transferases are important phase II detoxifying enzymes (Hayes et al. 2005). The cytosolic GSTs catalyze the conjugation of electrophiles to the reactive GSH (Keen and Jakoby 1978), forming a water soluble conjugate, which can be metabolized for excretion. The cGSTs are widely distributed in nature ranging from ancient archaea to plants and higher vertebrates suggesting a primordial crucial role in life (Oztetik and Cakir 2014) (Dixon et al. 2002) (Sheehan et al. 2001). They are

divided into three major superfamilies: cytosolic, mitochondrial and (MAPEG) membrane-associated proteins in eicosanoid and glutathione metabolism (Frova 2006). The bacterial fosfomycin-resistance proteins represent the distant GST family (Armstrong 2000).

Along with their classic role in reactive species (oxidative and carbonyl) stress metabolism and cellular detoxification of wide range of endogenous and xenobiotic compounds, cGSTs are also involved in metabolic pathways not associated with detoxification (e.g. biosynthesis of leukotrienes, prostaglandins, testosterone, and progesterone, as well as the degradation of tyrosine). The cGSTs of the alpha, pi, and mu classes have been shown to modulate signaling pathways that control cell proliferation, cell differentiation, and cell death by interacting with important signaling proteins in a non-enzymatic way (Laborde 2010). The detoxification of anticancer drugs by cGSTs poses a serious issue in drug development and therapy, being an extensive field of research (Lo and Ali-Osman 2007) (Ruzza et al. 2009) (Di Pietro et al. 2010).

The cytosolic GSTs are an important superfamily comprehending ~20 genes grouped into seven subfamilies in mammals (Sheehan et al. 2001) (Frova 2006). Some subfamilies are present throughout taxa and kingdoms, whereas others are specific to certain groups and species (e.g. additional subfamilies are present in plants, insects and bacteria) (McGonigle et al. 2000) (Soranzo et al. 2004) (Ranson et al. 1998) (Vuilleumier and Pagni 2002) (Frova 2006). Human cytosolic GSTs consist of seven subfamilies, with variable members within each: cGSTA (alpha) and cGSTM (mu) have five members each, cGSTO (omega) and cGSTT (theta) have two members each, and cGSTP (pi), cGSTZ (zeta) and cGSTS (sigma) have only one member. Less than 25% sequence identity exist between members of two different subfamilies, whereas members within each subfamily possess greater than 40% amino acid sequence identity (Wu and Dong 2012). Each cGST is divided into two major domains, domain I, also known as N terminal alpha/beta domain (thioredoxin-like fold, β1-α1-β2-α2-β3 – β4-α3) or G domain (for GSH binding), and domain II, also designated as an alpha helical C terminal domain (5 to 6 alpha helices) or H domain (binding hydrophobic substrate). All cytosolic GSTs are present in the form of dimers and involve interaction of domain I of one cGST subunit and domain II of other cGST subunit. The presence of cytosolic GSTs in homo or hetero dimer forms increases the diversity and catalytic activities of cGSTs (Fonseca et al. 2010). The dimeric structure also enhances protein stability and provides the active site with a proper structure for efficient catalysis (Wu and Dong 2012).

The survival of organism depends of its ability to cope with stress posed by endogenously compounds and/or environmental pressure, such as temperature, heavy metals, salinity, pesticides, pharmaceuticals, UV and reactive species produced from various stress reactive oxygen species (ROS) and reactive carbon species (RCS) (Lushchak 2011) (Yan et al. 2013). RCS can be more destructive than ROS and may have far-reaching damaging effects on target sites within or outside the cell membranes (Naudí et al. 2011). The cGSTs are a well characterized detoxification enzyme family among the enzymatic antioxidant system involved in the elimination of ROS produced in stress tolerances (Sharma et al. 2004) (Shi et al. 2014). The over expression of cGSTs is also suggested to increase the longevity by controlling the RCS (Naudí et al. 2013). The role of cGSTs in defense against harmful endogenous and/or exogenous compounds and reactive species oxidative and carbonyl stress suggests an important biological adaptation fundamental to species survival (Hayes and Pulford 1995) (Sheehan et al. 2001) (Cummins et al. 2013) (Schröder 2001) (Szalai et al. 2009) (Diao et al. 2011). The protective role of cGST gene family have had a major role in the successful adaptation against oxidative stress making cytosolic enzymes good examples of divergent evolution (Fonseca et al. 2010). Here, we performed a detailed characterization of cytosolic GSTs repertoire in 48 avian genomes for the elucidation of gene gain, gene loss and orthology and parology relationships of avian cGSTs. We performed extensive adaptive evolution analyses of avian cGSTs members found in genomic scans of 48 bird genomes to understand the role of positive selection in structural and functional diversification of the avian cGSTs and its implication in detoxification against various type of cellular stress and rapidly increasing xenobiotic diversity.

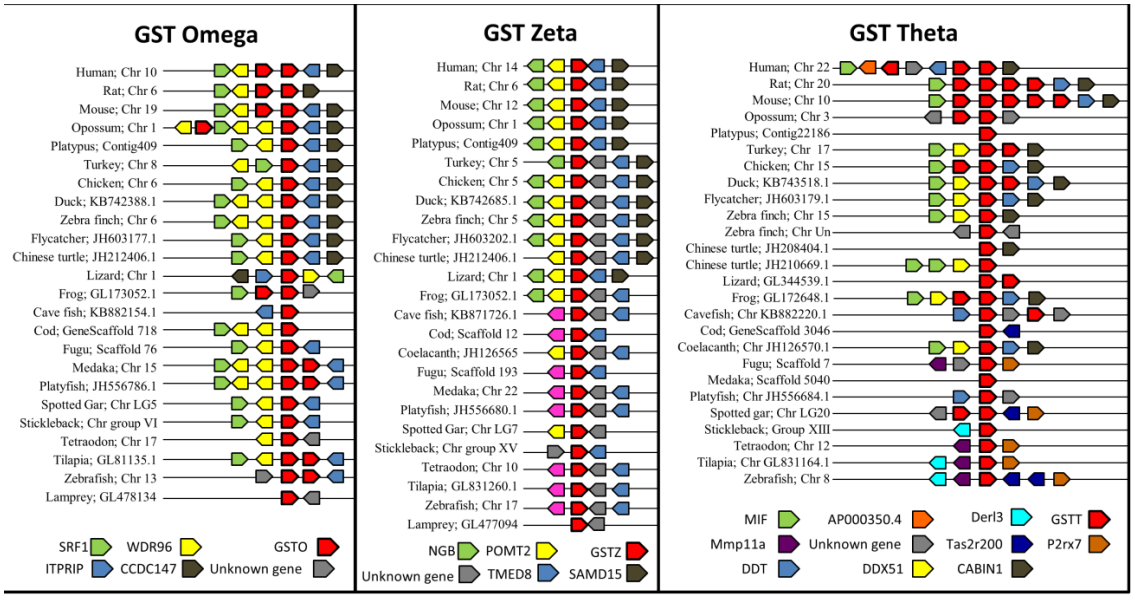
4.3 Results & Discussion

4.3.1 Genomic scan, synteny organization and gene gain

The comparative genomics studies allow us to understand the genome evolution in phylogenetically diverse species representing key stages in evolution (O'Brien et al. 1993). Gene duplication is an important force in genome evolution, which provides essential duplicates for functional diversity and evolution of gene and gene families driven species evolution. The genome sequencing projects provide direct information about events of gene duplication. In the human genome, the cGST genes occur in class-specific clusters on different chromosomes. We explored the arrangement of cGST genes in 48 bird genomes.We were able to retrieve the synteny information for all cGSTs except for cGSTM for which we found limited information (insufficient genome coverage in those regions). Avian orthologous cGSTs were present in the

homologous chromosome in chicken, turkey and songbird genomes (e.g. cGSTZ on chromosome 5, cGSTS on chromosome 4, cGSTA, cGSTT and cGSTO on chromosome 3, 15 and 6, respectively, in both chicken and zebrafinch, and that are homologous of the chromosomes MGA2, MGA17 and MGA8 in turkey (Zhang et al. 2011)) (Figure 4.1). Besides the finding of the conserved syntenic arrangement of cGSTs in birds and other vertebrates (Figure 4.1), we also found evidence of gene duplication within the alpha and theta class cGSTs in the avian lineage. We also found these duplicates in reptilian lineage (Figure 4.1). The duplicated members of the same class are clustered in tandem (Figure 4.1). The presence of gene duplicates increases functional diversity and the range of catalytic activity, providing further protection from deleterious agents and suggesting the role of gene duplication in the evolutionary diversification of the cGST gene family.

The gene search and synteny analysis shows that only the classes cGSTA and cGSTT are duplicated in birds and reptiles, whereas in mammals, along with cGSTA and cGSTT, the cGSTM and cGSTO are also duplicated. The duplicated members of the same class are arranged in tandem forming tight clusters. The syntenic analysis of five bird genomes supports the absence of cGSTP in birds, while cGSTP is present in reptiles (Figure 4.1). We also retrieved least number of sequences for GSTM from the 48 avian genomes (Supplementary Table 4.1) for which we could find syntenic information only in flycatcher genome but not in the other four birds used in the synteny analysis (possibly to do the lack of genomes coverage or due to species-specific genomic variation).



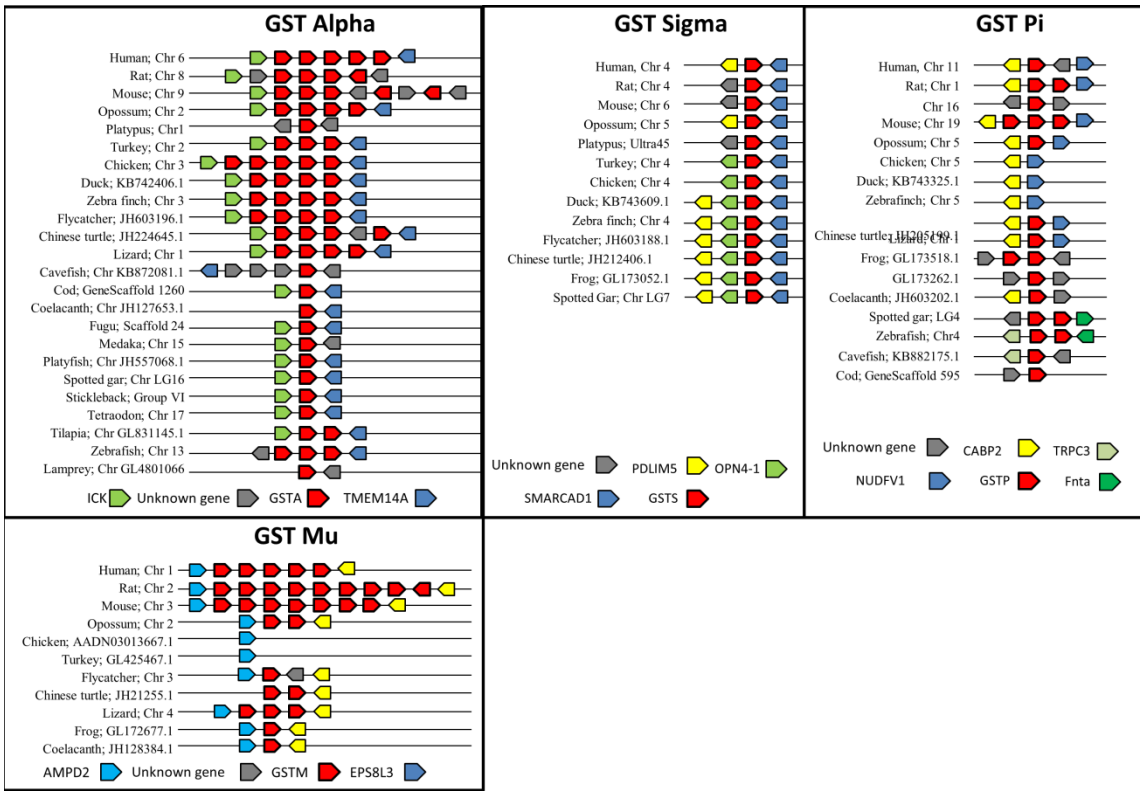


Figure 4.1: Synteny analysis results of cGST in vertebrates (The syntenic arrangement of cGSTs from five bird genomes and comparison with other vertebrates

4.3.2 Avain cGSTs subgenome and sequence conservation

The blast searches (Altschul et al. 1990a) and synteny analysis of cGSTs in the five bird genomes together with additional searches in 48 bird genomes (Supplementary **Table 4.1**) confirmed the presence of six (cGSTA, cGSTM, cGSTS, cGSTT, cGSTO

Table 4.1: The overall distribution of cGST in various groups as per present study see additional file for details

	Mu	Pi	Alpha	Sigma	Theta	Zeta	Omega
Mammals	X	X	X	X	X	X	X
Birds	X	Absenr	X	X	X	X	X
Reptiles	X	X	X	X	X	X	X
Amphibians	X	X	X	X	X	X	X
Fish	X	X	X	Absent	X	X	X
Mollusks	X	X	Absent	X	Absent	Absent	Absent
Artropods	X	X	Absent	X	X	X	X
Plants	Absent	Absent	Absent	Absent	X	X	Absent
Bacteria	Absent	Absent	X	X	X	X	X

and cGSTZ) out of seven cGSTs represented in mammals. The cGSTP was not found in birds, but searches in reptile genomes e.g. Anolis and Chinese turtle genome revealed presence of this gene in reptiles (Figure 4.1, Table 4.1 and Supplementary Table 4.2) suggesting that cGSTP was lost in the bird lineage as we could not find the gene in the 48 bird genomes analyzed.

Comparison of the amino acid sequences revealed less than 30% amino acid sequence identity between classes (cGSTA, cGSTS, cGSTM, cGSTP, cGSTT, cGSTZ and cGSTO) compared to 60% or more within duplicated members of the same class (cGSTA1 – cGSTA4 and cGSTT1 and cGSTT2) (Table 4.2). The Sequence alignments of avian cGSTs also show evolutionary signatures of some strictly conserved residues (Sheehan et al. 2001) (Frova 2006) (see Figure 4.2 and 4. 3).The closely related avian cGST classes (alpha, mu and sigma) show higher conservation, due to substrate promiscuity, as compared to distantly related cGST classes (theat, zeta and omega) (Achilonu et al. 2010) (Fonseca et al. 2010). Despite of variations found among the cGST classes, remarkable similarity is seen in the secondary structure of most of the cGST classes. We also found that closely related classes share common catalytic residues (e.g. Tyrosine in GSTA, GSTM and GSTS) (Figure 4.2). The physicochemical characteristics within the active site of each cGST subfamily provides the needed functional diversity of cGST multi gene family (Dirr et al. 1994) (Achilonu et al. 2010). The role of G-site is conserved for GSH binding. The C terminal domain is variable owing to its involvement in diverse substrate binding (Figure 4.2).

The secondary structure prediction using PSIpred shows the presence of an extra beta chain in C terminal domain of avian cGSTA, cGSTM, cGSTT and cGSTS (Supplementary Figure 4.1). Similar results were also obtained by Kim et al. 2013, overall this leads to 5 beta chains and 10 alpha helix (N-terminal domain contains a1-b3/b1–4 and C-terminus contains a4-a10/b5) for avian cGSTA. The extra beta chain (5th beta chain) in avian cGSTs further supported by known 3D structure of cGSTA (Figure 4.3) and other cGSTs from chicken (Supplementary Figure 4.2). We also found an extended N-terminal in cGSTO, and cGSTT and cGSTM have C-terminal loop together with an extra-long loop present only in cGSTT (Figure 4.2).

4.3.3 Phylogeny of avian and vertebrate GSTs

Bayesian and Maximum Likelihood approaches were used to see the phylogenetic relatedness of avian cGST classes. We used two datasets: (1) cGST from the five bird genomes used in the synteny analysis (Supplementary Table 4.1) and cGST from the five bird genomes used in the synteny analysis along with cGST from other vertebrates

(Supplementary Table 4.1 and Supplementary Figure 4.3).

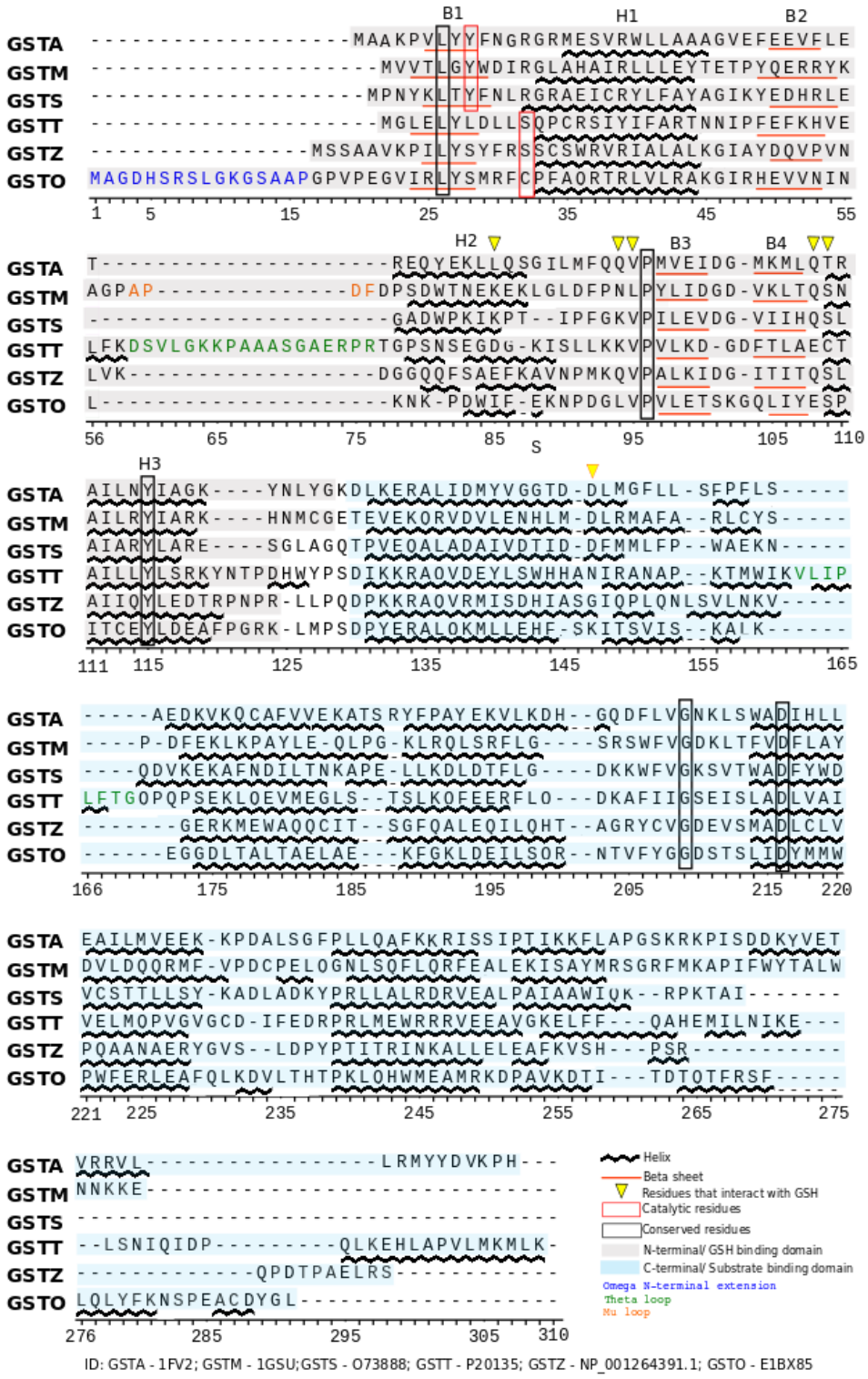


Figure 4.2: The 3D predicted structures of chicken cGST are showing the lack the beta chain. The catalytic active site present in each class is shown.

Table 4.2: Sequence identity between and within cGST classes of chicken sequences

ENSGALT00000026336	**GSTA1-1**	100%										
P20136	**GSTM**	20.90%	100%									
ENSGALT00000016938	**GSTS**	27.13%	26.63%	100%								
ENSGALT00000010254	**GSTT**	22.07%	17.27%	18.59%	100%							
ENSGALT00000013697	**GSTO**	18.01%	12.27%	20.10%	16.25%	100%						
ENSGALT00000016986	**GSTZ**	18.46%	11.36%	21.60%	21.23%	21.68%	100%					
ENSGALT00000008355	**GSTT**	19.81%	17.72%	19.59%	57.85%	18.75%	23%	100 %				
ENSGALT00000026339	**GSTA2**	70.72%	20.45%	27.63%	18.83%	17.04%	17.93%	17.04 %	100%			
ENSGALT00000031634	**GSTA1-3**	84.23%	22.27%	25.62%	20.27%	18.01%	16.21%	19.81 %	70.27 %	100%		
ENSGALT00000026333	**GSTA4**	63.96%	18.63%	28.14%	18.69%	15.65%	19.02%	16.08 %	70.40 %	62.61 %	100 %	
ENSGALT00000026335	**GSTA3**	66.21%	20%	26.63%	20.08%	18.30%	17.85%	16.96 %	68.60 %	67.11 %	68.30%	100%
		GSTA1-1	**GSTM**	**GSTS**	**GSTT**	**GSTO**	**GSTZ**	**GST T-LIKE**	**GSTA 2**	**GSTA 1-3**	**GST A4**	**GSTA 3**

The topological arrangement shows that cGSTs can be divided broadly into two major groups: one with cGSTA, cGSTM, cGSTS and cGSTP in vertebrates (cGSTP was not found in birds) and the other including cGSTT, cGSTZ and cGSTO in vertebrates. Theta, zeta and omega cGSTs are ancestral to alpha, sigma and mu classes. The phylogenetic relationship also holds true with respect to the conserved catalytic residues present in cGST classes (Figure 4.4). The catalytic residue activates GSH during catalysis and show evolutionary conservation in the catalytic site cGSTA, cGSTM and cGSTS have common catalytic site as tyrosine, whereas cGSTO have cysteine and cGSTT and cGSTZ have serine showing the shift in catalytic site among these groups of cGST classes.

4.3.4 Adaptive evolution

Earlier studies have shown that gene expansion and functional diversification of cGST played an important role in the adaptive evolution of these proteins in mammals but no such studies have been done on avian cGSTs. Here, we explored the adaptive role of cGST gene family from 48 bird genomes. Positive selection play an important role in functional diversification and is represented by omega ω(dN/dS) > 1 where dn and ds means non synonymous and synonymous substitution rates, respectively.

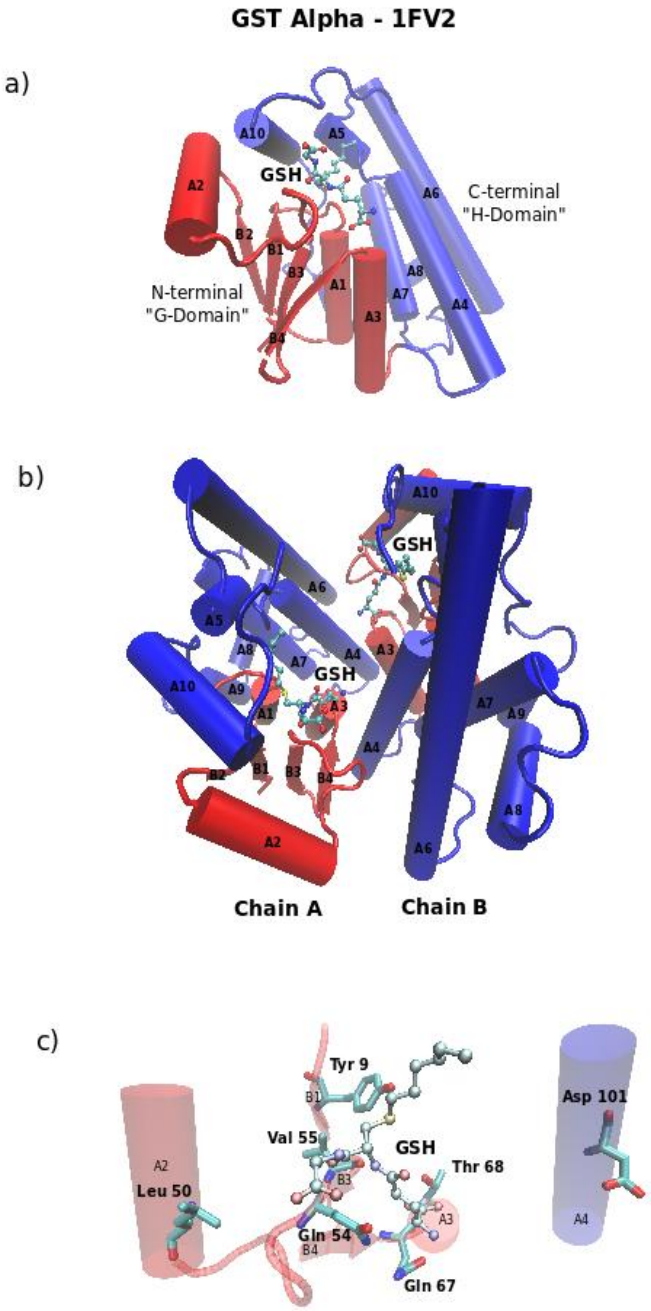


Figure 4.3:15The solved 3D structure of GSTA from chicken (PDB ID IVF2) a) the monomer showing the major domains with GSH b) dimeric structure of GST1-1 with domains and GSH c) The known catalytic and GSH binding sites are shown.

We used site models implemented in the PAML package for calculating variable ω ratios among sites. The site models allow the comparison of two nested site models, a neutral model that does allow for (ω < = 1) and an alternative model for positive

selection (ω > 1). We found that duplicated members of cGST classes are under strong positive selection with cGSTA1, cGSTA2 and cGSTA3 having highest omega and maximum number of sites under positive selection (4. 3 and Table 4. 4).

From the two members of cGSTT found in our study, only one had significant evidence of being evolving under positive selection. The other cGST classes found in birds (cGSTM, cGSTS, cGSTO, and cGSTZ) do not have duplicated members and of those only cGSTO and cGSTZ were under positive selection. Positive selection is the major driving force for the incorporation of useful changes in proteins, allowing the species to better adapt and survive either by functional innovation (gain of function) or subfunctionalization (sharing ancestral gene functions). Catalytic promiscuity plays an important role in the evolution of new functions (Tawfik and S 2010). Therefore, positive selection found in different avian cGSTs can be explained based on the important role played by GST in the protection from harmful stress conditions, complemented with the critical specific functions of each cGST. The detected positive selected sites are located in the substrate binding domain further supporting the role of positive selection in the fine tuning of the molecule to obtain protection against the wide variety of cellular stress (Table 4.4 and Figure 4.5).

The cGST alpha has the maximum number of isoforms in birds, with the highest number of representatives being found only in turkey (cGSTA1-1, cGSTA1-2, cGSTA1-3, cGSTA2, cGSTA3 and cGSTA4) but cGSTA1, cGSTA2, cGSTA3 and cGSTA4 were consistent in most bird genomes, which suggest that cGSTA underwent recent species-specific duplication in the turkey genome. The cGST alpha, mu and theta are involved in cellular reaction connected with stress conditions, such as the metabolism of products from oxidative stress reactions, thus having multiples gene copies increase the fitness due to improved elimination of harmful chemicals (Fonseca et al. 2010).

The cGST alpha are related with processing of small hydrophobic molecules (Wu and Dong 2012), but a non-detoxification function was also reported for rat cGSTA3 that have steroid isomerase activity in ovary and testis (Sheehan et al. 2001) (Wu and Dong 2012). The GSTA genes also have affinity for AFBO, an intermediated originated from Aflatoxin B1 activation by CYP450 complex (Kim et al. 2013). Aflatoxin B1 is a metabolite from some *Aspergillus* species and are frequently present in cereals and tree nuts that are the feeding base of several birds (Kim et al. 2013). This supports the importance of gene duplication and positive selection in the cGST alpha class. cGST mu also has a role in AFBO detoxification (Wu and Dong 2012) (Wang et al. 2000), but

in vitro studies using turkey cGST mu do not revealed activity in AFBO detoxification (Bunderson et al. 2013).

The cGST theta has affinity for several substrates, such as halogenated methanes and ethanes (like DCM), halogenated organic compounds (EDB, PNBC or PNPB) and also has sulfatase activity (Wu and Dong 2012). In addition, it was also reported that cGST theta can have some affinity for AFBO(Landi 2000). The positive selection in only one theta isoform suggests possibly the retention of important ancestral function by one of the genes and events of functional diversification by the other gene duplicates.

Lack of duplicated isoforms is observed in other cGSTs in birds. Some of those cGSTs have important functions, such as cGSTS (Hematopoietic prostaglandin D synthase) that is involved in the prostaglandin synthesis. Changes in cGSTS can have adverse effect on prostaglandin synthesis disturbing prostaglandins dependent functions, which may explain lack of duplicates in this class. By contrast, positive selection found in cGSTZ and cGSTO suggests a secondary but important involvement of these GSTs in protecting tissues against various stress conditions like reactive species stress and /or xenobiotics. The cGST omega is found duplicated in mammals and the ancestral enzyme is involved in ascorbate regeneration (capacity that is always well conserved in cGSTO2) and responsible for the maintenance of ascorbic acid levels in the brain (Wu and Dong 2012). The duplicated cGSTO on the other hand acquired the capacity to metabolize arsenic(Fonseca et al. 2010). The cGST zeta also catalyzes the biotransformation of several α-holoacids and is involved in important homeostatic reactions(Blackburn et al. 2006). Overall, the gene duplication, the catalytic promiscuity and the positive selection found in the substrate binding domain favored adaptive evolution by protecting against harmful internal and external stress conditions. Evaluation of the radical amino acid properties changes using TreeSAAP revealed that most of the positive selected sites retrieved by PAML also showed evidence of positive radical changes (+6 to +8) in amino acid properties (Table 4.4 and Figure 4.5). We further designated the sites having six and more unique changes as type I and type II sites (Table 4.4 and Figure 4.5). The positive radical changes can affect the functional and structural properties of the protein signifying their adaptive role.

Table 4.3: Positive selection results of avian GST sequences using nested site models comparison M1a and M2a, M7 and M8 in PAML

Gene	Model	Parameters	lnl	LRT test	deltaLRT	df	p-value	Significance level
GSTA1	M1a	w0 = 0.07336 p0 = 0.72989 w1 = 1.00000 p1 = 0.27011	-5932.933992					
	M2a	w0 = 0.07535 p0 = 0.69749 w1 = 1.00000 p1 = 0.25272 w2 = 4.21319 p2 = 0.04979	-5868.087047	M1a vs M2a	129.69389	2	6.88E-29	***
	M7	p = 0.20358 q = 0.50183	-5923.823749					
	M8	p0 = 0.93725 p = 0.26388 q = 0.78971 (p1 = 0.06275) w = 3.25411	-5857.679047	M7 vs M8	132.289404	2	1.88E-29	***
GSTA2	M1a	w0 = 0.06083 p0 = 0.78501 w1 = 1.00000 p1 = 0.21499	-5516.822492					
	M2a	w0 = 0.06359 p0 = 0.77054 w1 = 1.00000 p1 = 0.20646 w2 = 5.14075 p2 = 0.02300	-5459.027098	M1a vs M2a	115.590788	2	7.94E-26	***
	M7	p = 0.16944 q = 0.53005	-5513.683304					
	M8	p0 = 0.97664 p = 0.20538 q = 0.73274 (p1 = 0.02336) w = 4.61462	-5454.000425	M7 vs M8	119.365758	2	1.20E-26	***
GSTA3	M1a	w0 = 0.09399 p0 = 0.81375 w1 = 1.00000 p1 = 0.18625	-5816.961419					
	M2a	w0 = 0.09556 p0 = 0.79708 w1 = 1.00000 p1 = 0.18915 w2 = 4.93900 p2 = 0.01376	-5787.601448	M1a vs M2a	58.719942	2	1.77E-13	***
	M7	p = 0.22567 q = 0.67359	-5797.301416					
	M8	p0 = 0.98578 p = 0.25759 q = 0.84513 (p1 = 0.01422) w = 4.40390	-5765.23931	M7 vs M8	64.124212	2	1.19E-14	***
GSTA4	M1a	w0 = 0.06952 p0 = 0.83083 w1 = 1.00000 p1 = 0.16917	-4580.557585					
	M2a	w0 = 0.06952 p0 = 0.83083 w1 = 1.00000 p1 = 0.12177 w2 = 1.00000 p2 = 0.04740	-4580.557585	M1a vs M2a	0	2	1	NS
	M7	p = 0.22511 q = 0.90127	-4577.189098					
	M8	p0 = 0.93145 p = 0.35544 q = 2.36034 (p1 = 0.06855) w = 1.26824	-4571.985425	M7 vs M8	10.407346	2	0.005496	**
GSTM	M1a	w0 = 0.01362 p0 = 0.94214 w1 = 1.00000 p1 = 0.05786	-1783.341411					
	M2a	w0 = 0.01363 p0 = 0.94214 w1 = 1.00000 p1 = 0.03574 w2 = 1.00000 p2 = 0.02212	-1783.341411	M1a vs M2a	0	2	1	NS
	M7	p = 0.08116 q = 1.83387	-1768.735933					
	M8	p0 = 0.99999 p = 0.08117 q = 1.83415 (p1 = 0.00001) w = 1.00000	-1768.736464	M7 vs M8	0.001062	2	0.999469	NS
GSTS	M1a	w0 = 0.05819 p0 = 0.86425 w1 = 1.00000 p1 = 0.13575	-5071.357798					
	M2a	w0 = 0.05821 p0 = 0.86434 w1 = 1.00000 p1 = 0.09896 w2 = 1.00000 p2 = 0.03670	-5072.401113	M1a vs M2a	2.08663	2	0.352285	NS
	M7	p = 0.22076 q = 1.24522	-5056.561437					
	M8	p0 = 0.94757 p = 0.30278 q = 2.62904 (p1 = 0.05243) w = 1.00000	-5055.635983	M7 vs M8	1.850908	2	0.396351	NS
GSTT1L	M1a	w0 = 0.09143 p0 = 0.79800 w1 = 1.00000 p1 = 0.20200	-5845.953537					
	M2a	w0 = 0.09448 p0 = 0.79476 w1 = 1.00000 p1 = 0.18360 w2 = 2.35481 p2 = 0.02163	-5841.937016	M1a vs M2a	8.033042	2	0.018016	*
	M7	p = 0.31067 q = 0.98315	-5827.889827					
	M8	p0 = 0.96857 p = 0.40158 q = 1.66464 (p1 = 0.03143) w = 1.93589	-5815.817468	M7 vs M8	24.144718	2	5.72E-06	***
GSTT1L2	M1a	w0 = 0.03052 p0 = 0.91752 w1 = 1.00000 p1 = 0.08248	-4097.579884					
	M2a	w0 = 0.03052 p0 = 0.91752 w1 = 1.00000 p1 = 0.08248 w2 = 15.89914 p2 = 0.00000	-4097.579884	M1a vs M2a	0	2	1	NS
	M7	p = 0.14706 q = 1.44453	-4091.750895					
	M8	p0 = 0.96154 p = 0.21701 q = 3.47296 (p1 = 0.03846) w = 1.00000	-4090.676332	M7 vs M8	2.149126	2	0.341447	NS
GSTZ	M1a	w0 = 0.07366 p0 = 0.81148 w1 = 1.00000 p1 = 0.18852	-5858.203431					
	M2a	w0 = 0.07908 p0 = 0.80502 w1 = 1.00000 p1 = 0.16501 w2 = 3.11980 p2 = 0.02998	-5828.564743	M1a vs M2a	59.277376	2	1.34E-13	***
	M7	p = 0.19256 q = 0.57727	-5865.884961					
	M8	p0 = 0.96517 p = 0.25473 q = 0.97185 (p1 = 0.03483) w = 2.81044	-5829.033884	M7 vs M8	73.702154	2	9.90E-17	***
GSTO	M1a	w0 = 0.08781 p0 = 0.71438 w1 = 1.00000 p1 = 0.28562	-5050.515364					
	M2a	w0 = 0.09323 p0 = 0.71085 w1 = 1.00000 p1 = 0.23037 w2 = 1.86464 p2 = 0.05879	-5046.85327	M1a vs M2a	7.324188	2	0.025679	*
	M7	p = 0.23142 q = 0.51018	-5045.432859					
	M8	p0 = 0.84626 p = 0.41413 q = 1.92114 (p1 = 0.15374) w = 1.45627	-5033.019171	M7 vs M8	24.827376	2	4.06E-06	***

Level of significance (*p valve 0.01 to 0.05; ** 0.001 to 0.005; *** p < 0.001; ns: non-significant)

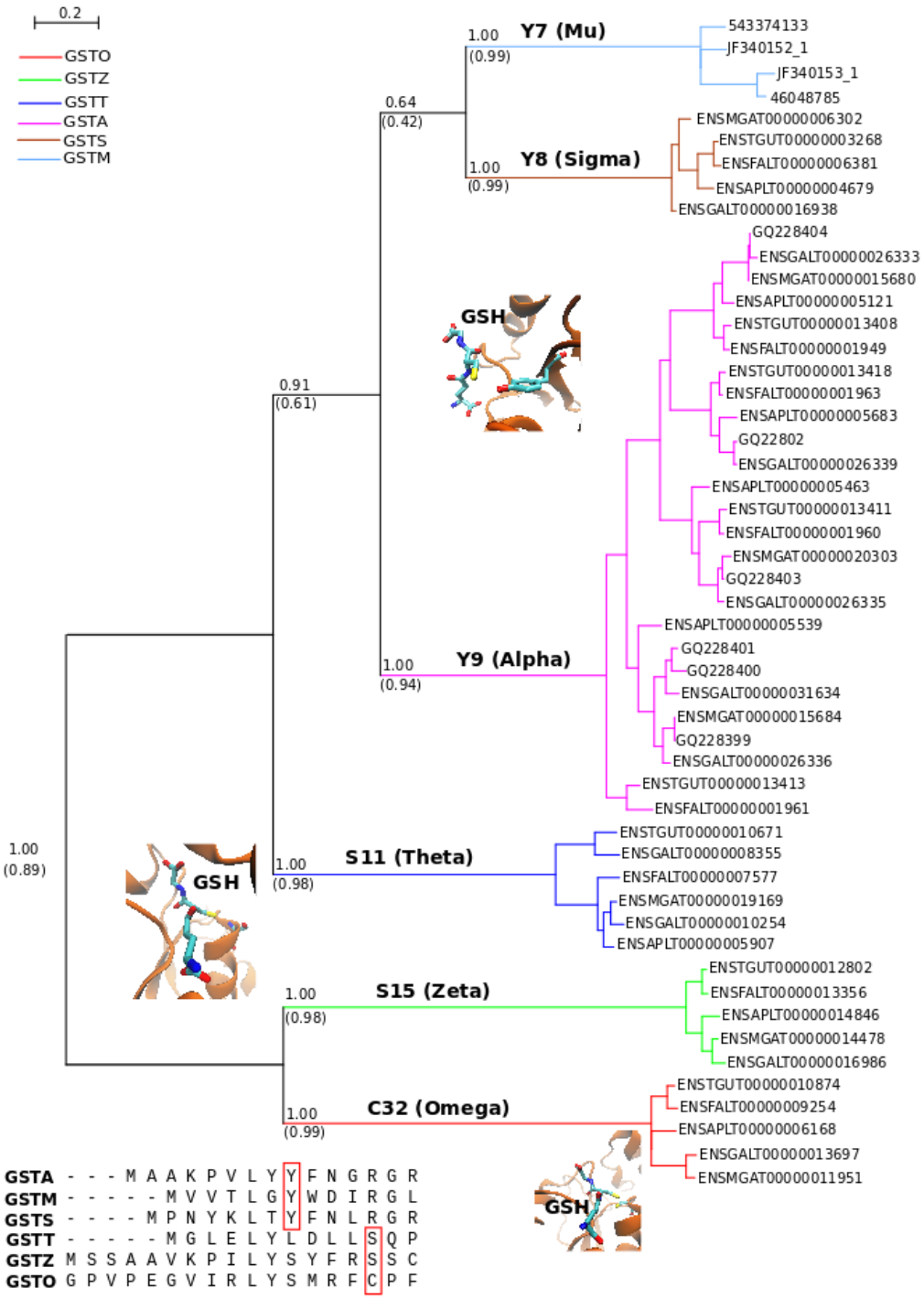


Figure 4.4: The Bayesian phylogeny of avian cGSTs using sequnces from five birds genomes with support for othologous and paralogous relationship . The valves above the branches are the bayesian posterior probability and the ML bootstrap support after 1000 replication are shown below the branches. The conserved catalytic site with evolutionary shift is shown for each class

Table 4.4: Positive selected sites with ω and Bayesian (BEB) analysis posterior probabilities shown for sites with PP > 0.95 in M8 that also have a PP > 0.90 in M2a. TreeSAAP analysis results present the total number of radical changes in amino acid properties and their assigned categories. Type I sites are shown in bold.

Gene	Sites	PAML		TreeSAAP properties			
		M2a	M8	Total	Chemical	Structural	Other
GSTA1 1vf1	**18 S**	0.923 4.092 +- 1.021	0.965* 3.512 +- 0.587	0	-	-	-
	72 L	1.000** 4.376 +- 0.540	1.000** 3.610 +- 0.336	6	R_F, R_a, H_p	B_l, M_V, V^0	-
	108 S	1.000** 4.376 +- 0.540	1.000** 3.610 +- 0.336	3	R_a, H_p, pH_i	-	-
	110 P	1.000** 4.376 +- 0.540	1.000** 3.610 +- 0.336	7	pH_i, R_a, H_p	H_C, V^0, K^0	M_W
	152 R	0.983* 4.317 +- 0.687	0.975* 3.541 +- 0.532	1	pH_i	-	-
	170 K	0.999** 4.374 +- 0.546	1.000** 3.609 +- 0.340	8	RF, pH_i, Pr, p	H_C, V^0, M_V	M_W
	208 S	1.000** 4.376 +- 0.542	1.000** 3.610 +- 0.338	2	p	K^0	-
GSTA2	**49 S**	1.000** 5.194 +- 0.606	1.000** 4.579 +- 0.561	3	P_r, R_a	B_l	-
	108 A	1.000** 5.194 +- 0.606	1.000** 4.579 +- 0.561	4	p, H, R_a	K^0	-
	208 L	0.999** 5.191 +- 0.615	1.000** 4.578 +- 0.563	3	R_a, H_p	K^0	-
	212 S	1.000** 5.194 +- 0.606	1.000** 4.579 +- 0.561	2	-	V^0, K^0	-
	215 N	1.000** 5.194 +- 0.606	1.000** 4.579 +- 0.561	7	R_F, H, R_a, H_p, P_r, p, pH_i	-	-
GSTA3	**10 V**	1.000** 4.821 +- 0.872	1.000** 3.262 +- 0.871	1	R_F	-	-
	49 S	1.000** 4.821 +- 0.873	1.000** 3.262 +- 0.871	2	P_r,p	-	-
	211 H	1.000** 4.822 +- 0.872	1.000** 3.262 +- 0.870	4	R_F, H, H_{nc}, R_a	-	-
GSTTL1	**134T**	0.907 2.322 +- 0.511	0.969* 1.678 +- 0.425	1	R_a	-	-
	153G	0.961* 2.397 +- 0.414	0.986* 1.692 +- 0.412	2	R_a	B_l	-
	235A	0.926 2.341 +- 0.481	0.982* 1.687 +- 0.414	1	R_a	-	-
GSTO	**137 L**	0.919 1.653 + 0.444	0.995** 1.497 + 0.048	3	P_r, p, R_a	-	-
GSTZ	**85R**	0.921 3.113 +- 0.742	0.985* 2.649 +- 0.432	1	H	-	-
	119S	0.999** 3.304 +- 0.432	1.000** 2.677 +- 0.383	8	R_a, pH_i, μ	B_l, H_C, M_V, V^0	M_W
	129M	0.960* 3.204 +- 0.611	0.995** 2.667 +- 0.399	2	P_r, p	-	-
	137T	1.000** 3.306 +- 0.427	1.000** 2.677 +- 0.382	1	pH_i	-	-
	172A	1.000** 3.306 +- 0.427	1.000** 2.677 +- 0.382	4	R_F, R_a	V^0, K^0	-

The number of positive selected sites is as per following reference sequences (GSTA1 = 1VF1, GSTA2 = ENSGALT00000026339, GSTA3 = ENSGALT00000026335, GSTTL1 = E1BUB6, GSTO = E1BX85, GSTZ = NP_001264391.1)

B_l: Bulkiness; R_a: Solvent accessible reduction ratio; H_c: Helical contact area; pH_i: Isoelectric point; M_v: Molecular volume; M_w: Molecular weight; V^0: Partial specific volume; H: Hydropathy; H_{nc}: Normal consensus hydrophobicity; P_r: Polar requirement; p: Polarity; R_F: Chromatographic index; H_p: Surrounding hydrophobicity; K^0: Compressibility; R_F: Chromatographic index; μ: Refractive index

4.4 Conclusions

Events of gene duplication and functional diversification have contributed to the dynamic evolution of cGST gene family involved in multiple functions including, detoxification of harmful endogenous and exogenous compounds to biosynthesis of leukotrienes, prostaglandins, testosterone, and progesterone, and degradation of tyrosine. The cGSTs thus play an important role in protection against harmful stress condition like reactive species, which are one of the major causes of disease and are reportedly involved in limiting longevity. Our study explored the evolution of cGST gene family within 48 bird species. We found that protection against reactive species stress (ROS and RCS) and xenobiotic most probably played an important role in the evolution of avian cGSTs. We found that cGST classes (cGSTA and cGSTT), which are known for their involvement in the protection from oxidative stress, were found duplicated in birds. Moreover, evidence of positive selection in these duplicated genes favoring the much-needed diversity for protection against reactive species along with complex variety of xenobiotic compounds, point towards their adaptive significance. We found cGSTS to be highly conserved possibly due to their critical and important role in prostaglandins synthesis. By contrast, positive selection found in cGSTZ and cGSTO also suggest the secondary role played by these genes in detoxification.

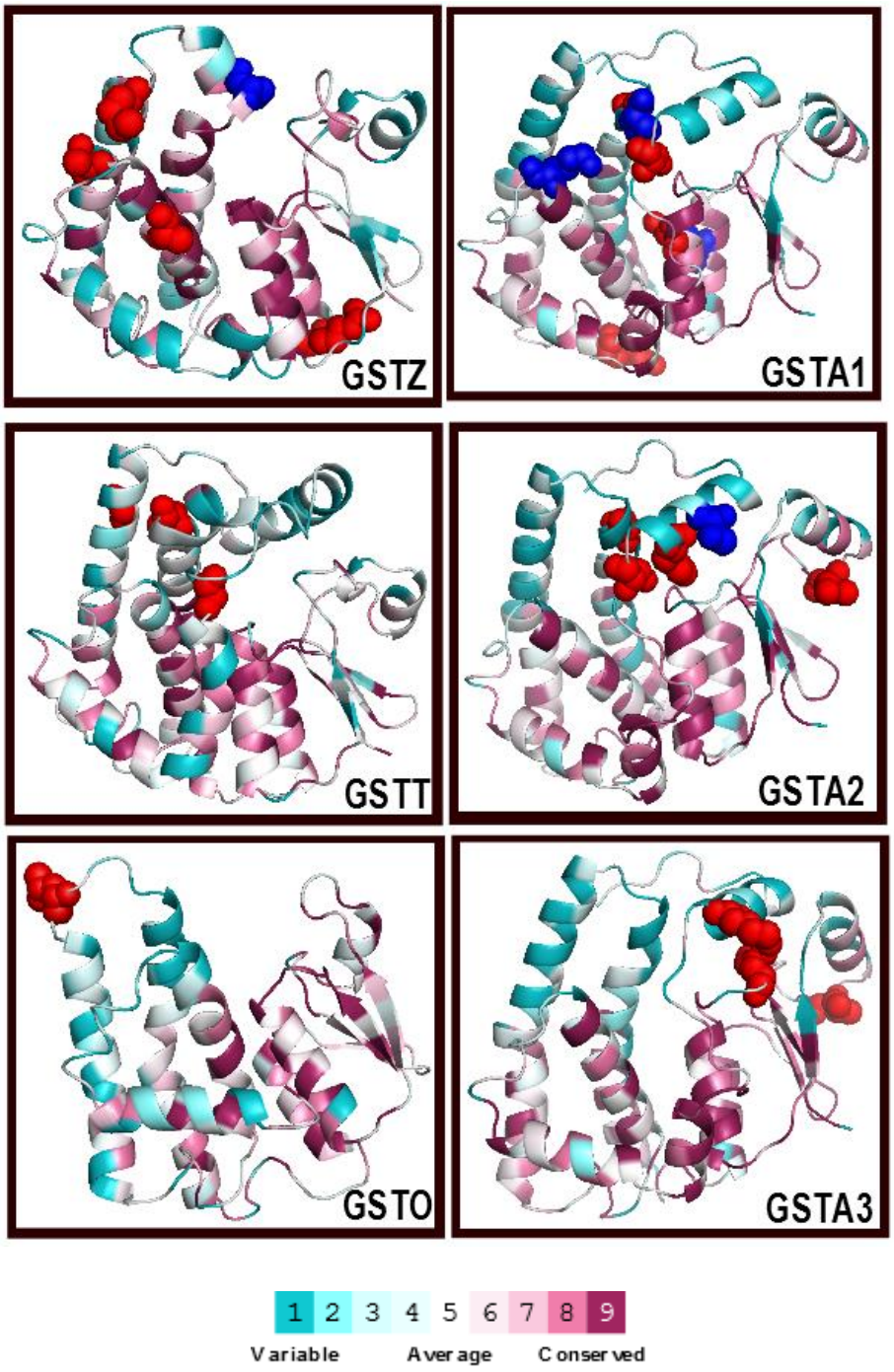


Figure 4.5: Positive selected sites found in present study are displayed in the predicted 3D structure of respective cGST and the color by spheres with type one sites shown by blue sphere and rest by red sphere. The sequences conservation was calculated in Consurf using the predicted 3D structure of the gene. The conservation grade is shown by color coding with turquoise-through-maroon

4.5 Materials and Methods

4.5.1 Identification of cytosolic GSTs in Birds

We performed blast searches (Altschul et al. 1990a) in 48 birds genoomes using the known mammalian and avian cytosolic cGSTs representing (cGSTA, cGSTM, cGSTP, cGSTS, cGSTO and cGSTZ). The sequences retrieved were used for adaptive evolutionary analyses (Supplementary Table 4.1). We also performed the synteny analysis to confirm the orthologous and paralogous relationship of avian cGSTs, and the sequences retrieved were used to positive selection analyses of avian cGSTs. The Complete information of the sequences used in this study is provided in (Supplementary Table 4.1).

4.5.2 Synteny and gene conservation

Synteny analysis was performed using Genomicus version 72.01(Muffato et al. 2010) (Louis et al. 2013) together with manual blast searches (Altschul et al. 1990a) for the genes and neighboring missed by Genomicus due to lack of annotation information. This helped in finding the location of each cGST gene together with the arrangement of neighboring genes to see the level of genomic conservation. The orthologous and paralogous relationship was also determined according to ensembl annotations. The Genomicus uses pair wise comparisons between species for identification of syntenic blocks assuming that the last common ancestor reflects accurately similar order and orientation of genes (Muffato et al. 2010) (Louis et al. 2013).

4.5.3 Phylogeny

The cGST sequences found from five birds genomes (Figure 4.1), were used to show the phylogenic relationship of avian cGST (4. 4). The phylogeny of vertebrate cGSTs is shown in Figure 4.4. The sequences were translated to amino acid and aligned using Muscle implemented in Seaview (Gouy et al. 2010). The alignment was visually inspected and manually edited to adjust the divergent sequences supported by their secondary structure relationships. The aligned sequences were back translated and checked for the level of saturation using all codon position along with 3rd codon position in DAMBE 5.3 (Xia 2013) (Xia et al. 2003). The ISS < ISS.c shows lack of saturation thus comparison between the index of substitution saturation (ISS) and critical value (ISS.c) shows the degree of saturation. ISS and ISSc were compared using symmetrical and asymmetrical topologies. Gene Conversion and recombination was tested using RDP and Geneconv (Martin et al. 2010) (Sawyer 1989). Jmodeltest (Darriba et al. 2012) inferred (TVM+I+G) based on AICc (Akaike information criterion corrected for finite sample size) the best substitution model, which was used for

phylogenetic reconstruction. Maximum likelihood (ML) analysis was performed in PhyML 3.0 (Guindon and Gascuel 2003) (Guindon et al. 2010) and run with 1,000 bootstrap replicates for both nucleotide and amino acid data. Bayesian analyses were carried for the nucleotide and amino acid datasets in MrBayes 3.2. ((Huelsenbeck and Ronquist 2001) (Ronquist and Huelsenbeck 2003)). We used number of substitutions nst = 6, rates = invgamma and T = 0.2 with two runs, each with five Markov chain Monte Carlo (MCMC) chains (one cold four incrementally heated chains were run). The analyses were run for 50 million generations to allow the standard deviation of split frequencies to reach a value below 0.01. Trees and associated model parameters were sampled every 1,000 generations. The first 25 % of the obtained trees were discarded as burnin. The cGSTs genes found from 48 birds genomes (Supplementary Table 4.1) were used for evaluation of level of positive selection using ML trees.

4.5.4 Adaptive evolution analyses of avian GSTs

The cGSTs genes and isoforms for each classes found in the 48 bird genomes (see (Supplementry Table 4.1) for details of sequences used for each isoform) were used for evaluating evidence of positive selection using likelihood ratio tests with the CODEML algorithm from the PAML4.7 package (Yang 2007)(Yang 1997). The PAML implements site specific models (M0, M1A, M2A, M3, M7 and M8). These site models allow the comparison of fit of two nested site-specific models: a neutral model that does allow for ($\omega <= 1$) and alternative model for positive selection ($\omega > 1$). We compared M1a (NearlyNeutral) versus M2a (PositiveSelection); M7 (Beta) versus M8 (Beta&ω) using likelihood ratio test (Nielsen and Yang 1998). The level of significance for LRTs was calculated using a chi-square approximation with twice the difference of log likelihood between the models ($2\Delta lnL$) with $\chi 2$ distribution, with a number of degrees of freedom calculated from the difference in number of parameters between the nested models M1a vs. M2a df = 2; M7 vs. M8 df = 2. Positive selected site were inferred by Posterior Bayesian analysis through the Bayes Empirical Bayesb (BEB) method (Yang et al. 2005).

We further employed protein level analysis using TreeSAAP (Woolley et al. 2003). TreeSAAP makes use of ancestral reconstruction and calculates the goodness-of-fit for changes in amino acid physiochemical properties. By default TreeSAAP make use of 31 properties and categorize the changes in amino acid properties into eight major categories, ranging from conservative to radical. We targeted the positive radical changes, .i.e. sites with magnitude of +6 and more. These radical changes are more likely to affect the structure and function of protein, to identify possible important adaptive changes. The positive selected radical sites were further divided into type I

and type II based on total number of unique radical changes, with type I having 6 or more unique changes and type II having less the six positive radical changes.

4.5.5 Sequence analysis and homology modeling

The sequence identity was calculated using SIAS online tool available at http://imed.med.ucm.es/Tools/sias.html. Secondary structures were predicted using the PSIPRED Protein Sequence Analysis Workbench at http://bioinf.cs.ucl.ac.uk/psipred/. We also used ESPript and ENDscript - http://endscript.ibcp.fr for prediction and presentation of conserved secondary structure based on 3D structure information of each GST classes from birds and mammals. The SWISS-MODEL template library (SMTL version 07-03-14, PDB release 28-02-2014) was searched with Blast and HHBlits for evolutionary related structures matching the target sequence (Altschul et al. 1990b) (Remmert et al. 2012) (Mariani et al. 2011: 9). The PDB template structures with high sequence identity of more than 60% were selected for reliable predicted structures with QMEAN4 valve above 0.55 (Schwede et al. 2003) (Benkert et al. 2011) (Arnold et al. 2006). VMD: visual molecular dynamics (Humphrey et al. 1996) and PyMol software (DeLano Scientific, SanCarlos, CA, USA) were used for structure visualization and displaying. ConSurf web server (Ashkenazy et al. 2010) allowed to generate evolutionary related conservation scores using the predicted structure to see conformation and functionally important region of the protein and displaying positively selected sites. Best hit for avian Theta (E1BUB6) template was 2c3n.1.A 56% identity.

Acknowledgements

IK was funded by a PhD grant (SFRH/BD/48518/2008) from Fundação para a Ciência e a Tecnologia (FCT). AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490).

5

Comparative evolutionary genomics of vertebrates TLR supergene family elucidates host-pathogen arms race in birds and supports the role of birds as viral vectors

5.1 Abstract

Background

The vertebrate's Toll-like receptors (TLRs) supergene family constitutes the first line of immune defense against diverse pathogens and provides a fascinating example of the host-pathogen evolutionary arms race. Here, we provided a comprehensive characterization of the evolutionary genomic dynamics of vertebrate TLRs using whole genome sequencing data from 72 species (mammals, birds, reptiles, amphibians and fishes). We further assessed the genomic diversification of avian TLRs (viral and non-viral TLRs) in 48 bird species employing state-of-the-art DNA and protein level analyses. Finally, we discussed our findings regarding host-pathogen interactions and adaptive evolution of various species/lineages of TLRs to diverse environmental conditions.

Results

We confirmed the presence of 26 TLRs across vertebrates. Most TLRs (TLR3, TLR5, TLR7, TLR8, TLR14, TLR21 and TLR22) were originated early in the evolution of vertebrates and before the diversification of Agnatha and Gnathostomata. TLRs followed multiple events of gene gain/loss leading to species/lineage-specific variations resulting in 20 to 13 subfamilies in fish and mammals, respectively. Significant evidence of positive selection was detected in all avian TLRs studied (1A, 1B, 2A, 2B, 3, 4, 5, 7 and 15) with positive selected (PS) sites comprehending 5%-11% of the overall protein length (omega values varying from 1.5 to 2.5). Both viral and non-viral avian TLRs were under high positive selection with TLR4 (non-viral) and TLR7 (viral) having the highest numbers of PS sites (20 and 23 sites with PP>0.99, respectively). Moreover, such PS sites showed radical changes in amino acid physiochemical properties, including type I radical changes likely affecting the structure and functionality of the TLR proteins. The rapid evolution of TLRs highlights the host-pathogen arms race leading to coevolution of ligands and receptors. Non-viral TLR4 had a high number of PS sites, which may favor their ability to cope with diversified ligands (e.g. lipopolysaccharide and lipoteichoic). The accelerated evolution of viral TLR7 genes suggests its adaptive role in the recognition of ssRNA (very high mutation rates due to lack of mismatch repair). Such strong selective pressure could result from the long-term coevolution of viruses and birds, which is insightful to understand the role of birds as natural virus reservoir and as vector of zoonotic pathogens.

Conclusions

Our results support the wider distribution of TLR supergene family across 72 vertebrate species, with varied rates of gene gain/loss and wide variable numbers of TLRs allowing the recognition of diverse pathogens. The 48 bird genomes provided strong support for the rapid evolution of both viral and non-viral avian TLRs and strengthen the hypothesis that the long-term coexistence of birds and viruses contributed to the strong selective pressure found in viral TLR immune genes. Overall the patterns of gene gain, gene loss and positive selection in the TLR gene family provided strong support for the evolutionary host-pathogen arms race.

Keywords:

Gene gain, gene loss, vertebrates, Toll-like receptors, immune response, host-pathogen, positive selection, gene homogenization.

5.2 Introduction

Toll-like receptors (TLRs) supergene family members are type-I transmembrane glycoproteins, belonging to pathogen recognition receptors (PRRs) class of proteins expressed in the cell membrane and intracellular vesicles like endoplasmic reticulum, endosomes, lysosomes, endolysosomes (Akira et al. 2006) (Kawai and Akira 2010). TLRs form the first line of immune defense system which recognize the diversity of pathogen associated molecular patterns (PAMPs) during the pathogens invasion, triggering the cascade of signaling pathways leading to adaptive immune response to eliminate the pathogens (Schnare et al. 2001)(Medzhitov 2001) (Iwasaki and Medzhitov 2010). The vertebrates TLRs studied till date are involved solely in immune response in contrast with the invertebrates TLR–like proteins, which are also associated with developmental functions (Imler and Hoffmann 2002).

The vertebrate TLR supergene family consists of large and variable number of genes due to gene gain and gene loss, which produces considerable variation within and between vertebrate species/lineages (mammals have 10 to 13 TLRs, while catfish have 20 genes belonging to 15 different TLR families). The TLRs identified till date are grouped into six major families with different number of TLRs members. The family TLR1 consists of TLRs1, 2, 6, 10, 14, 15, 18, 24 and 25, family TLR7 have three members (TLRs 7 to 9) and TLR11 family includes TLRs 11 to 13, 19 to 23 and 26, whereas families TLR3, TLR4, TLR5 are represented by a single member (Quiniou et al. 2013) (Zhang et al. 2013), (Roach et al. 2005).

The variety of TLRs allow to effectively detect diverse pathogenic ligands (O'Neill et al. 2013). The ligands of most of the TLRs have been identified. In mammals, TLR2 is

able to form heterodimers with TLR1, TLR6 and TLR10, and recognize di and triacylated lipoproteins (Guan et al. 2010) (Hasan et al. 2005) (Takeuchi et al. 2002). TLR4 binds with lipopolysaccharides (LPS) of Gram-negative bacteria and lipotheicoic acids of Gram-positive bacteria (Knapp et al. 2008)(Hoshino et al. 1999) (Kim et al. 2007) (Park et al. 2009), whereas TLR5 binds with flagellin protein present in the flagella (motility apparatus of microbial pathogens) (Hayashi et al. 2001). TLR3 recognizes dsRNA, TLR7 and TLR8 binds with ssRNA (Takeuchi and Akira 2007) (Liu et al. 2008) (Gantier et al. 2008) (Lund et al. 2004) and TLR9 binds with viral CpG-DNAs (Alexopoulou et al. 2001) (Diebold et al. 2004) (Heil et al. 2004) (Krug et al. 2004). In fishes, TLR22 also binds with dsRNA (Matsuo et al. 2008). The recent studies suggest the role of TLR11 and TLR12 in recognition of profilin present in protozoan pathogens (Yarovinsky et al. 2005) (Koblansky et al. 2013) and TLR13 in 23S rRNA (Shi et al. 2011) (Oldenburg et al. 2012). TLR15 and TLR21 in birds recognizes yeast-derived agonist and microbial CpG-DNAs, respectively (Boyd et al. 2012) (Keestra et al. 2010). The TLR21 also seems to replace the function of the missing TLR9 in birds (Brownlie et al. 2009). Avian TLRs are assumed to have similar ligands to those reported for mammalian TLRs (Alcaide and Edwards 2011). Each TLR gene consists of a highly conserved intracellular (cytoplasmic) Toll/interleukin-1 receptor (TIR) domain, which is responsible for signal transduction (Medzhitov et al. 1997), a conserved single transmembrane region and a variable extracellular domain (ECD), involved in the ligand recognition and dimerization. ECD consists of variable numbers (~16 to 28) of leucine-rich repeats (LRRs) motif (Matsushima et al. 2007).

Each LRR unit of ECD is 20-30 amino acid long with a beta sheet concave region composed by a “LxxLxLxxN” motif and a variable convex alpha helix region (Kajava 1998) (Kobe and Deisenhofer 1994) (Kobe and Kajava 2001) (Bell et al. 2003) (Matsushima et al. 2007) (Matsushima et al. 2010). The rich cysteine region in both N and C-terminal (LRR-NT and LRR-CT) of ECD form an horseshoe arc, which protects the hydrophobic core from exposure to solvents (Quiniou et al. 2013) (Kang and Lee 2011).

The diverse mechanism used by different TLRs paralogs for TLR-ligand recognition and binding, causes the formation of a m-shaped homo or heterodimeric complex, resulting in the activation of downstream signaling cascade by TIR domains (Jin et al. 2007) (Brodsky and Medzhitov 2007)(Kawai and Akira 2006) (Janeway and Medzhitov 2002) (Akira et al. 2006) (Jin and Lee 2008) (Botos et al. 2011) (Gantner et al. 2003) (Takeuchi et al. 2002). The TIR dimer is recognized by the TIR domain present in different signaling adaptor proteins such as MyD88, MAL, TRIF and TRAM (O’Neill and Bowie 2007), which trigger the cascade of signaling pathway resulting in the activation

of NFkB and the expression of various inflammatory and anti-pathogenic proteins (Doyle and O'Neill 2006) and the initiation of adaptive immune responses, which finally leads to the elimination of the invading pathogens (Iwasaki and Medzhitov 2010) (Medzhitov 2007) (Takeda and Akira 2004) (Kumar et al. 2009a).

