Comparative Genomics:

A CASE STUDY OF GENOME, CHROMOSOME AND GENE FAMILY EVOLUTION

HARDIP RAMESHBHAI PATEL

A THESIS SUBMITTED FOR THE DEGREE OF

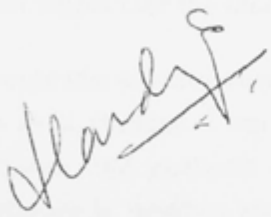
DOCTOR OF PHILOSOPHY

OF THE AUSTRALIAN NATIONAL UNIVERSITY

JANUARY 2010

Declaration

Except where specific reference is made to other sources, the work presented in this thesis is the work of the author. It has not been submitted, in whole or in part for any other degree.



Hardip Rameshbhai Patel

Acknowledgements

I would like to start by sincerely thanking my parents for believing in me and providing financial support for tuition fees. Without their constant encouragement and financial support, I would not have been able to accomplish this degree. I greatly appreciate their attitude toward my education and their unwavering support for my ambitions. I have great respect for the sacrifices they have made in helping me achieve my goals.

I would like to extend my sincere gratitude to Professor Jenny Graves for her acceptance of me as a student. I am thankful for her extraordinary supervision, constant support, patience and guidance over the past four years. My impressions of her unwavering enthusiasm, intellect and dedication to science will always be with me. It has been an immense pleasure to be associated with your science Jenny; Thank you for this amazing experience.

I would like to express my genuine gratitude to my supervisors Dr. Margaret Delbridge, Dr. Tancred Frickey and Dr. Edda Koina. Thank you for being excellent mentors. Marg, you have provided constant support and immense encouragement for the ideas that I presented to you during my research. You have helped me to become a better scientist. Tancred, I am extremely thankful for your patience when I approached you with my relentless questioning. Today, I can call myself "bioinfomatician" because of you. Thank you! Edda, thank you for your scientific expertise and warm personality that helped me get settled quickly in the lab and get past those initial jitters I had when I touched pipettes. To my surrogate supervisors Paul Waters, Janine Deakin and Tariq Ezaz, I thank you for your tremendous aid at different times during my project. I will always appreciate your friendship, advice and your willingness to help. To Denis O'Meally, I am grateful for your motivation and insightful discussions, which helped me during tough times.

I would like to express my sincere appreciation to my collaborator Ms. Chen Wei Wang and Amir Mohammadi who were patient and extremely helpful during the entire collaborative phase.

To members of the lab: thank you Paul Waters and Julie Chaumeil for beers, Denis O'Meally for drinking those beers with me, Hannah Bender for her lively personality, Vidushi Patel for curries, Tariq Ezaz, Tim Hore and Liz Murchison for intellectual conversations, Kristen Jordan-Rogers, Shafagh Al-Nadaf and Carly Smith for being good company for coffees. I would also like to thank Barbara Harris, Ke-jun Wei, Pat Meithke and Jan Elliott for their technical and administrative support. You guys made my life easier, thank you!

I would also like to thank all members of the lab not yet mentioned, for their help and friendship: Amber Alsop, Ruth Doherty, Nisrine El-Mogharbel, Jason Limnios, Daniel McMillan, Veronica Murtagh, Ali Livernois, Nerida Harley, and Claudia Rodrigues Delgado.

To my brother, sister and their spouses: thank you for accepting my commitment to my studies and for giving me the support to complete them. I really appreciate your encouragement and I value the impact you all have had on my life. To my daughter, who is not old enough to understand yet, thank you for your cute smiles and bubbly persona as it made the pain go away.

To my wife Anushka, I am extremely grateful for your love and patience, particularly during tough days, late nights and tight deadlines. I could not have done this without you and I am extremely lucky to have such a wonderful person to share my life with.