The gene gain and loss play an important role in the gene family evolution. The large and variable numbers of TLR genes allow the identification of a large variety of ligands present in diverse pathogens (bacterial, fungal, protozoan and virus). The important immunological function of TLRs, .i.e. protection of host from pathogens, requires them to evolve faster due to selective pressure of rapidly and ever evolving pathogens. This host-pathogen arms race makes TLRs important candidates for studying differential pathogen outcomes in diverse hosts. The functional variation of TLR genes favored by positive diversifying selection may improve the success of pathogen recognition. Recent studies have revealed important compensatory mechanisms of the TLR supergene family in both adaptive and innate response in the absence of major histocompatibility complex (MHC) II, CD4 and invariant chain (Ii) in Cod (Star et al. 2011), where TLR act as an alternative to MHCII, a well known conserved feature of the adaptive immune system of jawed vertebrates (Flajnik and Kasahara 2010) (Litman et al. 2010) (Star et al. 2011). Similarly, the coelacanth genome shows unique TLR gene family evolution due to its evolutionary proximity with the fish and tetrapods, along with its unique immune system lacking IgM (Amemiya et al. 2013). The innate immune receptors of coelacanth genome, had a mixture of mammalian and teleost specific TLRs related with the transitional position of coelacanth (Boudinot et al. 2014).

Here, we studied the evolutionary genomics of vertebrate TLRs using whole genome sequencing data from 72 species (mammals, birds, reptiles, amphibians and fishes). We further assessed the diversification of avian TLRs (viral and non-viral TLRs) in 48 bird genomes using state-of-the-art DNA and protein level analyses. Finally, we discussed our findings regarding host-pathogen interactions and adaptive evolution of various species/lineages of TLRs to diverse environmental conditions.

5.3 Results and Discussion

5.3.1 Genome Scan and phylogenetic resolution of vertebrates TLR supergene family

The comprehensive TLR data from diverse vertebrate species/lineages is very crucial for filling the important phylogenetic void and proper resolution of the vertebrate TLR superfamily. Our study was able to address the above issue and was able to present a clear phylogeny, homologous relationship and proper nomenclature of TLR supergene

family in diverse vertebrates (Figure 5.1), further supported with syntenic genomic information (Table 5.1).

The vertebrates TLR supergene family have been explored previously (Zhang et al. 2013), (Roach et al. 2005), but detailed information from the Sauropsida lineage, which includes birds and reptiles, was mostly absent. Here, we scanned in detail 48 birds (Jarvis et al. 2014) (Zhang et al. 2014), 8 reptiles (Table 5.1) and 4 fish genomes (Table 5.1), plus up to date genomic information from other vertebrates (mammals, fishes, amphibian and lamprey) providing a comprehensive picture of the TLR supergene family evolution (Table 5.1).

We explored the TLR variations in additional new genomes including the fishes Oryzias latipes, Oreochromis niloticus and Gasterosteus aculeatus (Table 5.1), eight diverge reptiles studied inhabit diverse environments (aquatic, semi-aquatic and terrestrial) under unique adaptive pressures (pathogens) including one lizard (Anolis carolinensis), one snake (Python molurus), three turtles (Pelodiscus sinensis, Chrysemys picta bellii and Chelonia mydus) and three crocodilians (Alligator mississipiensis, Gavialis gangeticus and Crocodylus porosus) and the genomic scan revealed unique distribution of TLRs. The TLRs have previously been reported in birds (Alcaide and Edwards 2011) (Temperley et al. 2008) and consisted of 8 subfamilies (TLR1-5, TLR7, TLR15 and TLR21). The present genomic scan of 48 bird genomes (Jarvis et al. 2014) (Zhang et al. 2014) belonging to diverse ecological niche did not revealed any new TLR.

The genomic data from diverse vertebrates (Figure 5.1 and Table 5.1) was used for the phylogentic reconstruction of the TLR supergene family. Phylogenetic analysis of TLR superfamily strongly supports the classification in six families with new and different subfamilies belonging to one of the six families, thereby increasing the total number of members (Table 5.1 and Figure 5.1). The family TLR1 (subfamilies TLR1, TLR2, TLR6, TLR10, TLR14, TLR15, TLR18, TLR24, TLR25 and TLR27), family TLR3, family TLR4, family TLR5, family TLR7, (subfamilies TLR7, TLR8 and TLR9) and family TLR11, (subfamilies TLR11-TLR13, TLR16, TLR19-TLR23 and TLR26). Phylogenetic relationship shows that TLR gene family evolution is shaped by differential rate of gene gain and loss. We found for the first time the TLR13, TLR18, TLR15, TLR21 and TLR22 in reptiles (Table 5.1). Thus, our results confirm that these TLRs are not restricted to particular vertebrate group as previously suggested (Quiniou et al. 2013) (Zhang et al. 2013) (Roach et al. 2005) (Alcaide and Edwards 2011) (Ishii et al. 2007) (Kasamatsu et al. 2010).

On other hand, some TLR subfamilies are limited to particular vertebrate group and species, specifically TLR6, TLR10, TLR11 and TLR12 that are present only in mammals, TLR25 is present only in jawed fishes and TLR24, TLR26 and TLR27 are species-specific, present in lamprey, cat fish and coelacanth, respectively. .The gene labeled as TLR13 in mammals shows that TLR13 clade have two diverse groups, one with coelacanth and mammals (labeled as 13M) and other with amphibian and reptiles (labeled as 13X) (Figure 5.1), supported by high bootstrap values and synteny analysis (Supplementary file 5.3 Figure (H1 and H2). Among fishes, TLR13 is coelacanth specific being absent in all other fishes.

Previous studies have also questioned the dubious naming of TLR14, TLR18 and TLR25, which all are phylogenetically closely related (Figure 5.1). We were able to resolve this ambiguity based on synteny and phylogenetic relationships. We support the TLR18 naming of zebrafish and catfish and our phylogeny and syntenic analysis suggests that TLR14 reported earlier in Fugu (Kasamatsu et al. 2010) belong to this clade and thus should be renamed as TLR18 (Supplementary file 5.3 Figure (I2). This further supports the absence of TLR14 from all jawed fish lineage. The genes previously referred as TLR18 in medaka, tilapia and cod (Zhang et al. 2013) are syntenic and form a monophyletic clade with TLR25 from cat fish (Zhang et al. 2013) and thus should be referred as TLR25 (Figure 5.1 and Supplementary file 5.3 Figure (P). The phylogeny also suggests that TLR14 and TLR25 are closely related and diversified after duplication from a common ancestral gene.

TLR14 was first originated in lamprey with duplication and diversification leading to TLR14D and TLR14A-C. TLR14A to C, are distantly related with TLR14D, and TLR14A to C are more close to fish TLR25 (Figure 5.1). The frog TLR14 also forms a distinct monophyletic clade possibly pointing towards distinct origin of the TLR14 in lamprey and frog (also visible in the phylogeny and the lack of shared synteny between TLR14 present in these two genomes). The searches for lamprey TLR7/8 A and B genes did not revealed any hits in other genomes, suggesting that TLR7/8A and TLR7/8B are lamprey specific present at the base of TLR7 and TLR8 clade. This supports that TLR7/8 may have given origin to TLR7 and TLR8 prior to the divergence of jaw fishes and tetrapods (Kasamatsu et al. 2010). The frog TLR12 (Kasamatsu et al. 2010) is more close to zebrafish TLR19 than the mammalian TLR12, and also lacks conserved syntenic relationship with both TLR12 and TLR19. Thus, we propose it to be retained as TLR16 (Roach et al. 2005). We also found partial TLR2 like (TLR2L) sequences in coelacanth, whose phylogenetic position was close to TLR2 (not shown). Thus, we refer it as TLR2L and placed it under TLR2. TLR15 earlier reported only in birds was

also found in reptiles. The TLR dataset obtained from an extended evaluation of additional newly sequenced genomes providing a much accurate classification and a well-supported TLR gene family phylogeny and overall improved evolutionary assessment of the vertebrate TLR gene family.

5.3.2 Comparative genomics, gene gain and loss in the evolution of vertebrate TLR superfamily

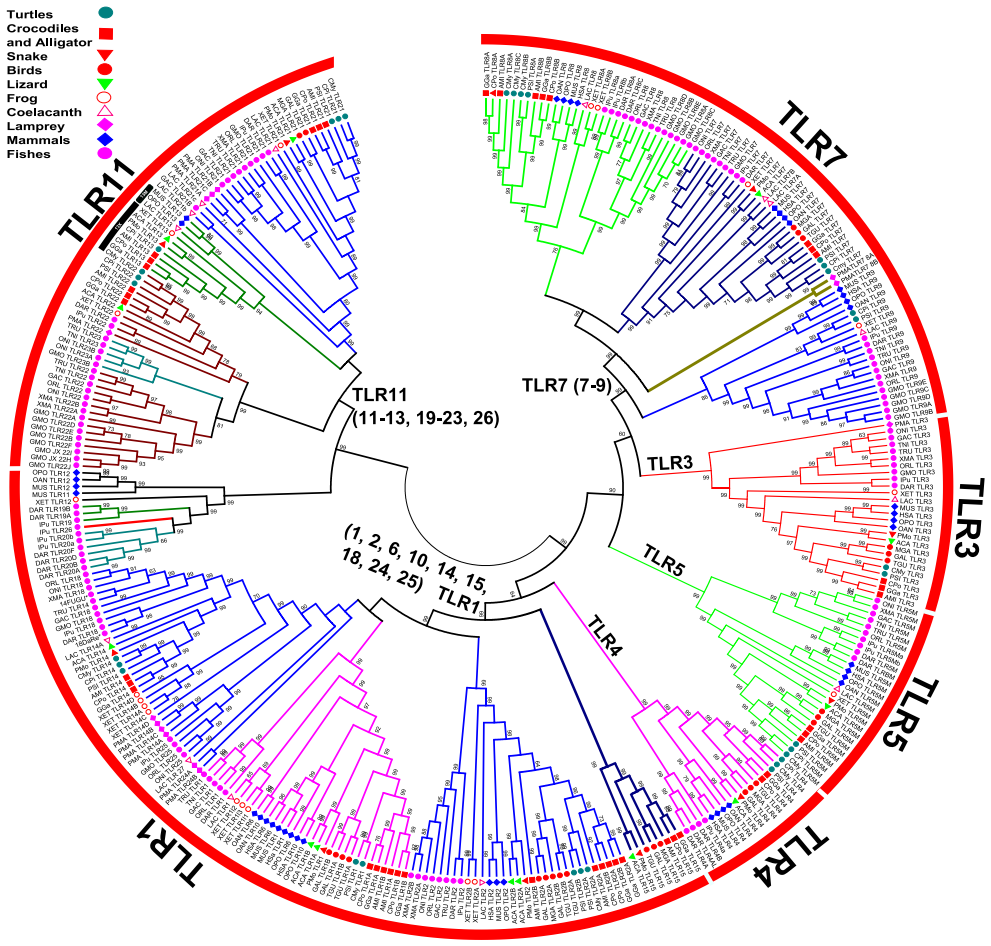


Figure 5.1: Phylogeny of vertebrate TLR gene family from 26 species (Table 5.1). The NJ tree was made in mega with 1000 bootstrap.

The dynamics of gene gain and gene loss plays an important role in evolution and diversification of gene families and can originate diverse gene family repertoires within and among vertebrate groups. These lineage/species-specific changes are important evolutionary phenomenon shaped by adaptive requirements. As discussed earlier, vertebrate TLR superfamily consist of family TLR1 (subfamily TLR1, TLR2, TLR6, TLR10, TLR14, TLR15, TLR18, TLR24, TLR25 and TLR27), family TLR3, family TLR4, family TLR5, family TLR7 (subfamily TLR7, TLR8 and TLR9) and family 11 (subfamily TLR11-TLR13, TLR16, TLR19, TLR20, TLR21, TLR22, TLR23, TLR26). Fishes have the highest number of TLRs, with only six subfamilies (TLR6, TLR10-TLR12, TLR15 and TLR16) missing in fishes (Jawed-Teleosts and Jawless-Lamprey). This shows that most of the subfamilies originated early in the fish lineage, and fishes also had the highest number of gene duplication events (Table 5.1 and Figure 5.2). Higher vertebrates have less TLR genes and have less duplication and more events of gene loss (Table 5.1 and Figure 5.2). This shows that gene loss and gain have played an important role in the evolution of TLR family in vertebrates with gene gain more prominent in lower vertebrates and gene loss more common in higher vertebrates (Table 5.1 and Figure 5.2) (e.g. among sauropsida, reptiles had expanded TLR repertoire with 13 subfamilies compared to 8 subfamilies found in birds (Table 5.1 and Figure 5.2). The fishes and frog have the highest number of genes among vertebrates (Table 5.1 and Figure 5.2). Further studies are needed to elucidate the role of ecological adaptation in shaping the TLR superfamily diversity (e.g. diversification of TLRs in aquatic and terrestrial habitats).

We found that TLR15 is not bird specific as it is also found in the reptilian genomes (Table 5.1). Similarly TLR21 is also found in reptiles and coelacanth together with other fishes and frog. The loss of TLR15 and TLR21 in the mammalian lineage possibly occurred after the divergence of sauropsida form synapsida lineage (Table 5.1 and Figure 5.2). The TLR18 and TLR22 were lost in both the avian and mammalian lineages possibly marking similar evolutionary phenomenon (Table 5.1 and Figure 5.2). Many TLRs have duplicated and have multiple copies (e.g. TLR1, TLR2, TLR4, TLR5, TLR7, TLR8, TLR9, TLR14, TLR19, TLR20, TLR21 TLR22, TLR23 and TLR24) (Table 5.1). Most of these duplications happened early during vertebrate TLR evolution in fish lineage (Table 5.1 and Figure 5.2) producing extra gene copies, except the TLR1 gene duplication, which occurred in frog.

In reptiles, we found duplication of the TLR1 with exception of turtle. TLR2 was also duplicated in all reptiles, TLR5 was duplicated only in lizard and TLR8 duplicated in all

reptiles with exception of lizard. However, birds have only two recent duplications, present in TLR1 and TLR2, whereas mammals lack any recent functional duplicates. Our results suggest that TLR1 followed independent gene duplication event and gave rise to TLR1A and TLR1B in birds, lizard and frog (Table 5.1 and 5. 2). Earlier studies in mammals suggested that TLR1 family (TLR1/6 and TLR10) in platypus and opossum originated after divergence of birds and the mammalian lineage, with further duplication event in TLR1/6 after divergence of Montremes/Theria to Laurasiatheria/Euarchontoglires giving rise to the TLR1 and TLR6 genes (Huang et al. 2011).

The genomic scan of fish genomes revealed tandem duplicated copies of TLR2A and TLR2B in two fishes (Table 5.1). Thus, TLR2 duplication occurred much early in evolution. TLR4 is absent in fishes with exception of catfish and zebrafish. In zebrafish we found three copies of TLR4 (all membrane bound) compared to earlier reports of two copies (Quiniou et al. 2013), whereas two TLR4 are reported in catfish, one is membrane bound and other is a soluble form (Zhang et al. 2013). We also find a partial copy of TLR4 to be present in frog as reported previously (Roach et al. 2005). Two forms of TLR5, one membrane-anchored (TLR5M) and one soluble (TLR5S; lacks the TIR domain) are present in vertebrates. TLR5M is found in all vertebrates, except in the cod genome, which lacks both TLR5 genes. From earlier studies, only fishes and amphibians have both members of TLR5 (TLR5M and TLR5S). We found both members in anolis (one was lacking the TIR domain and was present in different location). Thus, TLR5S is possibly not specific to aquatic fishes and frog, suggesting wider distribution and role of TLR5S in vertebrate species. The TLR5S is reported to be duplicated in stickleback (Table 5.1) and TLR5M is duplicated in Zebrafish and catfish (Quiniou et al. 2013) (Palti 2011) (Sullivan et al. 2009). Among fishes TLR5S is not found in zebrafish and coelacanth. Partial TLR5 sequence is also reported in lamprey (Kasamatsu et al. 2010). TLR7 is found duplicated only in coelacanth. TLR8 reported to be duplicated in fishes and frog, was also found duplicated in crocodiles and turtles but was lost in birds, anolis and snake.

The TLR7/8A and TLR7/8A are unique to the lamprey genome, and they form the root for TLR7 and TLR8 in the TLR phylogeny, and possibly gave birth to separate TLR7 and TLR8 genes (Kasamatsu et al. 2010). TLR11 is unique to mammals whereas TLR13 is present in coelacanth and tetrapods but is missing from rest of the fishes. Our study shows that the mammalian TLR13 (13M) (Figure 5.1) is syntenic to coelacanth, whereas TLR13X present in frog and reptiles are in syntenic arrangement, which means TLR13M and TLR13X are paralogs.

Table 5.1 - The TLR supergene family, showing gene gain, gene loss in different vertebrate species

S.No	Common Name	Scientific Name	TLR1										TLR3	TLR4	TLR5	TLR7			TLR11									
			1	2	6	10	14	15	18	24	25	27	3	4	5	7	8	9	11	12	13	16	19	20	21	22	23	26
1	Human	**H**omo **s**apiens	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
2	Mouse	*Mus musculus*	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
3	Opossum	*Monodelphis domestica*	0	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0
4	Platypus	*Ornithorhynchus anatinus*	0	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0	1	1	0	0	0	0	0	0	0
5	Chicken	*Gallus gallus*	1A-B	2 A-B	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0
6	Turkey	*Meleagris gallopavo*	1A-B	2 A-B	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0
7	Zebra Finch	*Taeniopygia guttata*	1A-B	2 A-B	0	0	0	1	0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	0	1	0	0	0
8	Saltwater crocodile	Crocodylus porosus*#	1A-B	2 A-B	0	0	0	1	1	0	0	0	1	1	1	1	8 A-B	0	0	0	1	0	0	0	1	1	0	0
9	Gharial	Gavialis gangeticus*#	1A-B	2 A-B	0	0	0	1	1	0	0	0	1	1	1	1	8 A-B	0	0	0	1	0	0	0	1	1	0	0
10	American alligator	Alligator mississipiensis*#	1A-B	2 A-B	0	0	0	1	1	0	0	0	1	0	1	1	8 A-B	0	0	0	1	0	0	0	1	1	0	0
11	Chinese softshell turtle	*Pelodiscus sinensis**	1	2 A-B	0	0	0	0	1	0	0	0	1	1	1	1	8 A-C	1	0	0	0	0	0	0	1	1	0	0
12	Green sea turtle	*Chelonia mydus**	1	2 A-B	0	0	0	0	0	0	0	0	1	1	1	1	8 A-C	0	0	0	0	0	0	0	1	1	0	0
13	Painted Turtle	*Chrysemys picta bellii**	1	2 A-B	0	0	0	0	1	0	0	0	1	1	1	1	8 A-C	1	0	0	1	0	0	0	1	1	0	0
14	Anole lizard	*Anolis carolinensis**	1A-B	2 A-B	0	0	0	1	1	0	0	0	1	1	5 A-B	1	0	0	0	0	1	0	0	0	1	1	0	0
15	Python	Python molurus*#	1	1	0	0	0	1	1	0	0	0	1	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0
16	Xenopus	*Xenopus tropicalis*	1 A-C	2 A-B	0	0	14 A-D	0	0	0	0	0	1	1	5 A-B	1	8 A-B	1	0	0	1	1	0	0	1	1	0	0
17	Coelacanth	*Latimeria chalumnae**	1	2 A-B	0	0	0	0	1	0	0	1	1	0	1	7 A-B	1	1	0	0	1	0	0	0	21 A-C	0	0	0
18	Platyfish	*Xiphophorus maculatus**	0	2 A-B	0	0	0	0	1	0	0	0	1	0	5 A-B	1	1	1	0	0	0	0	0	0	1	22 A-B	23 A-F	0
19	Medaka	*Oryzias latipes**	1	1	0	0	0	0	1	0	1	0	1	0	5 A-B	1	1	1	0	0	0	0	0	0	1	1	0	0
20	Tilapia	*Oreochromis niloticus**	0	2 A-B	0	0	0	0	1	0	1	0	1	0	5 A-B	1	8 A-B	1	0	0	0	0	0	0	1	1	23 A-B	0
21	Stickleback	*Gasterosteus aculeatus**	1	1	0	0	0	0	1	0	0	0	1	0	5 A-C	1	1	1	0	0	0	0	0	0	21 A-B	1	0	0
22	Fugu	*Takifugu rubripes*	1	1	0	0	0	0	1	0	0	0	1	0	5 A-B	1	1	1	0	0	0	0	0	0	1	1	1	0
23	Tetraodon	*Tetraodon nigroviridis*	1	1	0	0	0	0	1	0	0	0	1	0	5 A-B	1	1	1	0	0	0	0	0	0	1	1	1	0
24	Cod	*Gadus morhua*	0	0	0	0	0	0	1	0	1	0	1	0	0	1	8 A-F	9 A-E	0	0	0	0	0	0	1	22 A-L	23 A-B	0
25	Zebrafish	*Danio rerio*	1	1	0	0	0	0	1	0	0	0	1	4 A-C	5 A-B	1	8 A-C	1	0	0	0	0	19 A-B	20 A-F	1	1	0	0
26	Cat Fish	Ictalurus punctatus#	1	1	0	0	0	0	1	0	1	0	1	4 A-B	5 A-C	1	8 A-B	1	0	0	0	0	1	20 A-B	1	1	0	1
27	Lamprey	*Petromyzon marinus*	0	0	0	0	14 A-D	0	0	24 A-D	0	0	1	0	1	7/8	7/8	0	0	0	0	0	0	0	21 A-C	1	0	0

The detailed information for genomes with asterisk is provided for first time and for rest the updated information supported with synteny data is provided. The present and absence is shown with “0” showing absence and number “1” showing presence the number of duplicates are shown where numbers show subfamily name and alphabets show number of copies. Asterisk indicates the genomes used for the first time. The hash indicates genomes not used for synteny analysis.

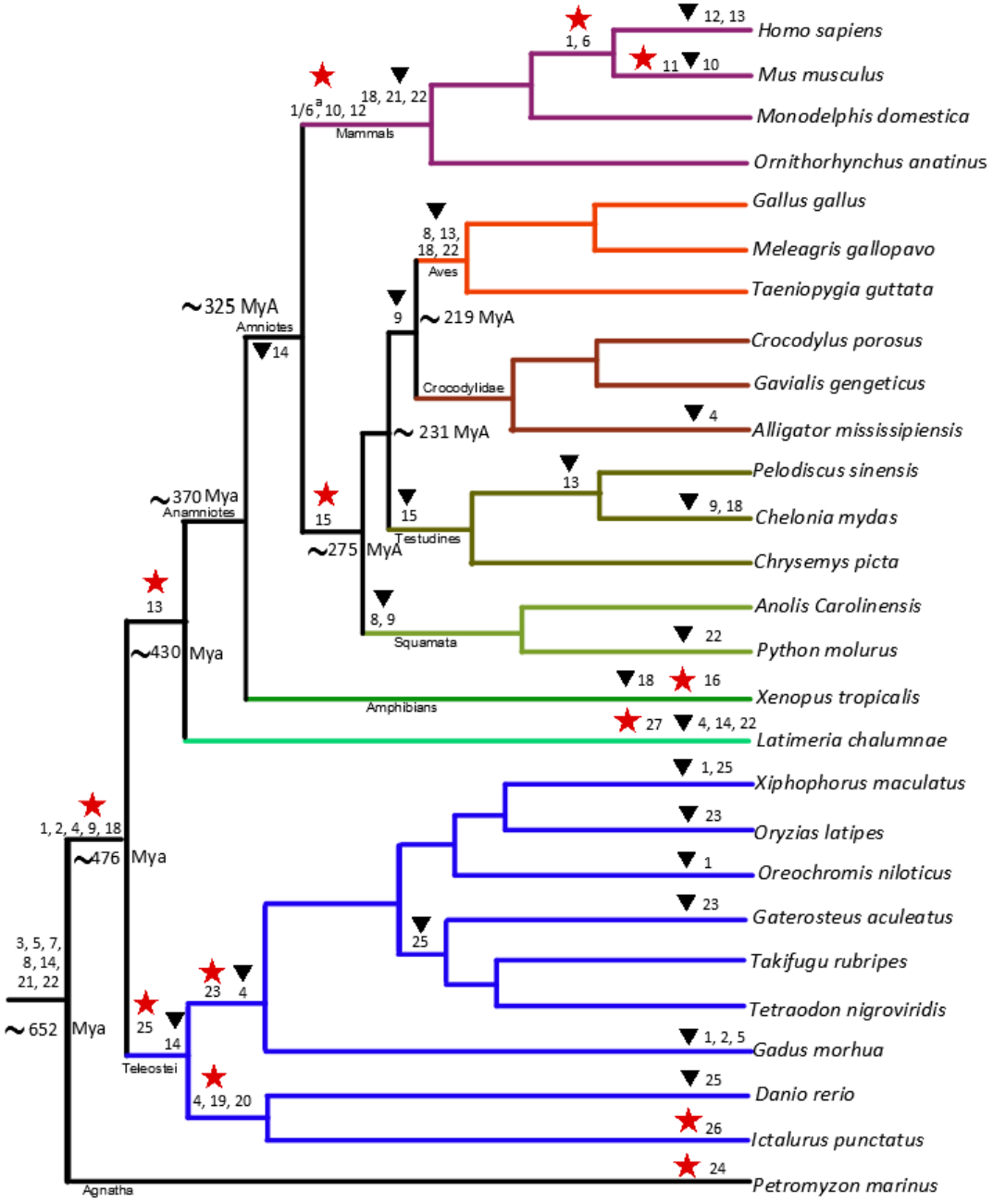


Figure 5.2: The molecular evolution of vertebrate TLR gene family with events of TLR subfamily gains and losses shown on the consensus phylogeny. The gains are represented by a star and the triangle shows loss of TLR subfamily. * The TLR10 and TLR1/6 lineages originated after divergence of Montremes and Theria is as per (Huang et al 2011)

In coelacanth there is only one TLR18 gene. We found a novel TLR which is related with TLR14/18 and is a coelacanth specific gene that we named TLR27 (Table 5.1).

TLR21 have three copies in coelacanth and we found that TLR21A is orthologous and syntenic to fishes TLR21, the other TLR21, .i.e. TLR21B clades with TLR21B of Stickleback and TLR21C form a separate clade. We found two copies of TLR22 in Platyfish whereas cod fish reportedly have 12 copies, which are under positive selection (Sundaram et al. 2012), two copies of TLR23 have been reported earlier in cod and we also found two copies in Tilapia and five copies in Platyfish genome.

The comparison for TLRs gene family repertoire shows that most of the TLRs originated in basal vertebrates (Lamprey and Fishes) and different TLRs families underwent different rates of gene gain and gene loss followed by different evolutionary fate due to diverse evolutionary pressures. Most of the subfamilies have lineage or species-specific genes (e.g. TLR6, TLR10, TLR11, TLR12 in mammals TLR13 tetrapod specific, TLR27 coelacanth specific, TLR15 sauropsida specific, TLR16 frog specific, TLR24 lamprey specific, TLR19, TLR20, TLR23, TLR25, and TLR26 fish specific). Among fishes coelacanth is connecting species between fishes and tetrapods, and this is supported by the presence of mammalian type TLR13 and fish specific TLR21A in the coelacanth genome. The extensive exploration of TLRs in the sequenced genomes helped to resolve the evolutionary history and provided greater insights into the distribution and diversification of vertebrate TLR supergene family.

5.3.3 Synteny analysis of TLR supergene family in vertebrates

The synteny analysis has been fundamental to understand the arrangements/organization and conservation of TLR genes, as well as to elucidate the homologous genes relationship. For the synteny analysis we scanned vertebrate genomes (Supplementary file 3) and retrieved the flaking genes for each TLR gene using extensive Blast searches (Altschul et al. 1990) complimented with Genomicus (Muffato et al. 2010) (Louis et al. 2013). The synteny analysis revealed interesting TLR gene relationships.

We found fishes and tetrapods lineage specific synteny conservation (synteny conserved within fishes and within tetrapods) with the exception of coelacanths. The coelacanth genome shows some shared features with both these groups suggesting its evolutionary proximity to both lineages. We found tandem arrangements in few subfamilies, e.g. subfamilies TLR7 and TLR8 belonging to family TLR7 and subfamilies TLR1, TLR6 and TLR10 members of family TLR1 are arranged in a tandem cluster (Supplementary file 5.3). Similarly duplicated members within subfamily were arranged

in tandem cluster, e.g. duplicated members of subfamilies TLR1, TLR2, TLR4, TLR5M, TLR5S, TLR8, TLR9, with two or multiple copies arranged in tandem.

Syntenic organization for TLR3 (Supplementary File 5.3, Figure-C), TLR5M (Supplementary File 5.3, Figure-E1), TLR7 and TLR8 (Supplementary file 5.3, Figure-F1), is evolutionary conserved across vertebrates from fishes to tetrapods with duplicated copies present in synteny (Supplementary file 5.3, Figure-F2 and F3), possibly supporting their functional importance. TLR1/6/10, TLR2 have conserved syntenic arrangement within tetrapods and coelacanth (Supplementary file 5.3, Figure-A1 and B1), which is different from other fish specific synteny (Supplementary file 5.3, Figure-A2 and B2). TLR4 shows conserved synteny in amniotes, which is totally different from the TLR4 organization found in frog and zebrafish (Supplementary file 5.3, Figure-D1 and D2). Fish specific TLR5S have conserved synteny and it is different from lizard and frog (Supplementary File 5.3, E2 and E3), TLR9 of frog, coelacanth, fishes and amniotes is having different flanking genes (Supplementary File 5.3, Figure-G1 to G3).

TLR13 is present in coelacanth and tetrapods and missing from all fish genomes studied. The mammals and coelacanth show shared synteny from TLR13 which is different from TLR13 organization present in frog and reptiles (Supplementary file 5.3, H2). Thus, there are two different TLR13 gene paralogs. The TLR14 has been previously reported in fugu and was found to be syntenic to TLR18 and thus named as TLR18. (Supplementary file 5.3, Figure-I1 and I2). Thus, TLR14 is only found in frog and lamprey. TLR15 is present in birds and reptiles and have conserved synteny (Supplementary file 5.3, Figure-J).

We also checked the arrangement of copies of TLR19 and TLR20 in zebrafish and found that TLR20 are arranged in two clusters located on the same chromosome possibly resulted from segmental duplication, with one cluster having four copies and other having two copies arranged in tandem, the two copies of TLR19 are also found in tandem. We found one member of TLR19 and TLR20 in cavefish genome, which was syntenic with zebrafish counterparts (Supplementary file 5.3 Figure-K and L). TLR21A present in jawed fishes, coelacanth and frog are syntenic (Supplementary file 5.3, Figure 2-M2), but this synteny is missing for birds TLR21 and also possibly from reptiles (Supplementary file 5.3, Figure-M1). We were not able to confirm the scenario in reptiles due to lack of missing data whereas fishes TLR21 is found in separate cluster (Supplementary file 5.3, Figure-M3). TLR22A-B found in platyfish and TLR22A, B and E reported in cod (Sundaram et al. 2012) are tandemly arranged and are

syntenic (Supplementary file 5.3, Figure-N1) the remaining extra copies of TLR22 are found in different location as referred previously (Sundaram et al. 2012). The TLR23A and TLR23B of tilapia and platyfish are found in tandem and are syntenic to fish TLR23 (Supplementary file 5.3, Figure-O). TLR25 of fishes have syntenic arrangement (Supplementary file 5.3, Figure-P). In the case of lamprey, the TLR7/8 A-B are not in tandem, and are found in different scaffold bordered with different genes as compared to other jawed vertebrates. TLR24A and B are tandem duplicates. The rest of the lamprey TLRs are found in small contigs and scaffolds with little or no information for synteny analysis.

The synteny analysis was able to highlight the genomic organization of TLRs which helped in understanding the level of homology of TLR gene present in different genomes. Lineage specific TLR e.g. fish specific and tetrapod specific TLRs showed highly conserved synteny with some variation found in coelacanth, which showed shared feature of both fish and tetrapods. Some TLR gene showed conserved synteny across vertebrates TLR3, TLR5M, TLR7 and TLR8 (conserved from mammals to fishes) (Supplementary file 5.3, Figure C, E1 and F1). The coelacanth TLR1 and TLR2 are having shared synteny with tetrapod's (Supplementary file 5.3, Figure A1 and B1), whereas TLR13 arrangement is same as mammals (Supplementary file 5.3, Figure H1) and only TLR21A is having shared synteny with fishes (Supplementary file 3, Figure M2). The findings point towards the evolutionary closeness of coelacanth with both fishes and tetrapods. Overall synteny analysis of widely distributed vertebrate TLR gene family shows conserved and differential role of TLRs shaped by gene gain and gene loss in diversification of the vertebrate TLR superfamily. Moreover, the rapidly evolving role of TLR supergene family suggests strong adaptive requirements for host-pathogen interaction due to diverse environmental conditions.

5.3.4 Gene conversion and recombination

The gene conversion and recombination can lead to false phylogenetic inference therefore we checked signals of gene conversion and recombination. The phylogenetic relationship show that duplicated avian TLR gene paralogs of TLR1 and TLR2 are closely related compared to their orthologous counterpart from other species, i.e. within species paralogs TLR1A is closely related to TLR1B and TLR2A with TLR2B. This is caused by concerted evolution possibly by gene conversion leading to homogenization of the paralogs thereby reducing the phylogenetic signals. The gene conversion is frequent in tandem duplicated gene paralogs located in close proximity. Gene conversion have earlier been reported in avian TLR1 and TLR2 (Huang et al. 2011). TLR1 and TLR2 are closely located and are separated by less physical distance

(~12kb and ~5Kb between TLR1 and TLR2 duplicates respectively) (Alcaide and Edwards 2011) (Temperley et al. 2008). The C terminal of TLR1A/B ranging from LRR14 to TIR domain is under gene conversion whereas in case of TLR2A/B, the N terminal region (N terminal to LRR8) and C terminal region (LRR15 till TIR) are homogenized) (Supplementary file 5.4, Figure A and B) (Alcaide and Edwards 2011) (Temperley et al. 2008). These homogenized regions may have important conserved function (Huang et al. 2011). The phylogeny made from conversion free region, i.e. N terminal region of TLR1A/B and central region TLR2A/B represent the species tree (Supplementary file 5.4, Figure B and D).

5.3.5 Comparative Domain architecture of vertebrate TLRs

TLRs consist of three major characteristic domains, extracellular domain (ECD), transmembrane domain (TM), and Toll/interleukin-I receptor domain (TIR) with the exception of TLR5S which lacks the characteristic TIR domain. The TM domain connects the ECD with cytoplasmic domain. The ECD is solenoid shaped structure involved in interaction with PAMPs present in varied pathogens and is also involved in formation of M shaped homo and/or hetero dimer which leads to signaling cascade by TIR activation (Kang and Lee 2011).

The ECD is made of variable number of leucine rich repeats (LRRs) as found in the predicted architecture of TLR from various lineages: chicken, lizard, Chinese soft shell turtle, coelacanth, stickleback (Supplementary file 5.5, Figure 1 to 5) and human (Kang and Lee 2011) . With each LRR being around 20-30 amino acid long and have a conserved LxxLxLxxN leucine rich motif and remaining variable region. The hydrophobic leucine residues constitute the conserved concave surface of parallel beta strands forming hydrophobic core where asparagine is involved in hydrogen bonding providing structural integrity (Kang and Lee 2011). Leucine can be replaced with other hydrophobic amino acids whereas asparagine can also be replaced with other hydrogen donor's likes threonine, serine, and cysteine. The variable "x" residues are responsible for the TLR function.

The exposed and convex surface of ECD is formed by the variable part of LRR repeat and this region is involved in PAMPs recognition. The cysteine clusters capping present in terminal LRRs (LRR-NT and LRR-CT) protect the terminal hydrophobic residues. The ECD can be further categorized into N terminal, central and C terminal subdomains as also seen in chicken, lizard, Chinese soft shell turtle, coelacanth and stickleback (Supplementary file 5.5, Figure 1 to 5).

The human TLR1, 2, 4, 6 and 10 have all three domains, whereas TLR 3, 5, 7, 8 and 9 have a single domain (Kang and Lee 2011). This categorization is due to the interrupted asparagine in LRRs of central domain, which results in structural flexibility whereas uniform LRRs repeats with continuous asparagine results in single domain architecture (Kang and Lee 2011). The comparison of the domain architecture of representative TLRs from chicken, Chinese soft shell turtle, Anolis lizard, coelacanth and stickleback was consistent with human TLRs. We also found single domain in TLR 3, 5, 7, 8, 9, 13, 15,18, 21, 22 and 27, whereas three domains are present in TLR 1, 2, 4. We found that some of the members deviated from their typical domain architecture, e.g. TLR1B chicken and TLR1 in coelacanth had single domain (Supplementary file 5.5, Figure 1 and 5). In case of TLR2, turtle and coelacanth showed single domain architecture (Supplementary file 5.5, Figure 2 and 5).

5.3.6 Rapid adaptive evolution of avian TLR supergene family

Positive selection is one of the hallmarks of Immune-defense-related genes (Nielsen et al. 2005) (Vallender and Lahn 2004) and especially those encoding recognition proteins, evolve under positive Darwinian selection (Sackton et al. 2007). There is growing evidence of positive selected sites in TLRs loci (Huang et al. 2011) (Alcaide and Edwards 2011) (Areal et al. 2011) (Sundaram et al. 2012). These positive selected sites can provide increased number of advantageous variations, which is important for pathogen recognition and host-pathogen arms-race, required for successful adaptation against changing environments and pathogens (Kosiol et al. 2008). We used different codon and protein level approaches as detailed below to find the positive selected sites and their possible effect on structural and functional diversification of respective TLRs (Figure 5.3 and Supplementary file 5.6 to 5.14). The homology modeling was used to confirm the important changes in TLR proteins (Figure 5.3 and Supplementary file 5.6 to 5.14).

We found variable number of positive selection sites among different TLR genes using M2a and M8 model. The positive selection model M2a and M8 implemented in PAML4.7 (Yang 1997) found significant signals of positive selection in all genes (Table 5.2). The reliability of sites was well supported by the fact that most positive selected sites, having PP>0.99 under M8 also had PP>0.90 in M2a. The M8 model found higher number of sites compared to more conservative model M2a (Table 5.2). We found positive selection in all avian TLRs studied with different percentage of positively selected sites for different genes ranging from 11% in TLR15 to 5% in TLR7. TLR4, TLR2A and TLR7 were having higher omega values of 2.6, 2.6 and 2.5, respectively, and TLR15 was found to have least value of 1.5.

Table 5.2: PAML results for nested site model comparisons for test of positive selection

Gene	No of Species	Model	Likelihood (l*nL*)	Parameters	2ΔlnL (LRT)	Significance (P-Valve)	No of PS Sites
TLR1A	25	M1A	-8578.362134	p: 0.65788 0.34212			
				w: 0.07820 1.00000			
		M2A	-8568.35735	p: 0.65059 0.30720 0.04221	20.009568	4.52E-05	9, 1*, 0**
				w: 0.08172 1.00000 2.43025			
		M7	-8574.975693	p = 0.24899 q = 0.48700			
		M8	-8557.092491	p0 = 0.93283 p = 0.32098 q = 0.81643	35.766404	1.71E-08	9, 5*, 1**
				(p1 = 0.06717) w = 1.97137			
TLR1B	29	M1A	-15198.74873	p: 0.64884 0.35116			
				w: 0.09042 1.00000			
		M2A	-15173.19182	p: 0.63406 0.31995 0.04599	51.113808	7.96E-12	12, 7*, 2**
				w: 0.09272 1.00000 2.46912			
		M7	-15185.50243	p = 0.25368 q = 0.45987			
		M8	-15147.32927	p0 = 0.91288 p = 0.34184 q = 0.83943	76.346322	2.64E-17	20, 14*, 4**
				(p1 = 0.08712) w = 1.88577			
TLR2A	36	M1A	-14595.10784	p: 0.58122 0.41878			
				w: 0.14194 1.00000			
		M2A	-14510.38966	p: 0.53089 0.40739 0.06172	169.436346	1.61E-37	18, 13*, 11**
				w: 0.13788 1.00000 3.10556			
		M7	-14590.69099	p = 0.34904 q = 0.40071			
		M8	-14500.54302	p0 = 0.91485 p = 0.40926 q = 0.52204	180.29595	7.07E-40	20, 16*, 11**
				(p1 = 0.08515) w = 2.56498			

TLR2B	29	M1A	-17453.34126	p: 0.70636 0.29364			
				w: 0.08820 1.00000			
		M2A	-17429.54293	p: 0.69833 0.27518 0.02650	47.596666	4.62E-11	10, 6*, 0**
				w: 0.09079 1.00000 2.72661			
		M7	-17439.57866	p = 0.24244 q = 0.52859			
		M8	-17403.82062	p0 = 0.94556 p = 0.29832 q = 0.80619	71.51608	2.95E-16	12, 10*, 5**
				(p1 = 0.05444) w = 1.98273			
TLR3	27	M1A	-17682.71542	p: 0.70932 0.29068			
				w: 0.10081 1.00000			
		M2A	-17667.59313	p: 0.70389 0.28584 0.01027	30.244572	2.71E-07	2 , 2*, 2**
				w: 0.10207 1.00000 3.25211			
		M7	-17680.6455	p = 0.29425 q = 0.62674			
		M8	-17654.95773	p0 = 0.94774 p = 0.38756 q = 1.03153	51.37553	6.98E-12	5, 2*, 2**
				(p1 = 0.05226) w = 1.87953			
TLR4	45	M1A	-27142.6224	p: 0.65430 0.34570			
				w: 0.08873 1.00000			
		M2A	-26989.66736	p: 0.63053 0.32499 0.04449	305.910084	3.74E-67	24, 19*, 16**
				w: 0.08946 1.00000 3.12384			
		M7	-27104.328	p = 0.24737 q = 0.45214			
		M8	-26948.5913	p0 = 0.94429 p = 0.28421 q = 0.58752	311.473402	2.31E-68	29, 27*, 20**
				(p1 = 0.05571) w = 2.60768			
TLR5	14	M1A	-10761.51484	p: 0.64778 0.35222			
				w: 0.07521 1.00000			
		M2A	-10754.44254	p: 0.65397 0.29395 0.05208	14.14459	0.000848284	0, 0*, 0**
				w: 0.08616 1.00000 2.27090			
		M7	-10767.19676	p = 0.16412 q = 0.25713			

		M8	-10751.93864	p0 = 0.88630 p = 0.31653 q = 0.82154	30.516246	2.36E-07	11, 2*, 0**
				(p1 = 0.11370) w = 1.85950			
TLR7	44	M1A	-31101.46341	p: 0.72733 0.27267			
				w: 0.06793 1.00000			
		M2A	-30961.25342	p: 0.71077 0.25222 0.03700	280.419986	1.28E-61	25, 24*, 19**
				w: 0.06917 1.00000 3.03247			
		M7	-31044.11967	p = 0.18953 q = 0.48287			
		M8	-30894.43886	p0 = 0.95423 p = 0.22385 q = 0.67658	299.36162	9.87E-66	30, 25*, 23**
				(p1 = 0.04577) w = 2.52048			
TLR15	45	M1A	-38867.70847	p: 0.66864 0.33136			
				w: 0.09826 1.00000			
		M2A	-38835.18559	p: 0.65901 0.31957 0.02142	65.045742	7.51E-15	8, 8*, 5**
				w: 0.09812 1.00000 2.29015			
		M7	-38651.19894	p = 0.32880 q = 0.73980			
		M8	-38608.52726	p0 = 0.94189 p = 0.39709 q = 1.16306	85.34337	2.94E-19	9, 8*, 3**
				(p1 = 0.05811) w = 1.50749			

The Number of coloum with number of posiitve selected sites shows the sites with PP>0.90, PP>0.95 and PP>0.99

Table 5.3: The positive selected sites identified by various methods

Gene	M8[a]	SLAC	FEL	REL	MEME	FUBAR	Integrated[b]	Total Common Sites[c] (X/Y)
TLR1A	258, 400**, 408**, 423, 450, 460, 483, 502, 591	388, 429, 566,	294, 342, 384, 388, 429, 460, 535, 566, 611,	297, 301, 384, 388, 400, 429, 438, 460, 463, 470, 550, 566, 611,	266, 284, 294, 346, 384, 388, 400, 411, 429, 460, 461, 463, 518, 535, 550, 559, 566, 571, 599, 608, 611,	388, 429, 460, 566,	266, 284, 294, 297, 301, 342, 346, 384, 388, 400, 411, 429, 438, 460, 461, 463, 470, 518, 535, 550, 559, 566, 571, 599, 608, 611; **408**	11/12
TKR1B	41,59, 88, 122, **148**, 149, 168, **175**, 248, 256, 277, 286, **298**, **308**, 318, 350, 357, 439, 486, 488,	38, 41, 119, 120, 122, 123, 148, 168, 175, 216, 232, 277, 398, 414,	11, 38, 41, 49,62,67,119,120,122, 123, 128, 148, 156, 167, 168, 175, 216, 232, 266, 277, 308, 371, 398, 408, 414, 627,	38, 41, 59, 119, 120, 122, 123, 148, 150, 168, 175, 181, 207, 216, 232, 248, 277, 286, 308, 311, 398, 414, 485,	26, 38, 41, 43, 44, 62, 92, 119, 120, 122, 123, 127, 144, 146, 148, 156, 167, 168, 175, 194, 206, 214, 216, 232, 266, 277, 308, 309, 311, 353, 366, 371, 398, 407, 408, 414, 432, 447, 639,	38, 41, 119, 120, 122,123, 148, 168, 175, 216, 232, 277, 308, 398, 414,	11, 26, 38, 41, 43, 44, 49, 59, 62, 67, 92, 119, 120, 122, 123, 127, 128, 144, 146, 148, 150, 156, 167, 168, 175, 181, 194, 206, 207, 214, 216, 232, 248, 266, 277, 286, 308, 309, 311, 353, 366, 371, 398, 407, 408, 414, 432, 447, 485, 627, 639; **298**	25/26
TLR2A	7**, 16, 59, 67, 108,** 129, **171, 206,** 220**,** 270**, 304,** 307**, 308, 311, 312, 338,** 363, 372, 393**,** 413	16, 63, 108, 292, 304, 308, 311, 315, 335, 393, 418, 419,	16, 28, 44, 45, 63, 74, 77, 108, 171, 209, 217, 235, 257, 292, 304, 308, 309, 311, 315, 335, 344, 356, 372, 393, 418, 419,	7, 16, 67, 108, 129, 171, 217, 235, 276, 280, 292, 306, 308, 311, 312, 335, 356, 372, 387, 392, 393, 413, 416, 418, 419,	16, 19, 28, 44, 63, 67, 77, 96, 108, 122, 125, 138, 171, 174, 187, 195, 209, 247, 250, 257, 264, 277, 280, 292, 294, 304, 308, 309, 311, 315, 322, 335, 344, 347, 349, 367, 372, 393, 418, 419, 425,	16, 108, 171, 292, 308, 311, 372, 392, 413, 418,	7, 16, 19, 28, 44, 45, 51, 63, 67, 74, 77, 96, 108, 122, 125, 129, 138, 171, 174, 187, 195, 209, 217, 235, 247, 250, 257, 264, 276, 277, 280, 292, 294, 304, 306, 308, 309, 311, 312, 315, 322, 335, 344, 347, 349, 356, 367, 372, 387, 392, 393, 413, 416, 418, 419, 425; **59, 206, 338**	31/34
TLR2B	50, 58**, 99, 162, 175,** 211, 260**, 295**, 297, 298, 329, **456**	89, 99, 137, 162, 176, 297, 328, 329, 390,	42, 48, 89, 99, 137, 149, 162, 176, 199, 200,219, 257, 274, 295, 297, 298, 299, 300, 302, 317, 329, 343, 390, 401, 412, 415,553, 600, 614, 625, 734,	89, 99, 162, 208, 295, 297, 298, 328, 329, 331, 343, 625,	25, 58, 68, 75, 89, 99, 137, 149, 162, 176, 189, 211, 226, 234, 239, 260, 274, 295, 297, 299, 302, 317, 329, 335, 343, 383, 390, 401, 415, 467, 496, 499, 500, 542, 550, 597, 625, 679,	89, 99, 162, 295, 297, 298, 328, 343, 625,	25, 42, 48, 58, 68, 75, 89, 99, 137, 149, 162, 176, 189, 199, 200, 208, 211, 219, 226, 234, 239, 257, 260, 274, 295, 297, 298, 299, 300, 302, 317, 328, 329, 331, 335, 343, 383, 390, 401, 412, 415, 467, 469, 496, 499, 500, 542, 550, 553, 597, 600, 603, 625, 679, 734; **175, 456**	23/25
TLR3	52, 166, **214, 237,** 703	25, 237, 307, 334, 703, 746,	9, 11, 19, 25, 30, 66, 68, 74, 214, 237, 263, 307, 334, 346, 370, 447, 468, 557, 698, 703, 744, 746, 815,	25, 30, 52, 74, 113, 137, 158, 166, 179, 180, 214, 237, 288, 307, 312, 326, 334, 343, 346, 447, 461, 463, 497, 556, 557, 619, 703, 746, 815	9, 25, 30, 38, 48, 68, 74, 93, 108, 192, 214, 234, 237, 263, 307, 334, 346, 349, 382, 393, 439, 447, 451, 461, , 473, 547, 557, 577, 605, 664, 677, 698, 703, 707, 715, 744, 746, 815	25, 52, 158, 214, 237, 334, 346, 703, 746, 815	9, 11, 19, 25, 30, 38, 48, 52, 66, 68, 74, 93, 108, 113,137, 158, 166, 179, 180, 192, 214, 234, 237, 263,288, 307, 312, 326, 334, 343, 346, 349, 370, 382, 393, 439, 447, 451, 461, 463, 468, 473, 497, 547, 556,557, 577, 605, 619, 664, 677, 698, , 703, 707, 715, 744, 746, 815	22
TLR4	**187, 245, 270,** 271, **274, 299, 302, 323, 352, 370, 375, 379, 380**, 387, **398,** 403**, 405, 406, 423,** 465, 522**, 595,** 624**, 627, 640,** 645**,** 650**, 655,** 834	106, 124, 146, 187, 245, 271, 301, 302, 352, 379, 380, 423, 467, 640, 654,	86, 95, 106, 119, 124, 127, 141, 146, 187, 204, 245, 270, 271, 277, 301, 302, 329, 352, 363, 379, 380, 403, 423, 467, 509, 596, 603, 606, 624, 639, 640, 654, 663, 732,	62, 86, 106, 124, 146, 187, 245, 270, 271, 301, 302,303, 323, 345, 352, 363,370, 379, 380, 387, 403, 423, 430, 438, 444, 467, 522, 596, 603, 624, 627, 640, 654, 703,	61, 63, 64, 85, 86, 95, 106, 115, 119, 124, 141, 146, 155, 180, 187, 223, 245, 270, 271, 273, 282, 297, 301, 302, 333, 345, 352, 363, 370, 379, 380, 397, 398, 403, 423, 435, 445, 467, 469, 474, 477, 493, 533, 548, 549, 562, 569, 570, 596, 597, 603, 606, 614, 624, 640, 654, 732, 780, 784, 827,	187, 271, 301, 302, 352, 379, 380, 403, 423, 467, 522, 596, 603, 624,	61, 62, 63, 64, 85, 86, 95, 106, 119, 124, 127, 141, 146, 155, 180, 187, 204, 223, 245, 270, 271, 273, 277, 282, 297, 301, 302, 303, 323, 329, 333, 345, 352, 363, 370, 379,380, 387, 397, 398, 403, 423, 430, 435, 438, 444, 445, 467, 469, 474, 477, 493, 509, 522, 533, 548, 549, 562, 569, 570, 596, 597, 603, 606, 614, 624, 627, 639, 640, 654, 663, 703, 732, 780, 784, 827; **274, 299, 375, 405, 406, 595, 655**	34/41

TLR5	20,106,130,132,147, **209**, 237, **281**,468,607,848	848	22, 24, 35, 53, 101, 108, 118, 130, 147, 173, 201, 226, 231, 258, 261, 264, 422, 466, 468, 656, 659, 833, 848,	20, 22, 33, 35, 106, 130, 132, 147, 237, 258, 261, 299, 468, 632, 646, 848,	13, 22, 24, 87, 130, 173, 181, 183, 196, 199, 201, 205, 217, 226, 231, 258, 261, 264, 265, 288, 378, 422, 424, 466, 501, 525, 625, 626, 632, 648, 650, 656, 659, 679, 833, 848, 859,	22, 35, 130, 261, 468, 848,	13, 20, 22, 24, 33, 35, 53, 87, 101, 106, 108, 118, 130, 132, 147, 173, 181, 183, 196, 199, 201, 205, 217, 226, 231, 237, 258, 261, 264, 265, 288, 299, 378, 422, 424, 466, 468, 501, 525, 625, 626, 632, 646, 648, 650, 656, 659, 679, 833, 848, 859; **209, 281**	24/26
TLR7	**65, 73, 118, 121, 122, 123, 148,** 152**, 156, 284, 334,** 360, **395,** 422,426, **503, 524,** 528, **549, 577, 677, 696, 704, 706, 722, 726,** 746, **747,** 758**, 920**	38, 56, 73, 121, 123, 156, 176, 395, 503, 521, 549, 577, 664, 681, 704, 722, 726, 751, 919, 920, 1049,	56, 73, 121, 123, 156, 169, 176, 229, 253, 279, 310, 395, 503, 521, 524, 528, 549, 577, 642, 664, 681, 701, 704, 706, 722, 726, 737, 751, 851, 857, 860, 919, 920, 951, 1049,	38, 73, 89, 95, 97, 121, 122, 123, 148, 152, 156, 176, 205, 279, 284, 334, 360, 366, 395, 398, 402, 422, 426, 494, 503, 524, 528, 549, 550, 573, 577, 678, 701, 704, 706, 712,722, 726, 746, 920,	38, 56, 73, 86, 95, 111,121, 123, 156, 167, 169, 174, 199, 229, 246, 279, 310, 313, 321, 361, 377, 395, 463, 465, 495, 503, 512, 521, 524, 528, 549, 577, 622, 624, 664, 681, 686, 698, 701, 704, 706, 708, 712, 722, 726, 737, 747, 751, 768, 792, 857, 895, 919, 920, 951, 1038, 1049,	73, 121, 123, 156, 205, 395, 503, 524, 528, 549, 577, 704, 706, 722, 726, 919, 920,	38, 56, 73, 86, 89, 95, 97, 111, 121, 122, 123, 148, 152, 156, 167, 169, 174, 176, 199, 205, 229, 246, 253, 279, 284, 310, 313, 321, 334, 360, 361, 366, 377, 395, 398, 402, 422, 426, 463, 465, 494, 495, 503, 512, 521, 524, 528, 549, 550, 573, 577, 622, 624, 642, 664, 678, 681, 686, 698, 701, 704, 706, 708, 712, 722, 726, 737, 746, 747, 751, 768, 792, 851, 857, 860, 895, 919, 920, 951, 1038, 1049; **65, 118, 677, 696**	44/48
TLR15	79, 119, **185**, 253, 262, **326**, 333, 353, **360**,	19, 26, 33, 65, 79, 89, 114, 136, 151, 169, 191, 203, 205, 268, 289, 292, 293, 339, 343, 359, 366, 413, 436, 458, 621, 623, 627	26, 33, 38, 65, 89, 102, 114, 136, 145, 151, 169, 191, 203, 259, 268, 289, 292, 293, 339, 343, 359, 366, 413, 436, 458, 621, 623, 627, 649, 656, 661, 725,	11, 13, 16, 17, 19, 26, 31, 33, 38, 48, 65, 79, 89, 102, 114, 120, 127, 128, 132, 136, 145, 149, 151, 158, 160, 164, 169, 170, 181, 191, 193, 194, 196, 200, 203, 205, 226, 249, 259, 267, 268, 289, 292, 293, 296, 315, 339, 343, 359, 363, 366, 382, 392, 413, 416, 436, 439, 450, 458, 463, 494, 495, 521, 523, 528, 550, 621, 623, 627, 649, 656, 661, 671, 676, 705, 712, 725, 813,	12, 26, 36, 54, 65, 89, 102, 105, 107, 114, 136, 143, 151, 159, 162, 169, 175, 186, 187, 191, 203, 225, 226, 229, 230, 232, 235, 259, 263, 268, 282, 285, 289, 292, 293, 329, 335, 339, 343, 359, 363, 366, 367, 383, 400, 413, 436, 458, 485, 530, 543, 550, 621, 623, 627, 634, 651, 661, 679, 815, 839, 862, 865 , 872, 873,	26, 89, 114, 136, 169, 191, 203, 259, 268, 289, 292, 293, 339, 343, 359, 366, 413, 436, 458, 621, 623, 661,	11, 12, 13, 16, 17, 19, 26, 31, 33, 36, 38, 48, 54, 65, 79, 89, 102, 105, 107, 114, 120, 127, 128, 132, 136, 143, 145, 149, 151, 158, 159, 160, 162, 164, 169, 170, 175, 181, 186, 187, 191, 193, 194, 196, 200, 203, 205, 225, 226, 229, 230, 232, 235, 249, 259, 263, 267, 268, 282, 285, 289, 292, 293, 296, 315, 329, 335, 339, 343, 359, 363, 366, 367, 382, 383, 392, 400, 413, 416, 436, 439, 450, 458, 463, 485, 494, 495, 521, 523, 528, 530, 543, 550, 621, 623, 627, 634, 649, 651, 656, 661, 671, 676, 679, 705, 712, 725, 813, 815, 839, 862, 865, 872, 873; **185, 326, 360**	38/41

a)- Sites detected under M8 model with PP>0.90 and the sites having PP>0.99 are shown in bold

The positive selected sites under SLAC, FEL, REL, MEME and FUBAR were indentified with default cutoff set in Datamonkey online server.

b) -Integrative approach includes all the sites identifed by M8 ,SLAC, FEL, REL, MEME and FUBAR. The sites detected by more than one method are underlined

c) - Common sites (**X/Y) X**: Total number of sites detected by two or more methods (underlined) **Y**: Total number of sites detected by two or more methods including sites with PP>99 in M8 site model (bold)

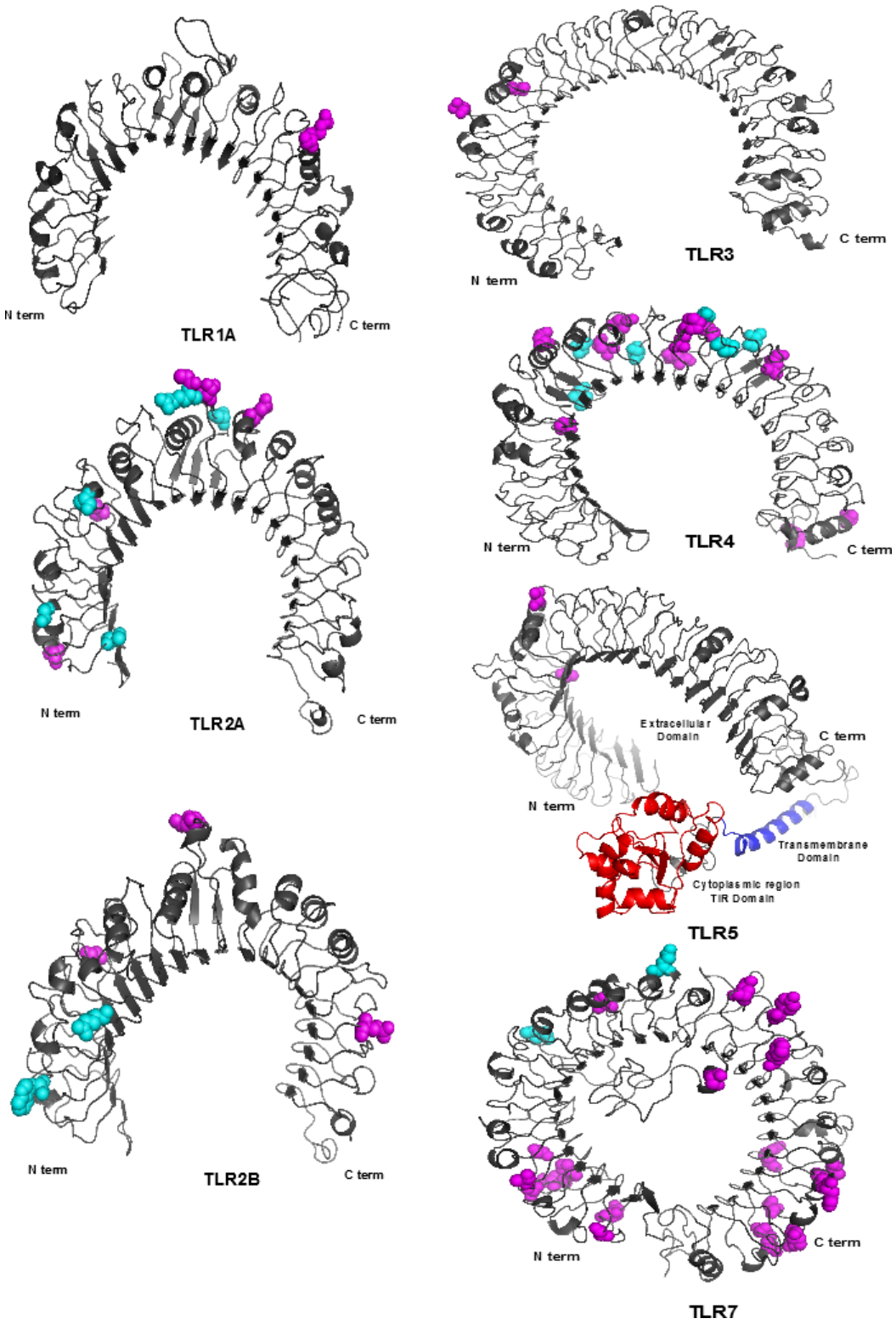


Figure 5.3: The positive selected sites with PP>0.99 in M8 which also had PP>90 in M2a are shown in the predicted structure of respective TLRs (except for TLR1B and TLR15 for which structure prediction was not significant) and are shown with magenta color. The site with cyan color represent type I changes detected by TreeSAAP.

Table 5.4: Positive selected sites detected with PP>0.99 in M8 model which also had PP>0.90 in M2a. The positive radical changes detected by TreeSAAP are shown with Type I changes underlined

Gene	Codon	Amino acid	M2a	M8	Total		Chemical		Structural		Other
TLR1A	*408*	R	2.467 +- 0.293*	2.200 +- 0.468**	1	1	R_F	0		0	
TLR1B	*148*	V	2.495 +- 0.123**	1.691 +- 0.398**	1	0		1	K^0	0	
	175	S	2.484 +- 0.176*	1.688 +- 0.401**	4	3	R_F, R_a, H_p	1	B_I	0	
	298	A	2.476 +- 0.207*	1.688 +- 0.400**							
	308	N	2.491 +- 0.141**	1.691 +- 0.398**	2	1	H	1	K^0	0	
TLR2A	*16*	Q	3.393 +- 0.312**	2.819 +- 0.468**	2	2	pH_i, H	0		0	
	59	P	3.374 +- 0.374**	2.808 +- 0.488**	2	2	R_a, H_p	0		0	
	67	V	3.393 +- 0.313**	2.819 +- 0.468**	8	3	pH_i, μ, R_a	4	M_V, V^0, B_I, H_C	1	M_W
	108	G	3.393 +- 0.311**	2.819 +- 0.467**	7	5	H, R_F, H_{nc}, H_t, R_a	2	B_I	0	
	171	Q	3.394 +- 0.309**	2.820 +- 0.466**	2	2	H_t, pH_i	0		0	
	206	S	3.388 +- 0.330**	2.815 +- 0.475**	9	4	R_F, μ, R_a, H_t	4	M_V, V^0, B_I, H_C	1	M_W
	304	A	3.394 +- 0.309**	2.820 +- 0.466**	9	7	R_F, H, R_a, H_p, pH_i, Pr, p	2	K^0, B_I	0	
	308	T	3.391 +- 0.317**	2.818 +- 0.469**	5	5	R_F, H, R_a, H_p, pH_i, Pr	0		0	
	311	A	3.394 +- 0.309**	2.820 +- 0.466**	1	1	Pr	0		0	
	312	R	3.391 +- 0.317**	2.818 +- 0.469**	6	4	pH_i, H, H_{nc}, R_a	2	V^0, H_C	0	
	338	E	3.371 +- 0.385**	2.807 +- 0.487**	2	2	pH_i, R_F	0		0	
TLR2B	*99*	W	2.510 +- 0.292*	2.407 +- 0.298**	6	5	H, H_{nc}, R_a, R_F, H_t	1	B_I,	0	
	162	Q	2.524 +- 0.254*	2.411 +- 0.289**	5	5	pH_i, p, R_F, H_{nc}, R_a	0		0	
	175	E	2.503 +- 0.309*	2.405 +- 0.303**	6	5	R_F, H, R_a, H_p, P_r	1	K^0	0	
	295	Q	2.506 +- 0.302*	2.406 +- 0.301**	2	2	pH_i, H	0		0	
	456	Q	2.514 +- 0.281*	2.409 +- 0.295**	5	3	R_a, H_p, pH_i	2	K^0, H_C	0	
TLR3	*214*	T	2.503 +- 0.087**	1.741 +- 0.430**	5	5	R_a, H_p, P_r, p, pH_i	0		0	
	237	R	2.500 +- 0.109**	1.736 +- 0.435**	3	3	R_F, H, H_{nc}	0		0	
TLR4	*187*	S	3.497 +- 0.062**	2.500 +- 0.019**	3	2	pHi, p	1	K^0	0	
	245	N	3.497 +- 0.059**	2.500 +- 0.017**	1	1	P_r	0		0	
	270	I	3.498 +- 0.044**	2.500 +- 0.001**	7	4	R_F, R_a, H_p, , pH_i	3	V^0, B_I, H_C	0	
	274	T	3.495 +- 0.102**	2.499 +- 0.041**	10	5	pH_i, μ, R_a, P_r, p	4	M_V, V^0, B_I, H_C	1	M_W
	299	E	3.498 +- 0.044**	2.500 +- 0.001**	3	2	pH_i, P_r	1	K^0	0	
	302	N	3.429 +- 0.411*	2.492 +- 0.111**	1	1	pH_i	0		0	
	323	N	3.498 +- 0.044**	2.500 +- 0.004**	9	5	pH_i, μ, R_F, H, H_{nc}	3	M_V, V^0, H_C	1	M_W
	352	E	3.486 +- 0.181**	2.498 +- 0.053**	2	2	P_r, p	0		0	
	370	D	3.496 +- 0.087**	2.500 +- 0.028**	3	1	pH_i	2	V^0, H_C	0	
	375	E	3.498 +- 0.056**	2.500 +- 0.021**	4	4	R_F, H_{nc}, R_a, H	0		0	
	379	G	3.498 +- 0.046**	2.500 +- 0.007**	6	5	R_F, H, R_a, H_p, pH_i	1	K^0	0	
	380	S	3.464 +- 0.291*	2.496 +- 0.080**	4	2	pH_i, P_r	2	V^0, H_C	0	
	398	T	3.498 +- 0.044**	2.500 +- 0.001**	8	3	pHi, μ, R_F	4	M_V, B_I, V^0, H_C	1	M_W
	405	P	3.498 +- 0.044**	2.500 +- 0.003**	5	4	R_F, H, H_{nc}, pH_i	1	K^0	0	
	406	R	3.498 +- 0.044**	2.500 +- 0.003**	8	3	μ, H, pH_i	4	M_V, B_I, V^0, H_C	1	M_W
	423	L	3.465 +- 0.289*	2.494 +- 0.094**							
	595	T	3.498 +- 0.051**	2.500 +- 0.015**	2	2	R_a, H_p	0		0	
	627	M	3.355 +- 0.582*	2.487 +- 0.143**							
	640	N	3.489 +- 0.158**	2.499 +- 0.048**	1	1	pH_i	0		0	

	655	G	3.485 +- 0.189**	2.498 +- 0.059**							
TLR5	*209*	T	2.143 +- 0.677	1.474 +- 0.154* #	3	3	P_r, p, H	0		0	
	281	T	2.089 +- 0.700	1.475 +- 0.151* #							
TLR7	*65*	K	3.460 +- 0.198**	2.500 +- 0.017**	1	1	pH_i	0		0	
	73	G	3.445 +- 0.271**	2.494 +- 0.094**	1	1	P_r	0		0	
	118	N	3.401 +- 0.422*	2.485 +- 0.154**	2	2	pH_i, H	0		0	
	121	I	3.458 +- 0.205**	2.499 +- 0.036**	3	2	R_a, H_p	1	K^0	0	
	122	T	3.456 +- 0.219**	2.498 +- 0.058**	3	3	R_F, H, H_{nc}	0		0	
	123	P	3.460 +- 0.197**	2.500 +- 0.001**	5	4	R_F, H, H_{nc}, R_a	1	K^0	0	
	148	S	3.439 +- 0.300**	2.487 +- 0.143**	1	0		1	K^0	0	
	156	A	3.427 +- 0.341*	2.492 +- 0.109**	3	2	R_a, H_p	1	K^0	0	
	284	I	3.405 +- 0.411*	2.487 +- 0.143**	7	4	R_F, R_a, H_p, pH_i	3	B_I, V^0, H_C	0	
	334	I	3.457 +- 0.212**	2.499 +- 0.041**							
	395	E	3.459 +- 0.198**	2.500 +- 0.020**	7	2	pH_i, μ	4	M_V, V^0, K^0, H_C	1	M_W
	503	S	3.444 +- 0.277**	2.495 +- 0.088**							
	524	Q	3.459 +- 0.203**	2.499 +- 0.037**	5	5	R_F, H_{nc}, R_a, pH_i, H	0		0	
	549	Y	3.460 +- 0.197**	2.500 +- 0.006**	3	2	pH_i, p	1	K^0	0	
	577	F	3.459 +- 0.201**	2.499 +- 0.028**	2	2	pH_i, p	0		0	
	677	P	3.453 +- 0.235**	2.497 +- 0.065**	2	1	H	1	K^0	0	
	696	R	3.460 +- 0.197**	2.500 +- 0.011**	1	1	pH_i	0		0	
	704	K	3.460 +- 0.197**	2.500 +- 0.001**	3	3	R_F, H, pH_i	0		0	
	706	H	3.458 +- 0.209**	2.499 +- 0.044**	1	0		1	K^0	0	
	722	T	3.404 +- 0.415*	2.486 +- 0.149**	2	2	R_F, R_a	0		0	
	726	R	3.459 +- 0.202**	2.499 +- 0.039**							
	747	R	3.459 +- 0.199**	2.500 +- 0.028**	4	4	R_F, H_{nc}, R_a, pH_i	0		0	
	920	T	3.457 +- 0.215**	2.498 +- 0.053**							
TLR15	*185*	K	2.499 +- 0.032**	1.498 +- 0.036**	3	2	R_a, pH_i	0	B_I	0	
	326	A	2.499 +- 0.034**	1.497 +- 0.040**	1	1	pH_i	0		0	
	360	P	2.500 +- 0.008**	1.500 +- 0.014**	3	2	H_t, H	1	K^0		

The ω values and Bayesian (BEB) analysis posterior probabilities are shown for sites with PP > 0.99 in M8 that also have a PP > 0.90 in M2a.TreeSAAP analysis results present the total number of radical changes in amino acid properties and their assigned categories.

Type I sites are shown are underlined. Properties symbols are as following: B_I: Bulkiness; H: Hydropathy; H_{nc}: Normal consensus hydrophobicity; H_p: Surrounding hydrophobicity; H_t: Thermodynamic transfer hydrophobicity; K^0: Compressibility; μ: Refractive index; M_v: Molecular volume; M_w: Molecular weight; P: Turn tendencies; p: Polarity; pH_i: Isoelectric point; P_r: Polar requirement; R_a: Solvent accessible reduction ratio; R_F: Chromatographic index; V^0:Partial specific volume; H_c: Helical contact area

The number of positive selected sites having highest (BEB) posterior probabilities (PP) also varied and maximum number of sites with high PP were found in TLR7 (Viral) and TLR4 (Non viral), which shows for the first time that both viral and non-viral TLR genes follow similar selective regime (Table 5.2). Model M8 detected high number of positive selected (PS) sites

(4.5% sites, Total 42 sites, 30 sites PP > 0.90, 25 sites PP > 0.95 and 23 sites PP > 0.99) in TLR7 and somewhat similar scenario was found in TLR4 (5.5% sites, Total 45 sites, 29 sites PP > 0.90, 27 sites PP > 0.95 and 20 sites PP > 0.99). TLR3 is having least number of sites with high PP and out of 5% positive selected sites in TLR3 only 5 were having PP>0.90. Overall these results show that number and strength of positive selection sites varies in TLR supergene family and both viral (TLR7) and non-viral (TLR4) TLRs evolved under strong positive selection.

To further compliment our results we used multiple approaches (SLAC, FEL, REL, MEME and FUBAR) implemented in HyPhy package (Table 5.3) (http://www.hyphy.org) (Pond et al. 2005) (http://www.datamonkey.org/) (Pond and Frost 2005a) (Delport et al. 2010) for detection of positive selected sites. These approaches revealed a high number of positive selected sites in viral and non-viral avian TLRs (Table 5.3). The sites detected by more than one method (concordant between two or more methods) were considered as robust candidate for positive selection (Table 5.3) (Wlasiuk and Nachman 2010).

The detected positive selected sites were used to explore the possible role of positive selection on structural and functional diversification of respective TLRs as discussed below. The TLR1A is known as orthologue of TLR10, while TLR1B is orthologous of TLR1/TLR6 and TLR2B of TLR2, whereas there is no functional mammalian orthologue of TLR2B. The TLRs form dimers and recognize the PAMPs, the heterodimer between TLR2/TLR1 and TLR10/TLR2 recognize triacyl lipoproteins (Jin et al. 2007) (Guan et al. 2010) and TLR6/TLR2 recognize diacyl lipopeptides (Kang et al. 2009). The avian TLR1/TLR2 form heterodimers and are activated by both diacyl (Malp-2) and triacyl (Pam3) lipopeptides, with the exception of TLR2a/ TLR1b, which are activated by Pam3 but not by Malp-2. In addition, TLR2a/TLRL1b is activated by peptidoglycan (Huang et al. 2011). The chicken TLR1-like proteins interact with TLR2-like proteins and recognize agonists identical to those in mammals by heterodimers between TLR2 and TLR1, 6 or 10. TLR1, TLR2 and TLR6 evolve under positive selection in mammals (Wlasiuk and Nachman 2010) (Areal et al. 2011). In birds we found 12 positively selected sites in TLR1A, 26 PS sites in TLR1B, 34 PS sites in TLR2A and 25 PS sites in TLR2B. The PS site 304 in TLR2A reported by (Huang et al. 2011) (Alcaide and Edwards 2011) and PS sites 293, 295 and 296 found in TLR2B by Edwards (corresponding to the chicken sequence) were also found in our study (Table 5.3), PS sites corresponding to 295, 297 and 298 in the zebra finch sequence.

Overall out of 12 sites in TLR1A (Supplementary file 6), 10 PS sites were present in LRRs, out of which 5 were in variable region and possibly involved in PAMP binding. Only one site each was present in LRR-CT and TM. All 26 sites in TLR1B (Supplementary file 7), were located in

ECD with 7 and 13 sites in conserved and variable region, respectively, and 4 sites in LRR-NT and 2 in LRR-CT. In TLR2A out of 34 sites (Supplementary file 8), 2 were in signal peptide, 4 in LRR-NT and out of the remaining 28 sites present in LRR, 17 sites were in variable region. All of 34 sites found in TLR2A, were present in ECD with 2 sites in signal peptide region, 4 sites in LRR-NT and the other 28 in LRRs (21 in variable region and remaining 7 in conserved region). We found 25 sites in TLR2B, (Supplementary file 9), all of them in LRRs with 14 in variable region and 11 in conserved region. The PS sites found in our study possibly could be related with wide range of ligands. In mammals TLR2 is also known to recognize a variety of compounds other than triacyl lipoproteins, which includes lipoteichoic acids, lipoarabinomannan, and zymosan (Takeda and Akira 2004).

Secondly the formation of combinatorial binding sites by selection of TLR1 or TLR6 as the dimerization partner can explain at least in part the broad ligand specificity and possibly similar mechanism could also be hold true for avian counterpart and explain the extent of PS sites found in TLR1 and TLR2 counterparts in birds. The comparison of chicken TLR2A and TLR2B genes with respect to the human TLR2 gene reveals that PS sites 280, 292, 304, 308, 309, 311, 312, 315, 335, 344, 356, 372, 392, 393 and 413 of TLR2A and PS sites 260, 274, 295, 297, 298, 299, 302, 317, 328, 329, 343, 390 and 401 of TLR2B map at or near to the ligand-binding domain and dimerization surface as identified from the crystal structure of the complex TLR2–TLR1 with the tri-acylated Pam3CSK4 lipoprotein in humans and mice (Jin et al. 2007). These results are also in accordance with large number of PS sites found in mammals in the same region (Wlasiuk and Nachman 2010).

TLR3 recognize dsRNA and prevent the spread of most viruses. TLR3-ECD binds with dsRNA at two sites located at opposite ends of the TLR3 horseshoe structure. The first dsRNA:TLR3 interaction site is located close to the C-terminus, on LRR19-LRR21 and the second dsRNA:TLR3 interaction site is located on the N-terminal end (LRR-NT-LRR3) (Liu et al. 2008). The intermolecular contact between the two C-terminal domain region of TLR3 coordinates and stabilizes the dimer by a series of protein-protein interactions. We found 22 PS sites in TLR3 (Supplementary file 10), two each in signal peptide, LRR-NT and TM domain, and three sites in TIR domain of the remaining 13 sites were present in LRRs, and 9 of these sites were in variable region. The PS sites found in TLR3 possibly help in recognition and protection against rapidly evolving viral RNA (Liu et al. 2008).

TLR4 forms heterodimer with myeloid differentiation factor 2 (MD-2) and recognizes diverse LPS molecules (Kim et al. 2007), (Park et al. 2009) along with components of yeast, trypanosoma, and even viruses (Kumar et al. 2009a) (Kumar et al. 2009b) (Wlasiuk and Nachman 2010). The ECD of TLR4 consists of three subdomains: N subdomain consists of

LRR-NT and LRR 1-6, the central subdomain range from LRR7-12 and C subdomain consist of LRR 13-22 and LRR-CT. LPS causes dimerization of the TLR4-MD-2 complex at central and/or the C-terminal, and interaction between TLR4 and LPS-MD-2 complex takes place at the concave surface between N and central sub domain (Kim et al. 2007). We found 41 PS sites in TLR4 (Supplementary file 11), three were present in TM domain and one in TIR domain. The remaining 37 sites are present in ECD, with 3 in LRR-CT and others 34 in LRRs with 23 in variable region and 11 in conserved region. The majority PS sites are present in regions involved in ligand binding and dimerization.

The TLR4 consist of 330 aa long variable region in ECD also known as middle region and within the middle region there is 82 aa which are hyper variable across species with species-specific changes (Hajjar et al. 2002). Majority of PS sites found in TLR4 were concentrated near to hyper variable region. The primary contact interface between TLR4 and MD-2 involves two chemically distinct regions, the A and B patches provided by the N-terminal and the central domains of TLR4 and main dimerization interface of TLR4 is located in central C-terminal domain. The presence of positive selected sites in TLR4–MD-2–LPS complex may support the remarkable versatility of the ligand recognition mechanisms employed by the TLR family which is essential for defense against diverse microbial infection (Park et al. 2009).

The bacterial flagellin is virulence factor recognized by TLR5 (Hayashi et al. 2001). The residues 174–401 in ECD of TLR5 are responsible for species-specific flagellin recognition (Andersen-Nissen et al. 2007). The alpha and epsilon Proteobacteria are able to evade TLR5 recognition by mutating key residues in the TLR5 recognition site (Andersen-Nissen et al. 2005) and it is suggested that positive selection in primate TLR5 may be related with coevolution between PRRs and their microbial ligands (Wlasiuk and Nachman 2010). Twenty six PS sites were found in TLR5 (Supplementary file 12), with 2 sites each in signal peptide, LRR-NT, TM and TIR domains, one site in LRR-CT and remaining 17 sites in LRRs of which seven were in variable region and 10 in conserved region. Seven of these PS sites were located in 228 amino acid region identified previously (Andersen-Nissen et al. 2007) and thus could play important role in species-specific flagellin recognition and defense.

TLR7 recognize single-stranded RNAs (Gantier et al. 2008). In TLR7 we found 4, 1 and 4 PS sites in LRR-NT, TM and TIR domain, respectively, and 39 sites were present in LRRs of which 33 sites were in variable region (Supplementary file 13). This variable region directly interacts with the ssRNA, thus high proportion of PS sites in viral TLR7 is related with host pathogen arms race. TLR7 is also known to be evolving under strong selection pressure in mammals (Wlasiuk and Nachman 2010) (Areal et al. 2011). Recent studies found TLR7 to be under positive selection in bats, which are natural reservoirs and carriers of numerous deadly

viruses (Wang et al. 2011). The study suggested that long-term coexistence of bats and viruses imposed strong selective pressures on the bat genome, and is reflected by strong positive selection in genes, e.g. TLR7 involved in first line of anti-viral defense, the innate immune system (G. Zhang et al. 2013). The rapid evolution of viral TLR and its associated role in maintaining and dissemination of viruses in bats points towards the similar role and rapid evolution of viral TLR7 is birds. The avian TLR7 are candidates for the detection of influenza (Lund et al. 2004) and involved in recognition of ssRNA (mutation rates very high due to lack of mismatch repair), points towards the long coexistence of viruses and birds and may help explain the role of birds as natural reservoir, and vector of zoonotic pathogen (Lund et al. 2004).

TLR15 recognize yeast-derived agonist (Boyd et al. 2012), we found 41 PS sites (Table 5.3) with 1, 3, 4, 2, and 1 PS sites in signal peptide, LRR-NT, LRR-CT, TM and TIR domain, respectively, and 30 sites in LRRs out of which 14 were present in variable region. None of the PS sites were located in the highly conserved three Box regions of the TIR domain (Slack et al. 2000) (Areal et al. 2011). The exact mechanism of action of TLR15 is yet unknown but presence of PS sites is in accordance with other TLRs involved in immune defense. All together we found maximum number of PS sites in ECD with few sites in signal peptide, TM and TIR domain possibly due to their conserved functions. The PS sites in LRR-CT and LRR-NT could be related with structure stabilization. Only seven PS sites were present in signal peptide, which may be related with proper localization of secretory proteins as suggested previously (Areal et al. 2011). Similarly, the 11 PS sites in transmembrane domain and 13 in TIR domain maybe suggestive of some flexibility for these amino acid changes (Areal et al. 2011).

These results were further supported by protein level approaches. The TreeSAAP (Woolley et al. 2003) allow the detection of positive radical changes in physiochemical properties of amino acids, which in turn can affect the structure and function of proteins. The TreeSAAP results complimented our previous results of positive selection. Most PS sites found under high PP>0.99 (Table 5.4 and Figure 5.3), showed highest number of positive radical physiochemical changes. Out of 71 sites with high PP>0.99 (Table 5.4), 59 PS sites showed positive radical changes in physiochemical properties (TreeSAAP categories 6, 7 and 8). We also found 15 PS sites with type I radical changes (sites with 6 or more positive radical changes). TLR4 (non-viral) and TLR7 (viral) showed maximum number of sites with PP>0.99 (Table 5.4). In TLR4 we found 20 such sites (Table 5.4 and Figure 5.3) among which 17 showed type II changes (sites with less than 6 positive radical changes) and 6 PS sites showed type I changes. Most of the type I changes were restricted to gene TLR2A (5 sites) and TLR4 (6 sites) (Table 5.4 and Figure 5.3). In TLR7, we found 23 sites with PP>0.99,

among which 19 showed type II changes and two PS sites showed type I changes (Table 5.4). The results showed majority of sites having PP>0.99 were found in ECD with only few (4 sites) found outside the ECD (Figure 5.3), e.g. site 16 from TLR2A, sites 640 and 655 from TLR4 and site 920 in TLR7 were found in signal peptide, TM domain and TIR domain of the respective gene.

The homology modeling further showed the role of the highly significant positive selected sites (Table 5.4) on the structural and functional diversification of TLRs. We confirmed that the majority of sites were restricted to ECD (Figure 5.3) a variable extracellular domain (ECD) involved in the ligand recognition and dimerization and thus important for structure and function of the proteins. This strongly supports the fact that host-pathogen arms race drives the rapid evolution of TLRs. The homology model prediction was not significant for TLR1B and TLR15.

The overall finding of rapid evolution of the TLR supergene family from newly sequenced bird genomes is very interesting, and is well supported by the fact that the immune genes are known to experience strong adaptive selection due to the rapid evolution of pathogens (Nielsen et al. 2005), i.e. the host-pathogen arms race increase the chance of protection against diverse pathogens. These results support that both mammalian and avian TLRs evolve under strong evolutionary pressure. Indeed, both viral TLR7 (G. Zhang et al. 2013) and non-viral TLR4 (Areal et al. 2011) evolve at rapid rate in both birds and mammals, and support their associated roles, e.g. the rapid evolution of non-viral TLR4 for diversified ligands (e.g. LPS-lipopolysaccharide and LTA- lipoteichoic) whereas viral (TLR7) in both bats (G. Zhang et al. 2013) and birds, possibly contributed in maintaining and disseminating numerous deadly viruses (G. Zhang et al. 2013).

5.4 Conclusions

Our study addressed two major objectives: the comparative evolutionary genomics of vertebrate TLR superfamily and extensive adaptive evolution of avian TLRs. The evolutionary genomics of TLR supergene family from diverse vertebrate (Table 5.1) revealed a resolved phylogeny, with precise gene gain, gene loss events, and synteny organization that confirmed homologous relationships required to accurately assess the molecular evolution of vertebrate TLR supergene family.

We found that TLR multigene family is broadly divided into 6 major families namely TLR1, TLR3, TLR4, TLR5, TLR7 and TLR11 and each family has different number of subfamilies with 10 subfamilies in TLR1 family, 10 subfamilies in family 11, 3 subfamilies in TLR7, and single subfamily in TLR3-5. We conclude that most of the TLR originated early during vertebrate evolution and gene duplication together with differential rate of gene gain and loss shaped the

TLR gene family evolution in vertebrates leading to species and linage-specific TLR variations. For example, subfamilies TLR19-TLR20 and TLR23-TLR26 are fish specific; whereas TLR15 was specific to sauropsida and TLR27 is specific to coelacanth. The subfamilies TLR6 and TLR10-TLR12 are mammal specific and subfamilies TLR18-20, TLR23, TLR25-26 are specific to teleost fishes and TLR24 is only found in lamprey. The higher duplication events found in fishes (subfamilies TLR2, TLR4, TLR5, TLR7, TLR8, TLR9, TLR14, TLR24, and TLR19 to TLR23) compared to tetrapods (TLR1, TLR2, TLR5, TLR8 and TLR14) and reduced TLR supergene family repertoire from fishes (20 subfamilies) to mammals (13 subfamilies) suggest the greater role of TLRs in fishes. This, which is further strengthen by lack of major histocompatibility complex (MHC) II, CD4 and invariant chain (Ii) in Cod (Star et al. 2011) and IgM (Amemiya et al. 2013) in coelacanth, suggesting the alternative and compensatory role of TLRs.

The synteny arrangement of TLR shows the fish and tetrapod specific genomic arrangements with exception of TLR subfamilies (e.g. TLR3, TLR5M, TLR7 and TLR8), which have conserved syntenic organization across all vertebrates. The coelacanth was an exception to this trend as its TLRs showed shared features with both tetrapods TLR1, TLR2 and TLR13 and fishes TLR21, suggesting their evolutionary proximity with both and/or due to its unique immune system which lacks IgM involved in adaptive immune system (Amemiya et al. 2013) (Boudinot et al. 2014).

The sequencing on many avian genomes (Jarvis et al. 2014) (Zhang et al. 2014) provided the unique opportunity to assess the extensive positive selection in avian TLRs. The multiple approaches used for positive selection of avian TLRs found in our study suggest that the host pathogen arms race have played an important role in rapid evolution of avian TLRs. We also found high positive selection in both viral and non-viral TLRs (Table 5.3) showing the both viral and non-viral TLR are evolving at rapid rate. Among non viral TLRs, we found TLR4 with maximum number of positive selected sites (20 sites with highest PP>0.99) and among viral TLR we found TLR7 with highest number of PS sites (23 sites with PP>0.99). TLR7 is involved in recognition of ssRNA (mutation rates very high due to lack of mismatch repair) and is candidate for the detection of influenza (Lund et al. 2004). The rapid evolution of TLRs possibly explains the host-pathogen arms race, which leads to rapid evolution of immune genes to adapt against the pathogens. The majority of PS sites were located in LRRs of ECD, which is mainly involved in PAMPs recognition. The large number of PS sites found in non-viral TLR4 points towards the wide diversity of pathogen they recognized (e.g. LPS-lipopolysaccharide and LTA- lipoteichoic). The finding of rapid evolution of viral TLR7 supports the host-pathogen arms race leading to co-evolution and possibly explains the strong selective pressure imposed by the long term coexistence of viruses and birds and may help

understanding the role of birds as natural reservoirs and vectors of zoonotic pathogens (Reed et al. 2003).

5.5 Materials and Methods

5.5.1 TLR gene finding and synteny analysis

We used representative sequences of vertebrate TLR1-26 as query and did exhaustive blast (Altschul et al. 1997) searches using intermediately stringent level of E-10 to retrieve all available TLRs from 8 reptiles and 4 fish (Table 5.1) .The details of all the sequences retrieved are provided in (Supplementary file 1). The hits were retrieved by extending 2000 bp at both ends and scaffolds of interest were then submitted to gene annotation program like GENSCAN (Burge and Karlin 1997)and a Hidden Markov Model gene prediction program MolQuest (http://www.molquest.com/) and FGENESH Softberry (http://linux1.softberry.com/berry.phtml) using parameters for chicken and anolis, to identify predicted gene sequence (Gillespie et al. 2013) (Wang et al. 2012). Finally, the predictions were verified by BLASTP against NCBI non-redundant protein sequence database. For adaptive evolution study of avian TLR, all 48 bird genomes CDS database were searched using blast and all significant TLR sequences were retrieved and used for adaptive analysis study (Supplementary file 2).

5.5.2 Phylogenetic analysis

The resulting orthologs were aligned using Muscle (Edgar 2004) implemented in Seaview (Gouy et al. 2010) and sequences were tested for nucleotide substitution saturation using DAMBE 5 (Xia 2013) by plotting number transition transversion against the genetic distance using F84 model (Huelsenbeck and Rannala 1997) which allows for different equilibrium base frequency and transition transversion rate bias for nucleotide substitution. Xia test (Xia et al. 2003) implemented in DAMBE5 (Xia 2013) was done to compare index score (ISS) with critical score (ISS.C) with 3rd and other codon positions to get estimate of saturation. The vertebrate TLR phylogetic tree was made in MEGA5 software using neighbor joining method, with 1000 bootstrap replication. For phylogenetic analysis of TLR1A, TLR1B, TLR2A and TLR2B sequences were aligned using Seaview and manually corrected and used to detect the best substitution model using jModelTest 2(Darriba et al. 2012) based on Akaike Information Criteria (AIC), which was used for Maximum Likelihood tree construction using PhyML(Guindon and Gascuel 2003) with 500 bootstrap replicates to check the robustness and reliability of tree (Felsenstein 1985). TLR genes used for adaptive analysis were aligned in Seaview (Gouy et al. 2010) using Muscle (Edgar 2004) and checked for saturation using DAMBE 5 (Xia 2013) as described above. The tree topology used for positive selection analysis follows the species tree provided by consortium (Supplementary file 4). (Jarvis et al. 2014) (Zhang et al. 2014).

5.5.3 Gene conversion

The sequences alignments were tested for recombination using GARD (Genetic Algorithm for Recombination Detection) (Pond et al. 2006) available online http://www.datamonkey.org, GENE-CONV (Sawyer 1989) and RDP (version 3) software for detection of gene conversion events with 1,000 permutation and bonferroni corrected p-value cutoff of $p<0.01$ and mismatch were allowed (/g1 =1). The gene conversion free region represent the species tree (Supplementary file 5.4 Figure 1 and 2).

5.5.4 Positive selection

In proteins different functional sites undergo through different selection pressures. If the changes are disadvantages (harmful) they are not inherited and thus are removed from population (Negative selection $\omega<1$), resulting in conservation of such sites and function of protein. On other side if changes are useful and help to better adapt to the environment they are positively selected and remain in population increasing the (Positive selection $\omega>1$). We analyzed TLRs (TLR1A, TLR1B, TLR2A, TLR2B, TLR3, TLR4, TLR5, TLR7, TLR15) except TLR21 from recently sequenced birds genomes (Supplementary file 2) for signals of diversifying positive selection using codon model implemented in PAML (Yang 1997) (Yang 2007) and Datamonkey (Pond and Frost 2005a) together with amino acid model in TreeSaap (Woolley et al. 2003). We employed different approaches to find signals of positive selection in avian TLR genes. Codeml in PAML package version 4.7 (Yang 1997) implements likelihood ratio test (LRT) for comparison of sophisticated nested site specific models calculated as twice the difference of log likelihood between the two models following chi square distribution with degree of freedom corresponding to the difference in number of parameters between the nested model i.e. null model (no selection) and alternate model (positive selection). The significant LRT means null model is rejected and sites are under positive selection. We compared M1a (Nearly Neutral)vs M2a (Positive Selection), and M7 (beta) vs M8 (beta& w) to find sites under positive selection. Bayes empirical Bayes (BEB) inferred the posterior probabilities of positive selected sites where higher PP meaning high confidence. Other than PAML site models we used Hyphy package (http://www.hyphy.org) (Pond et al. 2005) (http://www.datamonkey.org/) (Pond and Frost 2005a) (Delport et al. 2010) that provides different approaches (SLAC, FEL, REL, MEME and FUBAR) for detection of positive selected sites, including Single Likelihood Ancestral Counting (SLAC), Fixed Effects Likelihood (FEL), Random Effects Likelihood (REL), (Pond and Frost 2005b) Mixed Effects Model of Evolution (MEME) (Murrell et al. 2012), Fast Unconstrained Bayesian AppRoximation (FUBAR) (Murrell et al. 2013) and integrative approach. SLAC model uses ancestral sequences reconstruction, FEL calculates site by site dn/ds without assuming a prior distribution whereas REL assume a prior distribution across site, FUBAR ensures robustness against model misspecification, and MEME is most appropriate to detect episodic diversifying selection affecting individual codon

sites. Along with this, the integrative approach results incorporate all sites detected by SLAC, FEL, REL, FUBAR and MEME. The sites detected by two different methods are further supportive of positive selection.

Further support for our results was gained by complementary protein level approach implemented in TreeSAAP (Woolley et al. 2003). It uses ancestral sequence reconstruction to find the physiochemical properties change of amino acid replacement using 31 amino acid properties. The amino acid replacement can lead to conservative or radical change in physiochemical properties. The positive radical changes can lead to change in structure and/or function of protein and the number of radical changes at a site can be used as an indicator to show strength of positive selection. To facilitate interpretation of level of changes at a site we categorized the sites into two types. Sites having six or more radical changes were defined as type I and sites with less than six properties where defined as type II.

5.5.5 Domain architecture, Homology modelling and structure analysis

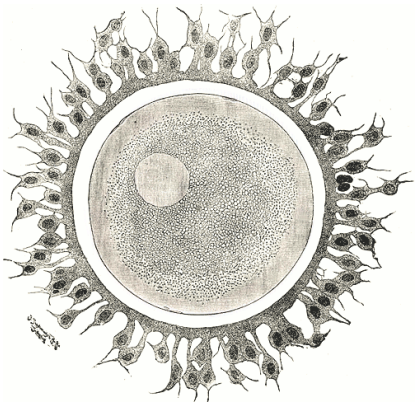
The LRRfinder was used to predict the domain architecture and define the protein domain locations of specific amino acid residues in TLR proteins (Supplementary file 6-14). This was also verified using the Uniprot protein database (http://www.uniprot.org webcite) whenever possible. The structure of each TLR was predicted using *CPHmodels* 3.2 protein homology modeling server, which resulted in significant modeled structure for complete region of TLR5 (Figure 5.3) and ECD region of TLR1A, TLR2A, TLR2B, TLR3, TLR4, and TLR7. No significant structure was predicted for TLR1B and TLR15. All the highly significant positive selected sites (Table 5.4) were displayed on the respective predicted structures to show the potential functional or structural significance of specific amino acid residues.

Acknowledgements

IK was funded by a PhD grant (SFRH/BD/48518/2008) from Fundação para a Ciência e a Tecnologia (FCT). AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490).

6

Comparative evolutionary genomics of the Zona Pellucida (ZP) gene family in vertebrates reveals gene expansion and adaptive evolution in the avian genomes



6.1 Abstract

Background

The Zona Pellucida (ZP) gene family is involved in egg envelope formation throughout the vertebrate lineage. These genes evolve rapidly, with differences occurring in the type and number of copies, causing lineage and species-specific gamete interactions that may originate reproductive barriers and lead to speciation. Contrary to monospermic mammals, birds and reptiles undergo physiological polyspermy, without detrimental consequence on later development.

Result

We found differences in type, number of genes and rapid evolution of vertebrate ZP genes. Our study explored for the first time the comparative evolutionary genomics and adaptive evolution of the ZP gene family among vertebrates and extensively in birds and mammals. We found that family ZPB and ZP3 have undergone fish specific expansion. The ZPB/4 duplicated in amniotes after divergence from amphibian and gave birth to ZP1 and ZP4 in amniotes. We found ZP2 in spotted gar and coelacanth thus affirming its presence in fishes. The ZPD is tetrapod specific present only in sauropsida and frog. The coelacanth ZP repertoire shows shared characteristic of tetrapods (ZPAX1, ZPY ZP3= ZP3-ENSLACG0000007941) and fishes (ZPB), suggesting evolutionary transition among lineages. The spotted gar genome (Actinopterygian) diversified from teleost before the whole genome duplication, also revealed by the ZP relatedness between fish and other tetrapods (e.g. the ZP2 in spotted gar showed shared synteny with tetrapods, whereas ZP2 was lost in teleosts). We detected lineage specific variations in ZPAX evolution with ZPAX1, ZPAX2 and ZPY found in tetrapods and coleocanth, whereas ZPAXA and ZPAXB are fish specific. The ZP1 and ZP2 showed significant omega variation among amniotes with highest omega values found in mammals, 0.31 and 0.51 for ZP1 and ZP2, respectively. The ZPAX2 had higher omega (ω = 0.30) compared to ZPY (ω = 0.20). The difference in omega estimate between ZPAX1 and ZPAX2 is not significant. The ZPAX1 in amniotes is evolving at faster rate than amniotes. Random site model suggest rapid evolution of ZP genes within avian and mammalian lineages. Among birds, ZP4 had highest omega (M8; ω = 3.84, n=9 PSC) and ZPD had the least (M8; ω =1.66, n = 21 PSC). Among mammals, ZP2 had highest omega (M8; ω =1.91 n = 26 PSC) and ZP4 had the least (M8; ω =3.84, n = 9 PSC). The positive selected sites detected by random site models showed positive radical changes in amino acid properties leading to functional and structural diversification of proteins with possible implications in ZP matrix formation and species specific gamete recognition.

Conclusion

The ZP gene family has undergone lineage and species specific changes as seen by gene gain, loss and functional diversification in all vertebrate lineages. The rapid evolution of ZPs in both monospermic mammals and physiological polyspermics birds and reptiles suggest that polyspermy avoidance is not the only driving force for the rapid evolution of ZPs, and is possibly caused by diverse but entwined evolutionary forces, which includes cryptic female choice, sperm competition, hybridization avoidance and sexual conflict and ultimately results in speciation.

6.2 Introduction

The egg is a haploid female gamete surrounded by an extracellular glycoproteinaceous matrix composed of Zona Pellucida (ZP) glycoproteins and referred as Zona Pellucida layer in mammals, previtelline membrane in birds and chorion in fishes. The ZPs are involved in sperm binding and possibly protection of oocyte and embryo (Smith et al. 2005). The rapid evolution of reproductive proteins mediating gamete interaction, gamete usage, storage, signal transduction and fertilization plays an important role in speciation. The gamete interaction is crucial step in sexual reproduction and the co-evolution of proteins involved in sperm-egg interaction can lead to reproductive isolation and the establishment of new species (Swanson and Vacquier 2002). The vertebrate egg envelope is formed by member of Zona Pellucida (ZP) gene family, and ZP gene family shows constant process of gene amplification and attrition across vertebrates, with consequent changes to the composition of the egg envelope (Hughes 2007).

The sperm-egg recognition involves multiple protein interactions and this is supported by the apparent plasticity in which ZP proteins are recognized by sperm (Castle 2002). This redundancy both guards against abrogation of fertilization due to single gene mutations and also provides a mechanism where in loss of envelope genes provides an impetus for modified sperm-egg recognition, reproductive isolation and thus speciation (Hughes 2007).

Earlier the number of ZP genes were thought to be associated with fertilization mechanism with high number of ZP genes present in external fertilization (as in amphibians or fishes) whereas less ZP genes in internal fertilization (as in mammals). However, the presence of high number of ZP genes in birds and reptiles does not support this theory (Goudet et al. 2008). Considerable differences are observed in type

and number of ZP gene members present across vertebrates. In human the Zona Pellucida gene family consist of four members, namely ZP1, ZP2, ZP3 and ZP4, whereas birds and frog have extra members ZPAX and ZPD (Goudet et al. 2008)(Hughes 2007).

The gene gain, loss and rapid evolution of ZP genes makes elucidation of their homologous relationship confusing and misleading, often resulting in nomenclature problems, with multiple names assigned for a single gene (Goudet et al. 2008)(Hughes 2007). The ZP glycoproteins have been classified and named based on various criteria e.g. molecular weight, charge, sequence length, sequence identity. This also resulted in multiple synonyms and confused nomenclature (Goudet et al. 2008) (Bleil and Wassarman 1980) (Harris et al. 1994) (e.g. the ZP1/ZPB1, ZP2/ZPA, ZP3/ZPC, ZP4/ZPB, ZPD/ZPX2 and ZPAX/ZPX1 and ZPY in frog). Various studies have tried to clarify this issue and grouped ZP genes into ZP1, ZP2, ZP3, ZP4, ZPAX and ZPD sub families (Smith et al. 2005) (Goudet et al. 2008)(Hughes 2007).

The function of ZP1 is similar in all species and it is mainly structural (Gupta et al. 2012) (Smith et al. 2005) forming disulfide bonds with ZP2/ZP3 filaments (Jovine et al. 2005) (Kubo et al. 2010). However, in humans it was suggested that ZP1 can also bind sperm and induce acrosomal exocytosis (Gupta et al. 2012). In chicken ZPC/ZP3 was identified as primary sperm receptor (Goudet et al. 2008), whereas the secondary sperm receptor function was attributed to ZPA/ZP2 (Hughes 2007). It was reported that ZPD, is the major component of mature egg envelope (Okumura et al. 2004), and dimeric ZP1 might be responsible for stimulate sperm activation (Okumura et al. 2004). Fishes have duplicated ZPB/ZP4, ZPC/ZP3 and ZPAX, with 2 or 3 copies in each of them (Meslin et al. 2012) (Sano et al. 2013) and fish ZPs are quit diverged as compared with mammalian ZP genes (Goudet et al. 2008). The ZPB/ZP4, ZPA/ZP2, ZPAX, ZPC/ZP3 and ZPD are involved in sperm-zona pellucida binding, whereas ZP1 is responsible for acrosome reaction (Meslin et al. 2012) (Mao and Yang 2013). The ZPC/ZP3 and ZPA/ZP2 are widespread in all vertebrates, these genes possibly have conserved functions (Goudet et al. 2008). ZP3/ZPC in addition to be the primary sperm receptor is also able to induce acrosomal reaction and ZP2/ZPA is the secondary sperm receptor and crucial to prevent polyspermy (Gupta et al. 2012). ZP2 mediates sperm/egg binding (Jovine et al. 2004) (Smith et al. 2005) and could also has function in prevention of polyspermy (Gupta et al. 2012). Less is known about ZP4, but this protein can bind sperm and induces acrosomal exocytosis in humans (Gupta et al. 2012). ZPD is the main constituting of mature chicken egg envelope and, in conjunct

with ZP1, stimulates sperm activation (Okumura et al. 2004). The function of ZPAX is still unknown and in *Xenopus laevis* this protein is inactive as sperm ligand (Vo and Hedrick 2000) and the chicken ZPAX has a structural function (Hughes 2007).

The ZP proteins act in synergism because the sum of sperm binding by individual isolated ZP proteins is much less than that in integral egg envelope (Hedrick 2008). Thus, possibly all ZPs have some function in induction of acrosome reaction, with exception of ZP2 (Gupta et al. 2012). The only exception to previously reported ZP functions is in fishes where, due to physiological characteristics, the role of ZPs appear to be only structural (Conner and Hughes 2003). In addition to be extremely important for sperm binding and fertilization, zona pellucida is also crucial for the correct embryo development (Conner et al. 2005). Generally ZP genes in chordates are expressed in oocytes but there are some exceptions (e.g. in some fishes ZP is expressed in ovary and/or liver) (Sano et al. 2013) (Berg et al. 2004) and in chicken ZP1 is expressed in liver and ZPC and ZPD are expressed in ovarian granulosa cells (Okumura et al. 2004) (Smith et al. 2005).

The greater complexity in the number and relationship of the vertebrate egg envelope forming Zona Pellucida gene family makes them important target of comparative and evolutionary genomics studies. Thus in depth exploration of gene family members from various genomes together with their genomic organization (synteny) supported with phylogenetic analysis can provide valuable information to solve the complex evolution of the ZP gene family. Therefore, the present study of gamete recognition proteins ZP gene family in newly sequenced vertebrate genomes, covering diverse variety of species and lineages (fish, amphibians, birds, reptiles and mammals) could elucidate the molecular evolution and the adaptive role of the ZP gene family in reproductive isolation and speciation.

6.3 Results and Discussion

6.3.1 Genomic scans and phylogenetic relationships of the Zona Pellucida gene family in vertebrates

We performed extensive genomic scans for vertebrates ZP gene family members in 26 diverse species covering major vertebrate's lineages of fishes, amphibians, birds, reptiles and mammals (Table 6.1, Figure 6.1). We also scanned 48 bird genomes (Jarvis et al. 2014) (Zhang et al. 2014) in order to find the ZP gene family members. The ZP gene family members found from genomic scans were used for phylogenetic

reconstruction. The genomic scans together with phylogenetic and synteny analyses provided high resolution of the ZP family repertoire distribution and evolution in vertebrate genomes and allowed assessing the phylogenetic relatedness and homologous relationships of ZP gene family members (Figure 6.1, Table 6.1 and Supplementary Table 6.1).

The phylogenetic analysis (Figure 6.1) confirmed that ZP gene family consists of 6 major subfamilies (ZP1, ZP2, ZP3, ZP4, ZPD and ZPAX). The overall phylogenetic tree suggests ZP3 to be the most ancestral. The phylogeny also shows that subfamilies ZP1, ZP2, ZPD and ZPAX form a closely related group, which is distantly related with subfamily ZP3 (Spargo and Hope 2003)(Goudet et al. 2008).

The phylogenetic relationship of ZP genes shows that ZPAX underwent independent species/lineage-specific duplication giving rise to ZPAX1, ZPAX2, ZPY, ZPAXA and ZPAXB members of ZPAX glycoproteins. Phylogenetic analysis shows that all these members are closely related and cluster together forming the ZPAX clade. The ZPAX originated before the divergence of Agnatha (*Petromyzon marinus*) and Gnathostomata around 540 million years ago (Mya) with the occurrence of ZPAX in lamprey and in spotted gar genomes (Table 6.1, Figure 6.2). The ZPAX duplication in teleosts, lead to ZPAXA and ZPAXB duplicates, which is in accordance with the whole teleost genome duplication event (Table 6.1, Figure 6.2). A different event (detailed later) gave rise to ZPAX1, ZPAX2 and ZPY in tetrapoda and coelacanth (Table 6.1, Figure 6.2). The ZPAX1 and ZPY first originated in coelacanth and syntenic to tetrapods, the ZPAX1 and ZPY are closely related in the phylogeny as compared to ZPAX2. The ZPAX2 in turn is closely related with fish ZPAX genes. The ZPAX2 first appeared in reptiles with conserved synteny in sauropsida and mammals. The ZPAX1 and ZPAX2 are present in the genomes in close proximity with conserved synteny (Figure 6.3), whereas ZPY is present in a distinct location (Supplementary File 6.2, Figure 4-c). These evolutionary events of ZPAX are well supported by the phylogeny and further corroborated by the syntenic organization.

The ZP phylogeny shows that the lamprey ZP2 and ZPB form a distinct clade, being the basal root of ZP1 and ZP2 genes (Figure 6.1), which suggests that basal chordates most likely shared a common ancestor (Xu et al. 2012). All lamprey ZP3 genes grouped together forming a monophyletic clade (Figure 6.1). The phylogeny also supports the origin of ZP1 and ZP4 from ZPB after tetrapod and fish divergence.

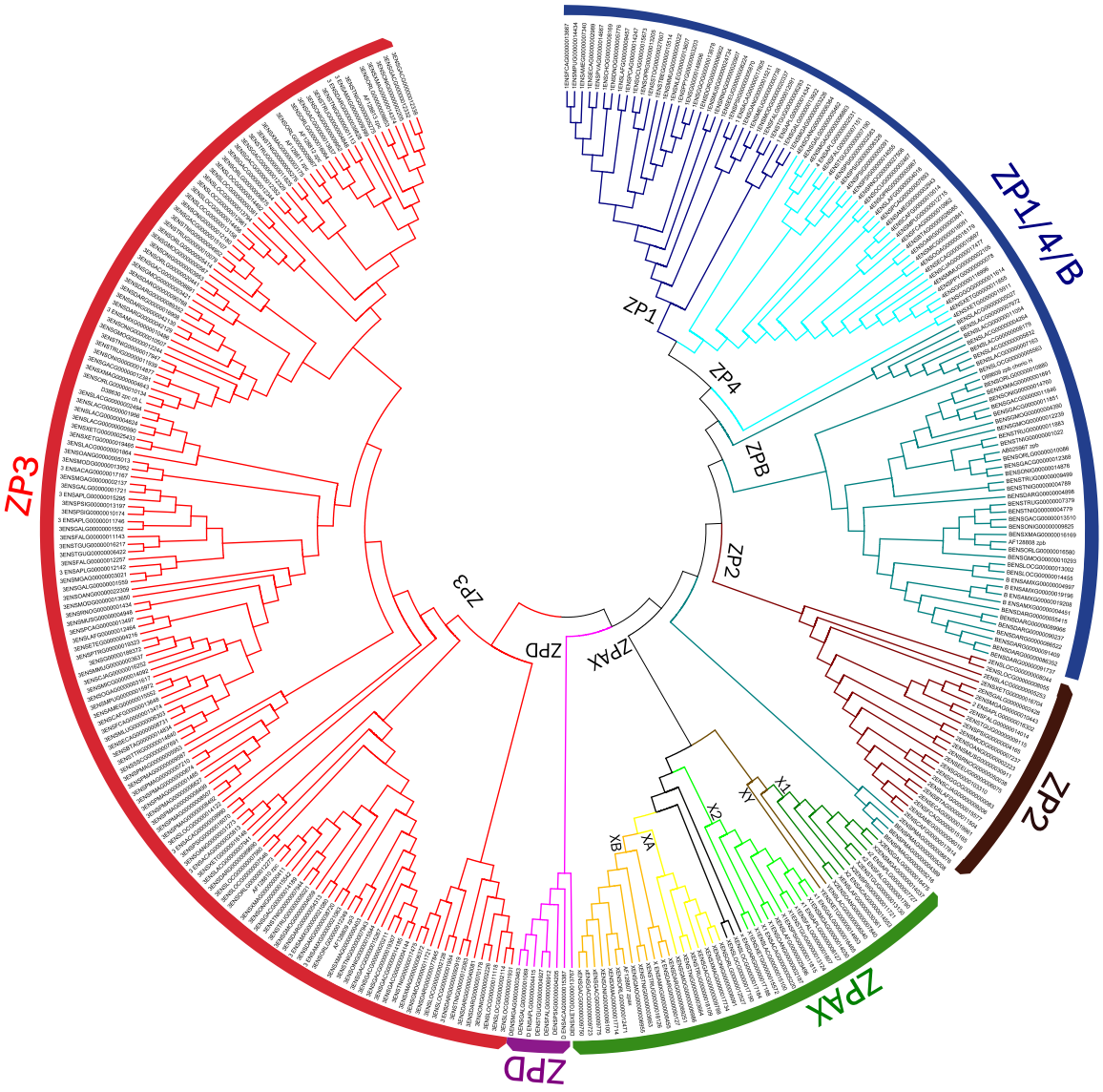


Figure 6.1: The phylogeny showing the molecular evolution of vertebrate ZP gene family.The amino acid sequences were used to construct the Neighbor Joining tree with 1000 bootstrap replication using MEGA5 software. All major clades have high bootstrap support of more than 90

6.3.2 ZP gene family repertoire in vertebrates

The dynamics of gene gain and loss plays an important role in the evolution of gene families. In this study we found that ZP genes in vertebrates form a multigene family, which can be divided into six major subfamilies: ZP1, ZP2, ZP3, ZP4, ZPD and ZPAX (Table 6.1).

Table 6.1: The gene gain and loss events of ZP gene family

		ZP1/4/B		ZP2/ZPA	ZPAX		ZPY	ZPD	ZP3/ZPC
		ZP1	ZP4		ZPAX1	ZPAX2			
Mammals	Humam	1	1	1	-	-	-	-	1
	Rat	1	1	1	-	-	-	-	1
	Mouse	1	-	1	-	-	-	-	1
	Elephant	1	1	1	1	1	-	-	1
	Opossum	1	-	1	-	-	-	-	2
	Platypus	1	1	2	1	1	-	-	3
Birds	Turkey	1	1	1	1	1	-	1	2
	Chicken	1	1	1	1	1	-	1	3
	Duck	1	1	1	1	1	-	1	3
	Zebrafinch	1	1	1	1	1	-	1	2
	Flycatcher	1	1	1	1	1	-	1	2
Reptiles	Chinese turtle	1	4	1	1	1	-	1	3
	Lizard	1	-	2[a]	1	1	1	1	3
		ZPB							
Amphibians	Frog	3		1	1	-	1	1	3
Fishes	Coelacanth	7		1	1	-	1	-	6
					ZPAXa	ZPAXb			
	Platyfish	2		-	1	1	-	-	6
	Medaka	4(1)		-	1	1	-	-	10(1)
	Tilapia	3		-	1	1	-	-	9
	Stickleback	4		-	1	3	-	-	17(2)
	Fugu	3		-	1	1	-	-	5
	Tetraodon	3		-	1	1	-	-	9
	Cod	3		-	1	1	-	-	7(2)
	Zebrafish	9(1)		-	1[b]	2	-	-	13(1)
	Cavefish	4		-	-	1	-	-	4
	Spotted gar	4		2	2[c]		-	-	14
	Lamprey	3		1	1[c]		-	-	9

ZPB/4 family is composed by ZPB genes present only in fishes and ZP4 genes present in tetrapods. ZPAX family possesses two big subfamilies. ZPAX subfamily in tetrapods and coelacanth could be classified in ZPAX1 and ZPAX2 whereas in fishes it could be classified in ZPAXa and ZPAXb. a) The two lizard genes are not syntenic and were classified as ZP2-Like genes.

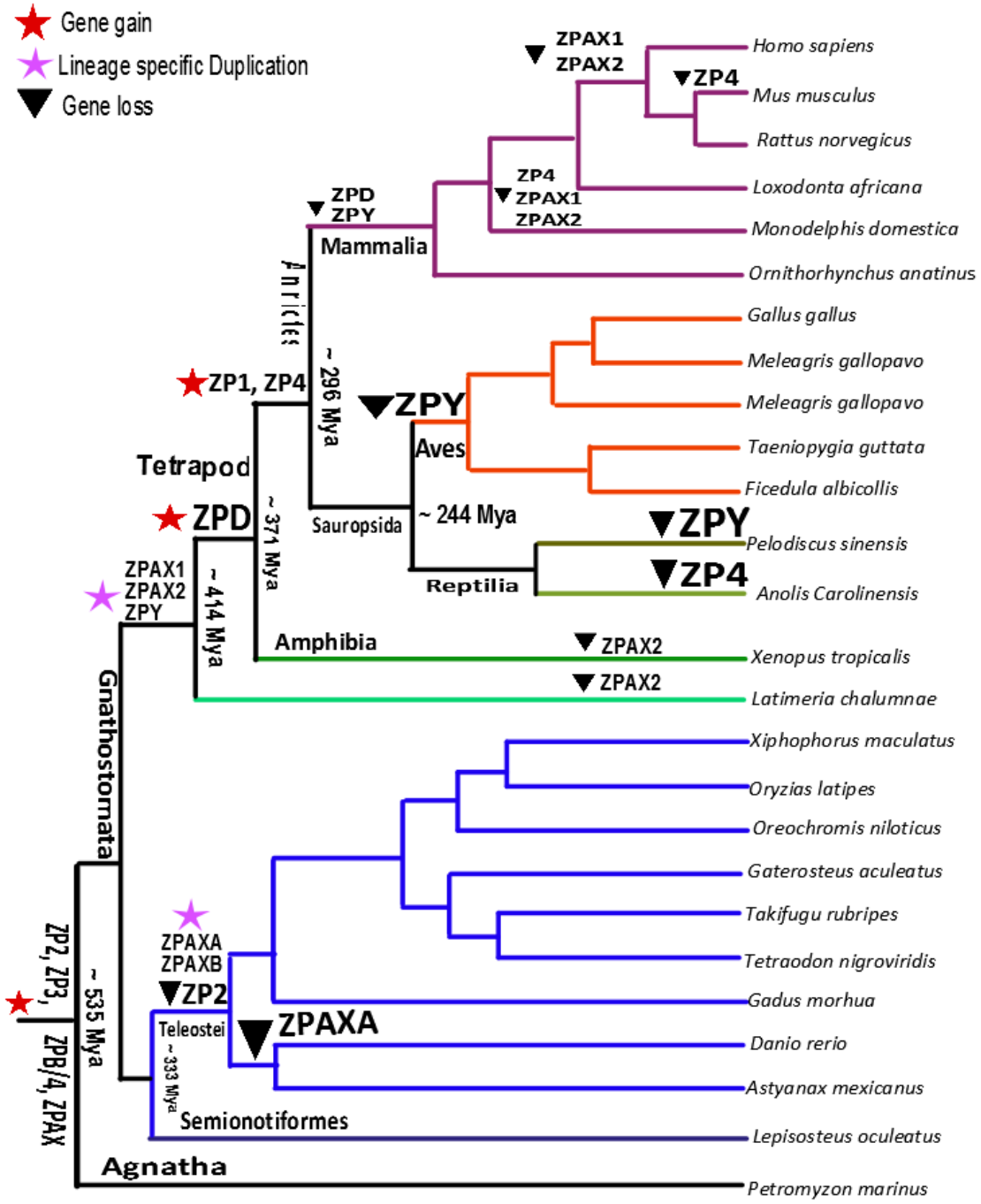


Figure 6.2: The gene gain and gene loss shaping the evolution of ZP subgenome in diverse vertebrate species

The subfamily ZPAX in turn underwent series of species and lineage-specific gene gain and loss leading to three classes, ZPAX1, ZPAX2 and ZPY in tetrapod, and ZPAXA and ZPAXB in fishes (Table 6.1, Figure 6.1 and Figure 6.2). The duplication of the

common ancestral gene ZPB in fishes gave rise to ZP1 and ZP4 (Spargo 2002). We found duplicated ZP1 and ZP2 genes in reptiles, suggesting that most likely duplication event took place in the lineage leading to the amniotes after their divergence from amphibians ~370 Mya leading to ZP1 and ZP4 in amniotes (birds, reptiles and mammals) (Figure 6.2). The species-specific distribution of ZP1 and ZP4 in mammals is caused by species-specific loss events of either ZP1 or ZP4 genes (Figure 6.2 and Table 6.1). Based on the loss of ZP1 or ZP4 genes, the mammals can be further divided into two groups, one lacking ZP4 (e.g. mouse) and other lacking ZP1 (e.g. dog and cat) (Goudet et al. 2008). Similary species-specific variations could explain the absence of ZP4 in Anolis but its presence in turtle. Independent species-specific ZPB/4 duplications occurred in lamprey, fishes, frog and turtle. The species-specific duplications of ZPB and ZP4 gave rise to two or more in-paralogs, e.g. ZPB have three copies in frog, and around two to 10 copies in fishes, whereas turtle have 4 copies of ZP4 (Table 6.1). Some of these duplicates also show tandem arrangements (Supplementary File 6.2, Figure 2e, 2i, 4b-1).

There is no earlier report of the presence of ZP2 in jawed fishes but our study found two copies of ZP2 in the spotted gar and one copy in coelacanth (Table 6.1 and Figure 2). We also found the presence of ZP2 in reptiles (Table 6.1 and Figure 6.2). The ZPD have not been reported previously in fishes and mammals and our search also did not found them. However, we were able to find ZPD in reptiles suggesting that ZPD in sauropsida and amphibian, likely originated ~430 Mya (Figure 6.2) before the divergence of amphibian and sauropsida lineage, possibly explaining a lineage specific role of ZPD.

The earlier studies have reported the ZPAX1 and ZPAX2 only in birds and ZPAX1 and ZPY in frog (Figure 6.2). However, we found the presence of ZPAX1, ZPAX2 and ZPY in reptiles and mammals (Figure 6.2). We also found ZPAX1 and ZPY in coelacanth. The ZPAX1 and ZPAX2 are also reported in the platypus genome, whereas ZPAX1 and ZPAX2 are pseudogeneized in other mammals (Smith et al. 2005) (Hughes 2007). We also found partial copies of ZPAX1 and ZPAX2 in the *Loxodonta africana* genome. The presence of ZPAX1, ZPAX2 and ZPY in tetrapods and coelacanth together with ZPAXA and ZPAXB further revealed the lineage-specific changes leading to fish and tetrapod specific ZPAX gene members (Table 6.1 and Figure 6.2). The ZP gene family repertoire of coelacanth genome shows shared characteristic with both tetrapods and fishes, e.g. like tetrapods ZPAX1 and ZPY are present in coelacanth secondly and one of the six ZP3-ENSLACG0000007941 genes show shared synteny with tetrapods

(Supplementary File 6.2, Figure 6c). Like fishes, coelacanth also shows presence of ZPB. This shared characteristic may be related with the water to land transition as suggested by the evolutionary proximity of coelacanth with tetrapods and fishes.

We found duplicated copies of ZPAXB gene in stickleback, zebra fish and two copies of ZPAX in the spotted gar (Figure 6.1, Figure 6.2, Table 6.1 and Supplementary file 6.2, Figure 4b-1). The ZP3 proteins are known as primary sperm receptors, widely distributed across vertebrates (Jovine et al. 2004), with high number of gene duplicates within fishes (4 to 19 copies) compared to tetrapods (1 to 3 copies). Among tetrapods, birds and reptiles have around three copies (e.g. ZP3a, ZP3b and ZP3c are present in birds) and among them ZP3a is closely related with mammalian ZP3 compared with the other more derived ZP3b and ZP3c. The gene duplication provides the important raw material for evolutionary modification and thus the duplicated genes found in vertebrate ZP gene family favored the neofunctionalization or subfunctionalization of the duplicated paralogs resulting in species adaptation and improved fitness.

6.3.3 Synteny organization of ZP genes

The syntenic organization of genes is useful in resolving and understanding the homologous relationship of genes. Here we performed the synteny analysis of ZP genes across vertebrates in order to understand their homologous relationships. The ZP1 and ZP4 were originated from a duplication event that took place in the common ancestor of amniotes. We found conserved synteny for ZP1 across all amniotic species (Supplementary File 6.2, Figure 1). By contrast, synteny of ZP1/4/B is not well conserved and shows independent lineage-specific synteny conserved within groups of mammals, birds, reptiles, frog, and fishes (Supplementary File 6.2, Figure 2a- 2k). The ZPB and ZP4 also underwent species-specific duplications resulting in multiple copies in reptiles, frog and fishes, which show different syntenic clusters. We further evaluated the gene arrangement of these duplicates and found that some ZPB and ZP4 are tandem duplicated (Supplementary File 6.2, Figure 2c-1, 2d-2, 2e).

There is no report of the presence of ZP2 in fishes though there are some reports of misnamed ZPB as ZP2 (Smith et al. 2005) (Hughes 2007) and based on this it was assumed that ZP2 first appeared in the tetrapod lineage after divergence from fishes. Recent studies have shown the presence of ZP2 in lamprey and we found ZP2 for the first time in jawed fishes, i.e. coelacanth and spotted gar. Synteny of ZP2 gene is conserved between spotted gar and tetrapods (Supplementary File 6. 6.2, Figure 3a).

The ZPAX genes present in fishes and tetrapods underwent multiple changes with differential gain and loss together with gene rearrangement giving rise to lineage-specific diversification, namely ZPAX1, ZPAX2 and ZPY in tetrapods, and ZPAXA and ZPAXB in fishes (Figure 6.3a and 6.3b). This is well supported with differential synteny arrangements of ZPAX genes: 1- amniotes, 2- coelacanth and frog, 3-teleost fishes, and 4- spotted gar (Actinopterygian fish), each with different level of shared synteny (Figure 6.3a). The teleost whole duplication lead to difficulty of ortholog assignment after lineage-specific loss of duplicated genes and the asymmetric evolution of gene duplicates (Amores et al. 2011) (Lu et al. 2012). The Actinopterygian fish genome (e.g. spotted gar) diversified from teleosts before the teleost-specific genome duplication and thus is helpful to understand the relatedness between fish and tetrapods. The present arrangement of ZPAX genes in these four groups (Figure 6.3a) suggest that the common ancestor of fish and tetrapods had at least two ZPAX genes, which underwent independent duplication and diversification leading to the present distribution of ZPAX genes. The most parsimonious explanation of this scenario based on the evolutionary changes from the transition of fish and tetrapods from a common ancestor is depicted in (Figure 6.3b).

This scenario suggest reversal event in ancestor arrangement (Figure 6.3b) resulting in a common scenario in all fishes, this also accompanied with the translocation of MSGN1 and GEN1 genes. This together with species-specific duplication of OSR1 gene created the present arrangement of ZPAX seen in spotted gar. The common fish ancestor leading to teleost underwent loss of one ancestral ZPAX together with translocation of its flaking genes (RDH14, NT5C1, OSR1, OSR1) and WDR35 gene (Figure 6.3b). The other ancestral ZPAX underwent independent duplication resulting in teleost specific ZPAXA and ZPAXB genes (Figure 6.3b) as seen presently in the majority of teleost fish genomes (Supplementary File 6.2, Figure 4b-1). In some teleost, e.g. stickleback and zebrafish, ZPAXB underwent species-specific duplication resulting in three and two copies, respectively, whereas in stickleback the duplicates were arranged in tandem, in zebrafish one of the duplicated ZPAXB was translocated (Supplementary File 6.2, Figure 4b). The phylogenetic positioning of ZPAX neighboring the KCNS3 and RHAG in zebrafish suggests species-specific diversification. The loss of ZPAXA in both zebrafish and cavefish could be explained by two scenarios: (1) the ZPAXA and ZPAXB appeared in the teleost ancestor and the loss happened in the branch leading to zebrafish

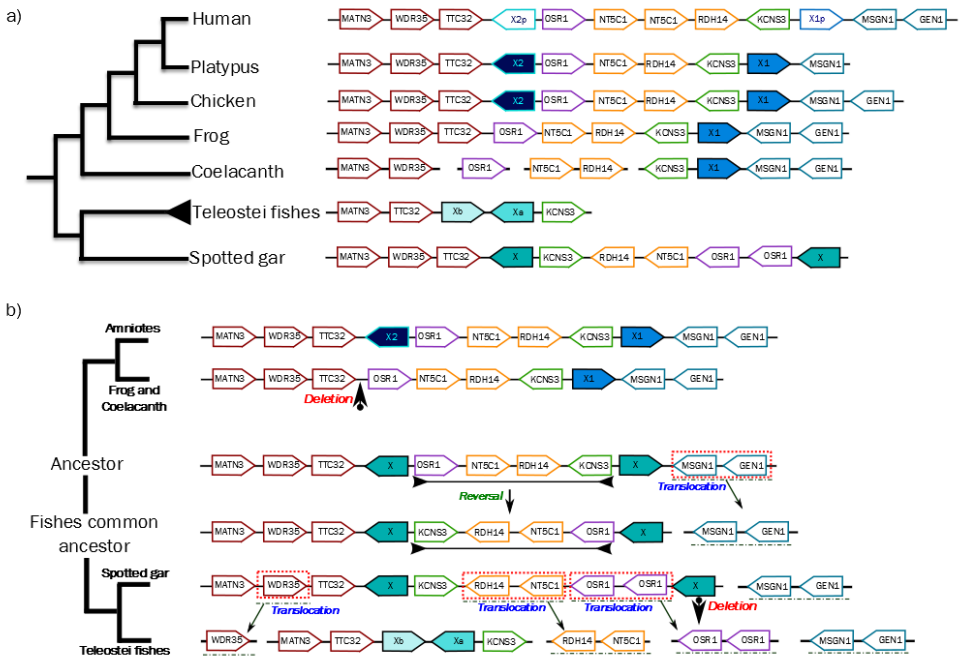


Figure 6.3: The molecular evolution and genomic rearrangement of ZPAX genes members

and cavefish as “KSNS” gene is replaced by “BON” gene in both species (Figure 2 and Supplementary File 6.2, Figure 4b); the ZPAXB gene duplicated and gave rise to ZPAXA after the divergence of zebrafish and cavefish from other teleosts. The ZPAXB gene shows intra chromosomal translocation in medaka (Supplementary File 6.2, Figure 4b).