Abstract

Comparisons between the genomes have provided invaluable resources and tools for the annotation of DNA sequences. Comparative genomics is widely used to identify function of DNA sequences, understand dynamics of the evolutionary processes and provide framework for genome organization. Particularly useful for comparative analyses are the genomes of distantly related organisms such as marsupials and human. When the DNA sequence data was scarce and expensive to acquire, comparative analyses of genomes were performed by comparative gene mapping studies. However, with the advent of cheaper and faster DNA sequencing technology, genome scale comparisons at a single nucleotide level were made possible. This cheaper and faster DNA sequencing has resulted in an avalanche of information available for comparative analysis.

A major challenge of comparative analysis is to explore genome organization and evolution by locating the markers on the chromosome. This has traditionally been achieved by genetic linkage and physical maps, which are generated by different experiments and not always easy to align. I have pioneered a strategy to characterize the low coverage genome sequences, which were subsequently used for identifying microsatellite markers for the tammar wallaby linkage map. My strategy allowed for easy integration of the physical map and linkage map to construct an integrated map of the tammar wallaby genome. Specifically, I first identified and used conserved blocks of synteny between opossum and human, which were reconstructed for tammar wallaby by overlapping sequence searches and assembly process. These reconstructed conserved blocks of synteny were then assigned to tammar wallaby chromosomes using the physical mapping data. Since tammar wallaby linkage groups have previously been assigned to chromosomes, the gap regions and corresponding conserved blocks of synteny were easily identified from the karyotype. Subsequently microsatellite markers were sought in the conserved blocks of synteny and tested for polymorphism in the mapping families. 26 targeted markers were finally used for linkage analysis. This systematic use and characterization of the low coverage genome sequences led to a great saving in cost and effort for obtaining high quality linkage map (150 markers), which will be used for the assembly of the low coverage sequences.

Once the assemblies of the genomes are produced, they can be used to infer fine scale karyotype rearrangements that occurred during the evolution of chromosomes. For instance, comparisons of the sex chromosomes between the three major groups of mammals (placental, marsupial and monotreme) has led to a new understanding of how human sex chromosomes evolved from two genome blocks, one representing a conserved therian X (XCR) and one a region added in the placental lineage (XAR). A recent study involving genome assembly comparisons, in which the human X chromosome genes were

compared with the chicken genome, concluded that the human X chromosome is composed of three evolutionary layers. This conclusion was inconsistent with the widely accepted model proposing only two evolutionary layers, the X-conserved region and the X-added region.

I therefore investigated the origins of the genome blocks making up the human X chromosome to identify the cause of this inconsistency and resolve it. My comparative analysis of the location and order of the human X chromosome homologous genes in rat, opossum and chicken genomes revealed that the problem arose because of inaccurate assignment of chicken paralogs as the orthologs. I identified the true orthologs of the human X genes in the chicken/zebrafinch EST database and showed that paralogs were incorrectly identified because the orthologs were missing from an incomplete chicken genome assembly. I then mapped the orthologs of human X genes in tammar wallaby by using the fluorescent *in situ* hybridization, and compared them with platypus, chicken, lizard and frog genomes to conclude that the human X chromosome is composed of only two evolutionary layers, the X-added region and the X-conserved region, as originally proposed.

The assignment of orthologous or paralogous relationships in order to track gene evolution is particularly difficult for genes that belong to large families. In the last part of my research I analysed a huge and rather unusual gene family, the olfactory receptor gene (ORG) family. Comparative analysis of ORGs is difficult because the family is extremely large (~1000 genes in mammals). No comprehensive analyses have yet been performed to identify and characterize members of this gene family in a systematic fashion since the advent of large-scale genome sequence data. Therefore, I performed exhaustive searches to first identify olfactory receptor genes in all vertebrates. There are approximately 1000 olfactory receptor genes in mammals and frog, 500 in birds and 150 in lizards and fish. I also classified this gene family in 101 evolutionarily related groups of genes to provide a framework for dissecting evolutionary pathways. I also proposed a systematic nomenclature for this gene family based on the classification. This specialist data mining and classification strategy for olfactory receptor genes will provide unique opportunities to advance our understanding of this gene family in the future.