Earlier studies have shown that ZPAX1 and ZPAX2 are pseudogenized in mammals and these pseudogenes are present in syntenic location (Hughes 2007). The functional copies of ZPAX1 and ZPAX2 are reported from platypus (Warren et al. 2008) and we also confirmed the presence of ZPAX1 and ZPAX2 in platypus together with partial copies of both these genes with conserved synteny in the *Loxodonta africana* genome . The ZPD subfamily also show conserved synteny (Supplementary File 6.2, Figure 5a). Whereas the ZP3 subfamily shows copy number variation across vertebrates and profuse variation in synteny (Supplementary File 6.2, Figure 6) especially in fishes with high copy number variation. The coelacanth ZP3 gene -ENSLACG0000007941 shows conserved synteny with reptiles and frog (Supplementary File 6.2, Figure 6c). The mammals, reptiles and birds with less gene duplicates show lineage-specific (conserved within mammals, birds, and reptiles) synteny (Supplementary File 6.2, Figure 6a, 6c and 6d).

6.3.4 The adaptive evolution

A previous study (Berlin et al. 2008) suggested the adaptive evolution of avian and mammalian ZPs, but only some genes were analyzed, namely ZP1, ZP2, ZP4 and ZPAX (with no distinction for ZPAX, ZPAX1 and ZPAX2) and with reduced species coverage (7-8 bird species and 7-10 mammals). Such study was able to reveal signatures of positive selection only in the avian ZP1 and ZP2, but not in ZP4 and ZPAX. In our study we have analyzed complete egg envelope ZP gene repertoire using in-depth positive selection analyses for avian, mammalian and reptiles ZPs (**Supplementary File 6.3)**. We also compared the dN/dS ratio among different tetrapod lineages using site, branch, branch site and clade model implemented in PAML (Yang 1997) (Yang 2007) (**Table 6.2 to 6.4)**. The use of maximum likelihood approaches implemented in HYPHY (SLAC, FEL, REL and FUBAR) (**Table 6.5**) (Delport et al. 2010) (Pond and Frost 2005a) (Murrell et al. 2013) further complemented our positive selection analyses. We also used a protein level approach implemented in TreeSAAP (Woolley et al. 2003) (**Table 6.6**) for the detection of positive radical changes in amino acid properties that can cause functional and structural protein changes.

6.3.5 Evidence of adaptive evolution in diverse vertebrate lineages using branch, branch site and clade models

Our study provided evidence that birds and reptiles have wider representation of ZP genes among vertebrates, which prompted us to compare the molecular and adaptive evolution of ZP genes in amniotes (Figure 6.4 a-c). Amniotes have common mode of internal fertilization, where mammals are considered to give birth to new ones compared to egg laying birds and reptiles. We compared omega estimate variation in different amniotes lineages using ZP1 and ZP2 genes (Figure 6.4a). ZPAX underwent frequent event of gene gain and loss resulting in the diversification of the ZPAX family, consequently we also checked the adaptive diversification of these genes in vertebrates to assess the omega variation in different lineage. For comparison of ZPAX1 and ZPAX2 we used two data sets: one with ZPAX1 and ZPAX2 (Figure 6.4b), and other with tetrapod ZPAX1, ZPAX2, ZPY together with fish specific ZPAXA and ZPAXB (Figure 6.4c). We created different partitions by dividing the dataset into four major groups for the majority of comparisons (**M** = mammals, **R** = reptiles, **B** = birds and **A** = ancestral branch connecting the sauropsida and mammals). The branch, branch site and clade model analysis was performed for comparative analysis of the evolutionary rates and the results are shown in Table 6.7 A and I.

.

Table 6.2: Results of Site Model implemented in PAML for avian ZPs

Gene	Model	Parameters	lnl	LRT test	deltaLRT	p-value	Total, ≥90, ≥95, ≥99
ZP1	M1	w0 = 0.12412 p0 = 0.71266 w1 = 1.00000 p1 = 0.28734	-19770.5579				
	M2	w0 = 0.12550 p0 = 0.70966 w1 = 1.00000 p1 = 0.28019 w2 = 3.11929 p2 = 0.01014	-19766.55766	M2a vs M1a	8.00047	0.018	**13, 1, 0, 0**
	M7	p = 0.44690 q = 1.03541	-19694.0901				
	M8	p0 = 0.97269 p = 0.50940 q = 1.34273 (p1 = 0.02731) w = 2.13842	-19678.62026	M8 vs M7	30.939682	0.000	**23, 4, 1, 0**
	M8a	p0 = 0.88517 p = 0.60097 q = 2.18796 (p1 = 0.11483) w = 1.00000	-19687.71161	M8vs M8a	18.183	0.000	
ZP2	M1	w0 = 0.14231 p0 = 0.65816 w1 = 1.00000 p1 = 0.34184	-22889.18584				
	M2	w0 = 0.14423 p0 = 0.64645 w1 = 1.00000 p1 = 0.32636 w2 = 2.50011 p2 = 0.02719	-22860.44281	M2a vs M1a	57.48606	0.000	**17, 6, 6, 2**
	M7	p = 0.37325 q = 0.59752	-22841.0				

	M8	p0 = 0.94490 p = 0.44286 q = 0.84115 (p1 = 0.05510) w = 1.90966	-22800.46136	M8 vs M7	81.029582	0.000	**26, 9, 6, 4**
	M8a	p0 = 0.78207 p = 0.63408 q = 2.32054	-22827.82507	M8vs M8a	54.727	0.000	
ZP3A	M1	w0 = 0.12247 p0 = 0.76740 w1 = 1.00000 p1 = 0.23260	-11036.34791				
	M2	w0 = 0.12785 p0 = 0.75121 w1 = 1.00000 p1 = 0.20381 w2 = 2.62967 p2 = 0.04498	-11002.17166	M2a vs M1a	68.352494	0.000	**14, 13, 11, 8**
	M7	p = 0.34895 q = 0.82794	-10986.21149				
	M8	p0 = 0.92605 p = 0.55156 q = 2.04725 (p1 = 0.07395) w = 1.91548	-10933.48228	M8 vs M7	105.458424	0.000	**14, 9, 8, 6**
	M8a	p0 = 0.87941 p = 0.65510 q = 3.36561 (p1 = 0.12059) w = 1.00000	-10964.90375	M8vs M8a	62.843	0.000	
ZP3B	M1	w0 = 0.09241 p0 = 0.76481 w1 = 1.00000 p1 = 0.23519	-10661.46819				
	M2	w0 = 0.09241 p0 = 0.76481 w1 = 1.00000 p1 = 0.18059 w2 = 1.00000 p2 = 0.05460	-10661.46819	M2a vs M1a	0	1.000	
	M7	p = 0.33573 q = 1.13061	-10580.1833				

	M8	p0 = 0.95024 p = 0.41998 q = 1.94259 (p1 = 0.04976) w = 1.20249	-10574.72083	M8 vs M7	10.924946	0.004	
	M8a	p0 = 0.92548 p = 0.44394 q = 2.32667 (p1 = 0.07452) w = 1.00000	-10575.4875	M8vs M8a	1.533	0.200	
ZP3C	M1	p: 0.93608 0.06392 w: 0.03904 1.00000	-7312.193267				
	M2	p: 0.93608 0.05552 0.00840 w: 0.03904 1.00000 1.00000	-7312.193267	M2a vs M1a	0	1.000	
	M7	p = 0.25472 q = 2.79001	-7246.465642				
	M8	p0 = 0.98165 p = 0.32092 q = 4.96885 (p1 = 0.01835) w = 1.03561	-7235.047968	M8 vs M7	22.83	0.000	
	M8a	p0 = 0.98101 p = 0.32232 q = 5.02588 (p1 = 0.01899) w = 1.00000	-7235.057535	M8vs M8a	0.019	0.890	
ZP4	M1	w0 = 0.14130 p0 = 0.64270 w1 = 1.00000 p1 = 0.35730	-13515.35788				
	M2	w0 = 0.14203 p0 = 0.63444 w1 = 1.00000 p1 = 0.35464 w2 = 4.30259 p2 = 0.01093	-13497.41244	M2a vs M1a	35.890872	0.000	**7, 5, 5, 3**

	M7	p = 0.41626 q = 0.62476	-13511.54469				
	M8	p0 = 0.98785 p = 0.44293 q = 0.68511 (p1 = 0.01215) w = 3.83672	-13490.18542	M8 vs M7	42.71854	0.000	**9, 6, 5, 3**
	M8a	p0 = 0.74753 p = 0.84308 q = 3.16082 (p1 = 0.25247) w = 1.00000	-13503.70899	M8vs M8a	27.047	0.000	
ZPX1	M1	w0 = 0.17065 p0 = 0.68824 w1 = 1.00000 p1 = 0.31176	-29612.86816				
	M2	w0 = 0.17477 p0 = 0.67827 w1 = 1.00000 p1 = 0.29937 w2 = 2.79445 p2 = 0.02236	-29571.0	M2a vs M1a	83.715614	0.000	**20, 11, 7, 5**
	M7	p = 0.54887 q = 0.89160	-29590.36187				
	M8	p0 = 0.96386 p = 0.66110 q = 1.22550 (p1 = 0.03614) w = 2.14678	-29528.90666	M8 vs M7	122.910418	0.000	**21, 16, 13, 7**
	M8a	p0 = 0.80986 p = 1.03618 q = 3.39749 (p1 = 0.19014) w = 1.00000	-29560.13715	M8vs M8a	62.461	0.000	
ZPX2	M1	w0 = 0.11648 p0 = 0.80971 w1 = 1.00000 p1 = 0.19029	-23574.41283				
	M2	w0 = 0.12019 p0 = 0.80459 w1 =	-	M2a vs	64.65	0.000	**16, 11, 7,**

		1.00000 p1 = 0.17560 w2 = 2.57296 p2 = 0.01981	23542.08503	M1a			**5**
	M7	p = 0.33122 q = 0.91870	-23533.33714				
	M8	p0 = 0.95646 p = 0.47579 q = 1.79943 (p1 = 0.04354) w = 1.87741	-23467.60892	M8 vs M7	131.5	0.000	**25, 15, 13, 7**
	M8a	p0 = 0.89156 p = 0.62936 q = 3.44126 (p1 = 0.10844) w = 1.00000	-23499.73102	M8vs M8a	64.244	0.000	
ZPD	M1	w0 = 0.14250 p0 = 0.68223 w1 = 1.00000 p1 = 0.31777	-15917.44617				
	M2	w0 = 0.14574 p0 = 0.67055 w1 = 1.00000 p1 = 0.30265 w2 = 2.45544 p2 = 0.02680	-15897.65737	M2a vs M1a	39.577588	0.000	**9, 6, 5, 2**
	M7	p = 0.49104 q = 0.90145	-15871.41886				
	M8	p0 = 0.93433 p = 0.64148 q = 1.54560 (p1 = 0.06567) w = 1.65768	-15839.39517	M8 vs M7	64.04739	0.000	**21, 9, 7, 2**
	M8a	p0 = 0.83974 p = 0.80369 q = 2.86885 (p1 = 0.16026) w = 1.00000	-15854.95352	M8vs M8a	31.116698	0.000	

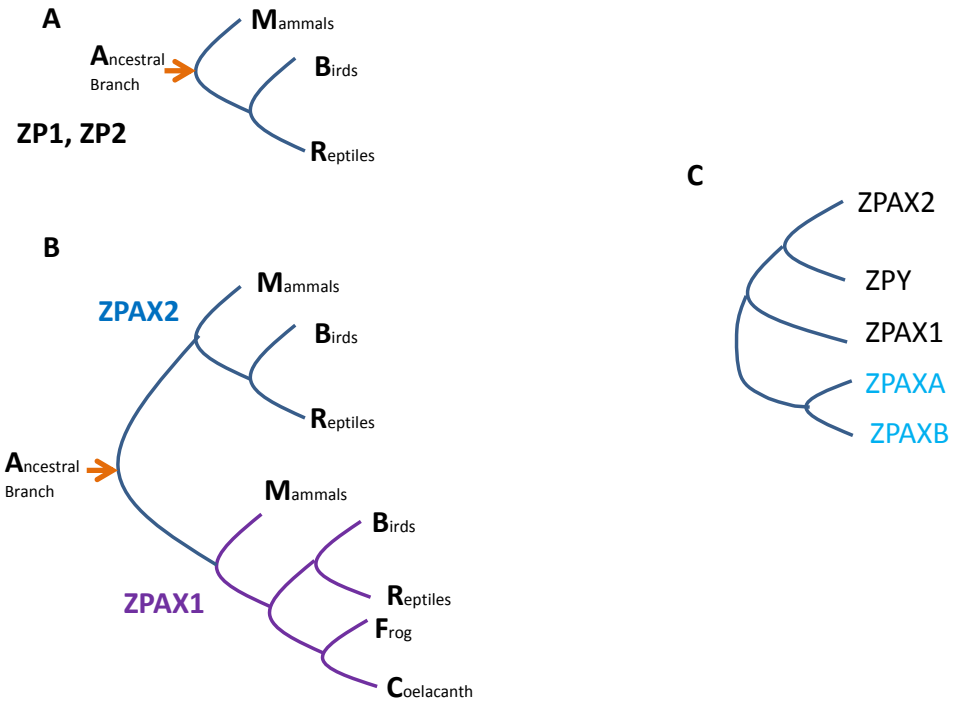


Figure 6.4: The different hypothesis tested to see dN/dS variation across different linages using branch, branch site and clade model for ZP genes

The branch model gives the dN/dS estimate for the selected branches of interest and help in understanding the variation along the branches. The branch specific omega for ZP1 gene for the four partitions show that dN/dS ratio was highest for mammals (w = 0.31) and LRTs comparison with three different null model suggest that significant difference in omega is present between mammals and birds, and bird and reptiles (LRT P <0.001). No significant difference was found between mammals and reptiles, and mammals and common ancestor Table 6.7a. The branch specific omega for ZP2 gene for the same four partitions show that dN/dS ratio was highest for mammals (w = 0.51) and LRTs comparison with three different null model suggest that significant difference in omega is present between all groups (LRT P <0.05) Table 6.7b. The comparison of ZPAX family in vertebrate (Figure 6.4c and Table 6.7) revealed highest omega in ancestral branch followed by ZPAXB in fishes, the LRT comparison also revealed that omega estimate is significantly different between ZPY and ZPAX2 genes Table 6.7c with

Table 6.3: Results of Site Model implemented in PAML for Mammalian ZPs

Gene	Model	Parameters	lnl	LRT test	deltaLRT	p-value	Total, ≥90, ≥95, ≥99
ZP1	M1	w0 = 0.14302 p0 = 0.60229 w1 = 1.00000 p1 = 0.39771	-20197.74358				
	M2	w0 = 0.14245 p0 = 0.58871 w1 = 1.00000 p1 = 0.39463 w2 = 3.00608 p2 = 0.01666	-20183.25957	M2a vs M1a	28.968016	0.000	**11, 6, 5, 2**
	M7	p = 0.52255 q = 0.92610	-20061.05825				
	M8	p0 = 0.97767 p = 0.56535 q = 1.08429 (p1 = 0.02233) w = 2.21975	-20043.72629	M8 vs M7	34.663922	0.000	**15, 6 ,4, 0**
	M8a	p0 = 0.87733 p = 0.65138 q = 1.73235 (p1 = 0.12267) w = 1.00000	-20056.20026	M8 vs M8a	24.947956	0.000	
ZP2	M1	w0 = 0.19631 p0 = 0.57129 w1 = 1.00000 p1 = 0.42871	-27349.42959				
	M2	w0 = 0.20199 p0 = 0.53194 w1 = 1.00000 p1 = 0.40622 w2 = 3.25182 p2 = 0.06184	-27239.91183	M2a vs M1a	219.03552	0.000	**47, 30, 24, 16**
	M7	p = 0.54137 q = 0.56031	-27316.95574				
	M8	p0 = 0.92475 p = 0.63904 q = 0.73572 (p1 = 0.07525) w = 2.69773	-27191.39876	M8 vs M7	251.11396	0.000	**49, 28, 24, 16**
	M8a	p0 = 0.70289 p = 1.01142 q = 2.66720 (p1 = 0.29711) w = 1.00000	-27293.52695	M8 vs M8a	204.25638	0.000	
ZP3	M1	w0 = 0.09377 p0 = 0.66830 w1 = 1.00000 p1 = 0.33170	-9262.528916				
	M2	w0 = 0.09640 p0 = 0.65515 w1 = 1.00000 p1 = 0.30629 w2 = 3.27939 p2 = 0.03856	-9242.314882	M2a vs M1a	40.428068	0.000	**14, 6, 4, 2**
	M7	p = 0.28102 q = 0.53177	-9249.595258				
	M8	p0 = 0.93071 p = 0.34822 q = 0.84156 (p1 = 0.06929) w = 2.35852	-9221.187117	M8 vs M7	56.816282	0.000	**18, 10 ,6 3**

	M8a	p0 = 0.75801 p = 0.54293 q = 3.10108 (p1 = 0.24199) w = 1.00000	-9240.163739	M8 vs M8a	37.953244	0.000	
ZP4	M1	w0 = 0.16331 p0 = 0.58255 w1 = 1.00000 p1 = 0.41745	-15316.59534				
	M2	w0 = 0.16331 p0 = 0.58255 w1 = 1.00000 p1 = 0.37863 w2 = 1.00000 p2 = 0.03882	-15316.59534	M2a vs M1a	0	1.000	
	M7	p = 0.55427 q = 0.72606	-15292.4916				
	M8	p0 = 0.89949 p = 0.67137 q = 1.16144 (p1 = 0.10051) w = 1.26487	-15287.61157	M8 vs M7	9.760058	0.008	
	M8a	p0 = 0.77771 p = 0.78582 q = 1.93799 (p1 = 0.22229) w = 1.00000	-15288.85957	M8 vs M8a	2.496004	0.114	

Table 6.4: Results of Site Model implemented in PAML for reptilian ZPs

Gene	Model	Parameters	lnl	Model comparison	LRT (2Δl)	*p*-value
ZP1	M1	w0 = 0.08963 p0 = 0.56862 w1 = 1.00000 p1 = 0.43138	-9864.099322			
	M2	w0 = 0.12681 p0 = 0.58906 w1 = 1.00000 p1 = 0.22503 w2 = 204.86958 p2 = 0.18591	-9841.43007	M2a vs. M1a	45.338504	0.000
	M7	p = 0.27911 q = 0.32632	-9877.092207			
	M8	p0 = 0.79060 p = 0.58349 q = 1.30399 (p1 = 0.20940) w = 186.89275	-9844.157621	M8 vs M7	65.869172	0.000
	M8a	p0 = 0.57099 p = 10.07853 q = 99.00000 (p1 = 0.42901) w = 1.00000	-9864.370143	M8 vs M8a	40.425044	0.000
ZP2	M1	w0 = 0.12509 p0 = 0.77357 w1 = 1.00000 p1 = 0.22643	-6438.807928			
	M2	w0 = 0.12692 p0 = 0.77727 w1 = 1.00000 p1 = 0.21822 w2 = 4.05224 p2 = 0.00451	-6438.765446	M1 vs. M2	0.084964	0.958
	M7	p = 0.50956 q = 1.24285	-6438.538528			
	M8	p0 = 0.96312 p = 0.70135 q = 2.11886 (p1 = 0.03688) w = 2.91100	-6436.412831	M7 vs. M8	4.251394	0.119
	M8a	p0 = 0.84559 p = 1.00494 q = 4.61444 (p1 = 0.15441) w = 1.00000	-6437.549996	M8a vs. M8	2.27433	0.132

Table 6.5: The results of different Maximum likelihood approaches under M8 model and SLAC, FEL, REL and FUBAR approaches

Gene	PAML	No of Species	HYPHY				
	M8	No of Species	SLAC	FEL	REL	FUBAR	Common and significant sites
ZP1	***11, 12***, 27, 35, 127, 205, 210, ***246***, 248, **249**, 358, 449, 475, **582, *591*** (15/**6**/3)	26	287, 296, **524, 591** (4/**2**/4)	34, **110**, 248, **287, 296,** 345, 431, **524, 591** (9/**5**/5)	248, **524, 591** (3/**2**/3)	**11,** 524, **591** (3/**2**/3)	***11, 12, 246,*** 248, **249, 287, 296, 524,** 582, ***591*** (10/**9**/6)
ZP2	3, 13, **17**, 31, 39, ***42,*** 56, 61, **127**, 129, ***130,*** 156, 166, **175,** 176, ***177, 178, 179, 180, 181, 200***, 204, 214, ***306,*** **344, *347,*** 446, 646, ***674, 677***, 678, ***679,*** 680, ***682, 683***, ***684, 685, 686, 687***, **688**, ***689, 691, 692, 693,*** **694, *695,*** 697, 707 (49/**30/** ***16*** / 36)	30	**42,** 56, 69, 81, 125, 154, 176, **180, 193,** 214, 231, 248, **263,** 646, 677 (15/**4**/15)	17, 23, 31, **42, 46, 56, 59, 67, 69, 81,** 118, **125**, 131, **154**, 162, **175,** 176, 177, **180, 193, 231**, 246, **248, 263, 305**, 436, 518, 632, **646, 677** (30/**18**/20)	**3, 13, 17, 31, 39, 42, 56, 61, 67,** 81, 100, **121,** 127, 129, **130, 132, 133,** 147, **156, 166**, 174, **175, 176, 177, 180, 181, 200,** 204, **214, 295, 306, 344, 347, 446, 503, 551,** 606, **640, 646**, 673, **674, 677, 680, 682, 686, 688, 693, 694, 695, 707, 717,** 737, **743** (53/**43**/38)	**42, 175,** 177, **180,** 263, **677** (6/**4**/6)	**3, 13, 17, 31, 39**, ***42,*** **46, 56,** 59, **61, 67, 69, 81**, **121, 125, 127,** 129, ***130,*** **132, 133, 154, 156, 166, 175, 176, *177, 178, 179, 180, 181***, **193**, ***200,*** 204, **214, 231, 248, 263, 295, 305, *306*, 344, *347*, 446, 503, 551**, 640, **646, *674, 677***, ***679***, **680, *682, 683***, ***684, 685, 686, 687***, **688**, ***689, 691***, ***692***, **693**, **694, *695***, **707, 717, 743** (67/**65**/44)
ZP3	2, 5, 7, ***14***, 17, ***26, 27, 28, 31***, **32**, 34, 84, **195,** 320, 342, **372, 374**, ***392*** (18/**10**/3)	18	82 (1/0/1)	24, **28**, 77, **82, 84,** 310, 323, 336, 340, **344**, 347, 372 (12/**4**/6)	**84,** 340, 344, **372** (4/**2**/4)	**28**,84 (2/**1**/2)	***14, 26***, **27, *28***, ***31, 32,*** **82, 84**, **195,** 340, **344, 372, 374, 392** (14/**13**/6)
ZP4	5, 6, 12, 20, 21, 46, 50, 79, 120, 333, 407, 464 (12/0/2)		134 (1/0/1)	**6**, 21, **73,** 84, **116, 134,** 247, 327, 344, **471**, 472, **505** (12/**6**/4)	134, 505 (2/0/2)	**6**, 21, 134 (3/1/3)	**6**, 21, **73**, **116, 134**, **471, 505** (7/**6**/4)
Chicken							
ZP1	138,149,233,**240**,292, 340,348,353,364,365, 396,**397**,399,405,**401**, 408,412,436,439, 448,461,741,***825*** (23/**4**/4)	**16**	3, 108, **365**, 366, **399, 661** (6/**3**/6)	3,16,60,100,**108,** 110,**156,209,292,** 319,**365,366,399,** 440,461,562,644, **661** (18/**8**/7)	268,288,**292,** 366,**399**,661 (6/**2**/4)	365,**399**,661 (3/**1**/3)	**3, 108, 156, 209, 240, 292, 365, 366**,**397**, **399,** 401, **461, 661, *825*** (14/1**4**/7)
ZP2	***26***,**54**,**118**,***132***,157, 158,159,***223***,229, 245,288,297,321, ***330***,***343***,365,379, 422,465,467,609, ***633***,647,**651**,673, 689 (26/**9**/9)	41	54, **132, 158**, 189, **254**, 266, **321**,343, 423, 429, **452, 633**, 647 (13/**6**/13)	25,28,**54,67**,78, 97,**100**,**132,146**, 148,158,166,**189, 254**,266,291,313, **321**,323,**343,423, 429,452,465**,624, **633,647,651**,663, 670,671 (31/**16**/18)	**54**,114,120,130, **132**,158,**166**,254, 266,297,298,**321, 323,343**,407,**423, 452**,458,480,624, 633,**647**,651 (23/**9**/16)	54, 132, 158, **254, 321, 343**, 452, 465, **633** (9/**4**/9)	***26,*** **54, 118**, ***132***, **158, 166, 189**, ***223***, **254, 266, 297, 321, 323**, ***330, 343***, **423, 429, 452, 465, 624**, ***633***, **647, 651** (23/**23**/19)
ZP3a	**89**,***91,93,95***, 98,***99***,103,***114***, 133,***168***,171, 199,***231,327***,(14/**9**/3)	41	**132**,185,288 (3/**1**/3)	61,131,**132**,133, 172,185,288,303 (8/**1**/5)	92,95,98,131, 132,133,185, 196,321, (9/**0**/6)	**132**, 133 (2/**1**/2)	**89**, ***91, 93***, ***95,*** **98**, ***99***, ***114,*** 131, **132**, 133, ***168,*** 185, ***231,*** 288, ***327*** (15/**11**/7)

ZP4	**4,*10***,12,***83,*** ***95***,173,456, ***502***,***540*** (10/**6**/4)	31	4,**25**,28,32, 83,217,**229**, 336,**436**,483, 540 (11/**3**/9)	2,**4,25**,28,**32**, 73,85,126,**217, 229,272,333**,336, **436**,469,**483** (16/**9**/8)	**4,9,10,25,28**, 37,40,**83,126, 216,217**,229, 238,**436,483**, 510,**512,540** (18/**13**/11)	**4**,25,436 (3/**1**/3)	**4,** 9, ***10***, **25, 28**, ***83***, ***95,*** **126, 216**, **217, 229, 436**, **483, *502,* 512**, ***540*** (16/**16**/11)
ZPD	7,***27***,29,***30,32,*** ***34,*****44**,61,76, 108,**167**,168, ***234,238***,261, ***270,***279,302, 370,384,405 (21/**9**/5)	44	**20, 34, 41**, 46, **108**, 159, 238, **342, 384** (9/**6**/8)	8, **20**, 22, **34**, 41, 46,108, 159, 238, **342, 351**, 384 (12/**4**/7)	**5,7,11,17,20,21, 22,34,36,**38,39, **40,41**,46,48,65, **66,82**,86,**108**,126, **129,134,151**,157, **159**,165,184,196, **209**,223,224,230, 238,240,271,**276, 289**,197,331,**342,** 345,**349,351,366, 367,371,372,374,** 379,394,**406,407** (53/**31**/9)	**20, 34**, 57, 108, 342, 384 (6/**3**/5)	**5, 7, 11, 17, 20, 21, 22**, ***27***, ***30, 32***, ***34*****,36, 40, 41, 44, 66, 82, 108, 129, 134, 151, 159, 167, 209**, ***234***, ***238***, ***270***, **276, 289, 342, 349**, **351, 366, 367, 371, 372, 374, 384, 406, 407** (40/**40**/11)
ZPAX1	***55*****,88**,183,***207***, **227,**241,***276***,***306***, ***309***,353,392,404, ***433*****,444,*****582,596***, ***652,662,699,723***, ***771*** (21/**16**/11)	45	79,**189**,190, 244,281, **306**,340,**378,582, 590,**689,**771** (12/**6**/11)	24,55,79,167, **189,190**,221, **244**,280,281, **304,306**,309, **340,378**,395, 409,467,**500,** 528,544,554, 557,**582,590, 689,**766,**771** (28/**12**/15)	**55,65,88,137,**189, **227,276,306,309, 404,419,420,**433, **444,473,554,582, 596,681**,757,**766, 771** (22/**19**/14)	189,**306**,378, **582**,590,771 (6/**2**/6)	***55,*** **65**, **88, 137, 189, 190,** ***207, 227***, **244, *276*, 281**, **304**, ***306, 309*****, 340, 378, 404, 419, 420**, ***433*****, 444, 473, 500, 554, *582,* 590**, ***596, 652***, ***662***, **681, 689,** ***699***, ***723*****, 766,** ***771*** (35/**35**/23)
ZPAX2	***34***,64,98,***227***,***229***, 293,***301***,317,***321***, **411,468**,472,476, 496,***682***,683,***711***, 719,***734,754,766,*** ***768,783,784***,795 (25/**15**/13)	46	**34, 64, 98, 227,** 229, **301, 321, 388, 411**, 500, **595**, 684, **700, 711, 734**, 754 (16/**12**/15)	**34,64,98,227,229, 264,301,321**,334, **388**,398,411,487, **500**,502,**550**,561, **593,595**,610,612, **700,711,734**,739, **754** (26/**17**/20)	**34,64,98,227,229, 264,293,301,317, 321,388,411,468, 472,496**,500,**502**, 511,534,**550,558, 561,595,610,612, 700,711,731,734, 737,754,761,766,** 768, (34/**30**/23)	**34, 64, 98, 227, 229, 301, 321**, 388,411,502, 595,**711,734 754** (14/**10**/13)	***34***, **64, 98**, ***227***, ***229,*** **264,** ***301, 321***, **388**, **411, 468**, 500, **502**, **550, 558, 561, 595, 610, 612,** ***682***, **700,** ***711***, **731**, ***734***, **737**, ***754,*** **761**, ***766***, ***768,*** ***783,*** ***784*** (31/**30**/23)

The total number of BEB inferred PSC under PAML M8 site model are categorized based on posterior probability, with PSC having PP ≥90 shown in bold, PSC with PP ≥95 in italics and PP≥99 as shaded in grey.The total PSC detected under with significance cutoff of 0.1 for SLAC, 0.1 for FEL, Bayes Factor 50 for REL and posterior probability ≥ 0.9 FUBAR. The PSC in both have higher significance cutoff of 0.05 for SLAC, 0.05 for FEL, Bayes Factor 80 for REL and posterior probability ≥ 0.95 FUBAR. PSC detected by two or more methods are underlined* Common PSC (underlined) and the PSC with higher significance (bold)

Table 6.6: The TreeSAAP results showing the positive radical changes in amino acid properties

			PAML		*TreeSAAP properties*						
	Gene	*Site*	M2a	M8	*Total*		*Chemical*		*Structural*		*Other*
Mammals	**ZP1**	11 Y	2.664 ± 0.432*	1.546 ± 0.263*	4	4	*H, H_{nc}, R_a, R_F*	0	-	0	-
		12 P	**2.672 ± 0.418**	1.547 ± 0.262*	6	4	*H_{nc}, p, pH_i, R_F*	2	*H_c, K^0*	0	-
		246 I	**2.671 ± 0.420**	1.548 ± 0.260*	2	2	*p, pH_i*	0	-	0	-
		591 T	2.647 ± 0.464*	1.536 ± 0.276*	3	3	*H_p, H_{nc}, pH_i*	0	-	0	-
	ZP2	130 G	**3.483 ± 0.204**	**2.836 ± 0.503**	4	3	*H, pH_i, R_F,*	1	*K^0*	0	-
		177 G	**3.491 ± 0.145**	**2.842 ± 0.491**	9	5	*H, H_p, P_r, R_a, R_F*	4	*B_i, K^0, M_v, V^0,*	0	-
		178 T	**3.498 ± 0.053**	**2.849 ± 0.479**	5	4	*H_p, pH_i, P_r, Ra*	1	*K^0*	0	-
		179 K	**3.497 ± 0.072**	**2.848 ± 0.480**	3	2	*pH_i, R_F*	1	*K^0*	0	-
		180 V	**3.482 ± 0.209**	**2.835 ± 0.502**	7	5	*H_p, P_r, pH_i, R_a, R_F*	2	*B_i, V^0*	0	-
		347 P	**3.481 ± 0.211**	**2.834 ± 0.506**	8	5	*H, H_{nc}, p, P_r, R_F*	3	*B_i, M_v, V^0*	0	-
		674 S	**3.495 ± 0.101**	**2.845 ± 0.487**	4	3	*H_p, pH_i, R_a*	1	*H_c*	0	-
		677 L	**3.498 ± 0.047**	**2.849 ± 0.478**	6	6	*H, H_p, p, P_r, R_a, R_F*	0	-	0	-
		682 S	**3.498 ± 0.052**	**2.849 ± 0.478**	6	6	*H, H_p, p, P_r, R_a, R_F*	0	-	0	-
		683 S	**3.498 ± 0.047**	**2.849 ± 0.479**	11	6	*H_p, H_t, μ, pH_i, R_a, R_F*	4	*B_i, H_c, M_v, V^0*	1	*M_w*
		685 E	**3.499 ± 0.036**	**2.850 ± 0.477**	4	3	*H, pH_i, R_F*	1	*K^0*	0	-
		686 K	**3.498 ± 0.045**	**2.849 ± 0.478**	7	6	*H, H_p, H_t, pH_i, R_a, R_F*	1	*K^0*	0	-
		687 S	**3.499 ± 0.039**	**2.849 ± 0.477**	11	6	*H_p, μ, pH_i, P_r, R_a, R_F*	4	*B_i, H_c, M_v, V^0*	1	*M_w*
		689 S	**3.497 ± 0.073**	**2.847 ± 0.481**	8	3	*μ, pH_i, R_a*	4	*B_i, H_c, M_v, V^0*	1	*M_w*
		691 T	**3.491 ± 0.143**	**2.841 ± 0.494**	10	4	*μ, pH_i, R_F, R_a*	5	*B_i, H_c, K^0, M_v, V^0*	1	*M_w*
		693 E	**3.488 ± 0.169**	**2.838 ± 0.500**	13	7	*H, H_p, H_t, μ, pH_i, R_a, R_F*	5	*B_i, H_c, K^0, M_v, V^0*	1	*M_w*
	ZP3	26 W	**2.522 ± 0.254**	**3.209 ± 0.577**	1	1	*p*	0	-	0	-
		27 L	**2.534 ± 0.214**	**3.235 ± 0.526**	6	6	*H, H_{nc}, p, pH_i, R_a, R_F*	0	-	0	-
		28 L	**2.530 ± 0.227**	**3.230 ± 0.537**	9	4	*μ, pH_i, P_r, R_F*	4	*B_i, H_c, M_v, V^0*	1	*M_w*

									-		
Birds	**ZP1**	825 A	2.397 ± 0.519	1.487 ± 0.154*	2	2	P_r, R_a	0	-	0	-
	ZP2	223M	**2.501 ± 0.025**	**2.311 ± 0.392**	2	2	H_p, R_a	0	-	0	-
		330A	2.485 ± 0.156*	**2.306 ±0.440**	0	0	-	0	-	0	-
		343T	2.482 ± 0.170*	**2.304 ± 0.403**	3	3	H_p, P_r, R_a	0	-	0	-
		633Q	**2.500 ± 0.029**	**2.311 ± 0.392**	3	3	H_{nc}, R_a, R_F	0	-	0	-
	ZP3a	91 T	**2.498 ± 0.044**	**2.508 ± 0.092**	0	0	-	0	-	0	-
		93 Q	**2.498 ± 0.044**	**2.508 ± 0.092**	6	6	H, H_{nc}, H_t, p, pH_i, R_F	0	-	0	-
		95 A	**2.495 ± 0.083**	**2.496 ± 0.162**	6	4	Hc, p, pH_i, P_r	1	V^0	1	M_w
		114 R	**2.497 ± 0.060**	**2.505 ± 0.115**	0	0	-	0	-	0	-
		168 A	**2.495 ± 0.080**	**2.498 ± 0.154**	0	0	-	0	-	0	-
		231 P	**2.495 ± 0.078**	**2.496 ± 0.161**	4	3	H_p, pH_i, R_a	1	H_c	0	-
	ZP4	10 V	**3.809 ± 0.742**	**2.437 ± 0.466**	1	1	P_r	0	-	0	-
		95 S	**3.815 ± 0.731**	**2.44 ± 0.462**	9	4	H_t, μ, R_a, R_F	4	B_l, H_c, M_v, V^0	0	M_w
		540 W	**3.821 ± 0.719**	**2.444 ± 0.454**	3	2	H_t, pH_i	1	K^0	0	-
	ZPD	27 V	2.487 ± 0.167*	**1.515 ± 0.151**	1	1	P_r	0	-	0	-
		34 G	**2.503 ± 0.069**	**1.518 ± 0.145**	2	2	pH_i, R_F	0	-	0	-
	ZPX1	55 H	**2.505 ± 0.191**	**2.492 ± 0.097**	4	4	H, H_{nc}, R_a, R_F	0	-	0	-
		276 S	**2.514 ± 0.149**	**2.494 ± 0.08**	0	0	-	0	-	0	-
		582 L	2.458 ± 0.319*	**2.479 ± 0.17**	2	2	p, P_r	0	-	0	-
		652 V	**2.517 ± 0.136**	**2.495 ± 0.074**	9	4	H_t, μ, R_a, R_F	4	B_l, H_c, M_v, V^0	1	M_w
		662 H	**2.517 ± 0.132**	**2.495 ± 0.073**	2	1	pH_i	1	K^0	0	-
		699 L	**2.517 ± 0.135**	**2.495 ± 0.074**	9	4	H_t, μ, R_a, R_F	4	B_l, H_c, M_v, V^0	1	M_w
		771 R	2.456 ± 0.324*	**2.48 ± 0.169**	1	1	p, pH_i	0	K^0	0	-
	ZPX2	229 V	**2.498 ± 0.055**	**1.846 ± 0.479**	0	0	-	0	-	0	-
		301 Q	**2.500 ± 0.015**	**1.848 ± 0.477**	2	2	H, pH_i	0	-	0	-
		711 I	**2.498 ± 0.056**	**1.847 ± 0.477**	0	0	-	0	-	0	-
		734 H	**2.500 ± 0.022**	**1.848 ± 0.477**	3	2	p, pH_i	1	K^0	0	-
		754 M	**2.488 ± 0.136**	**1.845 ± 0.479**	2	2	p, P_r	0	-	0	-
		768 T	2.465 ± 0.228*	**1.841 ± 0.484**	2	2	p, P_r	0	-	0	-
		784 T	2.484 ± 0.155*	**1.844 ± 0.486**	2	2	H_p, R_a	0	-	0	-

The sites with highest PP>0.99 in M8 are shown, with exception of ZP1 for which sites with PP>95 in M8 are shown due to lack of any site with PP>0.99. The sites with P>0.99 are in bold, sites with PP>0.95 are not bold and are marked with asterisk and site with PP>90 are not bold and without asterisk. TreeSAAP analysis results present the total number of radical changes in amino acid properties and their assigned categories. Type I sites are underlined. The Properties symbols representation is as given: **B_l**: Bulkiness; **H_p**: Surrounding hydrophobicity; **H_t**: Thermodynamic transfer hydrophobicity; **H_c**: Helical contact area; **H**: Hydropathy; **H_{nc}**: Normal consensus hydrophobicity; **K^0**: Compressibility; **M_v**: Molecular volume; **M_w**: Molecular weight; **μ**: Refractive index; **pH_i**: Isoelectric point; **p**: Polarity; **P_r**: Polar requirement; **R_a**: Solvent accessible reduction ratio; **R_F**: Chromatographic index; **V^0**: Partial specific volume.

The branch model dN/dS estimates for the selected branches of interest allowed to assess the variation along the branches. The branch specific omega for ZP1 gene for the four partitions showed that dN/dS ratio was highest for mammals (ω = 0.31) and LRTs comparison with three different null model suggest that significant difference in the omega is present between mammals and birds, and birds and reptiles (LRT P <0.001). No significant difference was found between mammals and reptiles, and mammals and common ancestor (**Table** 6.**7a**). The branch specific omega for ZP2 gene for the same four partitions shows that dN/dS ratio was highest for mammals (ω = 0.51) and LRTs comparison with three different null models suggest that significant difference in omega is present between all groups (LRT P <0.05) (Table 6.7b).

The comparison of ZPAX family in vertebrate (Figure 6.4c and Table 6.7) revealed highest omega in ancestral branch followed by ZPAXB in fishes, the LRT comparison also revealed that omega estimate is significantly different between ZPY and ZPAX2 genes (Table 6.7c) with ZPAX2 having higher omega (ω = 0.30) compared to ZPY (ω = 0.20). On the other hand the dataset comprising ZPAX1 and ZPAX2 genes (Figure 6.4b) was used for testing multiple hypotheses: 1- The difference in omega estimate between ZPAX1 and ZPAX2; **2-** If ZPAX1 evolve at different rates amniotes and anamniotes; **3-** If ZPAX1 evolve at different rates in birds and reptiles; 4**-** If ZPAX2 evolve at different rates in birds and reptiles. The difference in omega estimate between ZPAX1 and ZPAX2 is not significant. The ZPAX1 in anamniotes is evolving at slower rates than amniotes, the omega estimate was found significant higher in birds (ω =0.38) than reptiles for ZPAX1 (Table 6.7d), whereas no significant difference was observed between birds and reptiles for ZPAX2 (Table 6.7d).

The clade and branch site model allows variable omega between both the site and branches by prior division of the dataset into foreground and background lineages. The multi-clade model analysis was performed using the partitions (mammals, birds, reptiles and ancestral branch) (Figure 6.4 a-c) for ZP1, ZP2 and ZPAX genes. The results for ZP1 shows that the omega varies considerably between birds and mammals (p <0.05), reptiles and mammals (p <0.05), but difference between birds and reptiles was not significant (Table 6.7e). In ZP2, clade model does not find significant difference between mammals and birds but significant difference was observed between birds and mammals, and birds and reptiles (p<0.05) (Table 6.7f). The comparison of ZPAX1 and ZPAX2 partitions (Figure 6.4b) resulted in almost similar trend as branch models. The comparison between amniotes and frog for ZPAX1

revealed no significant difference in the omega estimate ($p<0.09$) (Table 6.7g). We used branch site model to detect signals of positive selection in ancestral branch connecting the sauropsida and mammalian lineages (Table 6.7 h and i). The results for ZP1 and ZP2 genes support the increased omega along the selected ancestral branch ($p < 0.05$) together with BEB inferred PSCs (Table 6.7h and 7i).

6.3.6 Random site model suggest rapid evolution of ZP genes within avian and mammalian lineages.

The results of random-sites analyses for the ZP1, ZP2, ZP3, ZP3a, ZP3b, ZP3c, ZP4, ZPAX1, ZPAX2 and ZPD gamete recognition genes from mammals and birds are shown in Table 6.2 - 6.6. The overall results support the important role of adaptive evolution in gamete recognition proteins. Birds have wider distribution of ZPs (ZP1, ZP2, ZP3, ZP4, ZPAX and ZPD) than mammals (ZP1-ZP4).

The comparisons of overall dN/dS ratio (ω = omega) and number of BEB inferred positive selected codon (PSC) show that among avian ZPs, ZP4 have highest omega (M8; ω = 3.84, n=9 PSC) and ZPD have the least (M8; ω =1.66, n = 21 PSC). Whereas highest number of BEB inferred PSC are present in ZP2 (M8; ω =1.91 n = 26 PSC) with ZP4 having the least (M8; ω =3.84, n = 9 PSC). The omega value for the 7 avian ZP genes found under positive selection ranged between 1.66 and 3.84 (ZP1 = 2.14, ZP2 = 1.91, ZP3a = 1.92, ZP4 = 3.8, ZPAX1= 2.15, ZPAX2 =1.88 and ZPD = 1.66) with an average of 2.21, the number of PSC range between 9 and 26 (ZP1 = 23, ZP2 = 26, ZP3a = 14, ZP4 = 9, ZPAX1 = 21, ZPAX2 = 25 and ZPD = 21) with an average of 20 PSC. The results also show that among four mammalian ZPs, three are evolving under positive selection (ZP1; ω = 2.22, n =15 PSC, ZP2; ω =2.70, n=49 PSC, and ZP3 ω =2.36, n = 18 PSC under M8 site model with ZP2 having the highest omega and maximum number of PSC (n = 49 PSC). The comparison of omega values and BEB inferred PSC, between mammalian and avian ZPs revealed different genes to have highest omega, i.e. ZP4 in birds and ZP2 in mammals. The high number of PSC in ZP2 in birds and mammals suggests possible involvement in gamete interaction. All avian ZPs with the exception of ZP3b and ZP3c are found to evolve under positive selection.

The use of multiple approaches (SLAC, FEL, REL and FUBAR) implemented in HYPHY package was used to compliment our previous findings (Table 6.5). These results are also supportive of the overall rapid evolution of ZP genes. These methods use different algorithms resulting in different number of PSC (for example SLAC and FUBAR being detected comparative less number of sites as compared to FEL and

REL). The grouping of the PSCs detected by two or more methods and also PSCs having high significance can further provide reliable estimates of PSCs (Table 6.5). Similar to PAML results, HYPHY also found maximum PSC in mammalian ZP2. Overall, we found 67 PSCs having higher significance and /or were common, with 65 sites found at higher significance and 44 common sites. In birds minimum PSCs are found for ZP3a (n = 14) and ZP4 (n = 10) and rest of the ZP genes under positive selection had almost same number n = 21-26 PSCs. The collective approach complemented the trend observed in the PAML results, with exception of ZP1, which showed deviation from the trend with fewer sites (Table 6.5). Like PAML site models minimum number of sites was found in ZP3a and ZP4 whereas number of PSCs in remaining ZP except ZP1 was around 31-53 (ZP2 = 31 sites by FEL; ZPD = 53 sites by REL).

The site model results were further complemented by employing amino acid based approach implemented in the TreeSAAP software to determine the positive radical changes in amino acid properties for PSCs with higher PP>0.99 (except ZP1 for which sites PP>0.95 were checked). The TreeSAAP analysis was able to find both type I and type II sites, which could result in functional and structural diversification of proteins (Table 6.6). All the mammalian PSCs showed positive radical changes in structural, functional and other properties with sites belonging to both type I and type II categories. ZP2 have maximum number of positive selected sites belonging to type I category 11 out of 16 having six or more changes. The total number of changes in ZP2 ranged from 3-11. Among birds most of the sites with higher PP had positive radical changes in amino acid properties except a few (e.g. site 229V and 711I in ZPAX2 do not showed any positive radical change). Out of 32 PSC in birds (Table 6.7), 25 were found any with radical changes and out which 5 sites belonged to type I category, with 2 sites each in ZP3a and ZPX1, one site in ZP4. The number of positive radical changes ranged from 1-9.

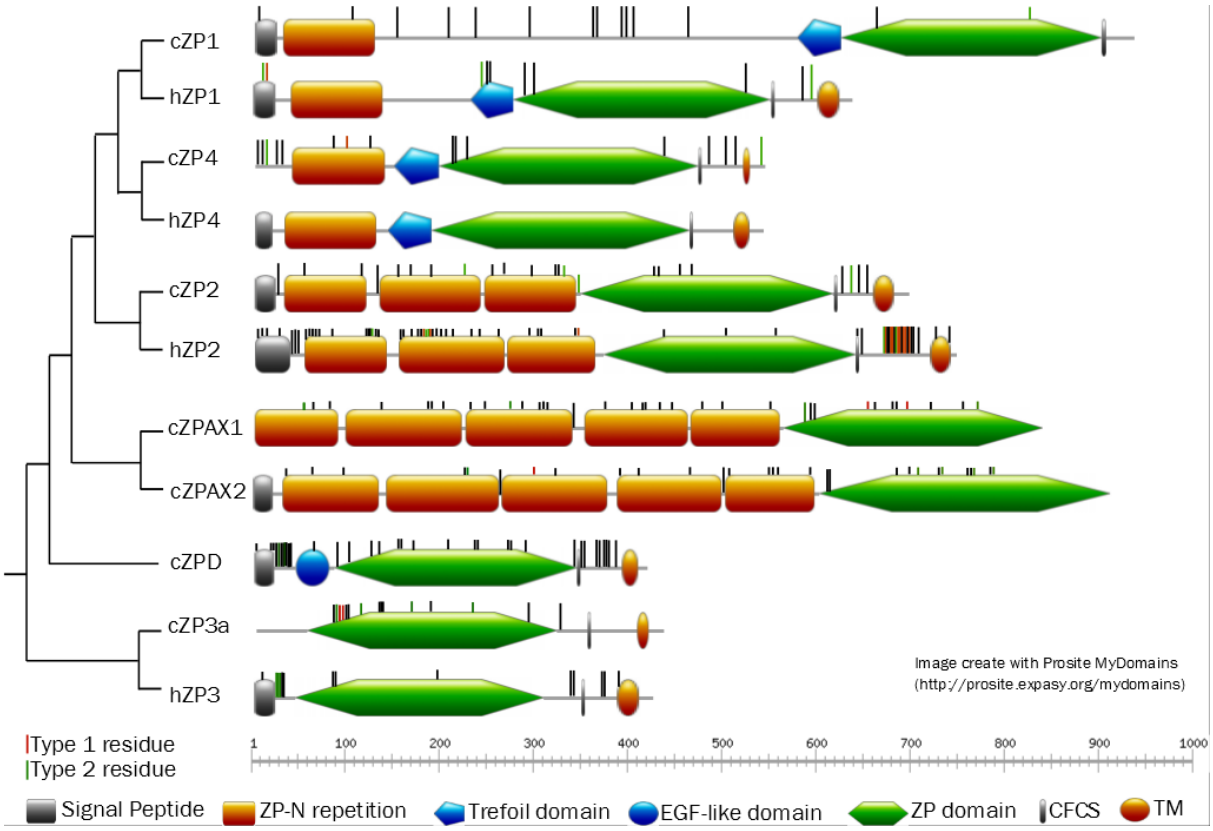


Figure 6.5: The domain architecture of ZP gene members showing the location of positively selected sites

6.3.7 ZP proteins architecture and Homology modelling

The comparison of primary, secondary and tertiary structure of ZP gene family members was done in order to infer the important regions of the ZP proteins that are influenced by positive selection found in our study and its possible role in functional and structural modifications of the proteins (Table 6.8, Figure 6.9, Supplementary File 6.3 to 6.15). The ZP proteins usually have a N-terminal signal sequence, a ZP domain, a trefoil domain, C-terminal propeptide that is cleaved by consensus furin cleavage site, and a transmembrane domain (Figure 6.5) (Han et al. 2010) (Plaza et al. 2010) (Claw and Swanson 2012). The ZP-domain is 260 amino acids long, with conserved Cys residues that define a ZP-domain signature. All ZP proteins, with exceptions of ZP1/ZPB, have a trans membrane domain that ensures the correct cleavage of the precursors (Williams and Wassarman 2001) (Jovine et al. 2005). Consensus Furin Cleavage Site (CFCS) is found in ZPD, ZP1 and ZPC/ZP3, the ZP4/ZPB proteins also have a trefoil/P-domain cells (Okumura et al. 2004) (Smith et al. 2005) that is a group of six conserved cysteines and other amino acids, in a trefoil arrangement may be important for carbohydrate binding and/or defense against proteolic degradation (Bork 1993) (Bork et al. 1996). The ZPD have an EGF domain (Goudet et al. 2008) which is a small domain of 30-40 amino acids with six conserved cysteines that can bind directly

receptors or mediate interactions via Ca2+-binding (Bork et al. 1996) (Wouters et al. 2005).

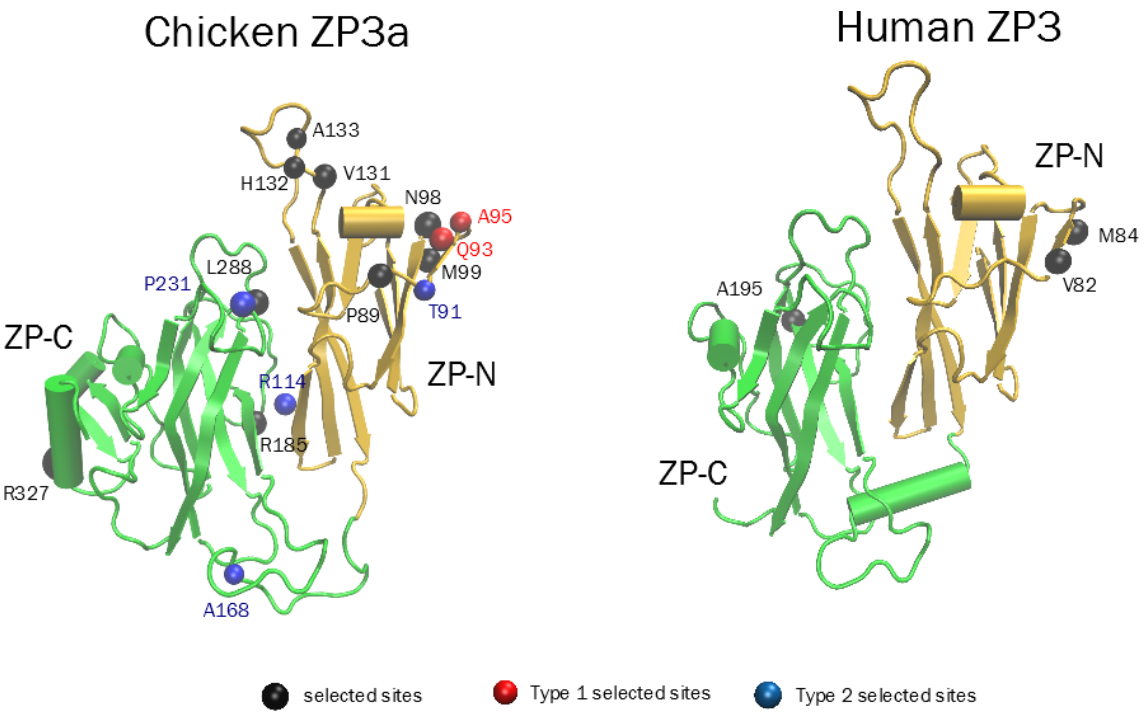


Figure 6.6: The homology modelling showing positive selected sites detected in ZP3 gene of mammals and birds. The chicken 3D4C PDB file was used for modelling

The ZP domain consist of two subdomains, ZP-N present at the N-terminal region of ZP domain and ZP-C at the C-terminus separated by a protease sensitive linker, each of them can exist as independent subdomains within proteins (Jovine et al. 2004) (Claw and Swanson 2012). ZP-N is thought to constitute a basic building block of ZP filaments ZP-C may mediate the specificity of interaction between subunits (Monné et al. 2008) (Sasanami et al. 2006)(Kanai et al. 2008)(Han et al. 2010) regulated by the external hydrophobic patch (EHP) inside the C-terminal and an internal hydrophobic patch (IHP) inside the ZP (Jovine et al. 2004). The ZP domain plays an important role in polymerization of ZPs. The multiple copies of ZP-N subdomains are present in some ZPs e.g. of ZP2 (Claw and Swanson 2012) and we also found many copies of ZP-N subdomains in ZPAX1 and ZPAX2 (Figure 6.5) with ZP1 and ZP4/ZPB have one ZP-N domain repetition and ZP2/ZPA and ZPAX have three and six ZP-N domain repetitions, respectively. The ZP-N domain acts like an independent structural domain and the number of repetitions could be also related with the specific biologic function of each

ZPs (Callebaut et al. 2007). We found many positive selected sites located in the ZP-N repeats and this could possibly be related with coevolution and polymerization of ZPs (Figure 6.5). The ZP-domain is responsible for polymerization, the formation of egg coat requires ZP3 (type I subunit) and at least one type II (ZP1/ZP2/ZP4-like) component (Jovine et al. 2005). In mammals, homodimers of ZP1 crosslink the filaments formed by ZP2/ZP3 (Claw and Swanson 2012) (Han et al. 2010) (Plaza et al. 2010). The 3D structure of ZP3 chicken (Han et al. 2010) shows biogenesis of ZP3 requires processing of N terminal signal peptide, formation of six intramolecular disulfide bonds and loss of C terminal propeptide that contains a polymerization blocking EHP and TM. This results in interaction between ZPN subdomain of one subunit of ZP3 with ZP3 subdomain of second subunit resulting in a homo dimer (Han et al. 2010).

The overall location of positive selected sites detected in our study show that most sites are present in the ZP-N repeats and the C terminal region of the proteins, especially in the ZP domain (Table 6.8, Figure 6.5 and Figure 6.6). The C terminal hyper variable region of ZP3 gene is known to be associated with species-specific variations and specificity of egg coat assembly (Han et al. 2010). The ZP-C subdomain is known to mediate the interaction between type I and type II ZP subunits (Han et al. 2010) (Okumura et al. 2007) (Monné et al. 2008) (Sasanami et al. 2006) and because different ZP3 disulphite connectivities are presented by cognate type II ZPs (Okumura et al. 2007) this suggest the tertiary structure of ZP-C subdomain of ZP3 determines the specificity of egg coat assembly (Figure 6.6) (Han et al. 2010).

Similarly the proximity of conserved O-glycan important for sperm binding and the hyper variable, positively selected C-terminal region of ZP3 (Figure 9 and Figure 10) possibly have a concerted role in the regulation of species-specific gamete recognition (Han et al. 2010) (Wassarman and Litscher 2008) (Swanson et al. 2001). The positive selected site found could increase the conformational flexibility of the C-terminal region especially of ZP3 and could clearly provide opportunities for protein-based recognition. The structural considerations also suggest how the sperm recognition function of ZP3 might have arisen during evolution as a specialization of its polymerization activity (Han et al. 2010) and this could also be explained with the coevolution of other ZP involved in the matrix formation (Table 6.8, Figure 6.5 and 6.6).

Table 6.7 a-i: The ZP model comparisons a). The Branch model tests for ZP1 from birds, reptiles, mammals and ancestral branch

Model	Parameter estimates	lnL	Model Comparisions	2 ΔlnL	P valve
M, R, B, **A**	w0 A = 0.20596;w1 R= 0.28343;w2 B= 0.17274;w3 M= 0.30870	-28047.07906			
M&A, R, B	**w0 MA = 0.30609**;w1 R = 0.27936;w2 B = 0.17285	-28048.22966	**M**, R, B, **A** vs **MA**, R, B	2.3012	0.12927409
M&R, B, A	w0 A = 0.20216;w1 B = 0.17241;w2 **MR= 0.30481**	-28047.56296	**M, R**, B, A vs **MR,** B, A	0.9678	0.325229466
M&B, R, A	w0 **MB=0.28741** w1 R=0.27921 w2 A=0.21262	-28065.38255	**M**, R, **B**, A vs **MB**, R, A	36.606976	1.44514E-09
B&R, M,A	w0 A= 0.22106;w1 M= 0.30863;w2 **BR = 0.22733**	-28055.23587	M, **R, B**, A vs **BR,** M,A	16.31362	5.36768E-05

b). The Branch model tests for ZP2 from birds, reptiles, mammals and ancestral branch

Model	Parameter estimates	lnL	Model Comparisions	2 ΔlnL	P valve
M, R, B, **A**	**w0 A= 0.25462;w1 B= 0.33766;w2R = 0.26582;w3 M = 0.51357**	-37573.91277			
M&A, R, B	**w0 MA = 0.50120;w1B = 0.33417;w2R = 0.26407**	-37582.25159	**M**, R, B, **A** vs **MA**, R, B	16.67765	4.42999E-05
M&R, B, A	**w0 A= 0.25200;w1 B = 0.32772;w2 MR= 0.46844**	-37609.03269	**M, R**, B, A vs **MR,** B, A	70.239852	5.25147E-17
M&B, R, A	**w0 A= 0.24840;w1 R = 0.25938;w2 MB= 0.48085**	-37590.15681	**M**, R, **B**, A vs **MB**, R, A	32.488092	1.19925E-08
B&R, M,A	**w0 A= 0.25546;w1 BR = 0.30264;w2 M= 0.51347**	-37576.67263	M, **R, B**, A vs **BR,** M,A	5.51973	0.018803161

c). Branch Model dN/dS variation between ZPAX2 and ZPY

Model	Parameter estimates	lnL	Model Comparisions	2 ΔlnL	P valve
A, ZPAXB, ZPAXA, ZPAX1, ZPAX2, ZPY	w0A = 0.58128;w1ZPAXB = 0.27361;w2 ZPAXA= 0.32575;w3 ZPAX1= 0.27424;**w4 ZPAX2= 0.29457;w5 ZPY= 0.19592**	-63117.39422	A, ZPAXB, ZPAXA, ZPAX1, ZPAX2, ZPY **vs**. A, ZPAXB, ZPAXA, ZPAX1, ZPAX2 & ZPY	12.076162	0.000511
A, ZPAXB, ZPAXA, ZPAX1, **ZPAX2 & ZPY**	w0A = 0.59081;w1ZPAXB = 0.27329;w2 ZPAXA = 0.32590;w3 ZPAX1 = 0.27433;**w4 ZPAX2 & ZPY = 0.27740**	-63123.4323			

d). Branch Model dN/dS variation in frog, reptilian and bird lineage					
	Parameter estimates	**lnL**	**Model Comparisions**	**2 ΔlnL**	**P valve**
ZPAX1 frog and tetrapods	w0A = 0.35237;w1X2 = 0.29305;w2 X1= 0.28740;w3 FROG = 0.18533	-37429.56621			
	w0 A= 0.35444;w1X2 = 0.29306;w2 X1 & FROG= 0.26850	-37436.82977	A, ZPAX2, ZPAX1, FROG **vs.** A, ZPAX2, **ZPAX1 & FROG**	14.52712	0.000138
ZPAX1 reptiles and birds	w0A = 0.33658;w1X2 = 0.29346;w2X1 = 0.22987;w3 R= 0.14800;w4 B= 0.38394	-37407.07039			
	w0 A= 0.35755;w1X2 = 0.29311;w2 X1= 0.22226;w3 B & R = 0.30472	-37430.12314	A, ZPAX2, ZPAX1, ZPAX1_reptile, ZPAX1_birds **vs.** A, ZPAX2, ZPAX1, **ZPAX1_reptile & ZPAX1_birds**	46.1055	1.12E-11
ZPAX2 reptiles and birds	w0 A= 0.34925; X1= 0.26842; X2= 0.24331;w3B = 0.32657;w4R = 0.28925	-37433.63613			
	w0 A= 0.34173;w2 X1= 0.26836; w1X2 = 0.24059;w3BR = 0.31204	-37434.16449	A, ZPAX1, ZPAX2, ZPAX2_reptile, ZPAX2_birds **vs.** A, ZPAX1, ZPAX2, **ZPAX2_reptile & ZPAX2_birds**	1.05672	0.303964

e). ZP1 multi clade comparisons for birds, reptiles, mammals and ancestral branch									
Model	**Parameter estimates**					**lnL**	**Model Comparisions**	**2 ΔlnL**	**P valve**
	wo	**po**	**p1**	**Omega**	**Prop.**				
M, R, B, A	0.04116	0.31739	0.19808	**M** = 0.35905 **R** = 0.28100 **B** = 0.25440 **A** = 0.22000	0.48453	-27250.46969			
M&A, R, B	0.04225	0.32107	0.19391	**MA = 0.35621 R =** 0.28187 **B** = 0.25594	0.48503	-27252.03363	**M, R, B, A vs MA, R, B**	**3.127868**	**0.076964332**
M&R, B, A	0.04132	0.31772	0.19749	**A** = 0.22039 **B** = 0.25435 **MR** = **0.34546**	0.48479	-27252.96539	**M, R, B, A vs MR , B ,A**	**4.991398**	**0.025473621**
M&B, R, A	0.03898	0.30709	0.20857	**MB = 0.33368 R =** 0.27305 A = 0.21811	0.48435	-27254.50273	**M, R, B, A vs MB, R, A**	**8.066074**	**0.004510173**
B&R, M, A	0.04073	0.3154	0.20024	**A** = 0.21919 **BR** = 0.26743 **M** = **0.35717**	0.48435	-27250.67655	**M, R, B, A vs BR, M, A**	**0.413722**	**0.520086385**
M1a	0.1615	0.64105	0.35895		-	-27468.94701			
M2a_rel	0.32412	0.48437	0.204	0.0403	0.31163	-27257.25704			

f). ZP2 multi clade comparisons for birds, reptiles, mammals and ancestral branch									
Model	**Parameter estimates**					**lnL**	**Model Comparisions**	**2 ΔlnL**	**P valve**
	wo	**po**	**p1**	**Omega**	**Prop.**				
M, R, B, A	0.05126	0.23812	0.26727	M = 0.34485 B = 0.37306 R = 0.27144 A = 0.42748	0.49461	-36801.93708			
M&A, R, B	0.05153	0.23808	0.26863	MA = 0.42072 B = 0.37399 A = 0.27297	0.49328	-36802.59672	M, R, B, A vs MA, R, B	1.319278	0.250722
M&R, B, A	0.05167	0.2368	0.2769	A = 0.35956 B = 0.38068 MR = 0.39059	0.4863	-36810.11871	M, R, B, A vs MR, B, A	16.363254	5.23E-05
M&B, R, A	0.05023	0.23541	0.27088	A = 0.34693 R = 0.27270 MB = 0.41362	0.49371	-36802.80523	M, R, B, A vs MB, R, A	1.736296	0.187609
B&R, M, A	0.05278	0.24097	0.26691	A= 0.34802 BR = 0.32265 M = 0.42883	0.49212	-36804.72224	M, R, B, A vs BR, B, A	5.570308	0.018268
M1a	0.20936	0.5523	0.4477			-36970.92277			
M2a_rel	0.38741	0.48554	0.27785	0.05171	0.23662	-36810.23839			

g). ZPAX multi clade comparisons for birds, reptiles, mammals and ancestral branch									
Model	**Parameter estimates**					**lnL**	**Model Comparisions**	**2 ΔlnL**	**P valve**
	wo	**po**	**p1**	**Omega**	**Prop.**				
A, ZPAX2, ZPAX1	0.06398	0.27649	0.1428	**A** = 0.63777 **X2** = 0.31620 **X1** = **0.33124**	0.58071	-36786.85719			
A, ZPAX2& ZPAX1	0.06373	0.2761	0.14253	**X1X2** = 0.32498/A0.63828	0.58137	-36787.03129	A, ZPAX2, ZPAX1 vs. A, ZPAX2& ZPAX1	**0.348188**	**0.555141**
A, ZPAX2, ZPAX1, ZPAX1_FROG	0.06285	0.27527	0.1421	A = 0.63694 X2 = 0.31735 X1 = 0.34240 F = 0.28207	0.58263	-36785.48752			
A, ZPAX2, ZPAX1 & ZPAX1_FROG	0.06398	0.27649	0.1428	**A** = 0.63777 **X2** = 0.31620 **X1** = **0.33124**	0.58071	-36786.85719	A, ZPAX2, ZPAX1, ZPAX1_FROG vs. A, ZPAX2, ZPAX1 & ZPAX1_FROG	**2.73934**	**0.097905**
A, ZPAX2, ZPAX1, ZPAX1_BIRD_REPTILES	0.0625	0.2747	0.14225	A = 0.64648 X2 = 0.31799 X1 = 0.29434 X1BR= 0.36035	0.58305	-36784.56624			
A, ZPAX2, ZPAX1 & ZPAX1_BIRD_REPTILES	0.06398	0.27649	0.1428	**A** = 0.63777 **X2** = 0.31620 **X1** = **0.33124**	0.58071	-36786.85719	A, ZPAX2, ZPAX1, ZPAX1_BIRD_REPTILES vs. A, ZPAX2, ZPAX1 & ZPAX1_BIRD_REPTILES	**4.581898**	**0.032311**
A, ZPAX1, ZPAX2, ZPAX2_BIRD_REPTILES	0.06399	0.2765	0.1428	A = 0.63780 X1 = 0.33125 X2 0.31634 X2BR = 0.31615	0.5807	-36786.85718			
A, ZPAX1, ZPAX2 & ZPAX2_BIRD_REPTILES	0.06398	0.27649	0.1428	**A** = 0.63777 X1 = 0.33124 **X2** = 0.31620	0.58071	-36786.85719	A, ZPAX1, ZPAX2, ZPAX2_BIRD_R	**2.2E-05**	**0.996258**

							EPTILES vs .A, ZPAX1, ZPAX2 & ZPAX2_BIRD_R EPTILES		
M1a	0.21039	0.69859	0.30141			-37016.28779			
M2a_rel	0.32755	0.5855	0.146	0.0588	0.2685	-36792.54474			

h). ZP1 Branch Site Model for ancestral branch

	(SC0) fg w	(SC0) p	(SC1) fg w	(SC1) p	(SC2a) fg w	(SC2a) p	(SC2b) p	**Lnl**	**LRT (2Δl)**	***p*-value**	BEB sites
Alternate	0.16089	0.61721	1	0.34455	62.10378	0.02454	0.0137	-27466			88P 0.573,251S 0.616,601P 0.806,603L 0.518
Null	0.1615	0.64105	1	0.35895	1	0	0	-27468.9	5.860746	0.015482	

i). ZP2 Branch Site Model for ancestral branch

	(SC0) fg w	(SC0) p	(SC1) fg w	(SC1) p	(SC2a) fg w	(SC2a) p	(SC2b) p	**Lnl**	**LRT (2Δl)**	***p*-value**	BEB sites
Alternate	0.20706	0.52124	1	0.42093	33.37513	0.03199	0.02584	-36960.7			135 Q 0.614, 139 P 0.850, 236 H 0.962*, 256 S 0.918, 326 S 0.919, 362 S 0.841, 383 Y 0.644, 386 Q 0.641, 412 Q 0.715, 539 K 0.679, 590 Y 0.518, 633 T 0.517, 667 S 0.690, 684 A 0.506
Null	0.20684	0.50692	1	0.40883	1	0.04664	0.03761	-36969.1	16.74496	4.28E-05	

Table 6.8: The location of positive selected sites in different region of ZP proteins.

Lineage	Gene	Signal peptide		ZP-N 1		ZP-N 2		ZP-N 3		ZP-N 4		ZP-N 5		Trefoil	EGF		ZP domain		CFCS	Propeptide		
																					TM	
Mammals	ZP1	2 (11, 12)	0	0	-	-	-	-	-	-	-	-	-	3 (246, 248, 249)	-	0	3 (287, 296, 524)	-	0	2 (582, 591)	0	0
	ZP2	4 (3, 13, 17, 31)	3 (39, 42, 46)	13 (56, 59, 61, 67, 69, 81, 121, 125, 127, 129, 130, 132, 133)	0	17 (154, 156, 166, 175, 176, 177, 178, 179, 180, 181, 193, 200, 204, 214, 231, 248, 263)	0	5 (295, 305, 306, 344, 347)	-	-	-	-	-	-	-	0	3 (446, 503, 551)	-	1 (640)	19 (646, 674, 677, 679, 680, 682, 683, 684, 685, 686, 687, 688, 689, 691, 692, 693, 694, 695, 707)	1 (717)	1 (743)
	ZP3	1 (14)	-	-	-	-	-	-	-	-	-	-	-	-	-	5 (26, 27, 28, 31, 32)	3 (82, 84, 195)	2 (340, 344)	0	2 (372, 374)	1 (392)	0
Birds	ZP1	1 (3)	0	1 (108)	-	-	-	-	-	-	-	-	10 (156, 209, 240, 292, 365, 366, 397, 399, 401, 461)	0	-	-	2 (661, 825)	-	0	0	-	-
	ZP2	0	1 (26)	2 (54, 118)	1 (132)	4 (158, 166, 189, 223)	0	6 (254, 266, 297, 321, 323, 330)	-	-	-	-	-	-	-	1 (343)	4 (423, 429, 452, 465)	-	0	4 (624, 633, 647, 651)	0	0
	ZP3a	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	14 (89, 91, 93, 95, 98, 99, 114, 131, 132, 133, 168, 185, 231, 288)	1 (327)	0	0	0	0
	ZP4	-	5 (4, 9, 10, 25, 28)	3 (83, 95, 126)	-	-	-	-	-	-	-	-	0	0	-	-	4 (216, 217, 229, 436)	-	0	3 (483, 502, 512)	0	1 (540)
	ZPD	6 (5, 7, 11, 17, 20, 21)	9 (22, 27, 30, 32, 34, 36, 40, 41, 44)	-	-	-	-	-	-	-	-	-	-	-	1 (66)	1 (82)	13 (108, 129, 134, 151, 159, 167, 209, 234, 238, 270, 276, 289, 342)	-	0	8 (349, 351, 366, 367, 371, 372, 374, 384)	2 (406, 407)	0
	ZPAX1	-	-	3 (55, 65, 88)	0	4 (137, 189, 190, 207)	0	7 (227, 244, 276, 281, 304, 306, 309)	1 (340)	6 (378, 404, 419, 420, 433, 444)	0	3 (473, 500, 554)	-	-	-	0	11 (582, 590, 596, 652, 662, 681, 689, 699, 723, 766, 771)	-	-	-	-	-
	ZPAX2	0	0	3 (34, 64, 98)	0	2 (227, 229)	1 (264)	2 (301, 321)	0	3 (388, 411, 468)	1 (500)	5 (502, 550, 558, 561, 595)	-	-	-	0	14 (610, 612, 682, 700, 711, 731, 734, 737, 754, 761, 766, 768, 783, 784)	-	-	-	-	-

The sites detected at higher significance cut off and common sites. Type 1 sites are shown in red and type II in green. Only unshaded regions represent the domain architecture, the shaded cell with dash represent that this region is not present in given ZP. The sites detected at higher significance cut off and common sites are shown as found in table above are shown. Type 1 sites are shown in red and type II in green.Only unshaded regions represent the domain architecture, the shaded cell with dash represent that this region is not present in given ZP

6.4 Conclusions

Successful embryonic development is enhanced when the zygote is protected from lethal microorganisms, or additional sperm. The animals use different mechanism to achieve this, e.g. blocking of subsequent sperm from ova. Selective pressures have resulted in an elaboration of diverse egg and sperm structures with variations in genes, e.g. we found species and lineage-specific variations in vertebrates with reduction in ZP gene family repertoire from fishes to mammals. We found comparative less ZP gene family members in amniotes compared to anmniotes, with gene gain playing important role in (sauropsida), birds and reptiles, and gene loss being more prominent feature in the mammalian lineage. The extra ZPs may play and important role in formation of ZP egg envelope matrix. The genomic scan and synteny analysis revealed species/lineage specific variations of ZP gene family with ZPAX1, ZPAX2 and ZPY in tetrapods and coeleocanth, whereas ZPAXA and ZPAXB are fish specific. The Sauropsida (birds and reptiles) have similar ZP family distribution. The ZP distribution in coelacanth shows shared features with tetrapods (ZPAX1, ZPY and ZP3-ENSLACG0000007941) and fishes ZPB, which support the evolutionary transition and proximity of coleocanth with both lineages. The finding of ZP2 in spotted gar and coeleocanth reveals that ZP2 is present in fishes and is not tetrapod specific.

The ZP1 and ZP2 showed significant omega variation among amniotes with highest omega in mammals, 0.31 and 0.51 for ZP1 and ZP2, respectively. The ZPAX2 having higher omega (ω = 0.30) compared to ZPY (ω = 0.20). The difference in omega estimate between ZPAX1 and ZPAX2 is not significant. The ZPAX1 in amniotes is evolving at faster rate than anamniotes. Random site model suggest rapid evolution of ZP genes within avian and mammalian lineages. Among birds, ZP4 have highest omega (M8; ω = 3.84, n=9 PSC) and ZPD have the least (M8; ω =1.66, n = 21 PSC), whereas among mammals ZP2 have highest (M8; ω =1.91 n = 26 PSC) with ZP4 having the least (M8; ω =3.84, n = 9 PSC). ZP1 and ZP4 genes mediate acrosome reaction through similar pathways, which could justify a possible subfunctionalization of those genes in some species. In ZP3 the positive selected sites are located in a region of the protein previously referred to interact with sperm, which might be crucial to species-specific sperm-egg interaction (Swanson et al. 2001). ZP3 proteins are widespread known as primary sperm receptor (Wassarman et al. 2004) (Goudet et al. 2008) with capacity to initiate signal transduction pathways to allow ZP penetration (Vo and Hedrick 2000), such as induction of sperm acrosome reaction (Smith et al. 2005) (Gupta et al. 2012). The presence of positive selected sites mostly located in the ZP-N repeats and ZP domain shows that these changes could be responsible for species-

specific variation leading to ZP polymerization and matrix formation. The positive selected sites detected by random site models showed positive radical changes in amino acid properties leading to functional and structural diversification of proteins with possible implications in ZP matrix formation and species-specific gamete recognition.

Many birds and reptiles show cryptic female choice which involve process like sperm ejection, sperm choice in polygynandrous mating system (Birkhead and Pizzari 2002), e.g. egg sperm interaction resulting in preference of sperm surface protein by egg coat proteins limiting the genetic contribution from non-compatible sperm (Claw and Swanson 2012) (Pizzari and Birkhead 2000) (Palumbi 1999). The post-copulatory sperm competition between sperm of different males provide important source of selection acting on improving the sperm quality, e.g. sperm mobility. The presence of rare bird hybrids supports hybridization avoidance and rapid evolution leading to reproductive barriers and speciation (Berlin et al. 2008) (Birkhead and Brillard 2007) (Birkhead 1998). The differences in reproductive goals of the sexes, create sexual conflict, which is a major force driving the co-evolutionary arms races in reproductive proteins (Edwards et al. 2005) (Gavrilets 2000) (Tregenza et al. 2000) and the variation in gene gain, gene loss and rapid evolution of egg envelope ZP proteins provide required variation for species isolation and speciation

The comparative genomics and adaptive evolution of egg envelope ZP subgenome reveals that gene gain, gene loss and positive selection play an important role in evolution of ZP subgenomes across vertebrates. The high rate of positive selection observed in monospermic mammals and physiological polyspermic birds and reptiles also points towards factors other than polyspermy responsible for driving the ZP diversification leading to speciation (Blount 1909) (Snook et al. 2011) (Elinson 1986)(Tarín 2000) (Wishart and Horrocks 2000). (Waddington et al. 1998) (Horrocks et al. 2000). (Iwao 2012). These factor may include crypic female choice, sexual conflict, sperm competition between sperm of different males provide important source of selection acting on improving the sperm quality, e.g. sperm mobility and hybridization avoidance (Swanson and Vacquier 2002). Our study suggest that the evolution of ZP proteins could be explained by the sexual conflict hypothesis, where female proteins involved in reproduction diversify rapidly and the positive selection in both monospermic and polyspermic lineage show that factors other than polyspermy avoidance are responsible, such as sensory exploitation, seminal fluid toxicity and mating-induced reduction in female lifespan (Smith et al. 2005) (Swanson et al. 2001).

6.5 Materials and Methods

6.5.1 Characterization of ZP subgenome genomic scan and synteny

We used representative sequences of ZP1-ZP4, ZPAX, ZPY and ZPD vertebrate ZPs as query and did exhaustive blast (Altschul et al. 1997) searches using intermediately stringent level of E-10 to retrieve all available ZPs from 26 vertebrate genomes (Table 6.1 Figure 6.1). Finally, the ZPs were verified by BLASTP against NCBI non-redundant protein sequence database. Adaptive evolutionary analyses of ZP genes were performed with all significant sequences retrieved from blast searches of 48 bird genomes database (Jarvis et al. 2014) (Zhang et al. 2014) (Supplementary file 16), mammalian and reptilian genomes (Supplementary additional file 17).

6.5.2 Phylogenetic analysis

The resulting orthologs were aligned using Muscle (Edgar 2004) implemented in Seaview (Gouy et al. 2010) and sequences were tested for nucleotide substitution saturation using DAMBE 5 (Xia 2013) by plotting the number of transition/transversion against the genetic distance using F84 model (Huelsenbeck and Rannala 1997), which allows for different equilibrium base frequency and transition transversion rate bias for nucleotide substitution. Xia test (Xia et al. 2003) implemented in DAMBE5 (Xia 2013) was done to compare index score (ISS) with critical score (ISS.C) for 3rd and other codon positions to estimate base saturation. The sequences alignments were tested for recombination using GARD (Genetic Algorithm for Recombination Detection) (Pond et al. 2006) available online http://www.datamonkey.org. The vertebrate ZP phylogenetic tree was made in MEGA5 software using the neighbor joining method, with 1000 bootstrap replications. For independent ZP ortholog topology assessment and positive selection, sequences were aligned using Seaview and manually corrected and the best substitution model was detected with jModelTest 2 (Darriba et al. 2012) based on the Akaike Information Criteria (AIC), followed by Maximum Likelihood tree construction using PhyML (Guindon and Gascuel 2003) with 500 bootstrap replicates to check the robustness and reliability of tree (Felsenstein 1985). The tree topology used for positive selection analysis in birds followed the species tree from Jarvis et al. (2014) (Supplementary file 4).

6.5.3 Positive selection

We performed in-depth positive selection analysis using multiple codon and protein level approaches approaches. The positive selection helps to better adapt to the environment by selecting the beneficial changes, (Positive selection $\omega>1$) whereas if the changes are disadvantages (harmful) they are not inherited and thus are removed from population (Negative selection $\omega<1$). We analyzed ZPs from mammalian, birds

and reptilian genomes (Supplementary file 16 and 17) for signals of diversifying positive selection using codon model implemented in PAML (Yang 1997) (Yang 2007) and Datamonkey (Pond and Frost 2005a) together with amino acid model in TreeSAAP (Woolley et al. 2003). We employed different approaches to find signals of positive selection in avian ZP genes. Codeml in PAML package version 4.7 (Yang 1997) implements likelihood ratio test (LRT) for comparison of sophisticated nested site specific models calculated as twice the difference of log likelihood between the two models following chi square distribution with degree of freedom corresponding to the difference in number of parameters between the nested model i.e. null model (no selection) and alternate model (positive selection). The significant LRT means null model is rejected and sites are under positive selection. We compared M1a (Nearly Neutral)vs M2a (Positive Selection), and M7 (beta) vs M8 (beta& w) to find sites under positive selection. Bayes empirical Bayes (BEB) inferred the posterior probabilities of positive selected sites where higher PP meaning high confidence. Other than PAML site models we used Hyphy package (http://www.hyphy.org) (Pond et al. 2005) (http://www.datamonkey.org/) (Pond and Frost 2005a) (Delport et al. 2010) that provides different approaches (SLAC, FEL, REL and FUBAR) for detection of positive selected sites, including Single Likelihood Ancestral Counting (SLAC), Fixed Effects Likelihood (FEL), Random Effects Likelihood (REL), (Pond and Frost 2005b) Fast Unconstrained Bayesian AppRoximation (FUBAR) (Murrell et al. 2013) and integrative approach. SLAC model uses ancestral sequences reconstruction, FEL calculates site by site dn/ds without assuming a prior distribution whereas REL assume a prior distribution across site, FUBAR ensures robustness against model misspecification. Along with this, the integrative approach results incorporate all sites detected by SLAC, FEL, REL and FUBAR. The sites detected by two different methods are further supportive of positive selection.

Further support for our results was gained by complementary protein level approach implemented in TreeSAAP (Woolley et al. 2003). It uses ancestral sequence reconstruction to find the physiochemical properties change of amino acid replacement using 31 amino acid properties. The amino acid replacement can lead to conservative or radical change in physiochemical properties. The positive radical changes can lead to change in structure and/or function of protein and the number of radical changes at a site can be used as an indicator to show strength of positive selection. To facilitate interpretation of level of changes at a site we categorized the sites into two types. Sites having six or more radical changes were defined as type I and sites with less than six properties where defined as type II.

The branch, branch site and clade model were used to estimate the dN/dS among branches and among sites and across branches both. The Figures 6.4a-c shows the different hypothesis tested. The free-ratio model (model = 1) assumes an independent dN/dS ratio for each branch. This model is very parameter-rich and its use is discouraged. The model = 2 allows you to have several dN/dS ratios for different branches (n-ratio) of interest specified by branch level e.g. in The branch models allow the dN/dS ratio to vary among branches in the phylogeny and are useful for detecting positive selection acting on particular lineages (Yang 1998) (Yang and Nielsen 1998) .For example the two ratio model estimates dN/dS ratio for the branch of interest (specified in tree by branch label) and it can be compared with null model one ratio model, and can also be used for neutrality test by comparing with model with fix omega =1.

The branch-site models (Zhang et al. 2005) allow to vary both among sites and across branches detects positive selection affecting a few sites along particular lineages (called foreground branches).The alternate model (model = 2 NSsites = 2) is compared with corresponding null model with (fix_omega = 1 and omega = 1). The posterior probabilities of positive selected sites are inferred by BEB.

Clade model C is specified by model = 3 Nssites = 2 while clade model D is specified by model = 3 NSsites = 3 using ncatG to specify the number of site classes (Bielawski and Yang 2004). The model C can be compared with the null models M1a (NearlyNeutral). M1a test for functional divergence among clades is prone to false positives under simple evolutionary conditions. The new null model M2a_rel (NSsites = 22 now specifies the site model M2a_rel) is proposed that better accounts for among-site variation in selective constraint (Weadick and B.S.W. Chang 2012). The clade model can be sued for testing multiple clade and are shown to be useful in inference of positive selection (Weadick and B.S. Chang 2012).

6.5.4 Domain architecture, Homology modelling and structure analysis

The domain architecture was based on the published literature additional file 3- 15. This was also verified using the Uniprot protein database (http://www.uniprot.org webcite) whenever possible. The know chicken ZP3, 3D structure PDB: 3D4C was used homology modeling of human ZP3 and chicken ZP3A using swiss model server http://swissmodel.expasy.org/ (Schwede et al. 2003) (Arnold et al. 2006). Positive selected residues were mapped onto the predicted structure using PyMOL (version 1.1) (http://www.pymol.org webcite).

Acknowledgements

IK was funded by a PhD grant (SFRH/BD/48518/2008) from Fundação para a Ciência e a Tecnologia (FCT). AA was partially supported by the European Regional Development Fund (ERDF) through the COMPETE - Operational Competitiveness Programme and national funds through FCT under the projects PEst-C/MAR/LA0015/2013, PTDC/AAC-AMB/104983/2008 (FCOMP-01-0124-FEDER-008610) and PTDC/AAC-AMB/121301/2010 (FCOMP-01-0124-FEDER-019490).

II. Discussion and conclusions

7

Discussion

The advent of genome sequence data is an important mile stone in evolutionary biology. Such genomic resources favored the advent of a variety of evolutionary approaches. The studies developed within this thesis applied various methodological approaches to characterize genes and gene families at the genome level, followed by detailed exploration of genetic changes at the DNA and protein level. Gene gain and gene loss are important indicators of gain or loss or function and/or functional diversification with changes in parent and/or offspring gene duplicates. Altogether, gene gain and loss, as well as changes at the sequence level can lead to increased fitness favoring adaptive evolution of species-specific changes.

In this thesis, genes and gene families involved in diverse functionalities relevant for the adaptive evolution of vertebrates were studied in detail using comparative evolutionary genomics approaches. Different genes and gene families faced varied levels of environmental pressure resulting in distinctive changes measured at the gene, gene families and genome level. The exploration of evolutionary changes ranging from genome to genes provides valuable insight to understand the evolutionary history of gene and genomes in diverse species and lineages (O'brien et al. 1983) (Malo et al. 2010) (Lande 1988) (Perry et al. 2012) (Chen et al. 2012).

Five major gene families involved in important functionalities have been selected for study (chapter 2 to chapter 6). The **chapter two** deals with KRTAP gene family involved in mammalian hair development and hair characteristic phenotypic variations. The KRTAP gene family consists of around 100 genes that are organized into 30 subfamilies with each having different number of gene members due to different rates of gene gain and gene loss varying across the different mammalian genomes. All these variations in the KRTAP gene family repertoire are further shaped by different genomic

forces like gene conversion and adaptive evolution. The adaptive evolution favors functional diversification and the origin of new subfamilies in species and lineage specific manner. By contrast, concerted evolution can lead to homogenization of gene members, which could be helpful in large scale expression of KRTAPs structural proteins involved in the hair/wool formation. The gene expansion found in hairy mammals like sloth and the loss of KRTAPs in hairless dolphin is related with their ecological adaptation, especially hair symbiotic relationships in sloth and fast swimming in dolphin.

The **chapter three** deals with Olfactory receptors (ORs), the largest multi gene family in vertebrates with more than 1000 genes present in mammals. Earlier studies have pointed out the relationship between the numbers of ORs and ecological adaptation. Here, the OR gene family repertoire was characterized in 48 avian and 2 reptilian genomes. Ecological factors especially the habit and the habitat determined the olfactory ability in birds and reptiles, expressed as the proportion of functional OR gene families. Moreover, the rapid expansion of ORs was followed by adaptive evolution point towards the ongoing natural selection and functional innovations shaping the olfactory ability in diverse avian species.

In **chapter four** the cGSTs gene family involved in detoxification was surveyed for signatures of selection, namely avian cGSTs. Differences in GST gene family repertoire, namely in the number of cGSTs members, were found in distinct lineages, e.g. cGST-P was lost in birds and reptiles. Positive selection was found in expanded members of the cGSTs classes, like cGSTA, where several sites in the proteins had radical changes in amino acid properties likely causing relevant structural and functional changes. The gene gain and gene loss, complemented with episodes of positive selection, provided overtime species and lineage specific adaptations possibly playing an important role in the protection against rapidly increasing environmental toxins and even defense against free radicals.

The chapter five describes the evolutionary analyses on the TLR gene family involved in the immune defense system of vertebrates. The comparative genomics of diverse vertebrates covering major lineages from fish to mammals allowed to unravel the differential gene gain and loss of the TLR gene family. Most of the TLR gene family members were originated early in the evolution, with most TLRs being found in basal vertebrate lineages. The TLR gene family evolution was shaped by different dynamics of gene gain and loss events, with most events of gene duplication occurring early in the fish lineage. The coelacanth TLRs showed some shared feature with both

tetrapods and fishes. TLR genes earlier thought to be restricted to certain lineages were found in new genomes studied, e.g. TLR15 earlier reported only in birds was found in reptiles. Evidence of positive selection was detected in all avian TLR genes studied (1A, 1B, 2A, 2B, 3, 4, 5, 7 and 15) and positive selected sites ranged from 5% to 11% with the omega value ranging from 1.5 to 2.5. Both viral and non-viral TLR genes were found to be under strong positive selection with the TLR4 (non-viral) and TLR7 (viral) having the highest number of positive selected sites (20 and 23 sites with high PP>0.99, respectively). Most of these positive selected sites also showed radical amino acid changes in physiochemical properties, including type I radical changes, which can possibly affect the structure and functionality of the proteins. The rapid evolution of TLR is related with the host pathogen arms race leading to coevolution of ligands and receptors. The high number of positive selected sites in non-viral TLR4 could be related with identification of diverse variety of ligands, e.g. LPS and LTA. The rapid evolution of the viral TLR7 gene suggests its adaptive role in the recognition of ssRNA (mutation rates very high due to lack of mismatch repair) and point towards the strong selective pressure imposed by the long term coexistence of viruses and birds, helping understanding the role of birds as natural reservoirs of vectors of zoonotic pathogens.

In **chapter six** comparative genomics and adaptive evolutionary analyses of the ZP gene family were performed among vertebrates, with particular emphasis in birds and mammals. We found that family ZPB/4 and ZP3 have undergone fish specific expansion. The ZPB/4 duplicated in amniotes after divergence from amphibian, originating the ZP1 and ZP4 in amniotes. ZP2 is present in the common ancestor of fishes and tetrapods, but was lost in teleosts. ZPD is tetrapod-specific with presence only in sauropsida and frog. The presence of tetrapod specific ZPY and ZPAX1 in coelacanth together with a well supported conserved synteny, points towards the evolutionary relatedness of coelacanth with the first fish that walked on land. The spotted gar genome, which is an Actinopterygian that diversified from teleosts before the teleosts whole genome duplication event is helpful in understanding the relatedness between fish and other tetrapods, e.g. the ZP2 in spotted gar showed shared synteny with tetrapods, whereas ZP2 was lost in teleosts. Genomic rearrangements were found in the evolution of ZPAX that shaped the lineage specific organization ZAPX. The, ZPAX1, ZPAX2 and ZPY members were found in tetrapods and coelacanth, whereas ZPAXA and ZPAXB were found in teleosts. The use of different complimentary methods to detect evidence of positive selection in diverse species supported the rapid evolution of ZP gene family members in both mammalian

and avian lineages together with radical changes in amino acid properties which can lead to functional and structural diversification of proteins, which could in turn be associated with ZP matrix formation and ultimately contribute to species-specific gamete recognition.

7.1 Role of gene gain and gene loss events in the evolution of gene families

Gene gain provides the important raw material for functional innovation and diversification (neo or sub functionalization), whereas gene loss could lead to loss of function (deletion or pseudogenizartion). Thus, characterization of gene family, which include the precise cataloging of gene gain and loss events is a critical step to understand the evolution of particular phenotypic variations in extant species. All known representative members of a target gene family were used as query and extensive blast searches were performed across genomes (Altschul et al. 1990a) to retrieve the complete repertoire (functional and nonfunctional genes) of the target gene family. The complete systematic characterization of gene family members is followed by the synteny analysis (Muffato et al. 2010a) (Louis et al. 2013) for confirmation of homologous relationships and characterization of ortholog and paralogs, a fundamental procedure for any evolutionary interpretation. The orthologs and paralogs were further subjected to in depth screening of changes at DNA and protein level, which allowed assessing the role of these changes in the evolution of phenotypic modifications.

The genes can evolve under different scenarios which may include concerted evolution resulting in homogenization of duplicated genes, positive selection leading to beneficial changes, purifying selection for removal of harmful changes and neutral changes increasing the polymorphism in population without changing the fitness. In concerted evolution, gene conversion and homologous recombination leads to the homogenization of duplicated genes (often arranged in tandem) that could lead to increased expression levels, as observed in the evolution of the KRTAP gene family. Moreover, most of the expanded subfamilies were found to be under positive selection e.g. KRTAP and OR gene families underwent rapid expansion, and functional divergence within these paralogs which leads to phenotypic modification as supported by the statistical analyses based on codon and proteins based approaches. The species and lineage specific gene gain and loss can change the functional output, e.g. proteins that undergo polymerization and show combinatorial complexity, the gene gain and/or loss could result in different outcome with changes in the polymerization process as seen in the KRTAP, TLR, ZP, cGSTs and OR gene family.

7.2 Role of DNA and protein level changes across sites and lineages

The use of highly sophisticated and complementary methods, provide good support for the findings. Sophisticated models and methods have been used for the dN/dS estimation to ensure accuracy of the comparative evolutionary genomics studies to characterize the genetic variation in genes and proteins along sites, across lineages (branch) and across braches and sites (branch site and clade). We estimated the dN/dS estimates within and between lineages (Yang 2007a) (Woolley et al. 2003) (Pond and Frost 2005a) (Murrell et al. 2013) (Pond et al. 2005). We found different level of signals of high dN/dS>1 positive selection in all gene families studied. In chapter 2 and 3 we found positive selection in rapidly expanded families of KRTAP and OR in species and lineage specific manner. Different families were found expanded and under positive selection in different lineages, e.g. OR 14 in birds and OR51 in reptiles, showing the different ecological requirements. In the TLR gene family involved in immune response, both viral and non-viral TLRs were found to be under strong positive selection and higher number of positive selected sites were found in viral TLRs suggesting the role of birds as viral vectors and reservoirs, which possibly resulted in host pathogen arms race and the coevolution of TLRs due to pathogen load. Comparison between mammals and birds show similar patterns of evolution with most TLRs having higher dN/dS estimates. Birds and reptiles are physiologically polyspermic whereas mammals are monospermic and comparison of dN/dS estimates across birds, reptiles and mammalian lineages suggested that physiologically polyspermy is not responsible of the rapid evolution of these reproductive proteins and other factors than polyspermy are involved in the rapid evolution of ZPs in birds. All those findings are strongly strengthened by the use of various complementary methods.

8

Conclusions

The overall findings of this thesis suggest that gene gain, gene loss, followed by changes in DNA and proteins play an important role in phenotypic modification and adaptation for better fitness and survival of an organism, often leading to speciation and ultimately contributing to biodiversity. In the first chapter, a general introduction was provided about the background and materials and methods for the comparative evolutionary genomics of gene and gene families leading to phenotypic modification and adaptation.

The subsequent chapters 2 to 6 provided an in depth exploration of five different gene families across varied species and lineages of vertebrates. Various approaches at the genome, gene and protein level were used for the comparative evolutionary genomics study of five gene families. The use of complementary methods provided strong support of the findings obtained.

The collective findings demonstrate the importance of rapidly increasing genome resources in studying the comparative genomics of genes, and most importantly the large and complex gene families, which would be impossible to perform without the availability of whole genome sequences. The usefulness of our study becomes much more prominent due to the varied range of vertebrate species analyzed covering almost all major vertebrate groups. The results from our study are very encouraging and show the whole genome sequencing data are helpful in understanding evolution of gene families (chapter 2-6).

The use of comparative genomics approach for gene family exploration showed that all these gene families (chapter 2-6) follow asymmetric evolution with different rates of

gene gain and gene loss at both species and lineage level. The changes in gene family are possible results of evolutionary pressure leading to useful phenotypic modification for better adaptive capacity of an organism, e.g. expanded repertoire of KRTAP gene family (chapter two) in hairy sloth is adaptive feature for hosting the symbionts, whereas gene loss in hairless dolphin is adaptive feature developed for fast frictionless swimming. The birds have reduced repertoire of functional olfactory receptors and thus birds possibly have comparatively reduced olfactory. The reduced sense of smell in birds is possibly supplemented by innovation on better developed vision system, which is also supported by the use of color display as mate choice together with lack of gene and gene families responsible for pheromone detection (vemero nasal receptors). We found that OR gene family OR14 in birds and OR51, OR52 in turtle are also under positive selection, which shows ongoing natural selection and adaptive innovation in gene families possibly is helpful for specialized odor detection or increased range of odor as per the changing environment. The functional OR gene families also determined the ecological adaptation in birds. OR gene families between birds and reptiles suggest the differential requirements of OR gene families. The cGSTs are important phase II metabolic detoxification isozymes and difference in cGSTs repertoire in vertebrates together with different rates dN/dS estimates in different cGSTs suggest towards the differential role of these genes in protection against the various xenobiotic to which an organism is exposed. The cGSTs are important for protection against xenobiotic and the presence of extra copies of cGSTs, e.g. GSTA in birds and lack of cGSTP in birds together with conserved cGSTs across vertebrates, shows both differential and conserved requirement of cGSTs. The positive selection in some cGSTs possibly supports the protective role of these genes against wide and rapidly increasing variety of environmental toxins. On other hand cGSTs like sigma do not show signals of positive selection possibly due to their highly conserved role in prostaglandin production. The TLR gene family diversity across vertebrates was related with changes in immune system, e.g. coelacanth lack IgM gene, but have tetrapod specific TLRs. Secondly, viral and non viral TLRs, both evolve under strong positive selection pointing towards coevolution and the host-pathogen arms race. The rapid evolution of viral TLRs in birds can possibly also explain their role as viral reservoirs and vectors. The ZP subgenome also varied across vertebrates, expanding in frog, birds and reptiles with the presence of extra copies of ZPs (ZPAX, ZPY and ZPD). Most of the ZPs are found under positive selection and this trend is common in all vertebrate lineages (monospermic mammals and polyspermic birds and reptiles), which also supports that factors other than positive selection are responsible for polyspermy avoidance

8.1. Future directions

The present study expands the horizon of genome analyses from gene to complete gene families, which is very useful to address the complete scenario of events as evolution of genes do not happen in isolation and there is dynamics of gene gain and gene loss together with functional divergence. Our study will help in expanding to future work on other species and lineages, e.g. we are expanding the exploration of KRTAP gene family to species from genus ovis (domestic and wild species), genus (*Camelus* Bactrian camel), genus *Vicugna*, genus *Capra* (goats). Such study will be very important given the economics importance, e.g. wool from cashmere goat have high commercial quality and quantity. The OR gene family together with vomeronasal and taste receptors will be expanded in other genomes especially to clarify evolutionary reasons that lead to delimitation of the fine boundary that exist between ORs and VNRs, along with their independent diversification and specialization. The TLR and ZP found in birds should be expanded to diverse reptilian species due to their phylogenetic closeness. This will be very useful in uncovering how these families are evolving between closely related linages and if the much similarity observed at level of gene numbers of TLR and ZP (sauropsida have almost same number of genes for TLR and ZP) is also extended at the gene level (positive selection), which can shed light on important evolutionary phenomenon (e.g. physiological polyspermy is common in both these lineages and both have same number of ZPs, which are also under positive selection). This thesis will encourage the future studies for maximum utilization of the rapidly available genomes sequences for comparative genomics and adaptive evolution of gene and gene families, which will help in better understanding the role of environmental pressure in shaping natural selection, adaptive landscape, fitness and speciation. The finding of such studies will play an important role in developing better strategies for biodiversity and conservation (Novembre et al. 2005) (Steinberger et al. 2000).

III. Bibliography

9. Bibliography

Achilonu I, Gildenhuys S, Fisher L, Burke J, Fanucchi S, Sewell BT, Fernandes M, Dirr HW. 2010. The role of a topologically conserved isoleucine in glutathione transferase structure, stability and function. Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun. 66:776–780.

Adipietro KA, Mainland JD, Matsunami H. 2012. Functional Evolution of Mammalian Odorant Receptors. PLoS Genet 8:e1002821.

Akira S, Uematsu S, Takeuchi O. 2006. Pathogen Recognition and Innate Immunity. Cell 124:783–801.

Alcaide M, Edwards SV. 2011. Molecular Evolution of the Toll-Like Receptor Multigene Family in Birds. Mol. Biol. Evol. 28:1703–1715.

Alexopoulou L, Holt AC, Medzhitov R, Flavell RA. 2001. Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. Nature 413:732–738.

Alibardi L, Dalla Valle L, Toffolo V, Toni M. 2006. Scale keratin in lizard epidermis reveals amino acid regions homologous with avian and mammalian epidermal proteins. Anat. Rec. A. Discov. Mol. Cell. Evol. Biol. 288:734–752.

Alibardi L, Jaeger K, Valle LD, Eckhart L. 2011. Ultrastructural localization of hair keratin homologs in the claw of the lizard Anolis carolinensis. J. Morphol. 272:363–370.

Alibardi L, Valle LD, Nardi A, Toni M. 2009. Evolution of hard proteins in the sauropsid integument in relation to the cornification of skin derivatives in amniotes. J. Anat. 214:560–586.

Alibardi L. 2003. Adaptation to the Land : The Skin of Reptiles in Comparison to That of Amphibians and Endotherm Amniotes. Biologia (Bratisl.) 41:12–41.

Alibardi L. 2004. Fine structure of marsupial hairs, with emphasis on trichohyalin and the structure of the inner root sheath. J. Morphol. 261:390–402.

Alibardi L. 2006. Structural and immunocytochemical characterization of keratinization in vertebrate epidermis and epidermal derivatives. Int. Rev. Cytol. 253:177–259.

Alibardi L. 2009. Embryonic keratinization in vertebrates in relation to land colonization. Acta Zool. 90:1–17.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990a. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990b. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Amemiya CT, Alföldi J, Lee AP, et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311–316.

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH. 2011. Genome Evolution and Meiotic Maps by Massively Parallel DNA Sequencing: Spotted Gar, an Outgroup for the Teleost Genome Duplication. Genetics 188:799–808.

Andersen-Nissen E, Smith KD, Bonneau R, Strong RK, Aderem A. 2007. A conserved surface on Toll-like receptor 5 recognizes bacterial flagellin. J. Exp. Med. 204:393–403.

Andersen-Nissen E, Smith KD, Strobe KL, Barrett SLR, Cookson BT, Logan SM, Aderem A. 2005. Evasion of Toll-like receptor 5 by flagellated bacteria. Proc. Natl. Acad. Sci. U. S. A. 102:9247–9252.

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst. Biol. 55:539–552.

Anisimova M, Kosiol C. 2009. Investigating protein-coding sequence evolution with probabilistic codon substitution models. Mol. Biol. Evol. 26:255–271.

Anisimova M, Liberles DA. 2007. The quest for natural selection in the age of comparative genomics. Heredity 99:567–579.

Anisimova M, Nielsen R, Yang Z. 2003. Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236.

Anon. http://www.ensembl.org/index.html.

Anon. http://www.ncbi.nlm.nih.gov/mapview/.

Areal H, Abrantes J, Esteves PJ. 2011. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. BMC Evol. Biol. 11:368.

Armstrong RN. 2000. Mechanistic Diversity in a Metalloenzyme Superfamily†. Biochemistry (Mosc.) 39:13625–13632.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006a. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinforma. Oxf. Engl. 22:195–201.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006b. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195–201.

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res. 38:W529–533.

Avery OT, MacLeod CM, McCarty M. 1944. Studies on the Chemical Nature of the Substance Inducing Transformation of Pneumococcal Types Induction of

Transformation by a Desoxyribonucleic Acid Fraction Isolated from Pneumococcus Type Iii. J. Exp. Med. 79:137–158.

Bahuguna A, Mukherjee SK. 2000. Use of SEM to recognise Tibetan antelope (Chiru) hair and blending in wool products. Sci. Justice 40:177–182.

Baldauf SL. 2003. Phylogeny for the faint of heart: a tutorial. Trends Genet. TIG 19:345–351.

Bargmann CI. 2006a. Comparative chemosensation from receptors to ecology. Nature 444:295–301.

Bargmann CI. 2006b. Comparative chemosensation from receptors to ecology. Nature 444:295–301.

Baxi KN, Dorries KM, Eisthen HL. 2006. Is the vomeronasal system really specialized for detecting pheromones? Trends Neurosci. 29:1–7.

Bell JK, Mullen GED, Leifer CA, Mazzoni A, Davies DR, Segal DM. 2003. Leucine-rich repeats and pathogen recognition in Toll-like receptors. Trends Immunol. 24:528–533.

Benkert P, Biasini M, Schwede T. 2011. Toward the estimation of the absolute quality of individual protein structure models. Bioinforma. Oxf. Engl. 27:343–350.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2013. GenBank. Nucleic Acids Res. 41:D36–42.

Benton R, Sachse S, Michnick SW, Vosshall LB. 2006. Atypical membrane topology and heteromeric function of Drosophila odorant receptors in vivo. PLoS Biol. 4:e20.

Berg AH, Westerlund L, Olsson PE. 2004. Regulation of Arctic char (Salvelinus alpinus) egg shell proteins and vitellogenin during reproduction and in response to 17β-estradiol and cortisol. Gen. Comp. Endocrinol. 135:276–285.

Berlin S, Qu L, Ellegren H. 2008. Adaptive evolution of gamete-recognition proteins in birds. J. Mol. Evol. 67:488–496.

Bielawski JP, Yang Z. 2004. A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. J. Mol. Evol. 59:121–132.

Bininda-Emonds ORP, Cardillo M, Jones KE, et al. 2007. The delayed rise of present-day mammals. Nature 446:507–512.

Birkhead TR, Brillard J-P. 2007. Reproductive isolation in birds: postcopulatory prezygotic barriers. Trends Ecol. Evol. 22:266–272.

Birkhead TR, Pizzari T. 2002. Postcopulatory sexual selection. Nat. Rev. Genet. 3:262–273.

Birkhead TR. 1998. Sperm competition in birds. Rev. Reprod. 3:123–129.

Blackburn AC, Matthaei KI, Lim C, Taylor MC, Cappello JY, Hayes JD, Anders MW, Board PG. 2006. Deficiency of glutathione transferase zeta causes oxidative

stress and activation of antioxidant response pathways. Mol. Pharmacol. 69:650–657.

Bleil JD, Wassarman PM. 1980. Structure and function of the zona pellucida: Identification and characterization of the proteins of the mouse oocyte's zona pellucida. Dev. Biol. 76:185–202.

Blount M. 1909. The early development of the Pigeon's egg, with especial reference to polyspermy and the origin of the periblast nuclei. J. Morphol. 20:1–64.

Bork P, Downing AK, Kieffer B, Campbell ID. 1996. Structure and distribution of modules in extracellular proteins. Q. Rev. Biophys. 29:119–167.

Bork P. 1993. A trefoil domain in the major rabbit zona pellucida protein. Protein Sci. Publ. Protein Soc. 2:669–670.

Botchkarev VA, Paus R. 2003. Molecular biology of hair morphogenesis: development and cycling. J. Exp. Zoolog. B Mol. Dev. Evol. 298:164–180.

Botos I, Segal DM, Davies DR. 2011. The Structural Biology of Toll-like Receptors. Structure 19:447–459.

Boyd AC, Peroval MY, Hammond JA, Prickett MD, Young JR, Smith AL. 2012. TLR15 Is Unique to Avian and Reptilian Lineages and Recognizes a Yeast-Derived Agonist. J. Immunol. 189:4930–4938.

Brodsky I, Medzhitov R. 2007. Two modes of ligand recognition by TLRs. Cell 130:979–981.

Brown TA. Genomes. In: How Genomes Evolve.NCBI Bookshelf. 2nd edition. Chapter 15. Oxford:Wiley-Liss. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21112/

Brownlie R, Zhu J, Allan B, Mutwiri GK, Babiuk LA, Potter A, Griebel P. 2009. Chicken TLR21 acts as a functional homologue to mammalian TLR9 in the recognition of CpG oligodeoxynucleotides. Mol. Immunol. 46:3163–3170.

Buck L, Axel R. 1991. A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell 65:175–187.

Bunderson BR, Kim JE, Croasdell A, Mendoza KM, Reed KM, Coulombe RA Jr. 2013. Heterologous expression and functional characterization of avian mu-class glutathione S-transferases. Comp. Biochem. Physiol. Toxicol. Pharmacol. CBP 158:109–116.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268:78–94.

Callebaut I, Mornon J-P, Monget P. 2007. Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinforma. Oxf. Engl. 23:1871–1874.

Castle PE. 2002. Could multiple low-affinity bonds mediate primary sperm-zona pellucida binding? Reproduction 124:29–32.

Chen IH, Kiang JH, Correa V, Lopez MI, Chen P-Y, McKittrick J, Meyers MA. 2011. Armadillo armor: Mechanical testing and micro-structural evaluation. J. Mech. Behav. Biomed. Mater. 4:713–722.

Chen L, Une Y, Higuchi K, Mori M. 2012. Cheetahs have 4 serum amyloid a genes evolved through repeated duplication events. J. Hered. 103:115–129.

Chuong C-M, Homberger DG. 2003. Development and evolution of the amniote integument: current landscape and future horizon. J. Exp. Zoolog. B Mol. Dev. Evol. 298:1–11.

Clarke KR. 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecol. 18:117–143.

Claw KG, Swanson WJ. 2012. Evolution of the egg: new findings and challenges. Annu. Rev. Genomics Hum. Genet. 13:109–125.

Conner SJ, Hughes DC. 2003. Analysis of fish ZP1/ZPB homologous genes--evidence for both genome duplication and species-specific amplification models of evolution. Reprod. Camb. Engl. 126:347–352.

Conner SJ, Lefièvre L, Hughes DC, Barratt CLR. 2005. Cracking the egg: increased complexity in the zona pellucida. Hum. Reprod. 20:1148–1152.

Cruz OD la, Blekhman R, Zhang X, Nicolae D, Firestein S, Gilad Y. 2009. A Signature of Evolutionary Constraint on a Subset of Ectopically Expressed Olfactory Receptor Genes. Mol. Biol. Evol. 26:491–494.

Cummins I, Wortley DJ, Sabbadin F, et al. 2013. Key role for a glutathione transferase in multiple-herbicide resistance in grass weeds. Proc. Natl. Acad. Sci. U. S. A. 110:5812–5817.

Czech-Damal NU, Liebschner A, Miersch L, Klauer G, Hanke FD, Marshall C, Dehnhardt G, Hanke W. 2011. Electroreception in the Guiana dolphin (Sotalia guianensis). Proc. R. Soc. B Biol. Sci. [Internet]. Available from: http://rspb.royalsocietypublishing.org/content/early/2011/07/21/rspb.2011.1127

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat. Methods 9:772.

Darwin C, Burndy Library donor D, Henry Sotheran Ltd. bookseller D, Edmonds & Remnants binder D. 1859. On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life. London : John Murray ... Available from: http://archive.org/details/onoriginofspec00darw

Dehara Y, Hashiguchi Y, Matsubara K, Yanai T, Kubo M, Kumazawa Y. 2012. Characterization of Squamate Olfactory Receptor Genes and Their Transcripts by the High-Throughput Sequencing Approach. Genome Biol. Evol. 4:602–616.

Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinforma. Oxf. Engl. 26:2455–2457.

Diao G, Wang Y, Wang C, Yang C. 2011. Cloning and Functional Characterization of a Novel Glutathione S-Transferase Gene from Limonium bicolor. Plant Mol. Biol. Report. 29:77–87.

Diebold SS, Kaisho T, Hemmi H, Akira S, Reis e Sousa C. 2004. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. Science 303:1529–1531.

Dirr H, Reinemer P, Huber R. 1994. X-ray crystal structures of cytosolic glutathione S-transferases. Eur. J. Biochem. 220:645–661.

Dixon DP, Lapthorn A, Edwards R. 2002. Plant glutathione transferases. Genome Biol. 3:1–10.

Doyle SL, O'Neill LAJ. 2006. Toll-like receptors: from the discovery of NFkappaB to new insights into transcriptional regulations in innate immunity. Biochem. Pharmacol. 72:1102–1113.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.

Eckhart L, Valle LD, Jaeger K, et al. 2008. Identification of reptilian genes encoding hair keratin-like proteins suggests a new scenario for the evolutionary origin of hair. Proc. Natl. Acad. Sci. U. S. A. 105:18419–18423.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edwards SV, Kingan SB, Calkins JD, Balakrishnan CN, Jennings WB, Swanson WJ, Sorenson MD. 2005. Speciation in birds: genes, geography, and sexual selection. Proc. Natl. Acad. Sci. U. S. A. 102 Suppl 1:6550–6557.

Eizirik E, Yuhki N, Johnson WE, Menotti-Raymond M, Hannah SS, O'Brien SJ. 2003. Molecular genetics and evolution of melanism in the cat family. Curr. Biol. CB 13:448–453.

Elinson RP. 1986. Fertilization in amphibians: the ancestry of the block to polyspermy. Int. Rev. Cytol. 101:59–100.

Escobar JS, Glémin S, Galtier N. 2011. GC-Biased Gene Conversion Impacts Ribosomal DNA Evolution in Vertebrates, Angiosperms, and Other Eukaryotes. Mol. Biol. Evol. 28:2561 –2575.

Felsenstein J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39:783.

Fish FE, Hui CA. 1991. Dolphin swimming–a review. Mammal Rev. 21:181–195.

Fleischmann RD, Adams MD, White O, et al. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269:496–512.

Flicek P, Amode MR, Barrell D, et al. 2011. Ensembl 2012. Nucleic Acids Res. [Internet]. Available from: http://nar.oxfordjournals.org/content/early/2011/11/15/nar.gkr991

Fonseca RR da, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. 2010. Molecular evolution and the role of oxidative stress in the expansion and functional diversification of cytosolic glutathione transferases. BMC Evol. Biol. 10:281.

Force A, Lynch M, Pickett FB, Amores A, Yan Y, Postlethwait J. 1999. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151:1531–1545.

Franbourg A, Hallegot P, Baltenneck F, Toutaina C, Leroy F. 2003. Current research on ethnic hair. J. Am. Acad. Dermatol. 48:S115–S119.

Freitag J, Beck A, Ludwig G, von Buchholtz L, Breer H. 1999. On the origin of the olfactory receptor family: receptor genes of the jawless fish (Lampetra fluviatilis). Gene 226:165–174.

Frova C. 2006. Glutathione transferases in the genomics era: new insights and perspectives. Biomol. Eng. 23:149–169.

Fujikawa H, Fujimoto A, Farooq M, Ito M, Shimomura Y. 2012. Characterization of the Human Hair Keratin–Associated Protein 2 (KRTAP2) Gene Family. J. Invest. Dermatol. 132:1806–1813.

Fumasoni I, Meani N, Rambaldi D, Scafetta G, Alcalay M, Ciccarelli FD. 2007. Family expansion and gene rearrangements contributed to the functional specialization of PRDM genes in vertebrates. BMC Evol. Biol. 7:187.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. TIG 23:273–277.

Ganley ARD, Kobayashi T. 2007. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res. 17:184–191.

Gantier MP, Tong S, Behlke MA, Xu D, Phipps S, Foster PS, Williams BRG. 2008. TLR7 is involved in sequence-specific sensing of single-stranded RNAs in human macrophages. J. Immunol. Baltim. Md 1950 180:2117–2124.

Gantner BN, Simmons RM, Canavera SJ, Akira S, Underhill DM. 2003. Collaborative induction of inflammatory responses by dectin-1 and Toll-like receptor 2. J. Exp. Med. 197:1107–1117.

Gautam P, Jha P, Kumar D, Tyagi S, Varma B, Dash D, Mukhopadhyay A, Mukerji M. 2012. Spectrum of large copy number variations in 26 diverse Indian populations: potential involvement in phenotypic diversity. Hum. Genet. 131:131–143.

Gavrilets S. 2000. Rapid evolution of reproductive barriers driven by sexual conflict. Nature 403:886–889.

George RD, McVicker G, Diederich R, Ng SB, MacKenzie AP, Swanson WJ, Shendure J, Thomas JH. 2011. Trans genomic capture and sequencing of primate exomes reveals new targets of positive selection. Genome Res. 21:1686–1694.

Gharib WH, Robinson-Rechavi M. 2013. The Branch-Site Test of Positive Selection Is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC. Mol. Biol. Evol. 30:1675–1686.

Gillespie MJ, Crowley TM, Haring VR, et al. 2013. Transcriptome analysis of pigeon milk production – role of cornification and triglyceride synthesis genes. BMC Genomics 14:169.

Glusman G, Bahar A, Sharon D, Pilpel Y, White J, Lancet D. 2000. The olfactory receptor gene superfamily: data mining, classification, and nomenclature. Mamm. Genome 11:1016–1023.

Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

Gong H, Zhou H, Dyer JM, Hickford JGH. 2011. Identification of the ovine KAP11-1 gene (KRTAP11-1) and genetic variation in its coding sequence. Mol. Biol. Rep. 38:5429–5433.

Gong H, Zhou H, Hickford JGH. 2011. Diversity of the glycine/tyrosine-rich keratin-associated protein 6 gene (KAP6) family in sheep. Mol. Biol. Rep. 38:31–35.

Gong H, Zhou H, Plowman JE, Dyer JM, Hickford JGH. 2010. Analysis of variation in the ovine ultra-high sulphur keratin-associated protein KAP5-4 gene using PCR-SSCP technique. Electrophoresis 31:3545–3547.

Gong H, Zhou H, Yu Z, Dyer J, Plowman JE, Hickford J. 2011. Identification of the ovine keratin-associated protein KAP1-2 gene (KRTAP1-2). Exp. Dermatol. 20:815–819.

Goudet G, Mugnier S, Callebaut I, Monget P. 2008. Phylogenetic analysis and identification of pseudogenes reveal a progressive loss of zona pellucida genes during evolution of vertebrates. Biol. Reprod. 78:796–806.

Gould S jay, Eldredge N. 1993. Punctuated equilibrium comes of age. Nature 366:223–227.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. 27:221–224.

Grada A, Weinbrecht K. 2013. Next-Generation Sequencing: Methodology and Application. J. Invest. Dermatol. 133:e11.

Gu X, Zou Y, Su Z, Huang W, Zhou Z, Arendsee Z, Zeng Y. 2013. An update of DIVERGE software for functional divergence analysis of protein family. Mol. Biol. Evol. 30:1713–1719.

Guan Y, Ranoa DRE, Jiang S, Mutha SK, Li X, Baudry J, Tapping RI. 2010. Human TLRs 10 and 1 share common mechanisms of innate immune sensing but not signaling. J. Immunol. Baltim. Md 1950 184:5094–5103.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Gupta SK, Bhandari B, Shrestha A, Biswal BK, Palaniappan C, Malhotra SS, Gupta N. 2012. Mammalian zona pellucida glycoproteins: structure and function during fertilization. Cell Tissue Res. 349:665–678.

Hajjar AM, Ernst RK, Tsai JH, Wilson CB, Miller SI. 2002. Human Toll-like receptor 4 recognizes host-specific LPS modifications. Nat. Immunol. 3:354–359.

Han L, Monné M, Okumura H, Schwend T, Cherry AL, Flot D, Matsuda T, Jovine L. 2010. Insights into Egg Coat Assembly and Egg-Sperm Interaction from the X-Ray Structure of Full-Length ZP3. Cell 143:404–415.

Han MV, Demuth JP, Mcgrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. Genome Res.:859–867.

Hardy MH. 1992. The secret life of the hair follicle. Trends Genet. 8:55–61.

Harris JD, Hibler DW, Fontenot GK, Hsu KT, Yurewicz EC, Sacco AG. 1994. Cloning and characterization of zona pellucida genes and cDNAs from a variety of mammalian species: the ZPA, ZPB and ZPC gene families. DNA Seq. J. DNA Seq. Mapp. 4:361–393.

Hartl DL. 2005. Genetics: Analysis of Genes and Genomes. Jones & Bartlett Learning

Hasan U, Chaffois C, Gaillard C, et al. 2005. Human TLR10 is a functional receptor, expressed by B cells and plasmacytoid dendritic cells, which activates gene transcription through MyD88. J. Immunol. Baltim. Md 1950 174:2942–2950.

Hayashi F, Smith KD, Ozinsky A, et al. 2001. The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. Nature 410:1099–1103.

Hayden S, Bekaert M, Crider TA, Mariani S, Murphy WJ, Teeling EC. 2010. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. 20:1–9.

Hayes JD, Flanagan JU, Jowsey IR. 2005. Glutathione transferases. Annu. Rev. Pharmacol. Toxicol. 45:51–88.

Hayes JD, Pulford DJ. 1995. The glutathione S-transferase supergene family: regulation of GST and the contribution of the isoenzymes to cancer chemoprotection and drug resistance. Crit. Rev. Biochem. Mol. Biol. 30:445–600.

Hedrick JL. 2008. Anuran and pig egg zona pellucida glycoproteins in fertilization and early development. Int. J. Dev. Biol. 52:683–701.

Heil F, Hemmi H, Hochrein H, Ampenberger F, Kirschning C, Akira S, Lipford G, Wagner H, Bauer S. 2004. Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. Science 303:1526–1529.

Henikoff S. 1997. Gene Families: The Taxonomy of Protein Paralogs and Chimeras. Science 278:609–614.

Hesse M, Zimek A, Weber K, Magin TM. 2004. Comprehensive analysis of keratin gene clusters in humans and rodents. Eur. J. Cell Biol. 83:19–26.

Hicks BD, Aubin DJS, Geraci JR, Brown WR. 1985. Epidermal Growth in the Bottlenose Dolphin, Tursiops truncatus. J. Invest. Dermatol. 85:60–63.

Higginbotham S, Wong WR, Linington RG, Spadafora C, Iturrado L, Arnold AE. 2014. Sloth Hair as a Novel Source of Fungi with Potent Anti-Parasitic, Anti-Cancer and Anti-Bacterial Bioactivity. PLoS ONE 9:e84549.

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677–681.

Hodgkin J. 1998. Seven types of pleiotropy. Int. J. Dev. Biol. 42:501–505.

Horrocks AJ, Stewart S, Jackson L, Wishart GJ. 2000. Induction of acrosomal exocytosis in chicken spermatozoa by inner perivitelline-derived N-linked glycans. Biochem. Biophys. Res. Commun. 278:84–89.

Hoshino K, Takeuchi O, Kawai T, Sanjo H, Ogawa T, Takeda Y, Takeda K, Akira S. 1999. Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. J. Immunol. Baltim. Md 1950 162:3749–3752.

Huang Y, Temperley N, Ren L, Smith J, Li N, Burt D. 2011. Molecular evolution of the vertebrate TLR1 gene family - a complex history of gene duplication, gene conversion, positive selection and co-evolution. BMC Evol. Biol. 11:149.

Huelsenbeck JP, Rannala B. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276:227–232.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinforma. Oxf. Engl. 17:754–755.

Hughes DC. 2007. ZP genes in avian species illustrate the dynamic evolution of the vertebrate egg envelope. Cytogenet. Genome Res. 117:86–91.

Humphrey W, Dalke A, Schulten K. 1996. VMD: visual molecular dynamics. J. Mol. Graph. 14:33–38, 27–28.

Imler JL, Hoffmann JA. 2002. Toll receptors in Drosophila: a family of molecules regulating development and immunity. Curr. Top. Microbiol. Immunol. 270:63–79.

Ishii A, Matsuo A, Sawa H, Tsujita T, Shida K, Matsumoto M, Seya T. 2007. Lamprey TLRs with properties distinct from those of the variable lymphocyte receptors. J. Immunol. Baltim. Md 1950 178:397–406.

Iwao Y. 2012. Egg activation in physiological polyspermy. Reprod. Camb. Engl. 144:11–22.

Iwasaki A, Medzhitov R. 2010. Regulation of adaptive immunity by the innate immune system. Science 327:291–295.

Janeway CA Jr, Medzhitov R. 2002. Innate immune recognition. Annu. Rev. Immunol. 20:197–216.

Jarvis et al. 2014.. Gilbert, and Guojie Zhang. 2014. Whole Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds. Science (accepted).

Jenkins BJ, Powell BC. 1994. Differential expression of genes encoding a cysteine-rich keratin family in the hair cuticle. J. Invest. Dermatol. 103:310–317.

Jenkins, J.2009. “Tursiops truncatus” (On-line), Animal Diversity Web. Accessed July 23, 2012 at http://animaldiversity.ummz.umich.edu/site/accounts/information/Tursiops_truncatus.html.

Jin MS, Kim SE, Heo JY, Lee ME, Kim HM, Paik S-G, Lee H, Lee J-O. 2007. Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell 130:1071–1082.

Jin MS, Lee J-O. 2008. Structures of TLR-ligand complexes. Curr. Opin. Immunol. 20:414–419.

Jovine L, Darie CC, Litscher ES, Wassarman PM. 2005. Zona pellucida domain proteins. Annu. Rev. Biochem. 74:83–114.

Jovine L, Qi H, Williams Z, Litscher ES, Wassarman PM. 2004. A duplicated motif controls assembly of zona pellucida domain proteins. Proc. Natl. Acad. Sci. U. S. A. 101:5922–5927.

Kajava AV. 1998. Structural diversity of leucine-rich repeat proteins. J. Mol. Biol. 277:519–527.

Kanai S, Kitayama T, Yonezawa N, Sawano Y, Tanokura M, Nakano M. 2008. Disulfide linkage patterns of pig zona pellucida glycoproteins ZP3 and ZP4. Mol. Reprod. Dev. 75:847–856.

Kang JY, Lee J-O. 2011. Structural biology of the Toll-like receptor family. Annu. Rev. Biochem. 80:917–941.

Kang JY, Nan X, Jin MS, et al. 2009. Recognition of lipopeptide patterns by Toll-like receptor 2-Toll-like receptor 6 heterodimer. Immunity 31:873–884.

Kariya N, Shimomura Y, Ito M. 2005. Size Polymorphisms in the Human Ultrahigh Sulfur Hair Keratin-Associated Protein 4, KAP4, Gene Family. J Invest Dermatol 124:1111–1118.

Kasamatsu J, Oshiumi H, Matsumoto M, Kasahara M, Seya T. 2010. Phylogenetic and expression analysis of lamprey toll-like receptors. Dev. Comp. Immunol. 34:855–865.

Kawai T, Akira S. 2006. TLR signaling. Cell Death Differ. 13:816–825.

Kawai T, Akira S. 2010. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nat. Immunol. 11:373–384.

Keen JH, Jakoby WB. 1978. Glutathione transferases. Catalysis of nucleophilic reactions of glutathione. J. Biol. Chem. 253:5654–5657.

Keestra AM, Zoete MR de, Bouwman LI, Putten JPM van. 2010. Chicken TLR21 Is an Innate CpG DNA Receptor Distinct from Mammalian TLR9. J. Immunol. 185:460–467.

Kim HM, Park BS, Kim J-I, et al. 2007. Crystal Structure of the TLR4-MD-2 Complex with Bound Endotoxin Antagonist Eritoran. Cell 130:906–917.

Kim JE, Bunderson BR, Croasdell A, Reed KM, Coulombe RA Jr. 2013. Alpha-class glutathione S-transferases in wild turkeys (Meleagris gallopavo): characterization and role in resistance to the carcinogenic mycotoxin aflatoxin B1. PloS One 8:e60662.

Kishida T, Hikida T. 2010. Degeneration patterns of the olfactory receptor genes in sea snakes. J. Evol. Biol. 23:302–310.

Kishida T, Kubota S, Shirayama Y, Fukami H. 2007. The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: evidence for reduction of the functional proportions in cetaceans. Biol. Lett. 3:428–430.

Knapp S, von Aulock S, Leendertse M, Haslinger I, Draing C, Golenbock DT, van der Poll T. 2008. Lipoteichoic acid-induced lung inflammation depends on TLR2 and the concerted action of TLR4 and the platelet-activating factor receptor. J. Immunol. Baltim. Md 1950 180:3478–3484.

Kobe B, Deisenhofer J. 1994. The leucine-rich repeat: a versatile binding motif. Trends Biochem. Sci. 19:415–421.

Kobe B, Kajava AV. 2001. The leucine-rich repeat as a protein recognition motif. Curr. Opin. Struct. Biol. 11:725–732.

Koblansky AA, Jankovic D, Oh H, et al. 2013. Recognition of Profilin by Toll-like Receptor 12 Is Critical for Host Resistance to Toxoplasma gondii. Immunity 38:119–130.

Koehn H, Clerens S, Deb-Choudhury S, Morton JD, Dyer JM, Plowman JE. 2010. The proteome of the wool cuticle. J. Proteome Res. 9:2920–2928.