Table of Contents

1 Introduction

Five kingdoms of life, accommodating unicellular organisms to few-celled micro and macroscopic organisms to structurally complex multicellular plants and animals, have lived on the earth in the last 3,500 million years. Charles Darwin appreciated the unity of this life and proposed the theory of evolution based on phenotypic observations. The theory stated that all living forms have descended from common parents with modification through natural selection. Modern day scientists have underpinned Charles Darwin's theory of evolution by understanding the common biochemical molecule of all living things; deoxyribonucleic acid (DNA). The use of molecular techniques in deciphering the DNA content of an organism has revolutionized our understanding of biology. It has given us new perspective to explore the molecular basis of disease susceptibility, evolution of a species, biochemical pathways, phenotype and others.

DNA sequencing technology was first introduced in the 1970s (Sanger and Coulson 1975, Wu and Taylor 1971). At first it was possible to deduce only an 8 – 10 nucleotide sequence, but this was rapidly advanced to obtain sequences as long as 60 - 300 nucleotides (Maxam and Gilbert 1977, Sanger *et al.* 1977). Subsequently the use of fluorescent-labeled dideoxynucleotides instead of radioactive labeled dideoxynucleotides (Prober *et al.* 1987), improvement in sequencing technology and automation of DNA sequencing paved the way for large-scale genome sequencing (reviewed in Hunkapiller *et al.* 1991). In the past two decades high-throughput automated sequencing was efficiently applied for obtaining full-length genome sequence of the human (Lander *et al.* 2001, Venter *et al.* 2001) and many other organisms, for understanding the evolution and functional aspects of the genome (Entrez Genome Database , National Human Genome Research Institute).

The revolution in our understanding and interpretation of biology had begun. Some of the underlying aims of the first, very expensive genome sequencing projects were to identify functional elements of the human genome, characterize disease-associated loci, and understand physiology and evolution at a molecular level. With the acceleration of sequencing technology, and a reduction of costs by several orders of magnitude, the question now becomes how we can capitalize on the DNA sequence data to ascertain trends in evolution and assign functions to individual nucleotides present in the genome. Comparative genomics is the obvious choice for elucidating evolutionary paths and annotation of function to the DNA sequence of organisms (reviewed in Clark 1999, O'Brien *et al.* 1999, O'Brien *et al.* 1997, Womack and Moll 1986). This thesis focuses on the use of modern comparative genomics tools and techniques to enhance the understanding of vertebrate genome evolution.

1.1 Comparative genomics: means to an end

Comparative genomics is the systematic study of homologous DNA segments, genes, chromosomes or genomes. It is an invaluable resource for understanding heritable characteristics, behaviour and phenotypes in different species. This is mainly achieved by mapping similarities and differences in DNA sequence between species on the phylogenetic tree of species. For example, rate of mutations in homologous DNA sequence can be used to determine species divergence/emergence times provided species tree is known (*e.g.* Bininda-Emonds *et al.* 2007). Likewise, absence of *XIST* locus in platypus and marsupials suggested that X inactivation in eutherian mammals by *XIST* has evolved only in eutherians since their divergence from marsupials and monotremes 148 MYA (Hore *et al.* 2007). Regardless of the purpose of comparative analysis, it is essential to first identify homologous DNA sequences between species for detailed analysis.

Homologous DNA can be sub-classified into two major evolutionary relationships called orthology and paralogy. Considering the example of homologous genes, orthologous genes in two or more species are the genes descended from a single gene in the last common ancestor (Figure 1A) (Fitch 1970). In other words, orthologous genes are related to each other by a speciation event. Comparative analysis of the genomes (*e.g.* Goodstadt *et al.* 2007), obtaining assisted assembly of the low-coverage genome (*e.g.* Pontius *et al.* 2007), studying patterns of natural selection (*e.g.* Kosiol *et al.* 2008), assigning function to genes (*e.g.* Sasson *et al.* 2006) and constructing species phylogenetic trees (*e.g.* Swingley *et al.* 2008), are some of the examples where orthology identification was essential and implemented in comparative genomics.