Kosiol C, Anisimova M. 2012. Selection on the protein-coding genome. Methods Mol. Biol. Clifton NJ 856:113–140.

Kosiol C, Vinař T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of Positive Selection in Six Mammalian Genomes. PLoS Genet 4:e1000144.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2011. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. Mol. Biol. Evol.

Krug A, Luker GD, Barchet W, Leib DA, Akira S, Colonna M. 2004. Herpes simplex virus type 1 activates murine natural interferon-producing cells through toll-like receptor 9. Blood 103:1433–1437.

Kubo H, Shiga K, Harada Y, Iwao Y. 2010. Analysis of a sperm surface molecule that binds to a vitelline envelope component of Xenopus laevis eggs. Mol. Reprod. Dev. 77:728–735.

Kumar H, Kawai T, Akira S. 2009a. Toll-like receptors and innate immunity. Biochem. Biophys. Res. Commun. 388:621–625.

Kumar H, Kawai T, Akira S. 2009b. Pathogen recognition in the innate immune response. Biochem. J. 420:1–16.

Laborde E. 2010. Glutathione transferases as mediators of signaling pathways involved in cell proliferation and cell death. Cell Death Differ. 17:1373–1380.

Lagerstrom MC, Hellstrom AR, Gloriam DE, Larsson TP, Schioth HB, Fredriksson R. 2006. The G Protein-Coupled Receptor Subset of the Chicken Genome. PLoS Comput. Biol. [Internet] 2. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1472694/

Lande R. 1988. Genetics and demography in biological conservation. Science 241:1455–1460.

Lander ES, Linton LM, Birren B, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Landi S. 2000. Mammalian class theta GST and differential susceptibility to carcinogens: a review. Mutat. Res. 463:247–283.

Lee YJ, Rice RH, Lee YM. 2006. Proteome analysis of human hair shaft: from protein identification to posttranslational modification. Mol. Cell. Proteomics MCP 5:789–800.

Leister D. 2004. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance gene. Trends Genet. TIG 20:116–122.

Lespinet O, Wolf YI, Koonin EV, Aravind L. 2002. The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. Genome Res.:1048–1059.

Lévai O, Feistel T, Breer H, Strotmann J. 2006. Cells in the vomeronasal organ express odorant receptors but project to the accessory olfactory bulb. J. Comp. Neurol. 498:476–490.

Lichtenstein G, Vilá B. 2003. Vicuna Use by Andean Communities: An Overview. Mt. Res. Dev. 23:198–201.

Liu H, Yue C, Zhang W, Zhu X, Yang G, Jia Z. 2011. Association of the KAP 8.1 gene polymorphisms with fibre traits in inner Mongolian cashmere goats. Asian - Australas. J. Anim. Sci. 24:1341 – 1347.

Liu L, Botos I, Wang Y, Leonard JN, Shiloach J, Segal DM, Davies DR. 2008. Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320:379–381.

Lo H-W, Ali-Osman F. 2007. Genetic polymorphism and function of glutathione S-transferases in tumor drug resistance. Curr. Opin. Pharmacol. 7:367–374.

Louis A, Muffato M, Roest Crollius H. 2013. Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res. 41:D700–D705.

Louis A, Muffato M, Roest Crollius H. 2013. Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res. 41:D700–705.

Löytynoja A, Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proc. Natl. Acad. Sci. U. S. A. 102:10557–10562.

Lu J, Peatman E, Tang H, Lewis J, Liu Z. 2012. Profiling of gene duplication patterns of sequenced teleost genomes: evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications. BMC Genomics 13:246.

Lund JM, Alexopoulou L, Sato A, Karow M, Adams NC, Gale NW, Iwasaki A, Flavell RA. 2004. Recognition of single-stranded RNA viruses by Toll-like receptor 7. Proc. Natl. Acad. Sci. U. S. A. 101:5598–5603.

Lushchak VI. 2011. Environmentally induced oxidative stress in aquatic animals. Aquat. Toxicol. 101:13–30.

Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154:459–473.

Maddison, W. P, D.R. Maddison. 2011. Mesquite: a modular system for evolutionary analysis. Version 2.75 http://mesquiteproject.org. Available from: http://mesquiteproject.org

Malnic B, Hirono J, Sato T, Buck LB. 1999. Combinatorial Receptor Codes for Odors. Cell 96:713–723.

Malo AF, Martinez-Pastor F, Alaks G, Dubach J, Lacy RC. 2010. Effects of Genetic Captive-Breeding Protocols on Sperm Quality and Fertility in the White-Footed Mouse. Biol. Reprod. 83:540–548.

Man O, Gilad Y, Lancet D. 2004. Prediction of the odorant binding site of olfactory receptor proteins by human–mouse comparisons. Protein Sci. 13:240–254.

Mao H-T, Yang W-X. 2013. Modes of acrosin functioning during fertilization. Gene 526:75–79.

Mariani V, Kiefer F, Schmidt T, Haas J, Schwede T. 2011. Assessment of template based protein structure predictions in CASP9. Proteins 79 Suppl 10:37–58.

Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res. 21:863–874.

Marotta M, Chen X, Inoshita A, et al. 2012. A common copy number breakpoint of ERBB2 amplification in breast cancer co-localizes with a complex block of segmental duplications. Breast Cancer Res. 14:R150.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. 2010. RDP3: A Flexible and Fast Computer Program for Analyzing Recombination. Bioinformatics 26:2462–2463.

Matsuo A, Oshiumi H, Tsujita T, Mitani H, Kasai H, Yoshimizu M, Matsumoto M, Seya T. 2008. Teleost TLR22 recognizes RNA duplex to induce IFN and protect cells from birnaviruses. J. Immunol. Baltim. Md 1950 181:3474–3485.

Matsushima N, Miyashita H, Mikami T, Kuroki Y. 2010. A nested leucine rich repeat (LRR) domain: the precursor of LRRs is a ten or eleven residue motif. BMC Microbiol. 10:235.

Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K, Kuroki Y. 2007. Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. BMC Genomics 8:124.

Mattila TM, Bokma F. 2008. Extant mammal body masses suggest punctuated equilibrium. Proc. Biol. Sci. 275:2195–2199.

Mauck B, Eysel U, Dehnhardt G. 2000. Selective heating of vibrissal follicles in seals (Phoca vitulina) and dolphins (Sotalia fluviatilis guianensis). J. Exp. Biol. 203:2125–2131.

Mazet F, Shimeld SM. 2002. Gene duplication and divergence in the early evolution of vertebrates. Curr. Opin. Genet. Dev. 12:393–396.

McGonigle B, Keeler SJ, Lau SM, Koeppe MK, O'Keefe DP. 2000. A genomics approach to the comprehensive analysis of the glutathione S-transferase gene family in soybean and maize. Plant Physiol. 124:1105–1120.

McKenzie GW, Abbott J, Zhou H, Fang Q, Merrick N, Forrest RH, Sedcole JR, Hickford JG. 2010. Genetic diversity of selected genes that are potentially economically important in feral sheep of New Zealand. Genet. Sel. Evol. GSE 42:43.

McLaren RJ, Rogers GR, Davies KP, Maddox JF, Montgomery GW. 1997. Linkage mapping of wool keratin and keratin-associated protein genes in sheep. Mamm. Genome Off. J. Int. Mamm. Genome Soc. 8:938–940.

Medzhitov R, Preston-Hurlburt P, Janeway CA Jr. 1997. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. Nature 388:394–397.

Medzhitov R. 2001. Toll-like receptors and innate immunity. Nat. Rev. Immunol. 1:135–145.

Medzhitov R. 2007. Recognition of microorganisms and activation of the immune response. Nature 449:819–826.

Meslin C, Mugnier S, Callebaut I, Laurin M, Pascal G, Poupon A, Goudet G, Monget P. 2012. Evolution of Genes Involved in Gamete Interaction: Evidence for Positive Selection, Duplications and Losses in Vertebrates. PLoS ONE 7:e44548.

Meyer W, Schmidt J, Busche R, Jacob R, Naim HY. 2012. Demonstration of free fatty acids in the integument of semi-aquatic and aquatic mammals. Acta Histochem. 114:145–150.

Millar SE. 2002. Molecular Mechanisms Regulating Hair Follicle Development. 118:216–225.

Monné M, Han L, Schwend T, Burendahl S, Jovine L. 2008. Crystal structure of the ZP-N domain of ZP3 reveals the core fold of animal egg coats. Nature 456:653–657.

Muffato M, Louis A, Poisnel C-E, Roest Crollius H. 2010a. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinforma. Oxf. Engl. 26:1119–1121.

Muffato M, Louis A, Poisnel C-E, Roest Crollius H. 2010b. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinforma. Oxf. Engl. 26:1119–1121.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Pond SLK, Scheffler K. 2013. FUBAR : A Fast, Unconstrained Bayesian AppRoximation for inferring selection. Mol. Biol. Evol.:mst030.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting Individual Sites Subject to Episodic Diversifying Selection. PLoS Genet 8:e1002764.

Muse SV, Gaut BS. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. 11:715–724.

Naudí A, Jové M, Ayala V, Portero-Otín M, Barja G, Pamplona R. 2011. Regulation of Membrane Unsaturation as Antioxidant Adaptive Mechanism in Long-lived Animal Species. Free Radic. Antioxid. 1:3–12.

Naudí A, Jové M, Ayala V, Portero-Otín M, Barja G, Pamplona R. 2013. Membrane lipid unsaturation as physiological adaptation to animal longevity. Front. Physiol. 4:372.

Nei M, Niimura Y, Nozawa M. 2008. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat. Rev. Genet. 9:951–963.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu. Rev. Genet. 39:121–152.

Néron B, Ménager H, Maufrais C, Joly N, Maupetit J, Letort S, Carrere S, Tuffery P, Letondal C. 2009. Mobyle: a new full web bioinformatics framework. Bioinforma. Oxf. Engl. 25:3005–3011.

Nguyen DT, Lee K, Choi H, et al. 2012. The complete swine olfactory subgenome: expansion of the olfactory gene repertoire in the pig genome. BMC Genomics 13:584.

Nielsen R, Bustamante C, Clark AG, et al. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biol 3:e170.

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

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

Niimura Y, Nei M. 2005. Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. Proc. Natl. Acad. Sci. 102:6039–6044.

Niimura Y, Nei M. 2006. Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. J. Hum. Genet. 51:505–517.

Niimura Y. 2009a. Evolutionary dynamics of olfactory receptor genes in chordates: interaction between environments and genomic contents. Hum. Genomics 4:107.

Niimura Y. 2009b. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. Genome Biol. Evol. 1:34–44.

Niimura Y. 2012. Olfactory Receptor Multigene Family in Vertebrates: From the Viewpoint of Evolutionary Genomics. Curr. Genomics 13:103–114.

Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome Res. 14:354–366.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302:205–217.

Novembre J, Galvani AP, Slatkin M. 2005. The geographic spread of the CCR5 Delta32 HIV-resistance allele. PLoS Biol. 3:e339.

O'brien SJ, Wildt DE, Goldman D, Merril CR, Bush M. 1983. The Cheetah Is Depauperate in Genetic Variation. Science 221:459–462.

O'Brien SJ, Womack JE, Lyons LA, Moore KJ, Jenkins NA, Copeland NG. 1993. Anchored reference loci for comparative genome mapping in mammals. Nat. Genet. 3:103–112.

O'Brien T, Semrad S, Serena A. 2010. Multiple congenital anomalies in an alpaca cria (Vicugna pacos). Vet. Rec. 166:759–760.

O'Neill LAJ, Bowie AG. 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nat. Rev. Immunol. 7:353–364.

O'Neill LAJ, Golenbock D, Bowie AG. 2013. The history of Toll-like receptors — redefining innate immunity. Nat. Rev. Immunol. 13:453–460.

Ohno S. 1970. Evolution by gene duplication. London; New York: Allen & Unwin; Springer-Verlag

OHNO S. 1970. Evolution by gene duplication. xv + 160 pp.

Okumura H, Aoki N, Sato C, Nadano D, Matsuda T. 2007. Heterocomplex formation and cell-surface accumulation of hen's serum zona pellucida B1 (ZPB1)with ZPC expressed by a mammalian cell line (COS-7): a possible initiating step of egg-envelope matrix construction. Biol. Reprod. 76:9–18.

Okumura H, Kohno Y, Iwata Y, Mori H, Aoki N, Sato C, Kitajima K, Nadano D, Matsuda T. 2004. A newly identified zona pellucida glycoprotein, ZPD, and dimeric ZP1 of chicken egg envelope are involved in sperm activation on sperm-egg interaction. Biochem. J. 384:191–199.

Oldenburg M, Krüger A, Ferstl R, et al. 2012. TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. Science 337:1111–1115.

Olender T, Fuchs T, Linhart C, Shamir R, Adams M, Kalush F, Khen M, Lancet D. 2004. The canine olfactory subgenome. Genomics 83:361–372.

Olender T, Lancet D, Nebert DW. 2008. Update on the olfactory receptor (OR) gene superfamily. Hum. Genomics 3:87.

Øyvind Hammer, r,David A. T. Harper, and Paul D. Ryan. 2001. PAST. Paleontological Statistics Software Package for Education and Data Analysis. Palaeontol Electronica. Available from: http://palaeo-electronica.org/2001_1/past/issue1_01.htm.

Oztetik E, Cakir A. 2014. New food for an old mouth: New enzyme for an ancient archaea. Enzyme Microb. Technol. 55:58–64.

Palmer E, Weddell G. 1964. THE RELATIONSHIP BETWEEN STRUCTURE, INNERVATION AND FUNCTION OF THE SKIN OF THE BOTTLE NOSE DOLPHIN (TU RSIOPS TRUNCATUS). Proc. Zool. Soc. Lond. 143:553–568.

Palti Y. 2011. Toll-like receptors in bony fish: from genomics to function. Dev. Comp. Immunol. 35:1263–1272.

Palumbi SR. 1999. All males are not created equal: fertility differences depend on gamete recognition polymorphisms in sea urchins. Proc. Natl. Acad. Sci. U. S. A. 96:12632–12637.

Pantalacci S, Chaumot A, Benoît G, Sadier A, Delsuc F, Douzery EJP, Laudet V. 2008. Conserved Features and Evolutionary Shifts of the EDA Signaling Pathway Involved in Vertebrate Skin Appendage Development. Mol. Biol. Evol. 25:912–928.

Park BS, Song DH, Kim HM, Choi B-S, Lee H, Lee J-O. 2009. The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. Nature 458:1191–1195.

Parry D a D, Smith T a, Rogers M a, Schweizer J. 2006. Human hair keratin-associated proteins: sequence regularities and structural implications. J. Struct. Biol. 155:361–369.

Parsons YM, Cooper DW, Piper LR. 1994. Evidence of linkage between high-glycine-tyrosine keratin gene loci and wool fibre diameter in a Merino half-sib family. Anim. Genet. 25:105–108.

Pearson H. 2006. Genetics: What is a gene? Nature 441:398–401.

Pearson WR, Wood T, Zhang Z, Miller W. 1997. Comparison of DNA sequences with protein sequences. Genomics 46:24–36.

Penn O, Privman E, Landan G, Graur D, Pupko T. 2010. An alignment confidence score capturing robustness to guide tree uncertainty. Mol. Biol. Evol. 27:1759–1767.

Pennisi E. 2007. DNA Study Forces Rethink of What It Means to Be a Gene. Science 316:1556–1557.

Perry GH, Melsted P, Marioni JC, et al. 2012. Comparative RNA sequencing reveals substantial genetic variation in endangered primates. Genome Res. 22:602–610.

Pietro G, Magno LAV, Rios-Santos F. 2010. Glutathione S-transferases: an overview in cancer research. Expert Opin. Drug Metab. Toxicol. 6:153–170.

Pizzari T, Birkhead TR. 2000. Female feral fowl eject sperm of subdominant males. Nature 405:787–789.

Plaza S, Chanut-Delalande H, Fernandes I, Wassarman PM, Payre F. 2010. From A to Z: apical structures and zona pellucida-domain proteins. Trends Cell Biol. 20:524–532.

Pond SLK, Frost SDW, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Pond SLK, Frost SDW. 2005a. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533.

Pond SLK, Frost SDW. 2005b. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. Mol. Biol. Evol. 22:1208–1222.

Pond SLK, Posada D, Gravenor MB, Woelk CH, Frost SDW. 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics 22:3096–3098.

Powell BC, Arthur J, Nesci A. 1995. Characterization of a gene encoding a cysteine-rich keratin associated protein synthesized late in rabbit hair follicle differentiation. Differ. Res. Biol. Divers. 58:227–232.

Powell BC, Nesci A, Rogers GE. 1991. Regulation of keratin gene expression in hair follicle differentiation. Ann. N. Y. Acad. Sci. 642:1–20.

Powell BC, Rogers GE. 1997. The role of keratin proteins and their genes in the growth, structure and properties of hair. Exs 78:59–148.

Pruett ND, Tkatchenko TV, Jave-Suarez L, Jacobs DF, Potter CS, Tkatchenko AV, Schweizer J, Awgulewitsch A. 2004. Krtap16, characterization of a new hair keratin-associated protein (KAP) gene complex on mouse chromosome 16 and evidence for regulation by Hoxc13. J. Biol. Chem. 279:51524–51533.

Purvis IW, Jeffery N. 2007. Genetics of fibre production in sheep and goats. Small Rumin. Res. 70:42–47.

Quignon P, Giraud M, Rimbault M, et al. 2005. The dog and rat olfactory receptor repertoires. Genome Biol. 6:R83.

Quiniou SMA, Boudinot P, Bengtén E. 2013. Comprehensive survey and genomic characterization of Toll-like receptors (TLRs) in channel catfish, Ictalurus punctatus: identification of novel fish TLRs. Immunogenetics.

Ranson H, Collins F, Hemingway J. 1998. The role of alternative mRNA splicing in generating heterogeneity within the Anopheles gambiae class I glutathione S-transferase family. Proc. Natl. Acad. Sci. 95:14284–14289.

Reed KD, Meece JK, Henkel JS, Shukla SK. 2003. Birds, Migration and Emerging Zoonoses: West Nile Virus, Lyme Disease, Influenza A and Enteropathogens. Clin. Med. Res. 1:5–12.

Remmert M, Biegert A, Hauser A, Söding J. 2012. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9:173–175.

Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, Hood LE, Aderem A. 2005. The evolution of vertebrate Toll-like receptors. Proc. Natl. Acad. Sci. U. S. A. 102:9577–9582.

Rogers GE. 2004. Hair follicle differentiation and regulation. Int. J. Dev. Biol. 48:163–170.

Rogers M a, Schweizer J. 2005. Human KAP genes, only the half of it? Extensive size polymorphisms in hair keratin-associated protein genes. J. Invest. Dermatol. 124:vii–ix.

Rogers M, Winter H, Langbein L, Bleiler R. 2004. The human type I keratin gene family: characterization of new hair follicle specific members and evaluation of the chromosome 17q21. 2 gene domain. Differ. Res. Biol. Divers. 72:527–540.

Rogers MA, Edler L, Winter H, Langbein L, Beckmann I, Schweizer J. 2005. Characterization of new members of the human type II keratin gene family and a general evaluation of the keratin gene domain on chromosome 12q13.13. J. Invest. Dermatol. 124:536–544.

Rogers MA, Langbein L, Praetzel-Wunder S, Giehl K. 2008. Characterization and expression analysis of the hair keratin associated protein KAP26.1. Br. J. Dermatol. 159:725–729.

Rogers MA, Langbein L, Winter H, Beckmann I, Praetzel S, Schweizer J. 2004. Hair keratin associated proteins: characterization of a second high sulfur KAP gene domain on human chromosome 21. J. Invest. Dermatol. 122:147–158.

Rogers MA, Langbein L, Winter H, Ehmann C, Praetzel S, Korn B, Schweizer J. 2001. Characterization of a cluster of human high/ultrahigh sulfur keratin-associated protein genes embedded in the type I keratin gene domain on chromosome 17q12-21. J. Biol. Chem. 276:19440–19451.

Rogers MA, Langbein L, Winter H, Ehmann C, Praetzel S, Schweizer J. 2002. Characterization of a first domain of human high glycine-tyrosine and high sulfur keratin-associated protein (KAP) genes on chromosome 21q22.1. J. Biol. Chem. 277:48993–49002.

Rogers MA, Winter H, Langbein L, Wolf C, Schweizer J. 2000. Characterization of a 300 kbp region of human DNA containing the type II hair keratin gene domain. J. Invest. Dermatol. 114:464–472.

Rogers MA, Winter H, Langbein L, Wollschläger A, Praetzel-Wunder S, Jave-Suarez LF, Schweizer J. 2007. Characterization of human KAP24.1, a cuticular hair keratin-associated protein with unusual amino-acid composition and repeat structure. J. Invest. Dermatol. 127:1197–1204.

Rogers MA, Winter H, Wolf C, Heck M, Schweizer J. Characterization of a 190-Kilobase Pair Domain of Human Type I Hair Keratin Genes. J. Biol. Chem. 273:26683–26691.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinforma. Oxf. Engl. 19:1572–1574.

Ruben JA, Jones TD. 2000. Selective Factors Associated with the Origin of Fur and Feathers. Am. Zool. 40:585–596.

Ruzza P, Rosato A, Rossi CR, Floreani M, Quintieri L. 2009. Glutathione transferases as targets for cancer therapy. Anticancer Agents Med. Chem. 9:763–777.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. Nat. Genet. 39:1461–1468.

Sahajpal V, Goyal S, Singh K, Thakur V. 2009. Dealing Wildlife Offences in India: Role of the Hair as Physical Evidence. Int. J. Trichology 1:18–26.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, Hutchison CA, Slocombe PM, Smith M. 1977. Nucleotide sequence of bacteriophage |[phi]|X174 DNA. Nature 265:687–695.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. 74:5463–5467.

Sano K, Kawaguchi M, Watanabe S, Nagakura Y, Hiraki T, Yasumasu S. 2013. Inferring the evolution of teleostean zp genes based on their sites of expression. J. Exp. Zoolog. B Mol. Dev. Evol. 320:332–343.

Sasanami T, Ohtsuki M, Ishiguro T, Matsushima K, Hiyama G, Kansaku N, Doi Y, Mori M. 2006. Zona Pellucida Domain of ZPB1 controls specific binding of ZPB1 and ZPC in Japanese quail (Coturnix japonica). Cells Tissues Organs 183:41–52.

Sato K, Pellegrino M, Nakagawa T, Nakagawa T, Vosshall LB, Touhara K. 2008. Insect olfactory receptors are heteromeric ligand-gated ion channels. Nature 452:1002–1006.

Sawyer S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. 6:526–538.

Schnare M, Barton GM, Holt AC, Takeda K, Akira S, Medzhitov R. 2001. Toll-like receptors control activation of adaptive immune responses. Nat. Immunol. 2:947–950.

Schneider MR, Schmidt-Ullrich R, Paus R. 2009. The hair follicle as a dynamic miniorgan. Curr. Biol. CB 19:R132–42.

Schröder P. 2001. The Role of Glutathione and Glutathione S-transferases in Plant Reaction and Adaptation to Xenobiotics. In: Grill D, Tausz M, Kok LJD, editors. Significance of Glutathione to Plant Adaptation to the Environment. Plant Ecophysiology. Springer Netherlands. p. 155–183. Available from: http://link.springer.com/chapter/10.1007/0-306-47644-4_7

Schwede T, Kopp J, Guex N, Peitsch MC. 2003a. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 31:3381–3385.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003b. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res. 31:3381–3385.

Schweizer J, Bowden PE, Coulombe P a, et al. 2006. New consensus nomenclature for mammalian keratins. J. Cell Biol. 174:169–174.

Scientists G 10K C of. 2009. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. J. Hered. 100:659–674.

Sharma R, Yang Y, Sharma A, Awasthi S, Awasthi YC. 2004. Antioxidant Role of Glutathione *S* -Transferases: Protection Against Oxidant Toxicity and Regulation of Stress-Mediated Apoptosis. Antioxid. Redox Signal. 6:289–300.

Sheehan D, Meade G, Foley VM, Dowd CA. 2001. Structure, function and evolution of glutathione transferases: implications for classification of non-mammalian members of an ancient enzyme superfamily. Biochem. J. 360:1–16.

Shi Y, Wang Q, Hou Y, Hong Y, Han X, Yi J, Qu J, Lu Y. 2014. Molecular cloning, expression and enzymatic characterization of glutathione S-transferase from Antarctic sea-ice bacteria Pseudoalteromonas sp. ANT506. Microbiol. Res. 169:179–184.

Shi Z, Cai Z, Sanchez A, Zhang T, Wen S, Wang J, Yang J, Fu S, Zhang D. 2011. A Novel Toll-like Receptor That Recognizes Vesicular Stomatitis Virus. J. Biol. Chem. 286:4517–4524.

Shibuya K, Kudoh J, Obayashi I, Shimizu A, Sasaki T, Minoshima S, Shimizu N. 2004. Comparative genomics of the keratin-associated protein (KAP) gene clusters in human, chimpanzee, and baboon. Mamm. Genome Off. J. Int. Mamm. Genome Soc. 15:179–192.

Shibuya K, Obayashi I, Asakawa S, Minoshima S, Kudoh J, Shimizu N. 2004. A cluster of 21 keratin-associated protein genes within introns of another gene on human chromosome 21q22.3. Genomics 83:679–693.

Shimomura Y, Aoki N, Rogers MA, Langbein L, Schweizer J, Ito M. 2002. hKAP1.6 and hKAP1.7, two novel human high sulfur keratin-associated proteins are expressed in the hair follicle cortex. J. Invest. Dermatol. 118:226–231.

Shimomura Y, Ito M. 2005. Human hair keratin-associated proteins. J. Investig. Dermatol. Symp. Proc. Soc. Investig. Dermatol. Inc Eur. Soc. Dermatol. Res. 10:230–233.

Slack JL, Schooley K, Bonnert TP, Mitcham JL, Qwarnstrom EE, Sims JE, Dower SK. 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.

Smith J, Paton IR, Hughes DC, Burt DW. 2005. Isolation and mapping the chicken zona pellucida genes: an insight into the evolution of orthologous genes in different species. Mol. Reprod. Dev. 70:133–145.

Snook RR, Hosken DJ, Karr TL. 2011. The biology and evolution of polyspermy: insights from cellular and functional studies of sperm and centrosomal behavior in the fertilized egg. Reproduction 142:779–792.

Soranzo N, Sari Gorla M, Mizzi L, De Toma G, Frova C. 2004. Organisation and structural evolution of the rice glutathione S-transferase gene family. Mol. Genet. Genomics MGG 271:511–521.

Spargo SC, Hope RM. 2003. Evolution and nomenclature of the zona pellucida gene family. Biol. Reprod. 68:358–362.

Spehr M, Gisselmann G, Poplawski A, Riffell JA, Wetzel CH, Zimmer RK, Hatt H. 2003. Identification of a Testicular Odorant Receptor Mediating Human Sperm Chemotaxis. Science 299:2054–2058.

Steiger SS, Fidler AE, Mueller JC, Kempenaers B. 2010. Evidence for adaptive evolution of olfactory receptor genes in 9 bird species. J. Hered. 101:325–333.

Steiger SS, Kuryshev VY, Stensmyr MC, Kempenaers B, Mueller JC. 2009. A comparison of reptilian and avian olfactory receptor gene repertoires: species-specific expansion of group gamma genes in birds. BMC Genomics 10:446.

Steinberger P, Andris-Widhopf J, Buhler B, Torbett BE, Barbas CF. 2000. Functional deletion of the CCR5 receptor by intracellular immunization produces cells that are refractory to CCR5-dependent HIV-1 infection and cell fusion. Proc. Natl. Acad. Sci. U. S. A. 97:805–810.

Steinert PM, North AC, Parry DA. 1994. Structural features of keratin intermediate filaments. J. Invest. Dermatol. 103:19S–24S.

Stenn KS, Paus R. 2001. Controls of hair follicle cycling. Physiol. Rev. 81:449–494.

Sullivan C, Charette J, Catchen J, Lage CR, Giasson G, Postlethwait JH, Millard PJ, Kim CH. 2009. The gene history of zebrafish tlr4a and tlr4b is predictive of their divergent functions. J. Immunol. Baltim. Md 1950 183:5896–5908.

Sun Y-B, Zhou W-P, Liu H-Q, Irwin DM, Shen Y-Y, Zhang Y-P. 2012. Genome-Wide Scans for Candidate Genes Involved to the Aquatic Adaptation of Dolphins. Genome Biol. Evol. [Internet]. Available from: http://gbe.oxfordjournals.org/content/early/2012/12/16/gbe.evs123

Sundaram AY, Kiron V, Dopazo J, Fernandes JM. 2012. Diversification of the expanded teleost-specific toll-like receptor family in Atlantic cod, Gadus morhua. BMC Evol. Biol. 12:256.

Suutari M, Majaneva M, Fewer DP, Voirin B, Aiello A, Friedl T, Chiarello AG, Blomster J. 2010. Molecular evidence for a diverse green algal community growing in the hair of sloths and a specific association with Trichophilus welckeri (Chlorophyta, Ulvophyceae). BMC Evol. Biol. 10:86.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. 3:137–144.

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. 98:2509–2514.

Szalai G, Kellős T, Galiba G, Kocsy G. 2009. Glutathione as an Antioxidant and Regulatory Molecule in Plants Under Abiotic Stress Conditions. J. Plant Growth Regul. 28:66–80.

Takeda K, Akira S. 2004. Microbial recognition by Toll-like receptors. J. Dermatol. Sci. 34:73–82.

Takeuchi O, Akira S. 2007. Recognition of viruses by innate immunity. Immunol. Rev. 220:214–224.

Takeuchi O, Sato S, Horiuchi T, Hoshino K, Takeda K, Dong Z, Modlin RL, Akira S. 2002. Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins. J. Immunol. Baltim. Md 1950 169:10–14.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.

Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24:1596–1599.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28:2731–2739.

Tarín JJ. 2000. Fertilization in Protozoa and Metazoan Animals: A Comparative Overview. In: Tarín JJ, MD AC, editors. Fertilization in Protozoa and Metazoan Animals. Springer Berlin Heidelberg. p. 277–314. Available from: http://link.springer.com/chapter/10.1007/978-3-642-58301-8_7

Tatusova TA, Madden TL. 1999. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol. Lett. 174:247–250.

Tawfik OK, S D. 2010. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. Annu. Rev. Biochem. 79:471–505.

Temperley ND, Berlin S, Paton IR, Griffin DK, Burt DW. 2008. Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss. BMC Genomics 9:62.

Thewissen JGM, Cooper LN, George JC, Bajpai S. 2009. From Land to Water: the Origin of Whales, Dolphins, and Porpoises. Evol. Educ. Outreach 2:272–288.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Res. 22:4673–4680.

Tregenza T, Butlin RK, Wedell N. 2000. Evolutionary biology: Sexual conflict and speciation. Nature 407:149–150.

Vallender EJ, Lahn BT. 2004. Positive selection on the human genome. Hum. Mol. Genet. 13:R245–R254.

Vandebergh W, Bossuyt F. 2011. Radiation and Functional Diversification of Alpha Keratins during Early Vertebrate Evolution. Mol. Biol. Evol. 2011.

Venter JC, Adams MD, Myers EW, et al. 2001. The sequence of the human genome. Science 291:1304–1351.

Vincent JFV. 2002. Survival of the cheapest. Mater. Today 5:28–41.

Vo LH, Hedrick JL. 2000. Independent and hetero-oligomeric-dependent sperm binding to egg envelope glycoprotein ZPC in Xenopus laevis. Biol. Reprod. 62:766–774.

Vuilleumier S, Pagni M. 2002. The elusive roles of bacterial glutathione S-transferases: new lessons from genomes. Appl. Microbiol. Biotechnol. 58:138–146.

Waddington D, Gribbin C, Sterling RJ, Sang HM, Perry MM. 1998. Chronology of events in the first cell cycle of the polyspermic egg of the domestic fowl (Gallus domesticus). Int. J. Dev. Biol. 42:625–628.

Walsh JB. 2001. Multigene Families : Evolution. Life Sci.:1–6.

Wang B, Ekblom R, Strand TM, Portela-Bens S, Höglund J. 2012. Sequencing of the core MHC region of black grouse (Tetrao tetrix) and comparative genomics of the galliform MHC. BMC Genomics 13:553.

Wang C, Bammler TK, Guo Y, Kelly EJ, Eaton DL. 2000. Mu-class GSTs are responsible for aflatoxin B(1)-8, 9-epoxide-conjugating activity in the nonhuman primate macaca fascicularis liver. Toxicol. Sci. Off. J. Soc. Toxicol. 56:26–36.

Wang L-F, Walker PJ, Poon LLM. 2011. Mass extinctions, biodiversity and mitochondrial function: are bats “special” as reservoirs for emerging viruses? Curr. Opin. Virol. 1:649–657.

Warren WC, Hillier LW, Graves JAM, et al. 2008. Genome analysis of the platypus reveals unique signatures of evolution. Nature 453:175–183.

Wassarman PM, Jovine L, Litscher ES. 2004. Mouse zona pellucida genes and glycoproteins. Cytogenet. Genome Res. 105:228–234.

Wassarman PM, Litscher ES. 2008. Mammalian fertilization:the eggs multifunctional zona pellucida. Int. J. Dev. Biol. 52:665–676.

Weadick CJ, Chang BS. 2012. Complex patterns of divergence among green-sensitive (RH2a) African cichlid opsins revealed by Clade model analyses. BMC Evol. Biol. 12:206.

Weadick CJ, Chang BSW. 2012. An improved likelihood ratio test for detecting site-specific functional divergence among clades of protein-coding genes. Mol. Biol. Evol. 29:1297–1300.

Whitten IH, Frank E. 2005. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco.

Wicher D, Schäfer R, Bauernfeind R, Stensmyr MC, Heller R, Heinemann SH, Hansson BS. 2008. Drosophila odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. Nature 452:1007–1011.

Williams Z, Wassarman PM. 2001. Secretion of mouse ZP3, the sperm receptor, requires cleavage of its polypeptide at a consensus furin cleavage-site. Biochemistry (Mosc.) 40:929–937.

Wishart GJ, Horrocks AJ. 2000. Fertilization in Birds. In: Tarín JJ, MD AC, editors. Fertilization in Protozoa and Metazoan Animals. Springer Berlin Heidelberg. p.

193–222. Available from: http://link.springer.com/chapter/10.1007/978-3-642-58301-8_5

Wlasiuk G, Nachman MW. 2010. Adaptation and constraint at Toll-like receptors in primates. Mol. Biol. Evol. 27:2172–2186.

Wong JL, Wessel GM. 2006. Defending the zygote: search for the ancestral animal block to polyspermy. Curr. Top. Dev. Biol. 72:1–151.

Wong WSW, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics 168:1041–1051.

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA. 2003. TreeSAAP: Selection on Amino Acid Properties using phylogenetic trees. Bioinformatics 19:671–672.

Wouters MA, Rigoutsos I, Chu CK, Feng LL, Sparrow DB, Dunwoodie SL. 2005. Evolution of distinct EGF domains with specific functions. Protein Sci. Publ. Protein Soc. 14:1091–1103.

Wu B, Dong D. 2012. Human cytosolic glutathione transferases: structure, function, and drug discovery. Trends Pharmacol. Sci. 33:656–668.

Wu D-D, Irwin DM, Zhang Y-P. 2008. Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol. Biol. 8:241.

Wu D-D, Irwin DM, Zhang Y-P. 2009. Correction: Molecular evolution of the keratin associated protein gene family in mammals, role in the evolution of mammalian hair. BMC Evol. Biol. 9:213.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26:1–7.

Xia X. 2013. DAMBE5: A Comprehensive Software Package for Data Analysis in Molecular Biology and Evolution. Mol. Biol. Evol.

Xu Q, Li G, Cao L, Wang Z, Ye H, Chen X, Yang X, Wang Y, Chen L. 2012. Proteomic characterization and evolutionary analyses of zona pellucida domain-containing proteins in the egg coat of the cephalochordate, Branchiostoma belcheri. BMC Evol. Biol. 12:239.

Yahagi S, Shibuya K, Obayashi I, Masaki H, Kurata Y, Kudoh J, Shimizu N. 2004. Identification of two novel clusters of ultrahigh-sulfur keratin-associated protein genes on human chromosome 11. Biochem. Biophys. Res. Commun. 318:655–664.

Yan H, Jia H, Gao H, Guo X, Xu B. 2013. Identification, genomic organization, and oxidative stress response of a sigma class glutathione S-transferase gene (AccGSTS1) in the honey bee, Apis cerana cerana. Cell Stress Chaperones 18:415–426.

Yang null, Bielawski null. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. 15:496–503.

Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. 46:409–418.

Yang Z, Wong WSW, Nielsen R. 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol. Biol. Evol. 22:1107–1118.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. CABIOS 13:555–556.

Yang Z. 1998a. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

Yang Z. 1998b. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. 51:423–432.

Yang Z. 2007a. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang Z. 2007b. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yarovinsky F, Zhang D, Andersen JF, et al. 2005. TLR11 activation of dendritic cells by a protozoan profilin-like protein. Science 308:1626–1629.

Young JM, Massa HF, Hsu L, Trask BJ. 2010. Extreme Variability Among Mammalian V1R Gene Families. Genome Res. 20:10–18.

Yu Z, Gordon SW, Nixon AJ, Bawden CS, Rogers MA, Wildermoth JE, Maqbool NJ, Pearson AJ. 2009. Expression patterns of keratin intermediate filament and keratin associated protein genes in wool follicles. Differ. Res. Biol. Divers. 77:307–316.

Zhang G, Cowled C, Shi Z, et al. 2013. Comparative Analysis of Bat Genomes Provides Insight into the Evolution of Flight and Immunity. Science 339:456–460.

Zhang J, Liu S, Rajendran KV, Sun L, Zhang Y, Sun F, Kucuktas H, Liu H, Liu Z. 2013. Pathogen recognition receptors in channel catfish: III phylogeny and expression analysis of Toll-like receptors. Dev. Comp. Immunol. 40:185–194.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. 22:2472–2479.

Zhang J. 2003a. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Zhang J. 2003b. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Zhang X, Firestein S. 2002. The olfactory receptor gene superfamily of the mouse. Nat. Neurosci. 5:124–133.

Zhang Y, Zhang X, O'Hare TH, et al. 2011. A comparative physical map reveals the pattern of chromosomal evolution between the turkey (Meleagris gallopavo) and chicken (Gallus gallus) genomes. BMC Genomics 12:447.

Zhou H, Gong H, Yan W, Luo Y, Hickford JGH. 2012. Identification and sequence analysis of the keratin-associated protein 24-1 (KAP24-1) gene homologue in sheep. Gene 511:62–65.

Zimek A, Weber K. 2005. Terrestrial vertebrates have two keratin gene clusters; striking differences in teleost fish. Eur. J. Cell Biol. 84:623–635.

Zhang et al. 2014. Avian Genome Consortium. 2014. Comparative Genomics Reveals Insights into Avian Genome Evolution and Adaptation. Science (accepted).

IV. Appendix

10

Appendix 1
(Supplementary materials for chapter 2)

Additional file 2.1 –Table 1 – The excel file shows the genomic coordinates of the KRTAP gene repertoires in 22 mammalian species studied. The Gene ID corresponds to the genomic location. (Available on request)

Additional file 2.2 –Figure 2 - The phylogeny of high cysteine KRTAP genes in 22 mammalian species. Neighbor-joining method with P-distance and interiors branch test with 1,000 replications (shown on the branches) was employed to build the trees. Figure 23 and 24 shows loss of, one to one orthologous relationship between two species due to concerted evolution. The KRTAP members are labeled with species abbreviation, Gene ID and KRTAP subfamily (Additional file 1) Figure 1-21 are in order, Gorilla, Pongo, Gibbon, Mormoset, Tarsies, Mouse lemur, Bushbaby, Treeshrew, Cavia, rabbit, Cow, Pig, Alpaca, Horse, Panda, Bat, Hedgehog, Elephant, Armadillo, Sloth and Wallaby. Figure 22 (Gorilla and Gibbon) and 23 (Gorilla and Cavia) shows reduced orthology with increase in divergence time. Figure 24 shows relationship between all HGT members in 22 genomes. (Available on request)

Additional file 2.3 –Table 3 - Gene pairs under significant gene conversion, as detected by GeneConv program (Available on request)

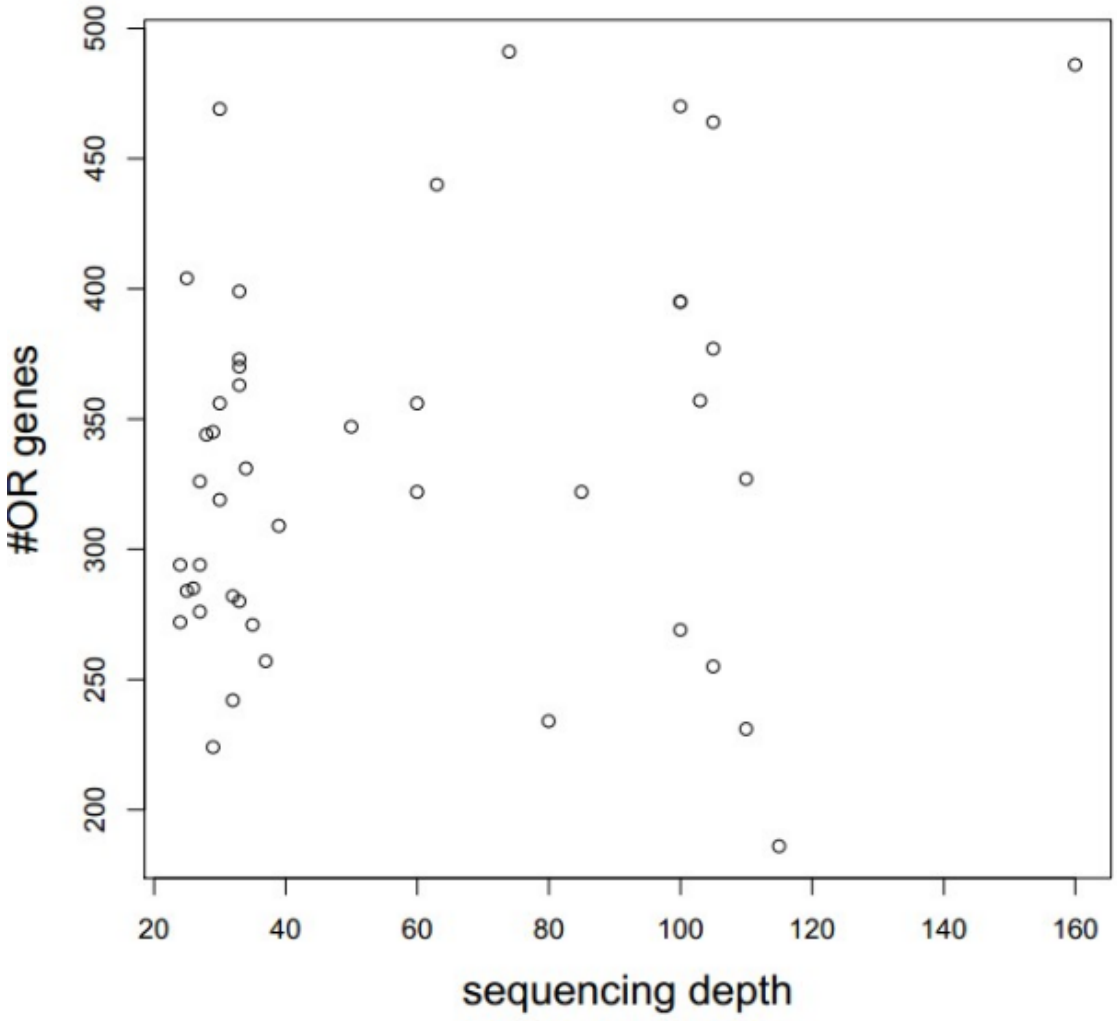
Additional file 2.4 –Table 4 - Results of RDP3 showing unique recombination events with statistical significance P value of less than 0.01 employing Bonferroni correction

Additional file 4 – Table 4 - Results of RDP3 showing unique recombination events with statistical significance P value of less than 0.01 employing Bonferroni correction

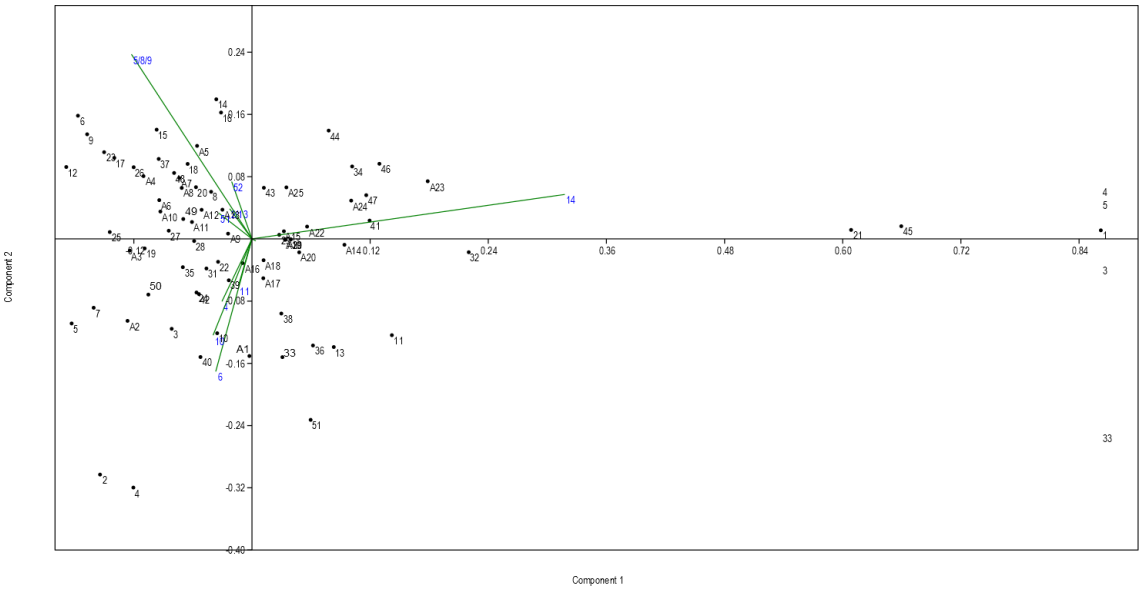
		Gorilla	Orangutan	Gibbon	Mormoset	Tarsier	Mouse lemur	Bushbaby	Treeshrew	Cavia	Rabbit	Dolphin	Cow	Pig	Alpaca	Horse	Panda	Bat	Hedgehog	Elephant	Armadillo	Sloth	Wallaby
HS-KRTAP	KRTAP1		1(8)	1(0)	0(0)	0(0)	0(0)	1(5)		2(3)	1(7)		1(6)	1(3)	0(0)	0(0)	0(0)	0(0)		2(6)		0(0)	
	KRTAP2	0(0)	0(0)	0(0)	0(0)			0(0)	0(0)	0(0)	0(0)		0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)	0(0)
	KRTAP3	0(0)	0(0)	0(0)	1(4)	0(0)	0(0)			1(4)	1(6)		0(0)	2(8)	0(0)	1(5)	0(0)	2(10)	1(6)	0(0)		0(0)	0(0)
	KRTAP4	2(9)	6(33)	5(28)	7(28)	2(6)	7(26)		5(23)	9(50)	18(101)		1(1)	5(9)	1(3)	6(49)	1(0)	2(11)	2(3)	7(21)	5(1)	3(9)	4(13)
	KRTAP5	1(1)	5(16)	5(19)	5(13)	3(14)	7(24)		5(46)	3(16)	13(31)		5(24)			3(10)	0(0)		2(11)	5(26)		4(10)	0(0)
	KRTAP9	3(21)	2(3)	1(3)	1(4)	2(5)	2(11)	0(0)	3(6)	0(0)	0(0)		2(8)	1(5)	0(0)	4(15)		2(11)	5(28)	0(0)	0(16)	7(54)	
	KRTAP10	5(16)	11(25)	9(28)	9(31)	2(6)	6(31)		9(21)	8(40)	4(14)		16(42)	8(29)		7(31)		4(10)		8(37)			
	KRTAP11																						
	KRTAP12	1(2)								0(0)			4(12)	2(11)		6(26)	1(10)		2(5)	6(16)		1(3)	4(16)
	KRTAP13	1(1)	0(0)	0(0)	1(3)		3(12)		1(0)	3(11)	5(13)		5(17)	1(3)	3(4)	6(23)	3(11)	3(13)	5(13)	3(7)	3(7)		
	KRTAP28	0(0)	0(0)	2(3)	1(1)	0(0)	0(0)		0(0)	2(4)	2(4)		0(0)	0(0)	1(2)	0(0)	0(0)	0(0)	0(0)	0(0)	1(2)	0(0)	3(6)
HGT-KRTAP	KRTAP6			1(3)	0(0)	0(0)	1(4)	0(0)	3(6)	0(0)	0(0)		0(0)	1(1)	0(0)		1(5)	0(0)	0(0)	0(0)			
	KRTAP7																						
	KRTAP8											0(0)		0(0)									
	KRTAP19	0(0)	0(0)	0(0)	0(0)	0(0)		1(2)	2(2)				1(1)			3(8)	2(9)	0(0)	1(5)	2(6)		0(0)	
	KRTAP20					0(0)		0(0)	0(0)	2(0)	2(10)		0(0)	0(0)	0(0)	0(0)	1(2)	0(0)		0(0)	2(13)	0(0)	0(5)
	KRTAP21	0(0)	0(0)		0(0)					0(0)	0(0)		1(2)			0(0)	2(4)	4(0)	1(5)	1(3)	1(4)	4(20)	

11

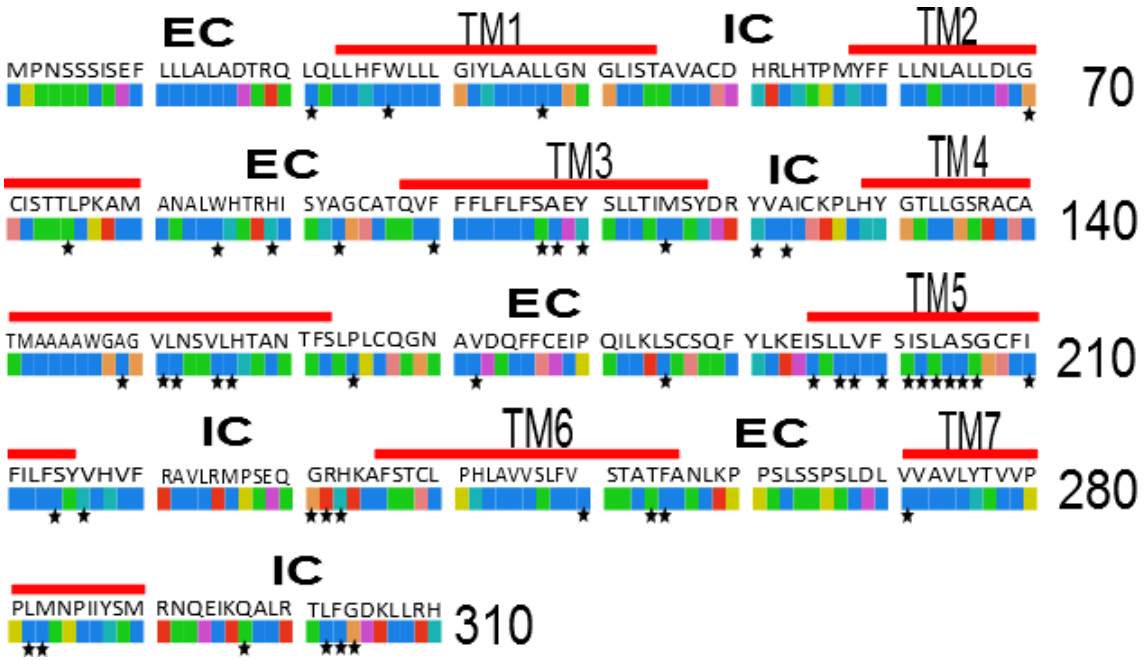
Appendix 2
(Supplementary materials for chapter 3)



Supplementary Figure 3.1. The numbers of identified OR genes ('Total' column of Table 1) versus the sequencing depths, not including the three published bird and reptilian genomes.



Supplementary Figure 3.2. The PCA showing the location of reconstructed ancestral states considering the ecogroups in Figure 3.4.



Supplementary Figure 3.3. Location of the positive selected sites (star) in the OR family 14 detected by more than two methods and found in eight bird species (see supplementary table 2).

Supplementary Table 3.1. The Number of intact, partial, and pseudo genes and sequencing depths for each genome analyzed.

Code	S.NO	Species	Partial	Pre-mature	Intact	Total Genes	Sequencing Depth
1	1	***Acanthisitta chloris***	125	86	13	**224**	29X
2	2	***Alligator mississippiensis***	234	361	406	**1001**	-
3	3	***Anas platyrhynchos***	138	160	49	**347**	50X
4	4	***Antrostomus carolinensis***	153	159	44	**356**	30X
5	5	***Apaloderma vittatum***	114	198	32	**344**	28X
6	6	***Aptenodytes forsteri***	176	151	29	**356**	60X
7	7	***Balearica regulorum***	158	186	55	**399**	33X
8	8	***Buceros rhinoceros***	133	125	13	**271**	35X
9	9	***Calypte anna***	156	149	22	**327**	110X
a	10	***Cariama cristata***	139	122	33	**294**	24X
b	11	***Cathartes aura***	173	184	47	**404**	25X
c	12	***Chaetura pelagica***	140	190	27	**357**	103X
d	13	***Charadrius vociferus***	167	172	56	**395**	100X
e	14	***Chelonia mydas***	209	507	250	**966**	-
f	15	***Chlamydotis undulata***	114	130	32	**276**	27X
g	16	***Colius striatus***	164	120	10	**294**	27X
h	17	***Columba livia***	150	239	51	**440**	63X
I	18	***Corvus brachyrhynchos***	108	110	16	**234**	80X
j	19	***Cuculus canorus***	131	105	33	**269**	100X
k	20	***Egretta garzetta***	163	228	100	**491**	74X
l	21	***Eurypyga helias***	130	123	27	**280**	33X
m	22	***Falco peregrinus***	144	288	32	**464**	105X
n	23	***Fulmarus glacialis***	187	157	29	**373**	33X
o	24	***Gallus gallus***	150	314	214	**678**	7X
p	25	***Gavia stellata***	163	167	40	**370**	33X
q	26	***Geospiza fortis***	83	92	11	**186**	115X
r	27	***Haliaeetus albicilla***	128	125	32	**285**	26X
s	28	***Haliaeetus leucocephalus***	72	157	35	**264**	Unknown
t	29	***Leptosomus discolor***	129	127	26	**282**	32X
u	30	***Manacus vitellinus***	124	98	9	**231**	110X
v	31	***Meleagris gallopavo***	124	160	31	**315**	17X
w	32	***Melopsittacus undulatus***	116	335	35	**486**	160X
x	33	***Merops nubicus***	126	110	21	**257**	37X
y	34	***Mesitornis unicolor***	151	175	19	**345**	29X
z	35	***Nestor notabilis***	117	96	29	**242**	32X
A	36	***Nipponia nippon***	173	166	38	**377**	105X
B	37	***Ophisthocomus hoazin***	182	224	64	**470**	100X
C	38	***Pelecanus crispus***	148	152	31	**331**	34X
D	39	***Phaethon lepturus***	137	141	31	**309**	39X
E	40	***Phalacrocorax carbo***	146	107	19	**272**	24X
F	41	***Phoenicopterus ruber***	158	156	49	**363**	33X
G	42	***Picoides pubescens***	125	107	23	**255**	105X
H	43	***Podiceps cristatus***	116	161	42	**319**	30X
I	44	***Pterocles gutturalis***	145	111	28	**284**	25X
J	45	***Pygoscelis adeliae***	164	137	21	**322**	60X
K	46	***Struthio camelus***	129	139	54	**322**	85X
L	47	***Taeniopygia guttata***	147	379	164	**690**	6X
M	48	***Tauraco erythrolophus***	178	250	41	**469**	30X
N	49	***Tinamus guttatus***	152	185	58	**395**	100X
O	50	***Tyto alba***	144	147	35	**326**	27X
		Totals	7233	8768	2606	**18607**	

Supplementary Table 3.2

Positive selected sites detected by five Maximum likelihood approaches, along with integrated sites, in expanded paralogous of OR14 in birds and OR 51, OR52 in turtle.

Species	Gene	No. of Sequences	Positive selected Sites					
			SLAC	FEL	REL	MEME	FUBAR	Integrative
Anas platyrhynchos	*OR14*	15	NO	16, **47**, **110**, 115, **156, 172, 282, 303**	96, **107**, **110, 156, 172, 282**	44, **47**, 50, 76, **107**, **110,** 141, 149, **156**, 157, **172**, 197, 205, 214, 268, **282, 303, 304**	**304**	16, 44, **47**, 50, 76, 96, **107**, **110**, 115, 141, 149, **156**, 157, **172**, 197, 205, 214, 268, **282, 303, 304,**
Balearica regulorum	*OR14*	11	NO	14, 71, 96, **115**, **131, 172, 201, 231**	**115**, **131, 172**, **201**, **231**, 297	52, 70, 75, 79, 93, **115, 131,** 141, 149, 156, **172**, 190, **201,** 205, **231**, 299	**172**	14, 52, 70, 71, 75, 79, 93, 96, **115, 131**, 141, 149, 156, **172**, 190, **201**, 205, **231**, 297, 299
Columba livia	*OR14*	14	NO	**75, 93, 110, 133, 172, 195, 197, 282, 283, 304**	**93, 105,** 106, **110, 195,** 203, **204**, 250, **282, 283, 304**	35, **75**, 87, **110, 133,** 145, 163, **172, 195, 197**, 198, 201, **204**, 214, 249, 258, **283**, 297, **304**	**93, 105, 110, 195,** 205**, 304**	35, **75**, 87, **93**, 105, 106, **110, 133**, 145, 163, **172, 195, 197**, 198, 201, 203, **204**, 205, 214, 249, 250, 258, **282, 283**, 297, **304**
Egretta garzetta	*OR14*	40	87, 190, **210, 214, 250**	**107, 110, 149, 152, 198, 203, 210, 214, 250, 255, 282**	**107, 110,** 111**, 149, 198, 203, 210, 282**	18, 21, 25, 73, 78, 93, **107**, 113, 145, **149, 152**, 155, **198, 203**, 204, **210, 214, 250**, 251, 253, **255**	**107, 203, 210, 214, 250, 282**	18, 21, 25, 73, 78, 87, 93, **107, 110**, 111, 113, 145, **149, 152**, 155, 190, **198, 203**, 204, **210, 214, 250**, 251, 253, **255, 282**
Gallus gallus	*OR14*	40	**70, 86, 89, 93, 131, 151, 164, 195, 203, 232, 250, 283**	**70, 86, 89, 93, 131, 151, 156, 164, 195, 203, 205, 206, 217, 232, 250, 254, 283, 297**	**86, 89, 93, 151, 164, 195, 200, 203, 217, 250, 254, 271, 283, 297**	50, **70, 86, 89, 93,** 99, 107, 110, **131, 151, 156,** 161, **164,** 192, **195**, 201, **203, 205, 206**, 214, **217, 232, 250, 254**, 265, **271, 283, 297**	**86, 89, 93, 131, 151, 164, 195, 200, 203, 206, 250, 254, 271, 283, 297**	50, **70, 86, 89, 93**, 99, 107, 110, **131, 151, 156**, 161, **164,** 192, **195, 200**, 201, **203, 205, 206**, 214, **217, 232, 250, 254,** 265, **271, 283, 297**
Ophisthocomus hoazin	*OR14*	20	**131, 202, 204, 206**	**37, 107, 131, 149, 200, 202, 204, 206, 250**	**200, 203, 206**	25, **37**, 71, **107, 131**, 134, **149,** 174, 198, **200, 202, 204, 206**, 230, **250**, 253, 272, 300, 302, 304	**202, 204**	25, **37**, 71, **107, 131,** 134, **149**, 174, 198, **200, 202**, 203, **204, 206**, 230, **250,** 253, 272, 300, 302, 304
Taeniopygia guttata	*OR14*	40	**108, 110, 201, 206, 233, 302**	47, 1**00, 108, 110, 185, 201, 202, 204, 206, 233, 302, 304**	**110, 201, 206, 255, 302**	83, **100**, 107, **108, 110**, 171, **185**, 186, 189, 199, **201, 202, 204, 206, 233**, 267, **302, 304**	89, **100, 108, 110,** 157, **201, 204, 206, 233, 255, 302**	47, 83, 89, **100**, 107, **108, 110**, 157, 171, **185**, 186, 189, 199, **201, 202, 204, 206, 233, 255,** 267, **302, 304**
Tinamus guttatus	*OR14*	15	**89**	**31, 47, 89, 149, 155, 197, 201, 214, 267,** 268	**197**, 202, 205, **254**	13, 14, **31**, 35, 38, 43, **47**, 50, 69, **89**, 101, 107, **149**, **155**, 159, 161, **197**, **201**, 204, **214**, 224, 243, **254**, **267**, 269, 295	**89**,**197**	13, 14, **31**, 35, 38, 43, **47**, 50, 69, **89**, 101, 107, **149**, **155**, 159, 161, **197**, **201**, 202, 204, 205, **214**, 224, 243, **254**, **267,** 268, 269, 295
Clenonia mydas	*OR51*	40	**10, 35, 55, 70, 78, 150, 169, 204, 206, 296,**	**10,**,**15,35**,**42, 55, 70,75,78,150 ,155, 158,163,169, 203,204 206,263**,**296**	**10,**,**35,70, 78,112 129, 150,** 157,**158, 163, 169,** 195, **203, 204, 206, 262**, 277, 281,	**10,**,**15,35,42**, **55**,**70,75,**91,110**,** 111,**112,129,150,** 154,**155,158,169,** 182,**203,204,206,** 221,**262**,**263**,**296**	**10,35,55,70,75, 78,150,169,203, 204,206,296,**	**10, 15,35, 42, 55,70,75,78,**91 110**,**,111,**112,129,150,**154, **155,**157,**158,163,169,**182, 195,**203,204,206,**221,**262**, **263,**277,281,**296,**
Clenonia mydas	*OR52*	40	**9, 24, 47, 74**, **96, 97, 109, 110, 135,**	**9, 24, 47, 96, 97, 109, 110,** 119, **135,** 141, **161, 168, 235**,	34, **47, 74**, 80, **96, 97, 109, 110, 111**, 158, **161, 168**, **221, 298**,	**9,** 12, **24, 47, 74,** 90, **96, 97, 109,110, 111, 135**, 150, **161,** 167, 168, 208, **221**, 231, 232,**235**,255, 257, 271,293, 294, **298**,	**9, 24, 47, 74, 96, 97, 109, 110, 135, 161,**	**9,** 12, **24,** 34, **47, 74,** 80, 90, **96, 97, 109, 110, 111,** 119,**135**, 141, 150, 158, **161**, 167, **168**, 208, **221**, 232, **235**, 257, 271, 293, 294, **298**,

The sites detected by more than two methods are in bold and underlined.

Supplementary Table 3.3 PARRIS results for evidence of positive selection in expanded OR14 family paralogs

Species	Gene	No. Of Seq	Null model, Log(L) No selection	Alternative model, Log(L) Positive selection	LRT	P value	Positive selection
Anas platyrhynchos	*OR14*	15	*-6442.53*	*-6442.53*	*0*	*1*	*No*
Balearica regulorum	*OR14*	11	*-5049.09*	*-5049.09*	*0*	*1*	*No*
Columba livia	*OR14*	14	*-5460.19*	*-5460.19*	*0*	*1*	*No*
Egretta garzetta	*OR14*	40	*-19877.3*	*-19872.3*	*10.00*	*0.00*	*Yes*
Gallus gallus	*OR14*	40	*-7969.27*	*-7961.50*	*15.55*	*0.00*	*Yes*
Ophisthocomus hoazin	*OR14*	20	*-8354.67*	*-8354.67*	*0*	*1*	*No*
Taeniopygia guttata	*OR14*	40	*-5676.27*	*-5669.36*	*13.82*	*0.00*	*Yes*
Tinamus guttatus	*OR14*	15	*-6547.56*	*-6547.56*	*0*	*1*	*No*
Clenonia mydas	*OR51*	40	*-11178.1*	*-11176.1*	*3.885*	*0.143*	*No*
Clenonia mydas	*OR52*	40	*-11394.8*	*-11386.3*	*17.125*	*0.00*	*Yes*

Supplementary Table 3.4 (A) Location of positive selected sites in 7TM domains as detected by SOSOI in OR51 gene paralogs of ***Clenonia mydas***

No.	Transmembrane region	N terminal	C terminal
1	SIPF***C***FMYV***I***SIVGNSVILFIIK	31	53
2	FLSML***A***LTDL***A***LS***I***TTIPTILGI	65	87
3	AQLFFIHLLQ***Y***IESSVLLLMAFD	102	124
4	L***V***SMLR***A***MV***L***ILPL***P***FLLKW***F***RY	149	171
5	LDSLLIFLSYVMILKTVLSIASH	212	234
6	LNTCVSHLCALLLFYTPEISLS***V***	241	263
7	ILLGYMALLLPPLMNPIVY***S***VR	277	298

Supplementary Table 3.4 (B) Location of positive selected sites in 7TM domains as detected by SOSOI in OR52 gene paralogs of ***Clenonia mydas***

No.	Transmembrane region	N terminal	C terminal
1	ISIPFSISYIIGLLGN***F***MLLFVV	31	53
2	MLALT***D***IAMSTFVVPKALCLFWF	69	91
3	MFFL***H***TVSIMQSAILVIMAFDRY	105	127
4	LVGLIKAVLFTLPMPLLL***S***RLPF	150	172
5	TFLVIGLDLTLIAL***S***YGLIIRAV	207	229
6	QKALNTCIAHIFVMLMYYLPGLF	239	261
7	PHIHIILNNLYLLVPPILNPIIY	274	296

Supplementary Table 3.5 Details of recombination breakpoint signals detected in OR paralogs using GARD

Species	Gene	No. Of Seq	Δc-AIC*	Number of breakpoints	Breakpoint location
Anas platyrhynchos	***OR14***	**15**	***240.499***	***4***	***74, 255, 334, 504***
Balearica regulorum	***OR14***	**11**	***156.479***	***2***	***195, 299***
Columba livia	***OR14***	**14**	***124.15***	***3***	***195, 320, 391***
Egretta garzetta	***OR14***	**40**	***748.467***	***2***	***195, 341***
Gallus gallus	***OR14***	**40**	***1028.48***	***2***	***366, 547***
Ophisthocomus hoazin	***OR14***	**20**	***192.023***	***2***	***119, 288***
Taeniopygia guttata	***OR14***	**40**	***363.344***	***3***	***211, 445***
Tinamus guttatus	***OR14***	**15**	***74.7407***	***2***	***484, 613***
Clenonia mydas	***OR51***	**40**	***454.853***	***3***	***208, 378***
Clenonia mydas	***OR52***	**40**	***172.761***	***1***	***321***

*Difference in Akaike information criteria of single tree (non-recombination model) and multi tree (recombination model) supports recombination. Note: Supplementry file for sequence will be provided on request.

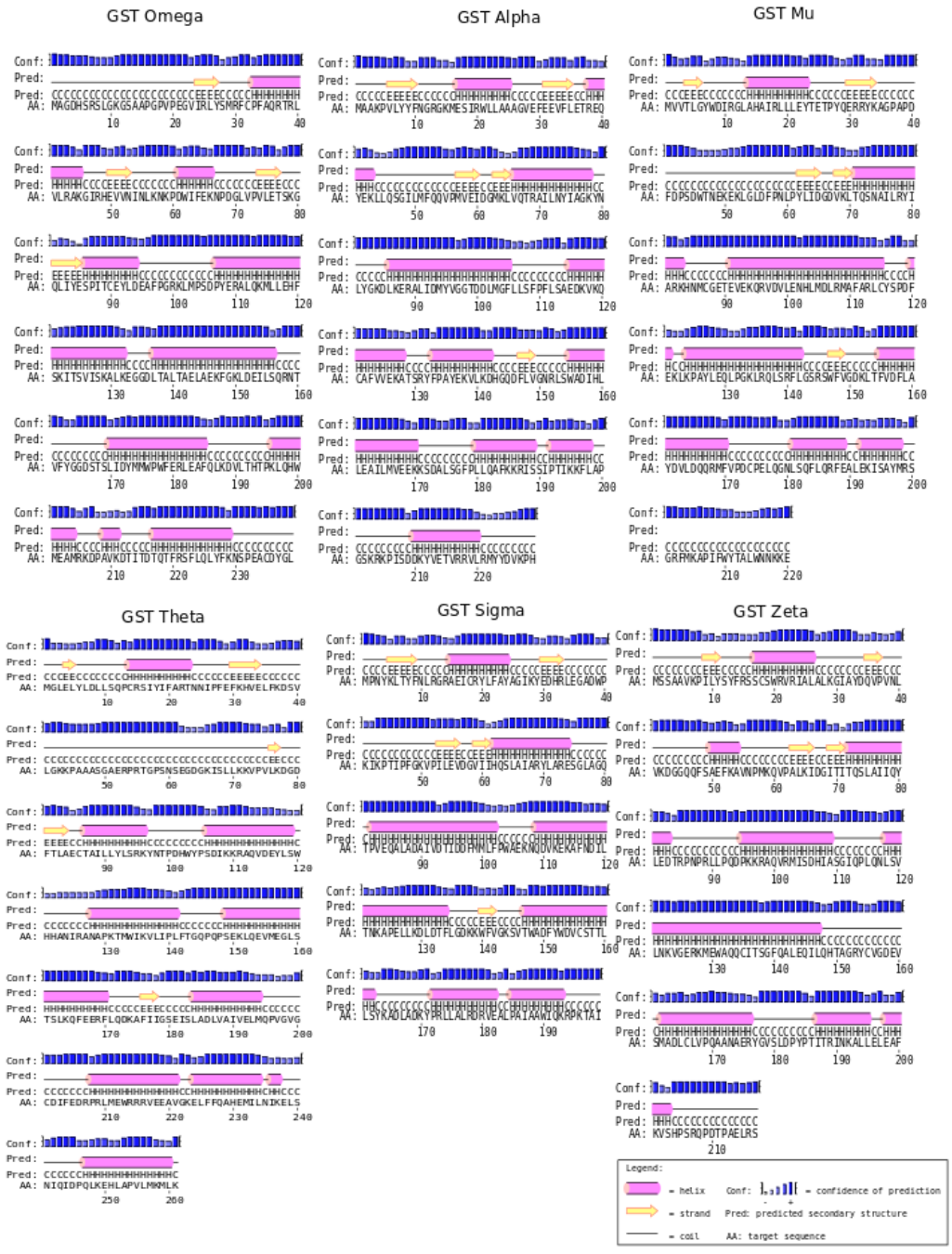
12

Appendix 3
(Supplementary materials for chapter 4

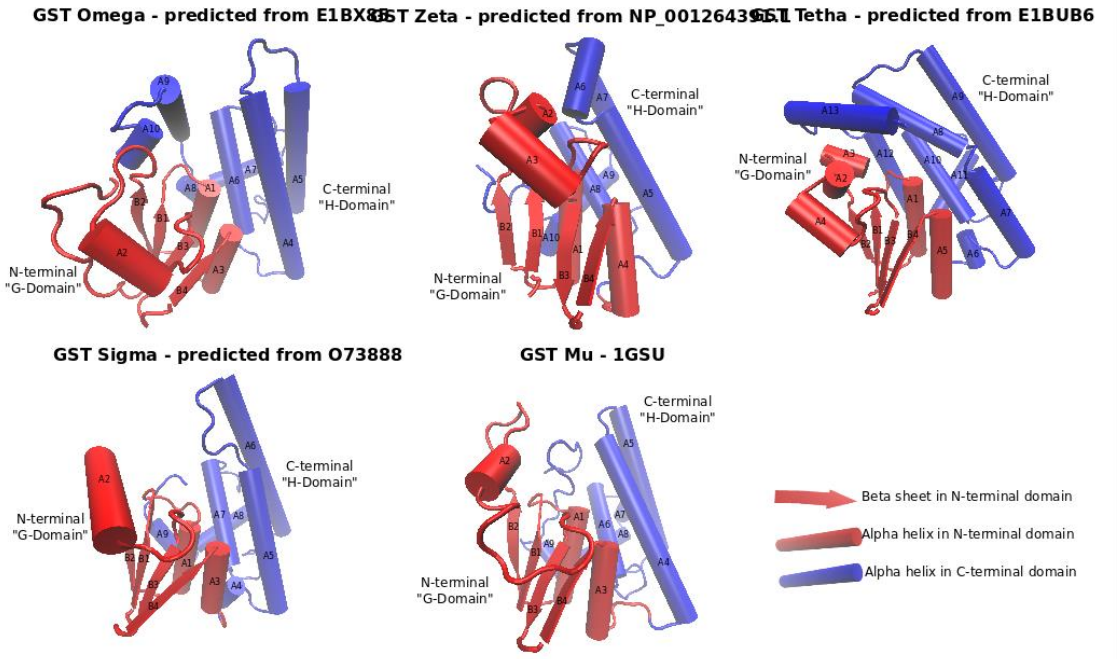
Additional files

Supplementary table 4.1 - The details of sequences used in present study (Availabe on request)

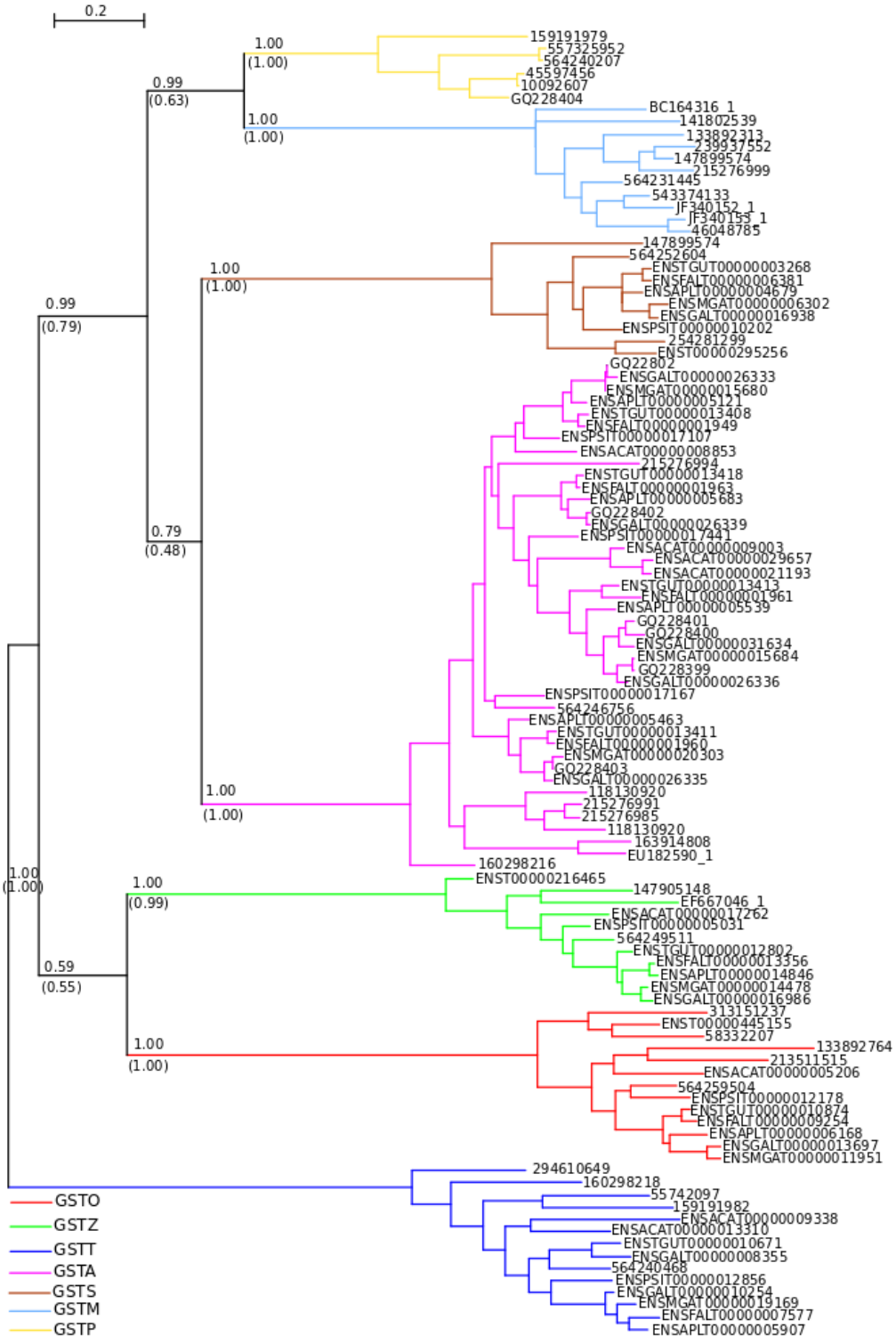
Supplementary table 4.2 -The details of sequences for Table 2 (Available on request)



Supplementary figure 4.1 - The secondary structure prediction results of PSIpred, shows the beta sheet accompanied with confidence of prediction



Supplementary figure 4.2 - The tertiary structure predicted structure of avian cGSTs



Supplementary figure 4.3 - The Vertebrate cGST phylogenetic tree produced by Bayesian method

Note:The Supplementary file with sequence are available on request.

13

Appendix 4
(Supplementary materials for chapter 5)

Supplementary file 5.1 - The Information for the sequences used for vertebrate TLR phylogeny shown in Figure 1 (Available on request)

Supplementary file 5.2 - The sequences information for TLRs used for positive selection analysis in birds (available on request)

Figure A1

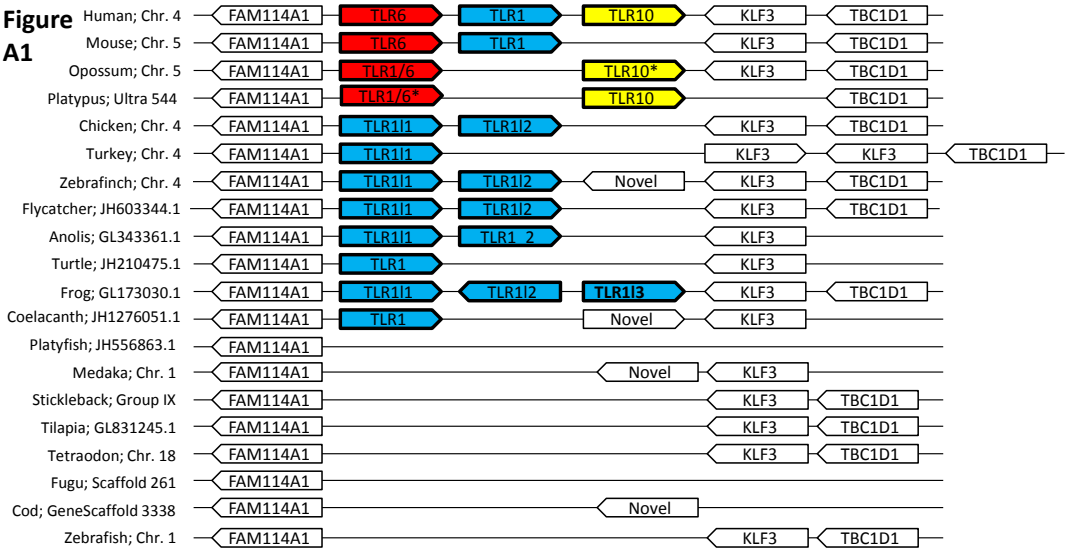
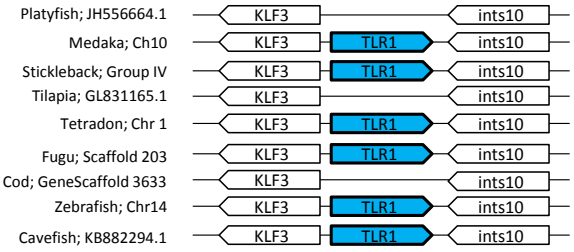


Figure A2



* Huang et al 2011

Figure B1

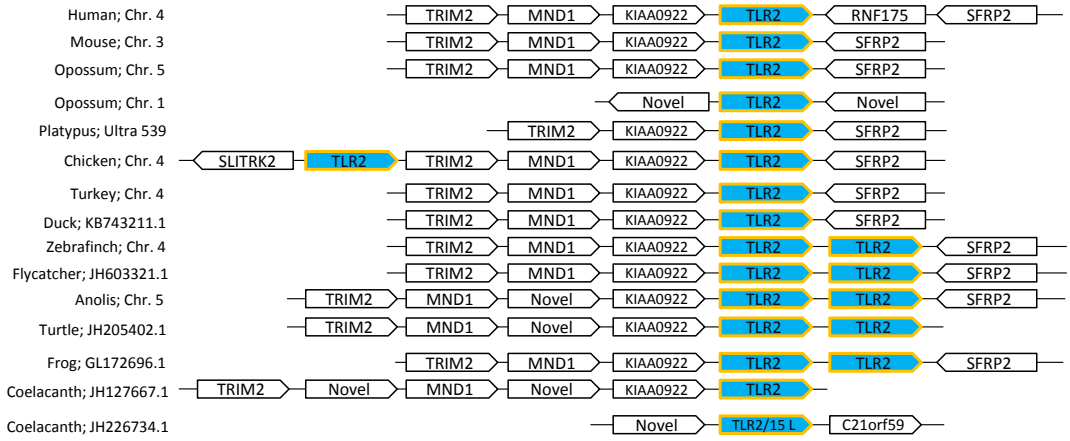


Figure B2

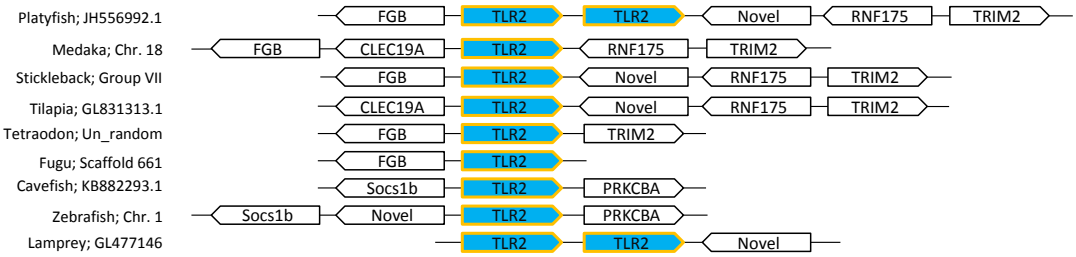


Figure C

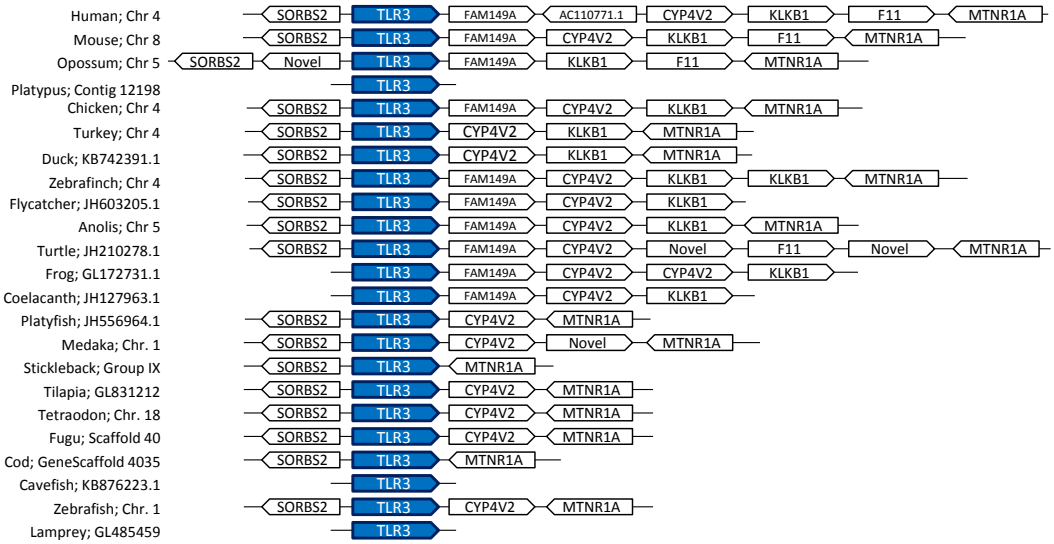


Figure D1

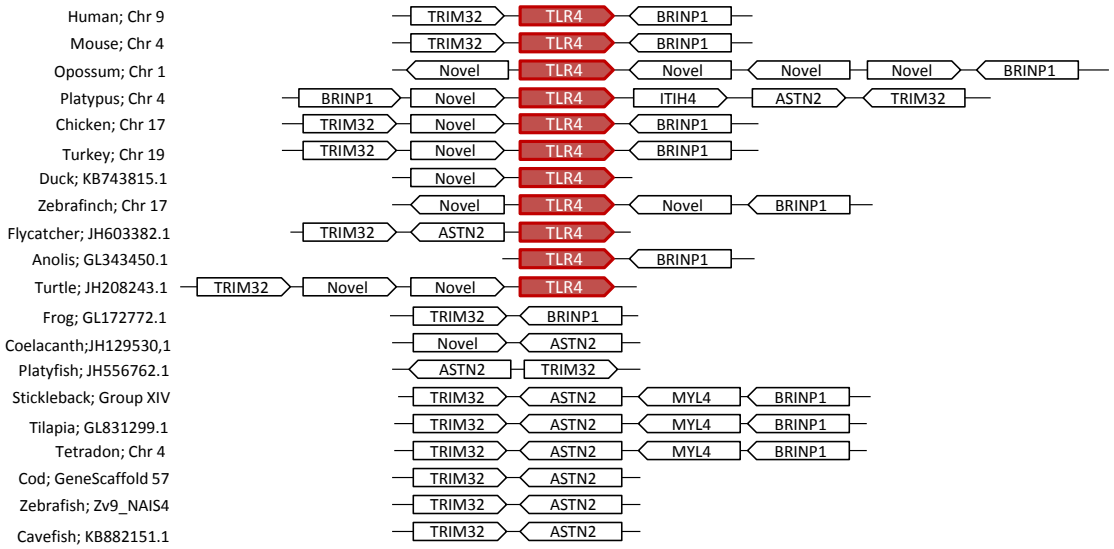


Figure D2

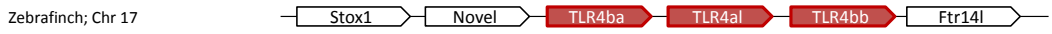


Figure E1

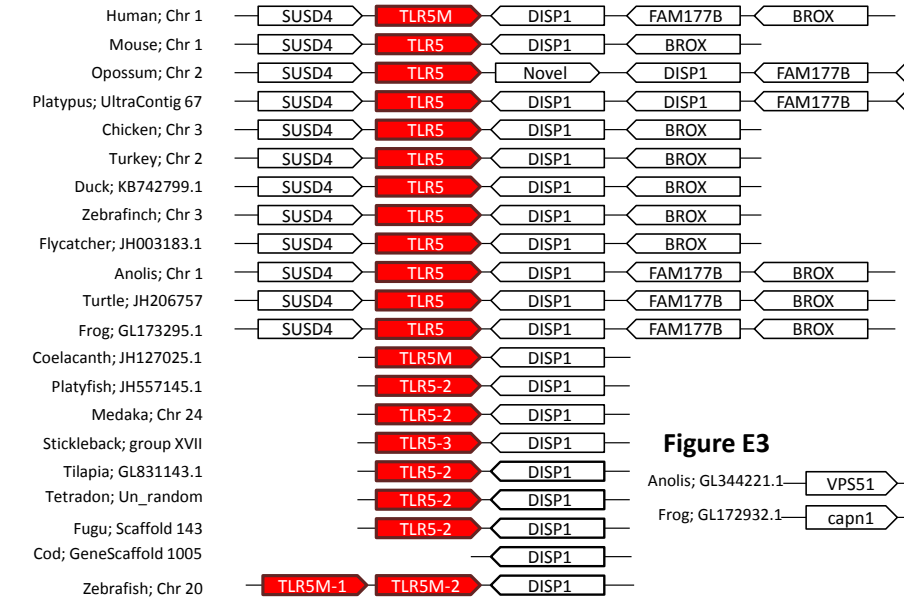


Figure E3

Figure E2

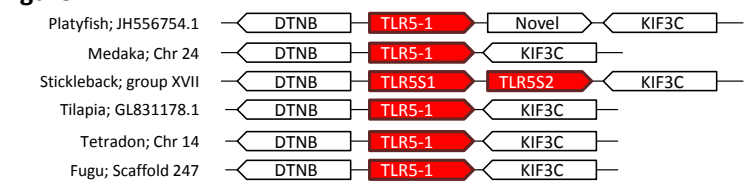


Figure F1

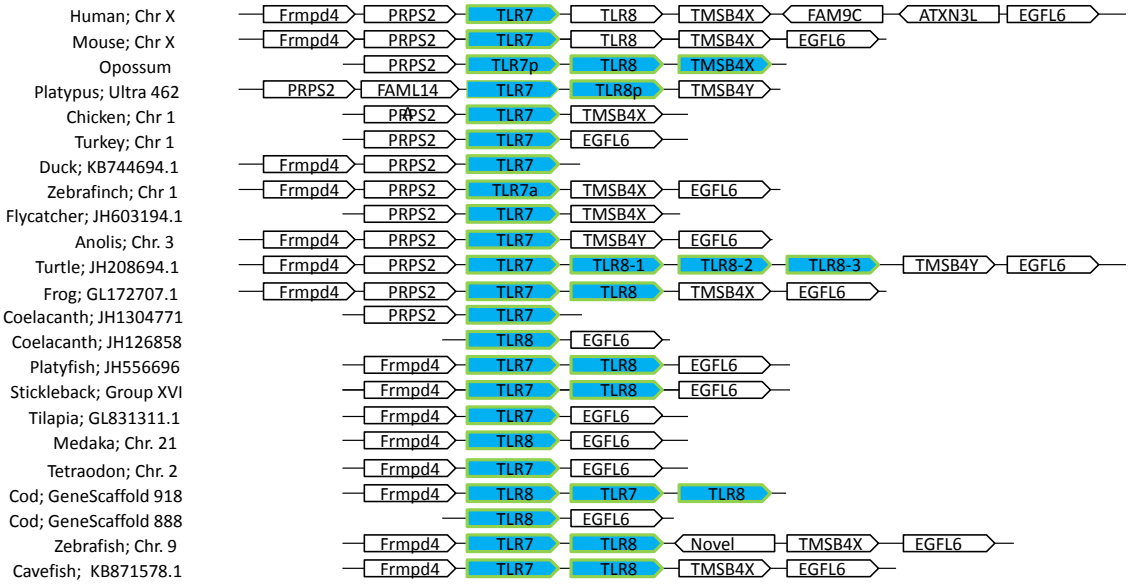


Figure F2

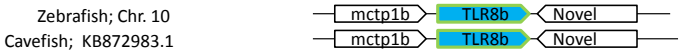


Figure F3

Figure G1

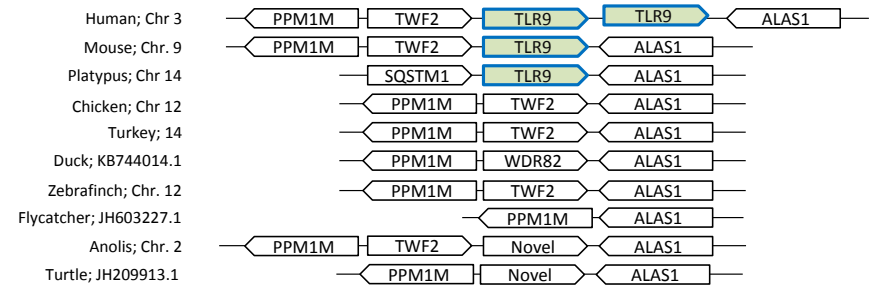


Figure G2

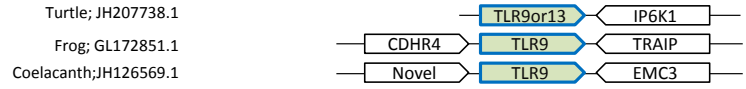


Figure G3

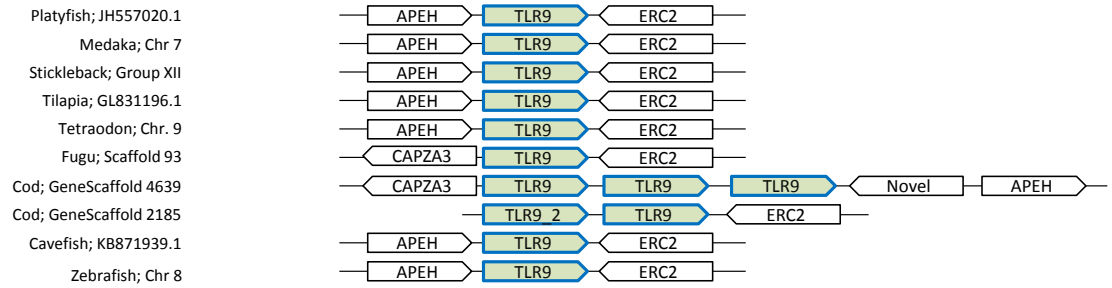


Figure H1

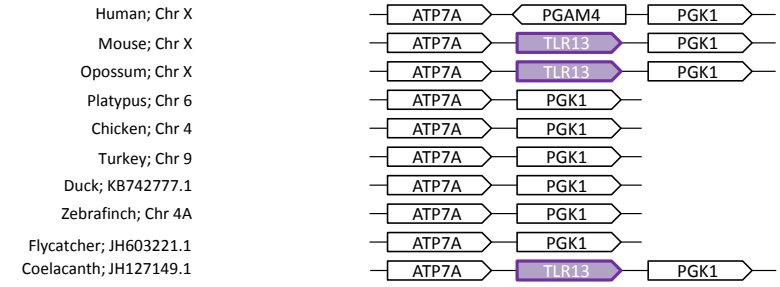


Figure H2

Figure I1

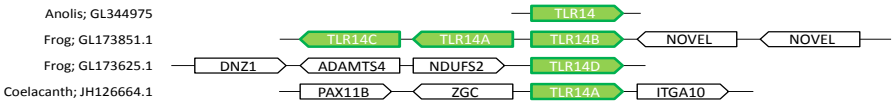


Figure I2

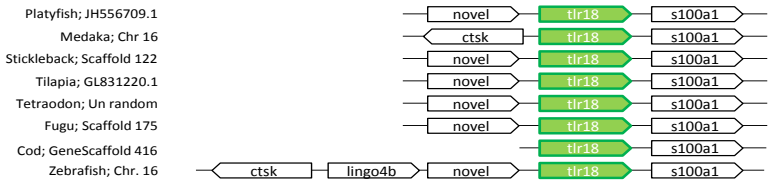


Figure I3

Figure J

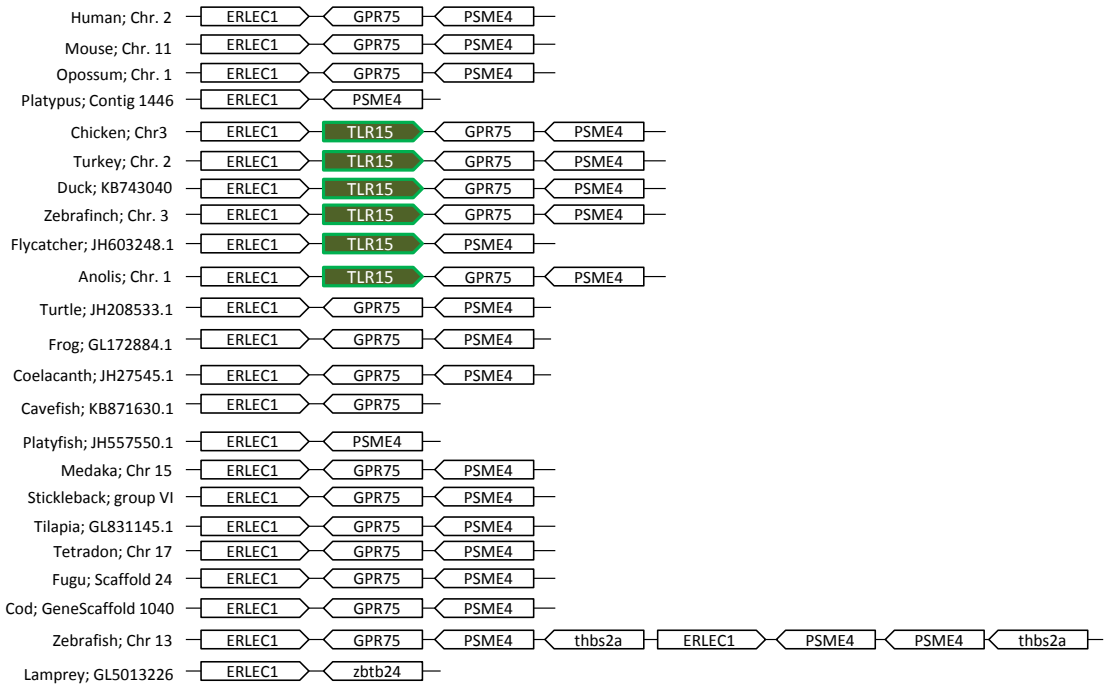


Figure K

Figure L

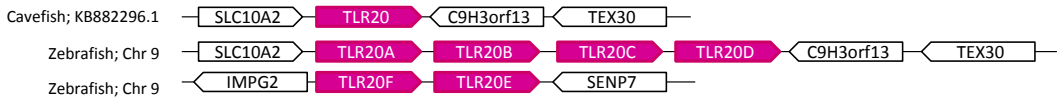


Figure M1

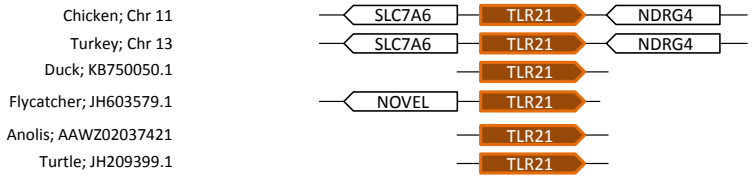


Figure M2

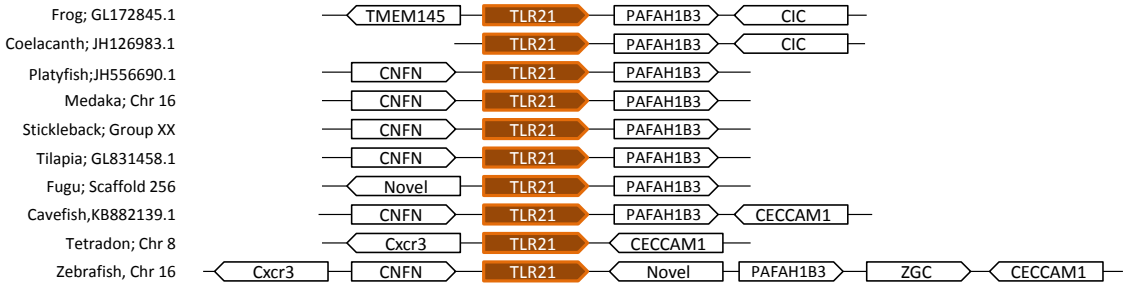


Figure M3

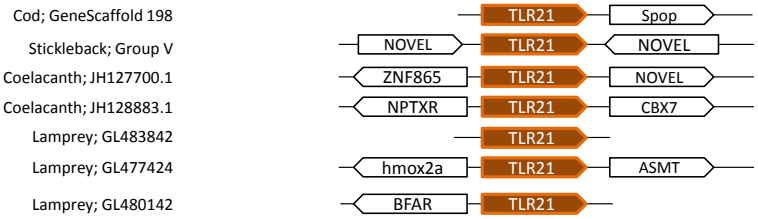


Figure N1

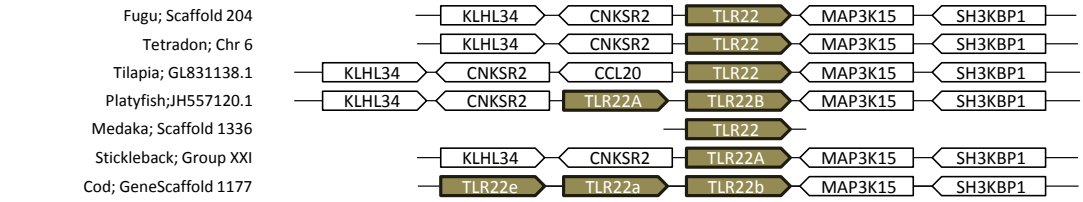


Figure N2

Figure O

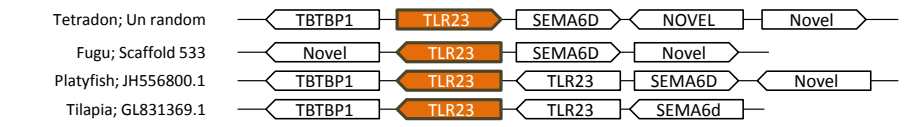
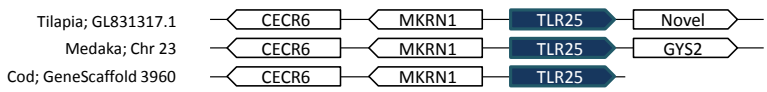
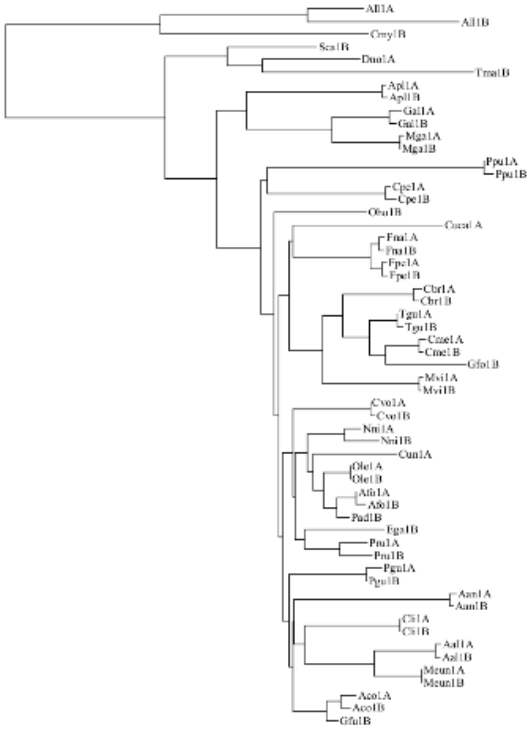


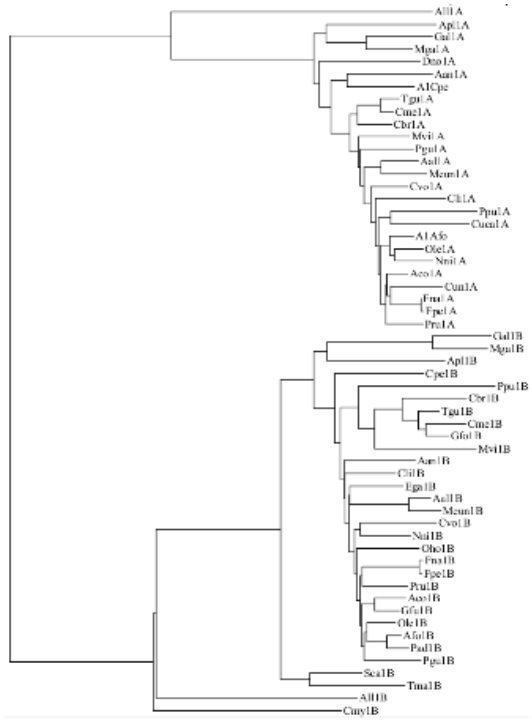
Figure P



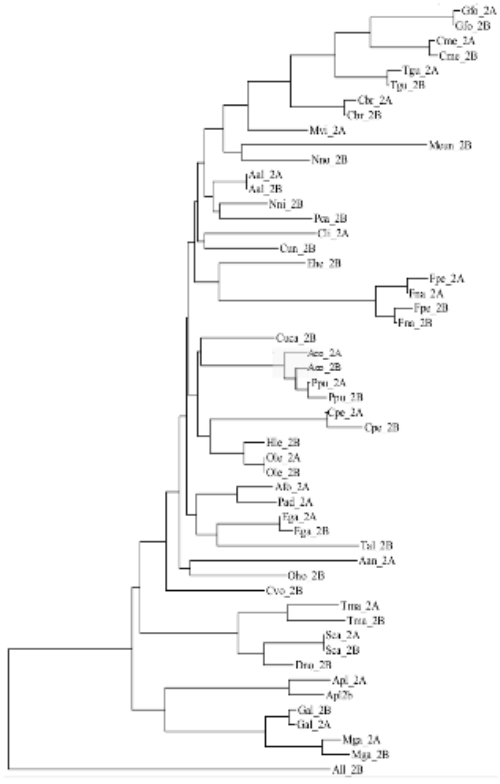
Supplementary file 5.3 figue A to P– Synteny analysis results of vertebrate TLRs



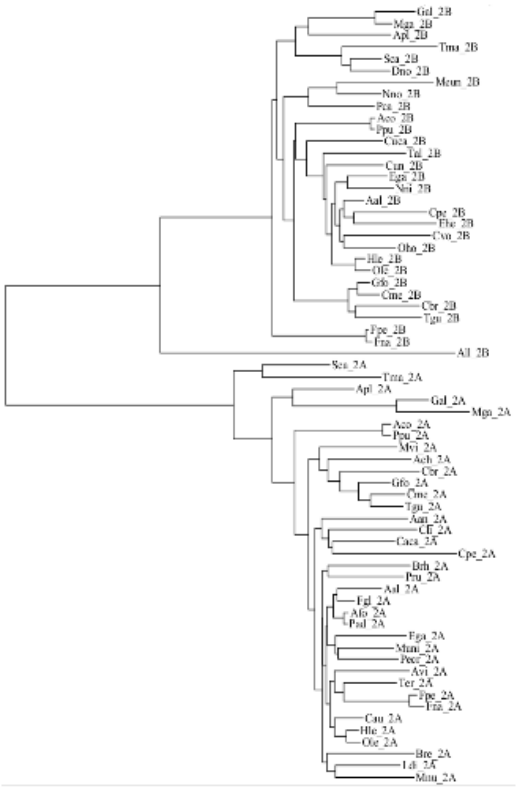
A - ML phylogeny of TLR1A and TLR1B showing gene conversion



B - ML phylogeny of TLR1A and TLR1B using N-terminal region free from gene conversion



C - ML phylogeny of TLR2A and TLR2B showing gene conversion



C - ML phylogeny of TLR2A and TLR2B susing central rigion terminal region free fromgene conversion

Supplementary file 5.4 Figure A to P - The phylogeny of TLR1A, TLR1B and TLR2A, TLR2B with regions under gene conversion and with regions free from gene conversion

Figure 1 **Chicken**

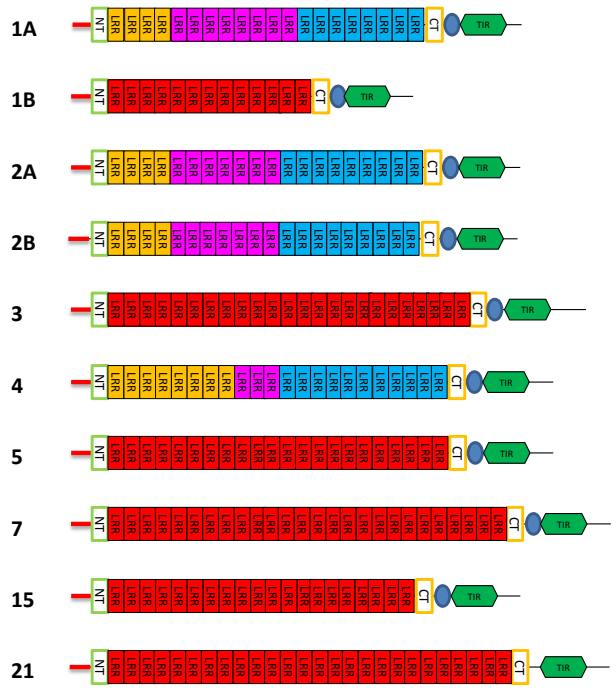


Figure 2 **Turtle**

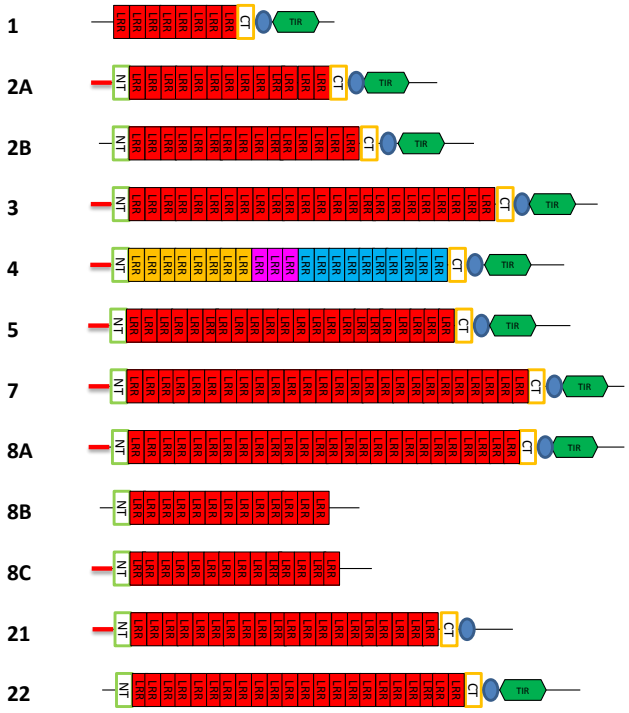


Figure 3 **Lizard**

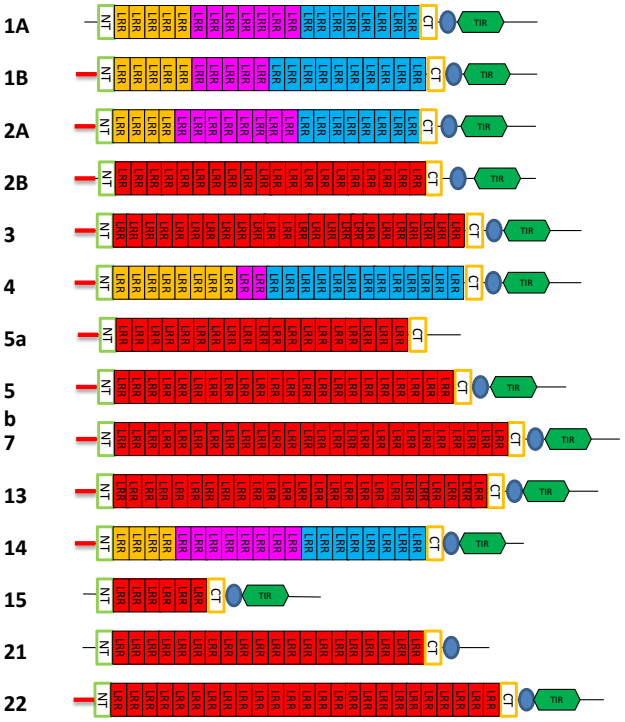
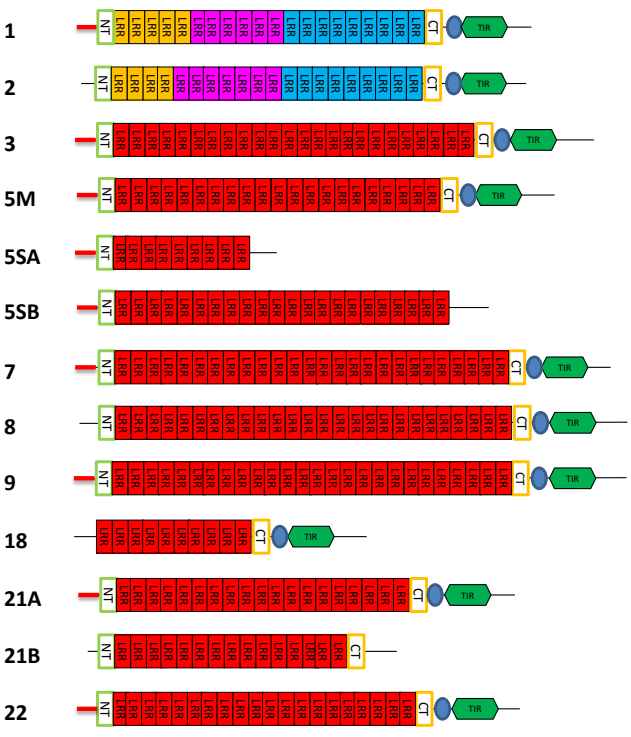
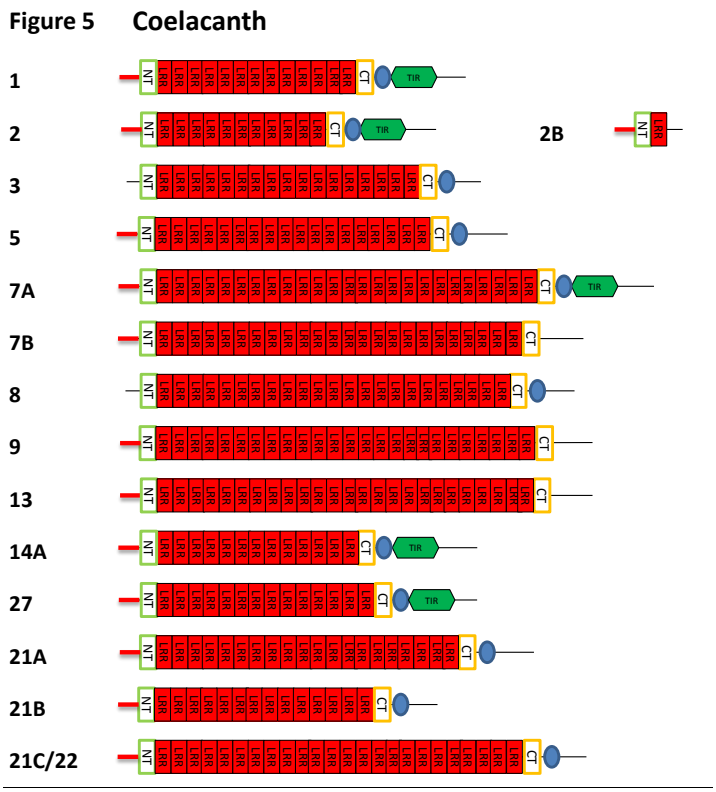


Figure 4 **Stickleback**





Supplementary file 5.5 Figure 1 to 5 - The domain architecture of TLRs from five diverse species that are Chicken, lizard, soft shell turtle, coelacanth and stickleback respectively

Supplementary file 5.6 to 5.14 . The location of positive selected sites with respect to domain architecture of respective TLRs (Labelled 6-14)-

Table 6

Domain characterization of TLR1A

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydroge residues are in red. The amino acids identified as under positive selection are in bold. In to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light

	TLR1A - Gallus gallus - NP_001007489.4		
Domain	Start	Stop	Sequence
Signal	1	24	MGSLTSIYVFACVFLSILWNNIQP
LRR-NT	25	52	TVENKITANYSGHLLTEVPKNIPVHTHI
LRR1	53	73	THILDLSHNSISEITNFRFTSLSD
LRR2	74	97	LQVLNLSHNLITELDFSAFMFNQD
LRR3	98	118	LEYLDLSHNNIWTAYCQLLAR
LRR4	119	143	LRHLDLSFNKFTVLPICQEFGIMFH
LRR5	144	166	LEYLGLSAMMIRRSDFRYVAHLQ
LRR6	167	189	LDTVFLTLEDFSLYEPLSLTALN
LRR7	190	217	TRSLHIVFATNQNFNFSLLYDGMSTSEK
LRR8	218	244	LKIVNLRYTLSHKDFPSPSLELQKKIK
LRR9	245	271	TTDLTLDTVDLEWTVILQIFLLVWDSS
LRR10	272	310	VEHLTVRNLIFR**G**PVVELTEYKHVPLLRSLEQLLSLGSS
LRR11	311	339	MKALTLERVRNKLYYFNQEILYRQFSEMN
LRR12	340	361	IDSLTIHDACMPHMLCPKKRSS
LRR13	362	385	FQYINFSRNALTDELFQNCDTL**A**N
LRR14	386	411	LK**I**LILHRNKFESL**S**KVSFMTS*R*MKS
LRR15	412	436	LRYLDMSSNLLRNSRAE**G**RCQWADS
LRR16	437	458	LAELDLSSNQLTEAVFECLPAN
LRR17	459	481	I**N**KV**D**LQNNQIANVPKGITELHS
LRR18	482	503	LQELNLASNRLADLPGCRAFTG
LRR19	504	527	LEILNIERNLILTPSADFFETCPS
LRR20	532	554	VKE**L**QAGQNPFKCSCELQ**D**FLRL
LRR-CT	555	591	ERQSGGKLSGW**P**EAYVCKYPEDLSGTQLEDFHLTELACNTT
Transmembrane	592	614	LLLVTALLLTLVLVAVVAF**P**CIY
TIR	615	818	LDVPWYVRMLWQWTQTKRRAWHDCPEERETALQFHAFISYSE RDSLWVKNELIPNLEKGEGCIQLCQHERNFIPGKSIVENIINCIEKS YKSIFVLSPNFVQSEWCHYELYFAHHRLFSENSNSLILILLEPIPSYV IPARYHKLKALMAKRTYLEWPKERSKHALFWANLRAVVNIKLPT SFETDEEQSDVTSTSSITQCLIK

Table 7

Domain characterization of TLR1B

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue and *type 1 residue* in dark blue.

	TLR1B - Gallus gallus - NP_001075178.3		
Domain	Start	Stop	Sequence
Signal	1	31	MTKNMRYLRNCFIYNCLFVFTFWDNIGLA
LRR-NT	32	76	NELFAS**V**PN**N**FLEDGLDKKNMSFPHSY**A**NN**Q**HYKADYGWV VIENT
LRR1	77	105	TESLSLSEIADDNVRKLITLLSKFRKGSR
LRR2	106	132	LRNLTLTNMSVDW**KD**I**IK**VLQVVWHSS
LRR3	133	158	IEYFNINNLTQLGNV*V*STRFDYS**K**TS
LRR4	159	187	MKAFAVNK**VL**ITDLYF*S*QDDIYNIFANMN
LRR5	188	209	IAALTIAESELIHMLCPSSDSP
LRR6	210	233	LRYINF**S**KNDLTDLLFQNCDKL**I**Q
LRR7	234	259	LETFILHRNKFESL**S**KVSFMTSRMKS
LRR8	260	284	LRYLDM**S**SNLLRNSRAE**G**RCQWADS
LRR9	285	306	L**A**ELDLSSNQLTE***A***VFECLPAN
LRR10	307	329	I*N*KV**D**LQNNQIASVPKGITELHS
LRR11	330	351	LQELNLASNRLADLPGCRAFTG
LRR12	352	375	LEILNIERNLILTPSADFF**E**TCPS
LRR13	376	398	VKELQAGQNPFKCSCELQDFLR**L**
LRR-CT	399	440	ERQSGGKL**S**GWPEA**Y**VCKYPEDLSGTQLKDFHLTELACNTTL
Transmembrane	441	463	LLVTALLLTLVLVAVVAFLCIYL
TIR	495	652	DVPWYVRMLWQWTQTKRRAWHDCPEERETALQFHAFISYSE RDSLWVKNELIPNLEKGEGCIQLCQHERNFIPGKSIVENIINCIEK SYKSIFVLSPNFVQSEWCHYELYFAHHKLFSENSNSLILILLEPIPPY VIPARYHKLKALMAKRTYLEWPKERSKHALFWANLRAAISINLS VADEQNRTEV

Table 8

Domain characterization of TLR2A

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue and *type 1 residue* in dark blue**.**

	TLR2A - Gallus gallus - NP_989609.1		
Domain	**Start**	**Stop**	**Sequence**
Signal	1	25	MFNQSK**Q**KPTMKLMW*Q*AWLIYTALA
LRR-NT	26	63	AH**L**PEEQALRQACLSCDA**T**QSCNCSFMGLDFIP*P*GLT**G**
LRR1	65	88	IT***V***LNLAHNRIK**L**IRTHDLQKAVN
LRR2	89	112	LRTLLLQSNQISSIDEDSF***G***SQGK
LRR3	113	136	LELLDLSNNSLAHLSP**V**WFGPLFS
LRR4	137	161	LQHLRIQGNSYSDLGESSPFSSLRN
LRR5	162	185	LSSLHLGNP*Q*FSIIRQGNFEGIVF
LRR6	186	209	LNTLRIDGDNLSQYEPGSLK***S***IR**K**
LRR7	210	233	INHMIIS**I**RRIDVFSAVIRDLLHS
LRR8	234	260	A**I**WLEVREIKLDIENEKLVQNST**L**PLT
LRR9	261	288	IQKLTFTGASFTDKYISQI**A**VLLKEIRS
LRR10	289	317	LRE**L**EAIDCVLEGKG***A***WDM***T***E***I******A******R***SK**Q**SS
LRR11	318	346	IETLSITNMTILDFYLF**F**DL*E*GIETQ**V**GK
LRR12	347	370	LKRLSIASS**K**VFMVPCRLARYFSS
LRR13	371	397	L**L**YLDFHDNLLVNNRLGETIC**ED**AWPS
LRR14	398	423	LQTLNLSKNSLKSLK**Q**AARY**IS**NLHK
LRR15	424	446	LINLDISENNFGEIPDMCEWPEN
LRR16	447	467	LKYLNLSSTQIPKLTTCIPST
LRR17	468	487	LEVLDVSANNLQDFGLQLPF
LRR18	488	509	LKELYLTKNHLKTLPEATDIPN
LRR19	510	533	LVAMSISRNKLNSFSKEEFESFKQ
LRR20	534	357	MELLDASANNFICSCEFLSFIHHE
LRR-CT	358	598	AGIAQVLVGWPESYICDSPLTVRGAQVGSVQLSLMECHR
Transmembrane	597	619	SLLVSLICTLVFLFILILVVVGY
TIR	620	793	KYHAVWYMRMTWAWLQAKRKPKRAPTKDICYDAFVSYS ENDSNWVENIMVQQLEQACPPFRLCLHKRDFVPGKWIV DNIIDSIEKSHKTLFVLSEHFVQSEWCKYELDFSHFRLFDEN NDVAILILLEPIQSQAIPKRFCKLRKIMNTKTYLEWPPDEEQ QQMFWENLKAALKS

Table 9

Domain characterization of TLR2B

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue and *type 1 residue* in dark blue.

	TLR2B - Zebrafinch - XP_002196402.1		
Domain	Start	Stop	Sequence
Signal	1	20	MTAHIWRVLAIYVILAASLS
LRR-NT	25	54	LKQACPSCDGSQLCNCSSMGLDFIPPGVTA
LRR1	56	79	ITVLNLAHNRIKRIQSQDLQQAVN
LRR2	80	103	LRALLLQSNKISSIDEDSF*W*SLEK
LRR3	104	127	LELLDLSNNSLAHLSPVWFGHLFS
LRR4	128	152	LQHLHLEGNSYRDLGQSSPFSSLKN
LRR5	153	176	LSSLHLGNP*Q*FSVIRHGNFEGI*E*L
LRR6	177	200	LHKLWIDGSNLSQYEQGSLKSIKQ
LRR7	201	224	INHMILNLRNGYIFSEIVRDLLHS
LRR8	225	251	VTWLEVRRIAFSIAAEMQVLRVMSSSF
LRR9	252	279	AKKISFRQTLLTDATVPEIVSILEDMPQ
LRR10	280	307	LVELELVDCRLLGTG*Q*WKMEIQAKKSQT
LRR11	308	336	LRILTIKKLSIEEFYLFTDLHSVEGLLSL
LRR12	337	360	LTRVTVQNTKVFLVPCRISQNLLS
LRR13	361	387	LVYLDLSANLLGDLSLEHSACQGGWPS
LRR14	388	413	LQALNLSQNSLSDLERTSKSLSHLGN
LRR15	414	436	LIVLDISQNNFGEIPDVCDWPKS
LRR16	437	457	LKYLNLSSTQIPKVTTCIP*Q*T
LRR17	458	477	LEVLDVSGNNLKEFGLRLPL
LRR18	478	499	LKELYLTRNQLKTLPGAAPIPN
LRR19	500	523	LVSLSVSRNKLNSFSKEEFESFRR
LRR20	524	547	MKLLDASGNNFICSCEFLSFIHHE
LRR-CT	548	586	AGISQVLVGWPDKYVCDSPLAVRGAQVGAVHLSLMECH R
Transmembrane	587	609	SLVVSLICVLVFLVILLLVAVGY
TIR	610	783	KYHMVWYLRMTWAWLQAKRKPKRAPPKDVCYDAFVSY SENDSDWVENTMVRELEQACPPFRLCLHKRDFVPGKWI VDNIIDSIEKSRKTLFVLSEHFVQSEWCKYELDFSHFRLFDE NNDAAILVLLEPIQSKAIPKRFCKLRKIMNTKTYLEWPLEEE QQQMFWFNLKIALRS

Table 10

Domain characterization of TLR3

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue.

	TLR3 – Gallus gallus- *NP_001011691.3*		
Domain	**Start**	**Stop**	**Sequence**
Signal	1	26	MGCSIPCW**N**SLSFRLVFVCLLCAS**V**G
LRR-NT	27	52	KQC**Q**IRNTMADCSHLKLTQIPSDLP**K**
LRR1	54	77	ITGLDISHNQLKKL**D**PENLT**E**YSN
LRR2	78	101	LIYLNAGYNIISKLKPGLCKNLPL
LRR3	102	125	LQILKLEHNQLHELPDGVFASCSN
LRR4	126	148	LTELNLGYNIIEVKNDPFKTLEN
LRR5	149	172	LNILDLSHN**H**LKSANLG**L**QQQLKN
LRR6	173	198	LRELVLYSNQITELNKEDLKFLSNTS
LRR7	199	222	LNSLDLSSNPLKEFH*T*GCLHAIGN
LRR8	223	249	LFGLILNNVELGEN*R*TKKLCTELSDTA
LRR9	250	275	IQNLSLSHVKLSH**I**NRLTLQGLQGTN
LRR10	276	299	LTVLNLSKNSLSVIEDDSFQWLSK
LRR11	300	323	LEYLNLE**D**NNIINVSSHLFYGLSS
LRR12	324	347	ITHLNLINSL**T**GKIEDFSFQWL**H**H
LRR13	348	371	LEYLIMDNNNFPRITTNMFTGLKN
LRR14	372	399	LKYLSLYNCNTNLQRITNKTFVSLANSS
LRR15	400	423	LQVLNLTKTRISTVESGAFSSLGQ
LRR16	424	448	LKILDLGLNEINQELTGHEFEGL**N**N
LRR17	449	472	IEYIYLSYNKNV**T**LRSESFIFVPS
LRR18	473	498	LRKLMLRKVGCNNLAISPSPFHPLRN
LRR19	499	522	LTVLDISNNNIANIKEDLFNGLHE
LRR20	523	554	LDILNLQHNNLARLWKCANPGGPVLFLKDVPN
LRR21	555	578	LH**I**ILNLKSNGFDEIPVHVFKGLHQ
LRR22	579	602	LKDLDLGSNNLNLLPATLFDDQTS
LRR23	603	627	LNTLNLQKNLITSVEENVFGPAFKS
LRR24	628	651	LRTLEMDFNPFDCTCESIAWFASW
LRR-CT	652	696	LNDTQAYIPGLQSQYICNTPPKYHGTLVLHFDTSAC KDSAPFKLL
Transmembrane	697	719	F**L**ITTT**V**VVMQFMFIVLLIHFEGW
TIR	720	896	RIAFYWNISINRILGFKELDRLPG**V**F**D**YDAYVIHAR KDTNWVLTNFTSLEENEQFQVKFCLEERDFEAGISE FEAIINCIRRSRKIIFIVTEHLL**Q**DPWCRKFKVHHAL QQAIEQSRDSIILIFLHNIQDYKLNHALCLRRGMFRS CCILNWPVQKERINAFHQQLMMALKSNSKVR

Table 11

Domain characterization of TLR4

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue and *type 1 residue* in dark blue.