Paralogs, in contrast to orthologs, are related to each other by a duplication event either prior to speciation in the common ancestor (outparalogs) or after speciation within a species (inparalogs) (Figure 1A) (Remm *et al.* 2001). The assignment of paralogy to genes in vertebrate genomes was essential to identify the whole genome duplication events in the common ancestor of vertebrates (Dehal and Boore 2005) and an additional round of duplication in the fish lineage (reviewed in Meyer and Van de Peer 2005). It is evident that orthology and paralogy assignments are crucial in comparative genomics to understand the evolution and organization of vertebrate genomes.

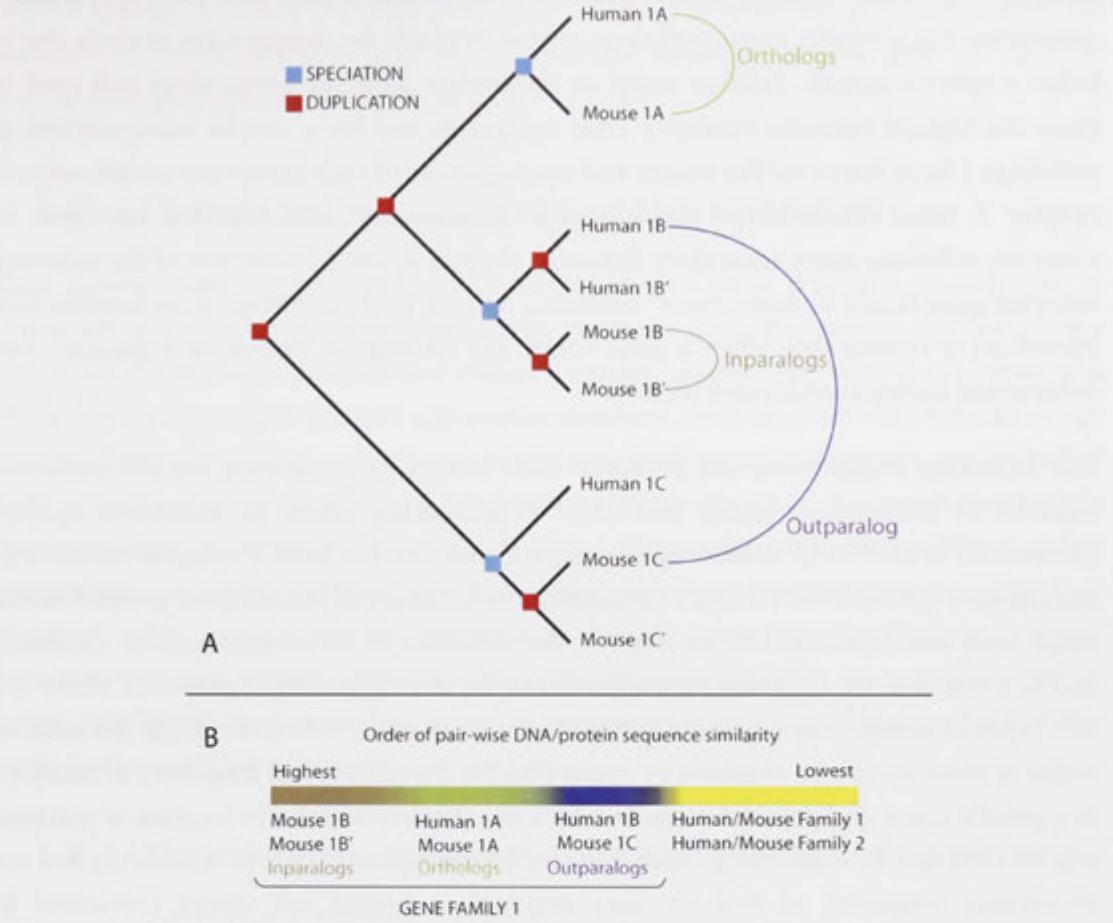


Figure 1 Example showing relationships of genes between human and mouse. (A) Gene 1A in human and mouse are separated by a speciation node suggesting that these genes are orthologs (green). Gene 1B and 1B' in mouse are related to each other by a duplication event suggesting that these genes are inparalogs (brown). Gene 1B in human and gene 1C in mouse are related to each other by a duplication event that occurred before the speciation event, suggesting that these genes are outparalogs. (B) Relative pair-wise sequence similarity of inparalogs is usually highest, followed by orthologs, followed by outparalogs.