	TLR4 – Gallus Gallus- *NP_001025864.1*		
Domain	**Start**	**Stop**	**Sequence**
Signal	1	30	MPSRAAPTALTLGVLLQLLLVLSLLAGCIP
LRR-NT	31	58	SPCLEVIPSTAFRCTGQNISGVPAEIPN
LRR1	60	83	TLDLDLSFNSLKLLSSNYFSSVPE
LRR2	84	107	LQ**F**LDLSRCHI**H**TIEDNSFVDL**Y**N
LRR3	108	131	LSTLILTANSL**Q**HLGL**A**AFHGLTS
LRR4	132	155	LKKLVLVET**S**ISSL**S**DLPIGHLNT
LRR5	156	180	LQELNLGHNNIASLKLPKYFANLTS
LRR6	181	208	LRHLSF*S*SNNITYISKGDLDALRETNRL
LRR7	209	231	NLTLVLSLNNIKYIQSGSFAKIH
LRR8	232	258	LGELILRSSFENL*N*AMHSSLQGLAGLQ
LRR9	259	288	VNRLIVGEFTN***/*L**K**I**T**AFQNGLLSGLCQVQ
LRR10	289	313	MQEFVLMCFR*E*F**E***N***DTDTLFDCIGN
LRR11	314	335	VTTIRLVDL***N***LETLSEVPMFSQ
LRR12	336	359	VKHLEWKRC**K**FQELPA*E*KLSLFKE
LRR13	360	383	LRV**L**RITKSK*D*LNGF*E*QKF***G******S***LTY
LRR14	384	409	LEV**V**DLSENRLSFL***T***CCSP**K**F*P**R***SPN
LRR15	410	432	LKHLNLSFNSDIS***L***TGEFANLRN
LRR16	433	457	LLYLDLQHTKLIHHGTYPVFLLLQK
LRR17	458	481	LIYLDISYT**K**THVMSHLIFHGLNS
LRR18	482	506	LQVLKMAGNSFENNTLTNNFENVRR
LRR19	507	530	LRILDISSCKLVWVD**Q**STFNALSE
LRR20	531	554	LKELIISNNKLLTFDPVTYKPLQA
LRR21	555	579	LTALDFSNNQMSFLSDSALEILPDS
LRR22	580	603	LVLLDISHNLFECSC***T***H**L**NFLKW**V**
LRR-CT	604	639	KE**K**QDLLQNKHSMICHTPAY***M***KN**M**SLSNFDMSSCHP
Transmembrane	640	662	***N***PTTVACSVTVLLA**A*****G***VFLFLIY
TIR	663	843	KYYFQLYYSLVLLSGCKHSAERGDIYDAFVIHSSKDQEWV MKELVEPLEEGKPPFQLCLYFRDFLPGVP**I**VTNIIQEGFLS SRNVIAVISADFLESKWCSFEFDIARSWQLVEGKAGIIMII LGEVDKTLLRQRLGLSRYLRRNTYLEWKNKEISRHIFWR QLTSVLLEGKKWNHEEIKLM

Table 12

Domain characterization of TLR5

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue.

	TLR5 – *Gallus Gallus- NP_001019757.1*		
Domain	**Start**	**Stop**	**Sequence**
Signal	1	22	MMLHQRLIIVFGIALAGDI**C**A**S**
LRR-NT	23	47	R**S**CYSEDQVSMY**N**SCNLTGVPPVPK
LRR1	49	72	TAKLFLTYNYIRQVTATSFPLLED
LRR2	73	98	LFLLEIGTQRVFPLYIGKEAFRNLPN
LRR3	99	122	LRVLDLG**F**NNILLLDLDSFAGLQR
LRR4	123	148	LTILRLF**Q**N**N**LGDSILEERYFQDL**R**S
LRR5	149	173	LEELDLSGNQITKLHPHPLFYNLT**I**
LRR6	174	199	LKAVNLKFNKISNLCESNLTSFQGKH
LRR7	200	229	F**S**FFSLSTN***T***LYKTDKMIWAKCPNPF**R**NIT
LRR8	230	256	F**N**SLDVS**E**NGWSTETVQYFCTAIKGTQ
LRR9	257	291	I**N**YL**S**FR**S**HTMGSGFGFNNLKNPD***T***DTFTGLARS D
LRR10	292	315	LHLLDISNGFIFSLNSLIFESLRN
LRR11	316	339	LEFLNLFRNKINQIQKQAFFGLEN
LRR12	340	363	LEILNLSSNLLGELYDYTFEGLHS
LRR13	364	387	IMYIDLQQNHIGMIGEKSFSNLVN
LRR14	388	406	LKIIDLRDNAIKKLPSFPH
LRR15	407	426	LTSAFLSDNKMMSVA**H**TAIV
LRR16	427	451	ATHIELERNWLANLGDLYVLFQVPG
LRR17	452	476	VQYLLLKQNRFSYC**V**K**H**VDAIENNQ
LRR18	477	505	LIYMDLGENMLQLVWERGLCLDVFRTLSK
LRR19	506	529	LQVLHLNNNYLSALPQEIFNGLTS
LRR20	530	551	LKRLNLASNLLSHLSLRVFPQS
LRR21	552	572	LTNLNLSGNQLFSPKPEVFMT
LRR22	573	596	LSILDITHNKYVCDCALKSLLVWL
LRR-CT	597	645	NETNVTLAGSESDRYCVYPPALAGVPVSFLTYDDC **D**EDELQQTLRFSVF
Transmembrane	646	668	VFLSVTLLMF**L**MS**T**IIFTRCRGI
TIR	669	861	CFVWYKTITKTLIGSHPPAADTSEYMYDAYLCYSK NDFEWVQNSLLKHLDSQYFDKNRFTLCFEERDFL PGEEHINNIRDAIWKSRKTICVVTRQFLKDGWCVE AFNFAQSRYFSDLKEVLIMVVVGSLSQYQLMKHK PIRIFLQRSRYLRWPEDYQDIGWFLD**N**LSSQILKEK KVQRN**V**SGIELQTIATVSH

Table 13 Domain characterization of TLR7

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue and *type 1 residue* in dark blue.

	TLR7 – Gallus gallus - *NP_001011688.1*		
Domain	**Start**	**Stop**	**Sequence**
Signal	1	36	MTNLSEVAAHRKMVHHARTSNALLFVLLFLFPMLLS
LRR-NT	38	73	**R**WFPKTLPCDVEAFESTV**R**VDCSDRRL*K*EVPRGIP*G*
LRR1	75	98	ATNLTLTINHIPRISPVSFT**Q**LEN
LRR2	99	136	LVEIDFRCNCVPPRLGPKD*N*VC*ITP*PSIENGSFAALSR
LRR3	137	157	LKSLYLDANQL*S*KIP**R**GLP*A*T
LRR4	158	181	LRLLSLEANNI**F**SIKKNT**F**SELRN
LRR5	182	213	IELLYLGQNCYYRNPCNVSFEIE**E**TAFLNLKN
LRR6	214	234	LTVLSLKSNNLTFIP**P**NLSST
LRR7	235	258	LKELYIYNNRIQEVQEHDLSNLYN
LRR8	259	298	LEILDLSGNCPRCYNAPYPC**T**PCPN***I***SIKIHSKAFYSLKK
LRR9	299	322	LRILRLHSNSL**Q**SIPSSWFKNIKN
LRR10	323	348	LKNLDLSQNFL***I***KEIGDAEFLKLIPS
LRR11	349	378	LVELDLSFNFE**L**QMYSPFLNLSKTFSCLSN
LRR12	379	405	LETLRIKGYVFKELRE***E***NLDPLLNLRN
LRR13	406	429	LTVLDLGTNFIKIADL**R**VFK**K**FRS
LRR14	430	504	LKI**I**DLSMNKISPSSGESNFYGFCSDHRITVEQYSRHVLQEMH YFRYDEYGRSCKSKDKEADSYQPLVNGDCM***S***Y
LRR15	505	528	GETLDLSRNNIFFVNS**I**DF*Q*DLS**F**
LRR16	529	553	LKCLNLSGNAISQTLNGSEF*Y*YLSG
LRR17	554	577	LKYLDFSNNRIDLLYSTAFKELK*F*
LRR18	578	607	LEILDLSNNKHYFLAEGVSHVLSFMKNLAY
LRR19	608	630	LKKLMMNENEISTSISTGMESQS
LRR20	631	661	LQTLEFRGNRLDIFWSDGKKEYLSFFKNLTN
LRR21	662	686	LE**Q**LDISPNMLNFLP*P*DVF**E**AMPPE
LRR22	687	710	LKILNLTSN*R*LHTF**N**WG*K*L*H*LLTK
LRR23	711	734	L**I**TLDLSNNLL*T*TVP***R***KLSNCTST
LRR24	735	758	LQ**E**LILRNNRIT*R*ITK**Y**FLRGAIQ
LRR25	759	784	LTYLDLSSNKIQIIKKSSFPENIINN
LRR26	785	807	LRMLLLHNNPFKCNCDAVWFVGW
LRR-CT	808	852	INQTQVAIPLLATDVTCAGPGAHKGRSLVFLDLNTCELDTSYF IM
Transmembrane	853	875	YALS**T**SAVLCLMMFAVMSHLYFW

Table 14

Domain characterization of TLR15

The conserved segment of each LRR is underlined. The conserved leucine or equivalent hydrophobic residues are in green and the conserved asparagine or equivalent hydrogen-donor residues are in red. The amino acids identified as under positive selection are in bold. In addition to that, *M8 sites under positive selection* are in italic, with *Type 2 residue* colored in light blue.

	TLR15 – Gallus gallus- *NP_001032924.1*		
Domain	**Start**	**Stop**	**Sequence**
Signal	1	22	MRILIGSLYFYFISFLFS**K**VNG
LRR-NT	23	55	FLT**Q**RTSPVS**S**FPFY**N**YSYLNLSSVSQAQAPKT
LRR1	56	79	ARALNFSYN**A**IEKITKRDFEGFH**V**
LRR2	80	103	LEVLDLSHN**H**IKDIEPGAFENL**L**S
LRR3	104	202	LVSVDLSFND**K**NLLVSGLAPHLKLIPTSGASG**P**SQIYMYFQ**K**SAEAA**L**E PSAPAELLPHLEDPPN**P**GNVNPRFRQRRTEEN*K*TSPPA**A**TLRPDLCG API
LRR4	203	233	**N**G**L**LDLSRTKLSNEELTAKLDAD**L**CQAQLGT
LRR5	234	260	VLEFNISHSDLEMDLLSLFILFLPM**K**D
LRR6	261	288	IQSVDAS**Y**NRITINNIDVEAICHFPFSN
LRR7	289	310	**F**SF**L**N**N**ISNNPINSLETVCLPAS
LRR8	311	334	ITVIDLSFTNISTIP*A*NFAKKLSK
LRR9	335	365	LERM**Y**VQG**N**QLIYTVRPENPSATPR**P***P*PG**T**VQ
LRR10	366	387	**I**SAISLVRNQAGTPIESLPES
LRR11	388	411	VKHLKVSNCSIVELPEWFANRMQE
LRR12	412	431	L**L**FLDLSSNRISMLPDLPIS
LRR13	432	454	LQQL**D**ISNSDIKIIPPRFKSLSN
LRR14	455	476	VTV**F**NIQNNKLTEMHPEYFPST
LRR15	477	498	LTTCDISKNKLKVLSLTKALEN
LRR16	499	520	LESLNVSGNLITRLEPACQLPS
LRR17	521	544	LTNLDSSHNLISELPDHLGQSLLM
LRR18	545	566	LKHFN**L**SGNKISFLQRGSLPAS
LRR19	567	590	LEELDISDNAITTIVQDTFGQLTS
LRR20	591	615	LSVLTVQGKHFFCNCDLYWFVNIYI
LRR-CT	616	653	RNPHL**Q**I**N**GKD**D**LRCSFPPDRRGSLVKSSNLTL**L**HCSL
Transmembrane	654	676	GI**Q**MAIT**A**CMAILVVLVLTGLCW
TIR	677	868	RFDGLWYVRMGWYWCMAKRRQYKKRPENKPFDAFISYSEHDADW TKEH**L**LKKLETDGFKICYHERDFKPGHPVLGNIFYCIENSHKVLFVLSPSF VNSCWCQYELYFAEHRVLDENQDSLIMVVLEDLPPDSVPQKFSKLRKL LKRKTYLKWSPEEHKQKIFWHQLAAVLKTTNEPLVRAENGPNEDVIE ME

14

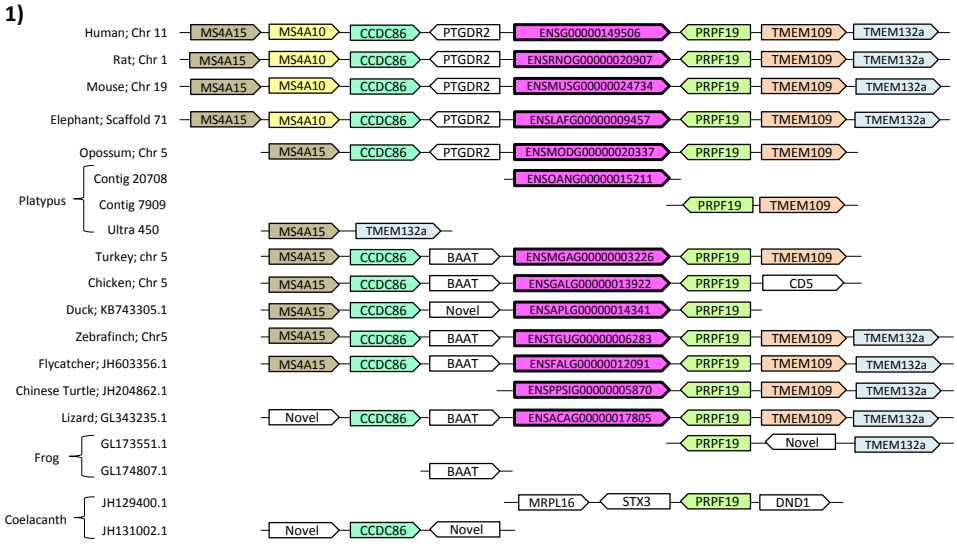
Appendix 5
(Supplementary materials for chapter 6)

Supplementary file 6.1- Sequence information used in phylogeny (Available of request)

Supplementary file 6.2- Genomic organization of ZP subgenome

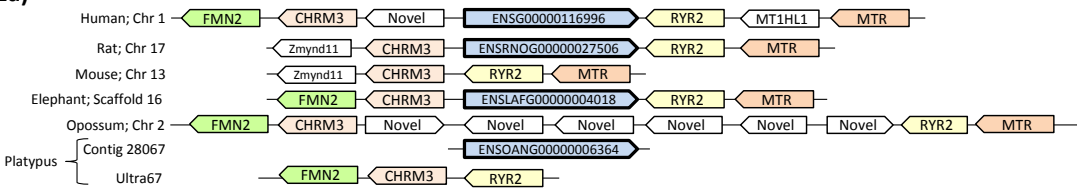
ZP1

1)



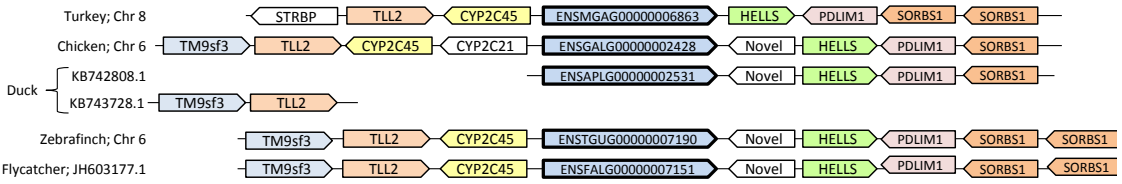
ZPB/4 Mammals

2a)



ZPB/4 Birds

2b)

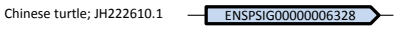


ZPB/4 Reptile

ZPB/4 Tetrapod

2c.1)

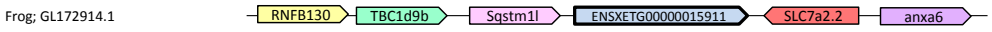
2c.2)



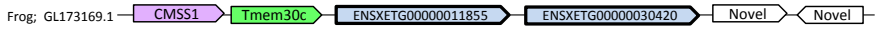
2c.3)

ZPB/4 Frog

2d.1)

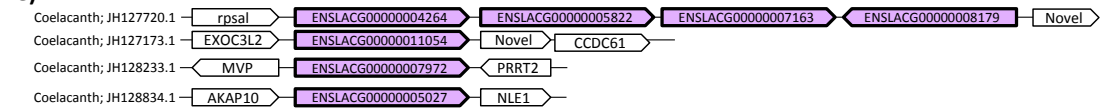


2d.2)

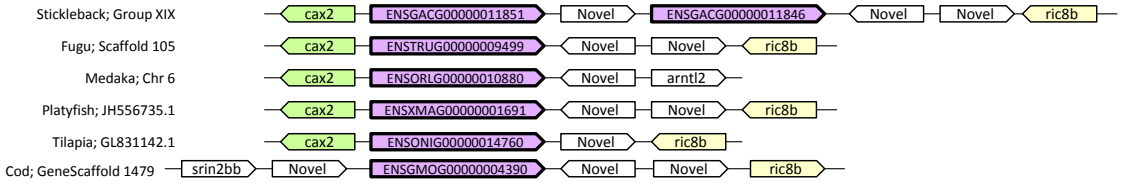


ZPB/4 Fishes

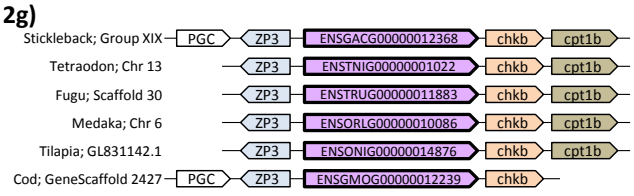
2e)



2f)

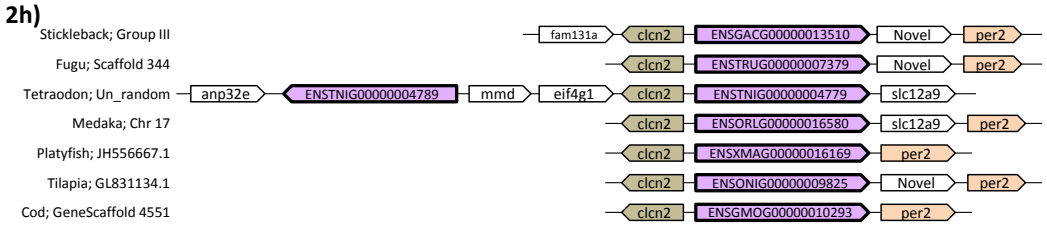


2g)

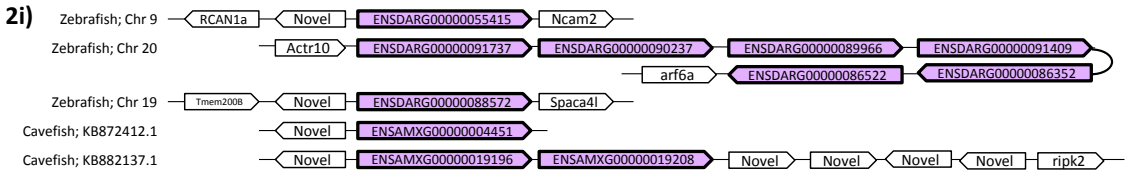


ZPB/4 Fishes

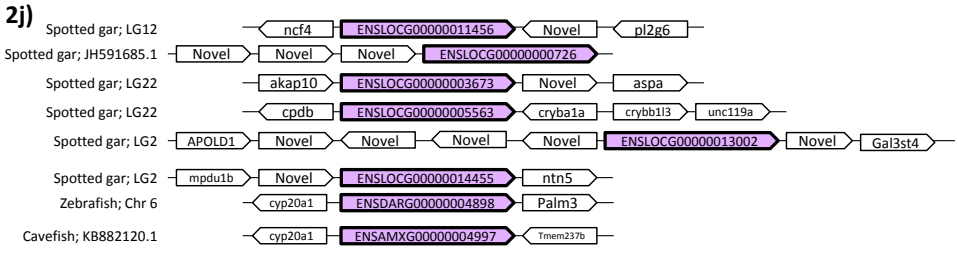
2h)



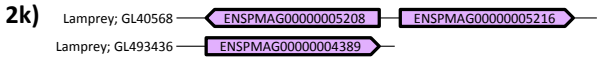
2i)



2j)

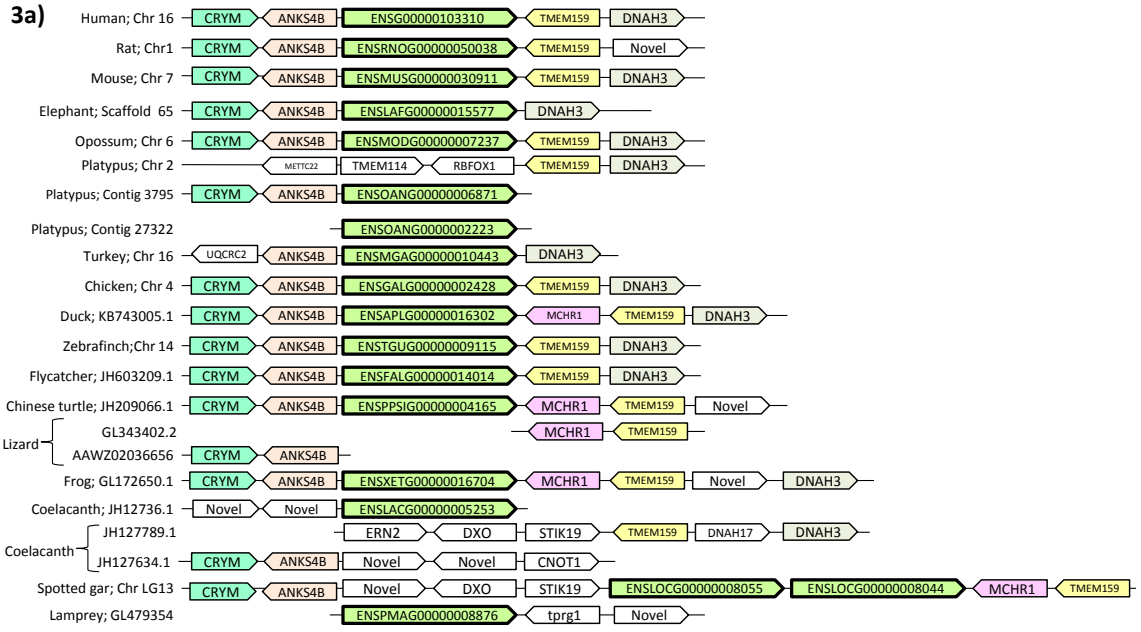


2k)



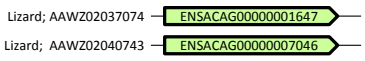
ZP2

3a)



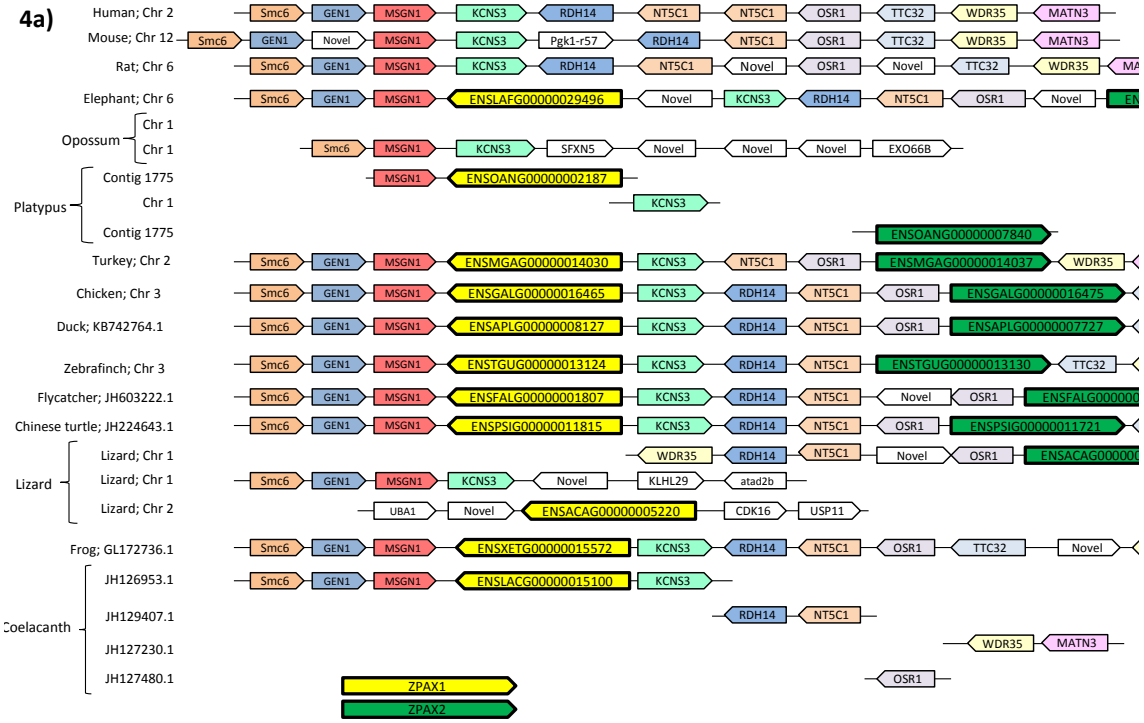
3b)

ZP2-like



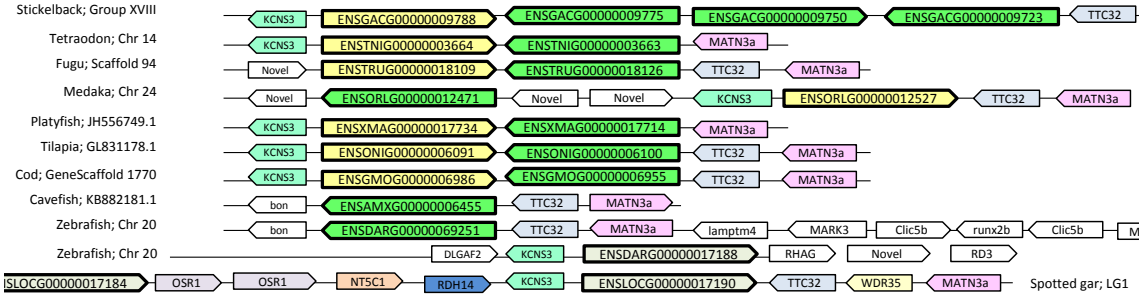
ZPAX1 and ZPAX2

4a)

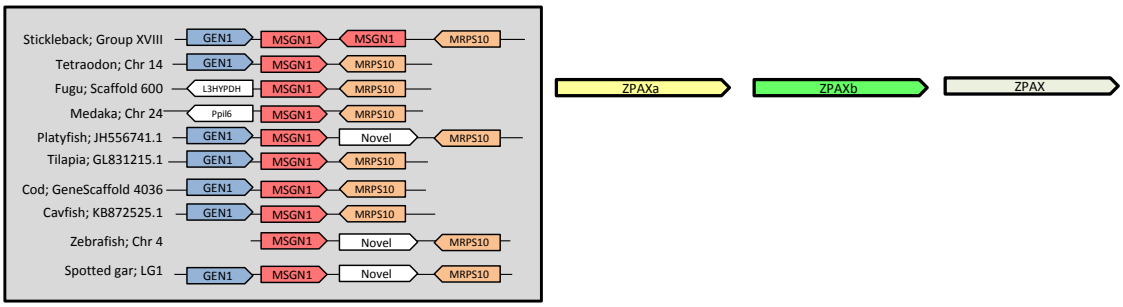


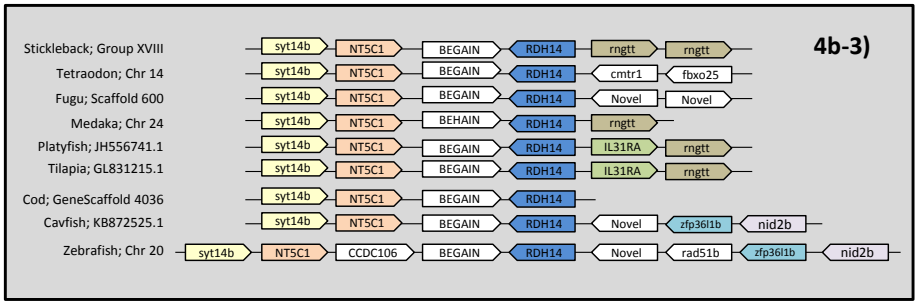
ZPAX, ZPAXa and ZPAXb

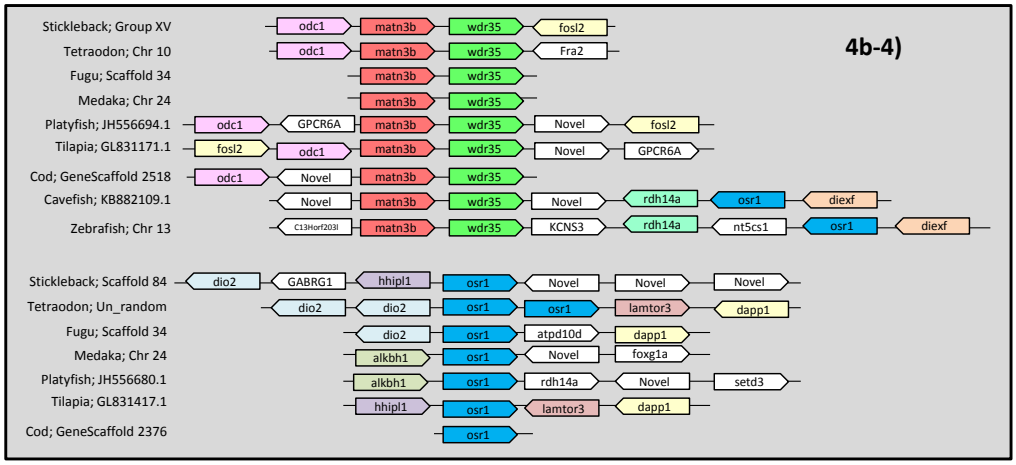
4b-1)



4b-2)

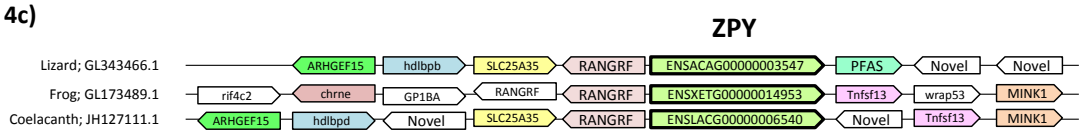


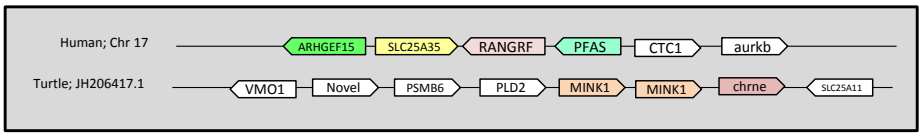

4b-3)
Stickleback; Group XVIII
Tetraodon; Chr 14
Fugu; Scaffold 600
Medaka; Chr 24
Platyfish; JH556741.1
Tilapia; GL831215.1
Cod; GeneScaffold 4036
Cavfish; KB872525.1
Zebrafish; Chr 20
syt14b
NT5C1
CCDC106
BEGAIN
BEHAIN
RDH14
rngtt
cmtr1
fbxo25
Novel
IL31RA
zfp36l1b
nid2b
rad51b


4b-4)
Stickleback; Group XV
Tetraodon; Chr 10
Fugu; Scaffold 34
Medaka; Chr 24
Platyfish; JH556694.1
Tilapia; GL831171.1
Cod; GeneScaffold 2518
Cavefish; KB882109.1
Zebrafish; Chr 13
Stickleback; Scaffold 84
Tetraodon; Un_random
Fugu; Scaffold 34
Medaka; Chr 24
Platyfish; JH556680.1
Tilapia; GL831417.1
Cod; GeneScaffold 2376
odc1
matn3b
wdr35
fosl2
Fra2
GPCR6A
C13Horf203l
KCNS3
rdh14a
nt5cs1
osr1
diexf
dio2
GABRG1
hhipl1
lamtor3
dapp1
atpd10d
foxg1a
alkbh1
rdh14a
setd3

ZPY

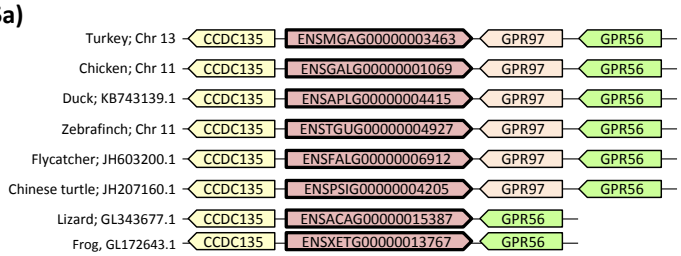
4c)

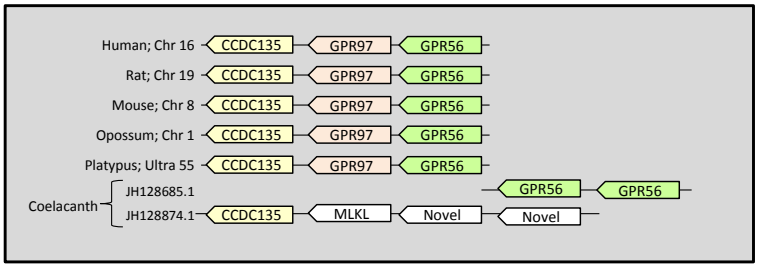

ZPY
Lizard; GL343466.1
Frog; GL173489.1
Coelacanth; JH127111.1
ARHGEF15
hdlbpb
hdlbpd
SLC25A35
RANGRF
ENSACAG00000003547
ENSXETG00000014953
ENSLACG00000006540
PFAS
Tnfsf13
Novel
wrap53
MINK1
rif4c2
chrne
GP1BA


Human; Chr 17
Turtle; JH206417.1
ARHGEF15
SLC25A35
RANGRF
PFAS
CTC1
aurkb
VMO1
Novel
PSMB6
PLD2
MINK1
chrne
SLC25A11

ZPD

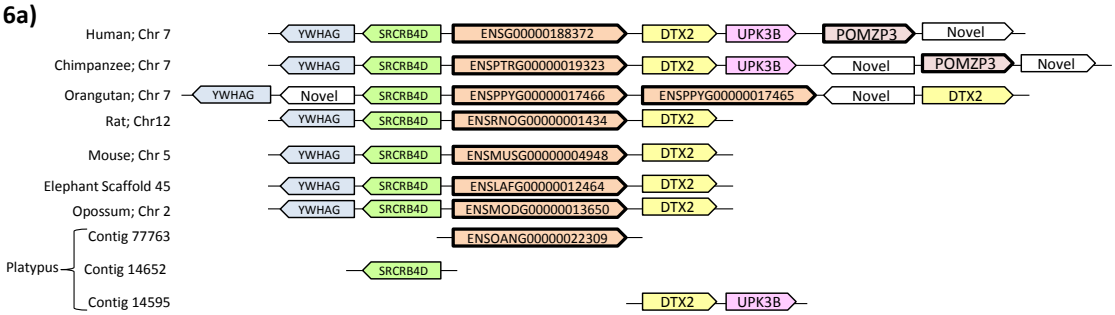
5a)


Turkey; Chr 13
CCDC135
ENSMGAG00000003463
GPR97
GPR56
Chicken; Chr 11
CCDC135
ENSGALG00000001069
GPR97
GPR56
Duck; KB743139.1
CCDC135
ENSAPLG00000004415
GPR97
GPR56
Zebrafinch; Chr 11
CCDC135
ENSTGUG00000004927
GPR97
GPR56
Flycatcher; JH603200.1
CCDC135
ENSFALG00000006912
GPR97
GPR56
Chinese turtle; JH207160.1
CCDC135
ENSPSIG00000004205
GPR97
GPR56
Lizard; GL343677.1
CCDC135
ENSACAG00000015387
GPR56
Frog, GL172643.1
CCDC135
ENSXETG00000013767
GPR56


Human; Chr 16
CCDC135
GPR97
GPR56
Rat; Chr 19
CCDC135
GPR97
GPR56
Mouse; Chr 8
CCDC135
GPR97
GPR56
Opossum; Chr 1
CCDC135
GPR97
GPR56
Platypus; Ultra 55
CCDC135
GPR97
GPR56
Coelacanth
JH128685.1
GPR56
GPR56
JH128874.1
CCDC135
MLKL
Novel
Novel

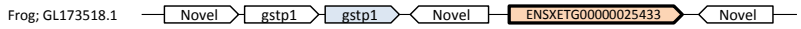
ZP3

6a)

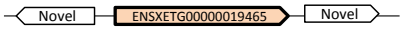

Human; Chr 7
YWHAG
SRCRB4D
ENSG00000188372
DTX2
UPK3B
POMZP3
Novel
Chimpanzee; Chr 7
YWHAG
SRCRB4D
ENSPTRG00000019323
DTX2
UPK3B
Novel
POMZP3
Novel
Orangutan; Chr 7
YWHAG
Novel
SRCRB4D
ENSPPYG00000017466
ENSPPYG00000017465
Novel
DTX2
Rat; Chr12
YWHAG
SRCRB4D
ENSRNOG00000001434
DTX2
Mouse; Chr 5
YWHAG
SRCRB4D
ENSMUSG00000004948
DTX2
Elephant Scaffold 45
YWHAG
SRCRB4D
ENSLAFG00000012464
DTX2
Opossum; Chr 2
YWHAG
SRCRB4D
ENSMODG00000013650
DTX2
Platypus
Contig 77763
ENSOANG00000022309
Contig 14652
SRCRB4D
Contig 14595
DTX2
UPK3B

ZP3

6b.1)



6b.2)

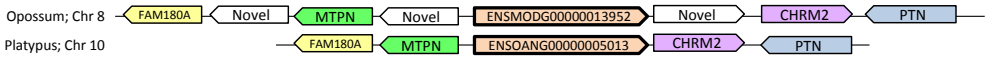


6b.3)

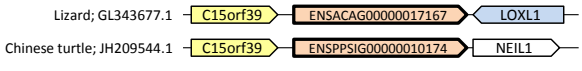
6b.4)

ZP3

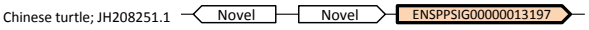
6c.1)



6c.2)

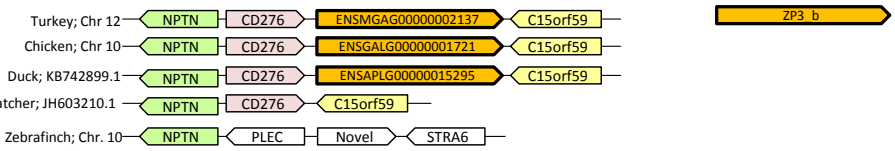


6c.3)



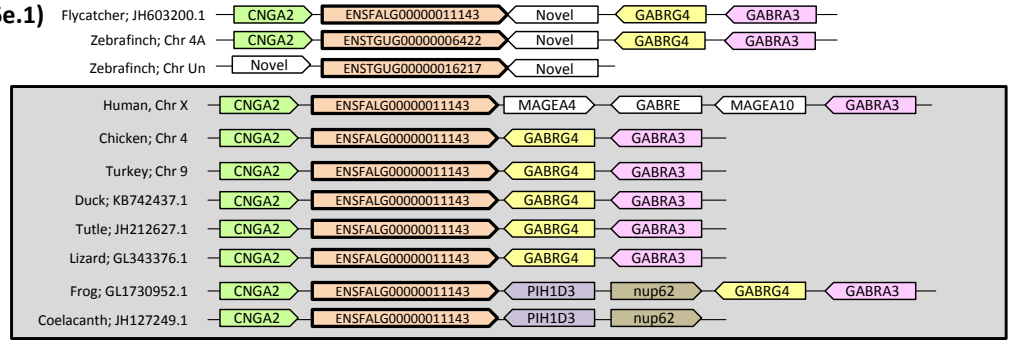
ZP3

6d)

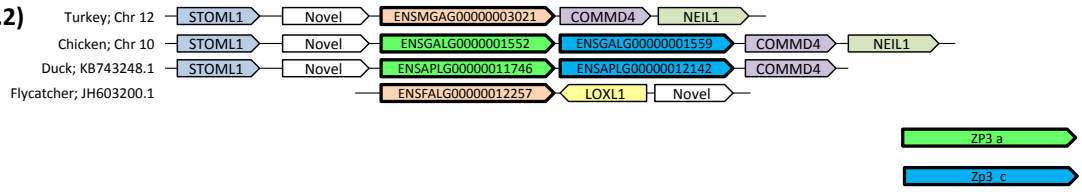


ZP3

6e.1)

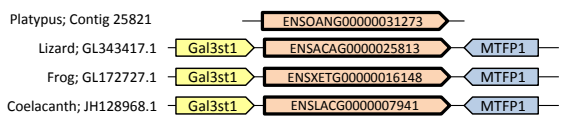


6e.2)

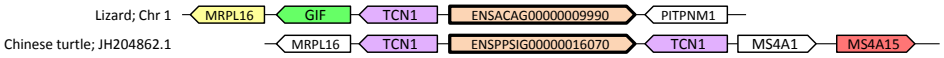


ZP3

6f)

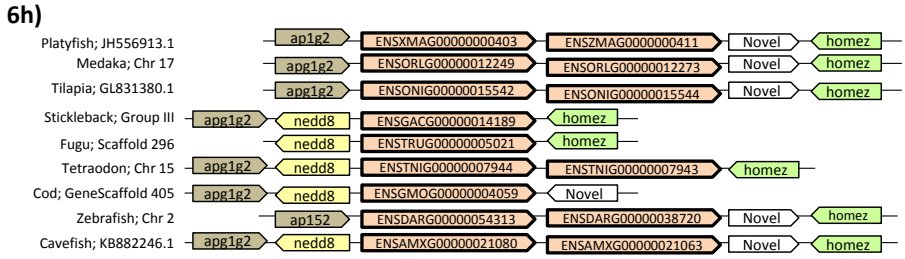


6g)

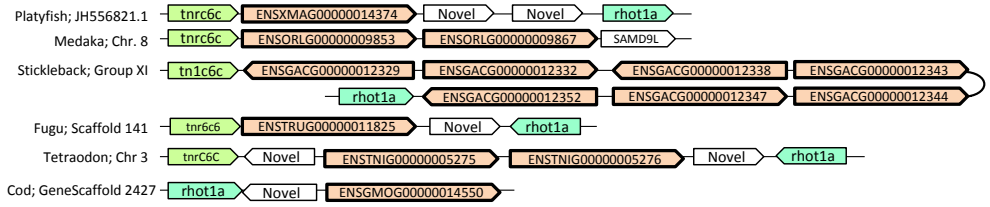


ZP3

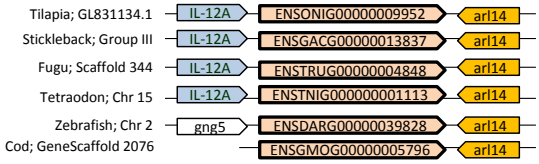
6h)



6i)

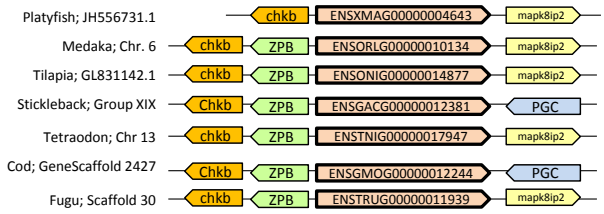


6j)

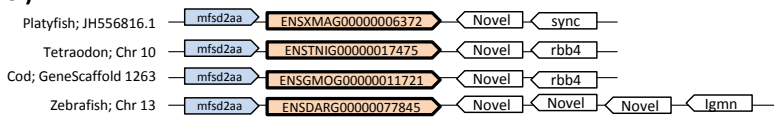


ZP3

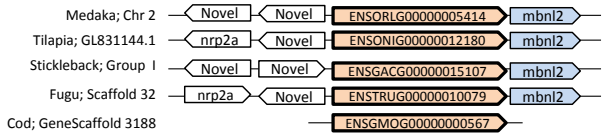
6k)



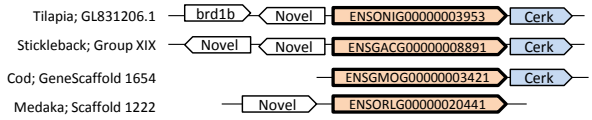
6l)



6m)



6n)



Supplementary file 6.3-6.12- Domain architecture of ZP proteins

Supplementary file 6.3-Human ZP1 ENSG00000149506

Signal peptide		1	24	MAGGSATTWG***YP****VALLLLVATLGL	[1]
		25	40	GRWLQPDPGLPGLHSY	
ZP-N		41	137	DCGIKGMQLLVFPRPGQTLRFKVVDEFGNRFDVNNCSICYHWVTSRPQEPAVFSADYRGCH VLEKDGRFHLRVFMEAVLPNGRVDVAQDATLICPKP	[2]
		138	231	DPSRTLDSQLAPPAMFSVSTPQTLSFLPTSGHTSQGSGHAFPSPLDPGHSSVHPTPALPSPG PGPTLATLAQPHWGTLEHWDVNKRDYIGTHLS	
trefoil		232	277	QEQCQVASGHLPC**/**VR**R**TSKEACQQAGCCYDNTREVPCYYGNTATV	[3,4]
ZP domain		278	550	QCFRDGYF**V**LVVSQEMA**L**THRITLANIHLAYAPTSCSPTQHTEAFVVFYFPLTHCGTTMQVAG DQLIYENWLVSGIHIQKGPQGSITRDSTFQLHVRCVFNASDFLPIQASIFPPPSPAPMTQPGPL RLELRIAKDETFSSYYGEDDYPIVRLLREPVHVEVRLLQRTDPNLVLLLHQCWGAPSANPFQQ PQWPILSDGCPFKGDSYRTQMVALDGATPFQSHYQRFTVATFALLDSGSQRALRG**L**VYLFCS TSACHTSGLETCSTACSTGTT	[2,4]
CFCS		551	554	RQRR	[5]
P r o p e p t i d e		555	599	SSGHRNDTARPQDIVSSPGPVGFEDS**Y**GQEPTLGP***T***DSNGNSSLR	
	TM	600	623	PLLWAVLLLPAVALVLGFGVFVGL	[6, 7]
		624	637	SQTWAQKLWESNRQ	

X - bold	under strong selection
X - bold and green	under strong selection by M8
X - underlined	common site
X - bold and blue	under strong selection by M8 *
X - bold and dark blue	under strong selection by M8 **
X - italics	Type 2 site
X - asterics*	Type 1 site

1. Peterson, T.N., et al., SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 2011. 8: p. 785-786.

2. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

3. Bork, P., A trefoil domain in the major rabbit zona pellucida protein. Protein Science, 1993. 2: p. 669-670.

4. Bausek, N., The major chicken egg envelope protein ZP1 is different from ZPB and is synthesized in the liver. Journal of Biological Chemistry, 2000.

5. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

6. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

7. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.4 Human ZP2 Q05996

Signal Peptide	1	38	MACRQRGGSWSPSGWFNAGWSTYRSISLFFALVTSGNS	Uniprot database
	39	53	IDVSQLVNPAFPGTV	
ZP-N	54	140	TCDEREITVEFPSSPGTKKWHASVVDPLGLDMPNCTYILDPEKLTLRATYDNCTRRVHGGHQMTIRVM NNSAALRHGAVMYQFFCPA	[1]
	141	153	MQVEETQGLSAST	
ZP-N	154	265	ICQKDFMSFSLPRVFSGLADDSKG*TKV*QMGWSIEVGDGARAKTLTLPEAMKEGFSLLIDNHRMTFHV PFNATGVTHYVQGNSHLYMVSLKLTFISPGQKVIFSSQAICAP	[1]
	266	268	DPV	
ZP-N	269	362	TCNATHMTLTIPEFPGKLKSVSFENQNIDVSQLHDNGIDLEATNGMKLHFSKTLLKTKLSEKCLLHQFYL ASLKLTFLLRP*ETVSMVIYPECLCES	[1]
	363	370	PVSIVTGE	
ZP domain	371	638	LCTQDGFMDVEVYSYQTQPALDLGTLRVGNSSCQPVFEAQSQGLVRFHIPLNGCGTRYKFEDDKVVY ENEIHALWTDFPPSKISRDSEFRMTVKCSYSRNDMLLNINVESLTPPVASVKLGPFTLILQSYPDNSYQQ PYGENEYPLVRFLRQPIYMEVRVLNRDDPNIKLVLDDCWATSTMDPDSFPQWNVVVDGCAYDLDNYQ TTFHPVGSSVTHPDHYQRFDMKAFAFVSEAHVLSSLVYFHCSALICNRLSPDSPLCSVTCPVSS	[1]
CFCS	639	642	RHRR	[2]
propeptide	643	716	ATGATEAEKMTVSLPGPILLLSDDSSFRGVGSSDL*KASGS*S*GEK*S*RS*ET*GE*EVGSRGAMDTKG HKTAGDVGSKA	
TM	717	738	VAAVAAFAGVVATLGFIYYLYE	[3, 4]
	739	745	KRTVSNH	

1. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

2. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

3. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

4. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.5 Human ZP3 ENSG00000188372

Signal Peptide		1	22	MELSYRLFICLLLWGSTELCYP	1
		23	44	QPLWLLQGGASHPETSVQPVLV	
ZP domain		45	308	ECQEATLMVMVSKDLFGTGKLIRAADLTLGPEACEPLVSMDTEDVVRFEVGLHECGNSMQVTD DALVYSTFLLHDPRPVGNLSIVRTNRAEIPIECRYPRQGNVSSQAILPTWLPFRTTVFSEEKLTFSLR LMEENWNAEKRSPTFHLGDAAHLQAEIHTGSHVPLRLFVDHCVATPTPDQNASPYHTIVDFHG CLVDGLTDASSAFKVPRPGPDTLQFTVDVFHFANDSRNMIYITCHLKVTLAEQDPDELNKACSFS KPSNS	2
		309	348	WFPVEGSADICQCCNKGDCGTPSHSRRQPHVMSQWSRSAS	
CFCS		349	352	RNRR	3
Propeptide		353	385	HVTEEADVTVGPLIFLDRRGDHEVEQWALPSDT	
	TM	386	409	SVVLLGVGLAVVVSLTLTAVILVL	4, 5
		410	424	TRRCRTASHPVSASE	

1. Peterson, T.N., et al., SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 2011. 8: p. 785-786.

2. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

3. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

4. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

5. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.6 - Chicken ZP1 ENSGALG00000013922

Signal peptide	1	24	MG**R**SRSLLLPLLLLLPAGLPSGLA	1
	25	30	LLQYHY	
ZP-N	31	127	DCGDFGMQLLAYPTRGRTVHFKVLDEFGTRFEVANCSICMHWLNTGEDGGLIFSAGYEGCHVLVKDGRYVLRVQLEE**M**LLSGVVAASY EVNMTCPRP	2
	128	577	AGYEILRDEKVHGHQRPDRGNGALSSHG**V**NVLIPRPRPGLLQHTAHSALAIPRPQLPPGAPVEQSHSMAHTQSILGHSELQ**H**QPQPGTG HLRPQPQNQPGMIHASGQTQMGVLRPGLQSQNQPGMVHAGGQIHPGVLRPGLQSQNQHGLLNVGSQTQPGVLRPG**L**QSQNQQG LVRPGSETQPGVLRPGLQSSNQHGLAQPGGQTQLGVLRPGLQSQNQHGLARPGGQSQPGVLRPGL**QS**SNQHGLVRPGSESQPGVLH PGLQSPNQHGL**L**H**P**G**G**QSQPGVLRPGLQSQSQHGLLHPGGQSQPGALRPGLQLQNQPGLVHAGSQTQAGLFHSGL**Q**PLNQPSLVRP GLQPGLMHTSTHTQAGFVRPGLQPQSQLGMLLPSLQSHAQGSLLRPSLQSQAGLLQPSQPRPGLLRPGLPSRPGLVSPGLQSQAQPGLL HPTALFYPSAGAGEPLT	
Trefoil domain	578	623	REQCQVAVGRLSCVSPPGRDACLQAGCCFDDTDRATPCYYGNTATV	3, 4
ZP domain	624	899	QCLPEGHFVLVVPRGLSAQPYNLDSVRLASTQPGCQP**T**QTTDAFVLFHFPVTQCGTTVQVIEDRLVYENQLISTIDVQPGPRGSVTRDSV YILHARCIYNATELLPLSLEVAVPPTAAPLAQPGPLQLQLRIATDESYSSYYPDADYPLVKVLRDPIYVEVRLLQKTDPNLVLVLHQCWAAPS TSPAAEPQWPILVDGCPF***A***GDNYRTQLVPVGPATLQLPFPSHYQRFAISTFAFVDSPSMVVLEGEVYILCSASVCHLSQPEPCRPSCQVAV PS	2,4
CFCS	900	904	RARR	5
Propeptide	905	934	AAADRKAADILGTVTSRGRIVLPQGPAAGRR	

1. Peterson, T.N., et al., SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 2011. 8: p. 785-786.

2. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

3. Bork, P., A trefoil domain in the major rabbit zona pellucida protein. Protein Science, 1993. 2: p. 669-670.

4. Bausek, N., The major chicken egg envelope protein ZP1 is different from ZPB and is synthesized in the liver. Journal of Biological Chemistry, 2000.

5. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

Supplementary file 6.7 Chicken ZP2 NM_001039098.1

Signal Peptide		1	23	MRGRLLLLLLFGFLLFLAPGASG	1
		24	31	EWDLSESM	
ZP-N		32	119	TCLQDRLELELPRELGNYTWHVRAVDVSGEEMMSCEHAVDYEKLLLSALLVNCTSLEHGQYQLRLLLLLNGTAGEERNVTYSAH CSAA	2
		120	133	HGDEIIAPLFVGET	
ZP-N		134	241	NCTKDSMAVTFPGPSLSDEHLVQVAVLTGTLTIDDGIKVHQLSLGEAMQHGYSFLADGHHLVFQAAFTATGVVSYKHNHKALY TAALKLMYGPPEHRLTVESRMLCVP	2
		242	244	GPV	
ZP-N		245	341	FCNTTHMTVAIPAFPGTLMAVAVEDETIPMDQLQDKGITLKTTVGVELHVSRRVLKSTLHGESCPRVQSYLSSLKLTFHFHEETV AMVMHPQCPCDQ	2
		342	347	LTPIAA	
ZP domain		348	615	ACTRDGYMDFEVLAGSTTPPLVLDTLRLRDPTCKPASRSPLNDRAWFHVPLSGCGTRYWLEGEKIMYENEVRALRSDSVLHRIS RDSEFRLAVLCSFSNGDASVSVRVDNPPPLAASTNQGPLSLILLSYPEDSYRQPYHDDQYPIVRYLQQPIFMEVQVLNRNDPNLY LQLDDCWATALEDPTSLPQWNIVVDGCEYEQDSYRTVFHPVGHGVSYPNYRQRLEVKAFAFVSGDKALPGLVYFHCSVLICSRF QLDSPLCTARCPRLP	2
CFCS		616	619	RRKR	3
Propeptide		620	656	GSGMLGASSVVSLQGPVLLVPHGWAAARGGTLLSKVV	
	TM	657	679	WAAVTATAVGVFSLTAIMLLFMD	4, 5
		680	695	LLKCLKRRALMANVVY	

1. Peterson, T.N., et al., SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 2011. 8: p. 785-786.

2. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

3. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

4. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

5. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.8 Chicken ZP3a ENSGGALG00000001552

		1	54	MGAAYSSWGFLRGETELWGAQALGQPHVFSQPSPWGWVDVSQLQAASPLHPVSV	
ZP domain		55	319	WCQEAQVVVTVHRDLFGTGRLVRAADLTLGTAACPATAQ*NA*AENMVTFVAGLHECGSTLRVTPEALIYSTSLNYSP VHAGNPVIIRTSPAVVPIECRYPRRSNVSSHAIQPTWAPFHSTLSSEQKLLFSLRLMNDDWSTERASAVFQLGEVLRMQA SVSVGNHAPLRLFVDSCVATPSPDRGSSPHYAFIDFSGCMVDGRLDDTTSTFISPRPRLDVLQFAVDVFKFAEDSSSLLYIT CHLKVSPASQPPDPQNKACSFHKPSGL	[1]
		320	351	WAPVEGTRAVCSCCETQSCGTARRSLQPFTPS	
CFCS		352	355	RQRR	[2]
Propeptide		356	404	GRSRRERPSGPTLGEADVMLGPLLIHSHRQHPAGRVPSAGAGSHVGLAA	
	TM	405	417	GLAVLLLAALGAL	[3, 4]
		418	433	AARQAQQSCLNRVLNF	

1. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

2. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

3. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

4. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.9 Chicken ZP4 ENSGGALG00000005462

		1	39	MGVVGQAMAVFGAVFLGLLGPFALVVGTWSRPFADPGLL	
ZP-N		40	138	ACGQGSLQLTLPSGWEGNASFVLTAWDTEGKAHALQNDSGCGLWVSDALDGSRVVS*VSYTSCYVFGWDGNYFIIVGLEGTD AAGQKVVHEEKLFMCPAD	1
		139	148	LPALDAPSSS	
Trefoil domain		149	195	VCSAVRSQDRLPCASLPISQGDCEVRGCCYNPRDKVKTCYYGNTVTA	2
ZP domain		196	470	HCTPDGQFSIAVSRDVTLPPVILDSVHLASGRSAGCIPVVKNNAFVVYQFPLSACGTTFQVTGDQAVYENELVASRDVKTGSLGS VTRDSTFRLHVRCSYAITGTFVPLSVQVFTLPPLPAVSQPGPLSLELRVASDERYSSYYTDNDYPVVKALRDPIYIEVRILQRTDPDLV LVLHHCWATPSINPHQQTQWPVLVNGCPYAGDNYQTQLVPLSTASGLLFPSHYQRFTLYTFTFVDSASQEVLSGLVYLHCSASV CHRSVQESCANTCPARA	1
CFCS		471	474	RGKR	3
Propeptide		475	518	SAEHTLKDSASRVSSKGPVIFLQDELRRVADVNDFRAAAASWAL	
	TM	519	525	GFAAVAAGAVLGMVLVAA	4, 5
		526	542	VLWWRK	

1. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

2. Bork, P., A trefoil domain in the major rabbit zona pellucida protein. Protein Science, 1993. 2: p. 669-670.

3. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

4. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

5. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.10 Chicken ZPD ENSGGALG00000040439

Signal Peptide		1	21	MEGT**V**T**Y**LLL**F**SALRL**A**GC**E**G	1
		22	44	**N**KSEL*V*SPHNSR*G*R**F**ALR**A**K**K**RSS	
EGF-like domain		45	80	DACVPNPCQHHGGCQVIEDRP**I**CSCKPGFTGAFCQD	2
		81	85	V**V**LKL	
ZP domain		86	343	ACEEEHMKMMVRKEVFELLKIP**R**ELVHLKNQACKVSEREEEGE**M**FFAA**T**LTGENHTACGSVIQQN**S**S HVSYSN**I**IETGREA**H**RGVISRSFQLEVHFSCVYAYEQVVKMPFALTPVDKLVQFMV**R**EGHFNVSMRLY KTASYLEPYDLLT**A**AVP**I**TDTLYVMLKIEGQHQLRYFLLSVEDCWATPS**A**DPYQD**V**LHELIEQGCPHD**E**T VTYLNAIGESTTAKFSFQMFQFVGYPKVFLHCRVRLCLPDGPEPCAKQCPT**L**W	3,4
CFCS		344	347	RSKR	5
Propeptide		348	390	A**L**A**D**DYNKIVSYGPIHLL**AA**PSL**RV**E**S**HHPRADQQE**L**KGPSLW	
	TM	391	407	LPGILILLCVLGVLT**MA**	6, 7
		408	418	AAAVSRRRRMV	

1. Peterson, T.N., et al., SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 2011. 8: p. 785-786.

2. Wouters, M.A., et al., Evolution of distinct EGF domains with specific functions. Protein Sci, 2005. 14(4): p. 1091-103.

3. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

4. Okumura, H., et al., A newly identified zona pellucida glycoprotein, ZPD, and dimeric ZP1 of chicken egg envelope are involved in sperm activation on sperm–egg interaction. Biochemical Journal, 2004. 384: p. 191–199.

5. Jovine, L., et al., A duplicated motif controls assembly of zona pellucida domain proteins. PNAS, 2004. 101(16): p. 5922-5927.

6. Cserzo, M., et al., On filtering false positive transmembrane protein predictions. Protein Engineering, 2002. 15: p. 745-752.

7. Cserzo, M., et al., TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. Bioinformatics, 2004. 20(1): p. 136-137.

Supplementary file 6.11 Chicken ZPAX1 ENSGGALG00000016465

ZP-N	1	89	MEHFKDKYLSFSAVDQSGIAWELDEALASQCGYTITYSSRNSIVFRASALSCHS*H*LEKDVFTVT**V**KIKASH TSDMKNATTYLRSASCPY	[1]
	90	97	RPWSPREL	
ZP-N	98	221	VCETNYMEVSARRDVPQTEKDIILSEPEDWILSYPKAKA**G**EASVWQILFHQPEEKRALLVSDAWRAGYG LNSTETRILLRVPYNTAHIQLV**KA**QGITFSAVRSSTFYKQQWMILMVDTAVACPV	[1]
	222	224	DGV	
ZP-N	225	338	NYTNKTIIWTVPKYFQALC**A**GATDFKDVLVEAGVNLHKLSAEEMASRKYVL*S*NDIN**T**ITMKIPIGAEGGS YKTSVSSGK**H**G**A**KY**S**INLFLEHQWEDNKWGLTKYTIIKEIETPF	[1]
	339	350	E**Q**VELAVTNNLN	
ZP-N	351	460	LSARLMNVTVGMFLLDVELVNLTIEGT**T**VTVPEAIQHGYLTYEIQYANGSKIY**V**IQVSFDAPGIKKEY**VI**DD TREYTLNVTL**K**FIILPTRDTF**S**VPIITVSAVKDAVLPS	[1]
	461	463	ARG	
ZP-N	464	558	FCDENDFHL**I**ITHGNVDQNWLPFISEQHLVPEVAQE**D**YYSLNDNGTHLTVSVPFLSSLVDYKDIHISGVM ASLHLTLKDGITLANKKDFS**I**SCRF	[1]
	559	562	PPSE	
ZP domain	563	837	LIQCLPNGTVVITAIKLVR*L*ADLDTSL**L**VLRDK**Q**CKPSLVTKKTATFKFNVNTCGTSRKFNSTSITYENDILYF RPGNDIPVYQLRFVC*V**YTIKHSADV*H*YENKKNLPPSIKPGFDSL**D**LSLKLFK**E**KSYSEPYQE*L**EYPVVKYL REALYFEVELLQPADPRLELNLEDCWATNSQSQDSLPRWPILINGCERSEDSYRTVFH**E**VNYS*R*RVKFPQ HLKRFEVTVFTFVQGTALLQMQLYLHCSVVICSTTPLPSDVICQRGCNPGTQRLGEHADF	[1]

1. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.

Supplementary file 6.12 - Chicken ZPAX2 ENSGGALG00000016475

Signal peptide	1	21	MGYSLCRIWMLFLLFCVGEER	[1]
	22	31	QMAPPGLVQS	
ZP-N	32	133	SCHSRIFWMKLNKLLLQGKFFQLEINDPYTGP**V**PLDDKLASRCGYVLSEDVWGN PVFRASVLACHV**V**NEADELFSLTVNIKISSFASMRAAVTYTYPMFCSY	[2]
	134	141	SSWASREI	
ZP-N	142	261	VCEENYMEVSVKTDVPAVSNDYTVAWMSALPETQNVAYQVWQLMFVSPSGR KRILVSDAAKLGYSFNNTLSRVYLRAPYHSNESE*I*S*V*VSGVDMNMITSTSMYRQR WLLLLIDMTVSCPL	[2]
	262	264	DG**T**	
ZP-N	265	376	SFTDTTLTWTVPSVIPTLVLQESTFLSKTILMGVDG*Q*HIVNPDKNNYLLEHNKTHI GITIPIGAEGGKLKSSISGGVYGIIYSIDLFLEHTWTDADWQTTKYTVIKSITTPF	[2]
	377	387	MPRIPTVINNT	
ZP-N	388	498	**L**PEEKIFNVAFGHFLPDVSLVAI**A**IGNVPFTLREAQHHGYKIYETPFSNGTKEFILEV SFDDPYVLKEYVNRNETKYTLL**V**NYTISTGPEMIPYYHSAEVECVIADIEIPE	[2]
	499	501	A**V**G	
ZP-N	502	596	**Y**CDEGNLYLAIPGFGLHQYWNLYLGTKLLNRHTANTNGYLATTNATHL**I**LQIPLF A**V**GV**I**YEEVSFQKIKARFDVALRKVRTMETLQTFSVSC**N**F	[2]
	597	602	NSPAFI	
ZP domain	603	911	LCYPDGT**V**I**I**SAQMKTVPGIDMSRTQLRDSSCKPKEYNKGHAFFKFHVTTCGTSV RFEGDHIVYENEISYEKETLQGQG**H**STITRDPDYRLTVLCYY**R**AKETVMLGAF*I*SK PSASHPSGSGTVVPRSN**S**AV*H*RRIRQALNVVSRVSKSESF*M*DFYEPN**V**VILK**R**P*T* ESVFLEVELKDESP*D*TELYLDNCWVTGSLDFNSTPRWNITVDGCEINGSEYVAVF CSVAASSRVRHPSHFKRLAVRTLTHRLEQVYVHCSVAACSAANTLPGIPCRGQCS PSTERNAFPGHNSAHLQGYVLAGPVWIVESDLR	[2]

1. Peterson, T.N., et al., SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 2011. 8: p. 785-786.

2. Callebaut, I., J.P. Mornon, and P. Monget, Isolated ZP-N domains constitute the N-terminal extensions of Zona Pellucida proteins. Bioinformatics, 2007. 23(15): p. 1871-4.