The availability of genomic sequences has enabled the large-scale assignment of orthologs and paralogs between numerous species (Dehal and Boore 2006, Li *et al.* 2006, Li *et al.* 2003, O'Brien *et al.* 2005, Tatusov *et al.* 2003, Vilella *et al.* 2009). At the DNA/protein sequence level, inparalogs generally exhibit the highest pair-wise sequence similarity since they have evolved most recently through duplication within a given species (Figure 1B). In pairwise comparisons of two or more species, orthologs normally exhibit higher pairwise similarity compared to outparalogs. It should be noted that the pairwise similarity of inparalogs would be higher than orthologs. The pairwise similarity between the members of a gene family is normally higher than pairwise similarity with members of any other gene family. This property of relative pairwise similarity of members of the gene family can be used to identify orthologs and paralogs from genome sequence data. *However, care should be taken while using pairwise similarity as the only measure for ortholog/paralog assignment since species-specific gene loss, gene family expansions, gene*

divergence, gene conversion and incomplete genomic sequence data can result in erroneous annotation. For example, if true orthologs are not available for comparative analysis due to either a species-specific deletion event or incomplete data set, outparalogs will tend to show the highest pairwise similarity after inparalogs and these can be misconstrued as orthologs. I have discussed the causes and consequences of such erroneous assignments in chapter 3, titled "Evolution of the human X chromosome", and provided strategies to annotate orthologs more accurately. Similarly, chapter 4, titled "Evolution of the olfactory receptor gene family in vertebrates" discusses on how to classify large gene families into evolutionary relationship when a gene family has undergone significant expansion and contraction during evolutionary history.

The homology assignments can be made more robust by comparing the chromosomal location of homologous genes and their neighbouring genes in individual species (Andersson *et al.* 1996). In that regards, genetic mapping has been a valuable resource of comparative genomics to facilitate comparing the locations of homologous genes. Genetic maps have also been used to understand the dynamics of vertebrate genome evolution and as a resource for assigning chromosomes to the draft assembly of genomes. There are two types of genetic maps; genetic linkage maps (Sturtevant 1913), in which the relative order of markers can be obtained by measuring the recombination frequency of markers in a genetic cross, and physical maps, in which the absolute or relative location of markers can be obtained by either fluorescence *in situ* hybridization (FISH) or measuring the co-occurrence frequency of two or more markers in hybrid cell clones (reviewed in Chowdhary and Raudsepp 2005, Moran and James 2005). Genetic maps have been widely used in comparative genomics to elucidate major evolutionary trends in vertebrate genomes.

Apart from genetic maps, chromosome painting is also used for comparative analysis at the cytogenetic level (Breen *et al.* 1999, Glas *et al.* 1999, Goureau *et al.* 1996, Rens *et al.* 1999, Shetty *et al.* 1999, Yang *et al.* 2003). Chromosome painting has specifically been very useful in understanding the genome evolution and organization of closely related species for which genomic sequence data is not available (Rens *et al.* 2003). Flow-sorted or micro-dissected chromosomes, or chromosome enriched DNA libraries, are used as fluorescence-labeled probes for hybridization onto the fixed metaphase chromosomes of the species of interest (target DNA). The chromosomal segments of the target species that are homologous to the probe DNA will hybridize to reveal patterns of conservation between probe and target DNA. In recent years, a more sophisticated form of chromosome wide comparisons has been made possible, owing to genomic sequence availability, called chromosome e-painting (Kemkemer *et al.* 2006). In chromosome e-painting, the orthologous genes (instead of chromosomal DNA probes) are mapped to a reference species genomic sequence (instead of the metaphase chromosomes) by using sequence alignment tools. Large blocks of conserved synteny are then inferred using tools like the

GRIMM synteny software (Pevzner and Tesler 2003a). Chromosome painting is useful for highlighting major chromosomal rearrangements between species and its need in comparative genomics cannot be neglected.

1.2 Comparative analysis of the genomes

Both genetic maps and chromosome painting studies have played a crucial role in comparative genomics for understanding the trends in genome evolution and the correct definition of homology. Comparisons between species usually begin with the selection of a set of markers (genes, conserved non-coding elements, microsatellites and others) in one species. Corresponding homologous markers and their genomic positions are identified from genetic maps in the other species. Based on the locations of the genetic markers in the two species, a comparative map can be drawn. This comparative map can be useful in ascertaining conserved blocks of synteny between two species which then can be further explored for gene family evolution, repeat expansion/contraction, major insertion/deletion events, genome rearrangement events and the proto-ancestral karyotype. The following section discusses vertebrate genome evolution in light of the comparative mapping of homologous markers.

1.2.1 Homologous syntenic blocks: pieces of vertebrate genomes' jigsaw puzzle

A homologous syntenic block (HSB) is a modern terminology for referring to the segment on a chromosome of one species that is conserved in almost similar order in all species for which these HSBs are defined (Murphy *et al.* 2005). This definition of homologous syntenic blocks will be used for remainder of this thesis. One of the earliest comparisons of genetic linkage maps uncovered 36 HSBs between human and mouse (Nadeau and Taylor 1984). The distribution of length of the observed HSBs between human and mouse fitted the exponential curve, as one would expect from the random distribution of HSB length. This led to popular theory of "random breakage model" of chromosome evolution. Comparisons of genetic maps of the cow and cat aligned against the human genetic map revealed relatively high conservation of synteny. This contrasted with human-mouse comparisons, where synteny disruption is more frequently observed (Lyons *et al.* 1997, O'Brien and Nash 1982, Womack and Moll 1986). Multi-species chromosome painting including primates, carnivores, artiodactyls and perissodactyls also revealed large blocks of conserved synteny between various mammalian species (Wienberg and Stanyon 1997). Unfortunately, the analysis of the distribution of HSB length using data from multiple species was not possible since genetic maps of cow and cat were not dense enough to identify all homologous syntenic blocks, big or small, and chromosome painting could not reveal small or intra-chromosomal rearrangements.

However, the availability of full-length DNA sequence from human and mouse showed the need to revisit the "random breakage model" (Pevzner and Tesler 2003a; 2003b). Detailed comparison of full-length human and mouse genomes revealed 281 homologous syntenic blocks. The distribution of length of the newly discovered 281 homologous syntenic blocks did not fit the exponential distribution curve (the basis of the random breakage model) because numerous small HSBs that were discovered were either discarded as statistical noise or not discovered when random breakage model was proposed. Therefore alternative "fragile breakage model" of genome evolution was proposed (Pevzner and Tesler 2003a; 2003b). In the fragile breakage model, the genome consists of short fragile regions with higher propensity towards breakage and large solid regions with lower propensity towards breakage. The probability of a breakpoint in the short fragile region follows a *Poisson* process and the probability of breakage in the solid regions is zero in this model. The probability distribution of the fragile regions and solid regions in the genome was in agreement with the number of HSBs and breakpoints regions discovered between human and mouse genome (Pevzner and Tesler 2003b). This fragile breakage model was further put to the test in other mammalian and bird species.

The fragile breakage model of genome evolution suggests a dedicated set of loci (breakpoint regions) in the genome that have a relatively higher propensity towards breakage during evolution compared to the solid regions (homologous syntenic blocks). Therefore, the fragile regions should be common between species and reused for karyotype rearrangements during evolution and solid regions should appear as conserved blocks of synteny in multiple species with little probability of breakage in them. The comparison of cow BAC-end sequence maps with that of the human and mouse genomes revealed that cow, human and mouse genomes have evolved as mosaics of homologous syntenic blocks (solid regions), thus supporting the fragile breakage model (Everts-van der Wind *et al.* 2005, Larkin *et al.* 2003, Wind *et al.* 2005). Comparing 3,204 homologous markers covering approximately 91% of the human genome, 201 homologous syntenic blocks were identified between cow and human. However, more than a two species comparison is essential for identifying breakpoint region reuse. Comparative analysis of the human genome with that of the rat, mouse, cat, dog, cow and pig genomic maps revealed 256, 267, 111, 205, 232 and 144 homologous syntenic blocks in each species respectively (Murphy *et al.* 2005). This multi-species comparison also revealed that of all the breakpoint regions (fragile regions) observed between human and other species, at least 20% of them are reused (same breakpoint region in more than once species), which supports the fragile breakage model. More recently, a comprehensive analysis was performed based on the genomic sequences of eutherian mammals, opossum and chicken (Kemkemer *et al.* 2009a, Larkin *et al.* 2009). Rigorous testing of random breakage model (null hypothesis) against observed breakpoint regions (test hypothesis) confirmed that the observed breakpoint regions in amniotes are not randomly distributed in the genome but they are confined to "fragile regions" of the genome, thus supporting the fragile

breakage model. The homologous syntenic blocks are observed between fish and human as well (Barbazuk *et al.* 2000, Sasaki *et al.* 2007). This strongly suggests that vertebrate genomes are made up of homologous syntenic blocks, which might be considered as jigsaw pieces that have different arrangements in different species.

1.2.2 Drivers of vertebrate genome evolution

The fragile regions in the genome are more prone to rearrangements and they are reused in karyotype changes during vertebrate evolution (Larkin *et al.* 2009, Murphy *et al.* 2005). What are the characteristic features of these fragile regions and are they involved in causing breaks in the karyotype? Examination of the chicken chromosome 28 sequence, BAC clone sequences from the gibbon containing an evolutionary breakpoint (the region where at least one species have karyotype rearrangement), and the full-length alignment of the mouse genome against the human genome suggest that there was enrichment of lineage specific genes and segmental duplications in the breakpoint regions (Bailey *et al.* 2004, Carbone *et al.* 2009, Gordon *et al.* 2007). Similarly transposable elements, tandem repeats, L1 LINE repeats, *Alu* repeats, endogenous retrovirus sequences, cancer breakpoints, CpG islands, gene density, mutation density (SNPs, indels, copy number variation), and GC nucleotide density are also enriched in the fragile regions of the genomes (Larkin *et al.* 2009, Longo *et al.* 2009, Murphy *et al.* 2005, Ruiz-Herrera *et al.* 2006, Schibler *et al.* 2006). Detailed analysis of the location of breakpoint regions with respect to protein coding genes shows that these breakpoint regions are more likely to occur in intergenic regions compared to regions within genes (Lemaitre *et al.* 2009). This is because if the breakpoints disrupt coding genes, the genes may be rendered non-functional and therefore, the breakpoints are selected against in the regions coding for genes. The characteristic features enriched in the breakpoint regions are merely associations of these features with the breakpoint regions. It is not known if these features cause the breakpoints in the genome or they are consequence of the breakpoints in the genome. Further studies are required to confirm the causes and consequences of breakpoints in the vertebrate genomes.

1.3 Scope of the present study

Comparative gene mapping, chromosome painting, survey sequencing and genomic DNA sequencing techniques have all helped deduce the evolutionary paths and annotate newly sequenced genomes in the most efficient manner. We have come a long way in the genomic sequence era, however, large numbers of vertebrate genomes are sequenced at a lower 2-fold coverage only and almost all of them remain without a good assembly representing contiguous stretches of chromosomes. The understanding of the origin and evolution of parts of genomes has remained ambiguous and unexplained because of the

poor assembly (*e.g.* gene dense regions on the human X chromosome). Moreover, large gene families have always been avoided for systematic analysis and annotations (*e.g.* the olfactory receptor gene family). Comparative genomics can be efficiently applied to address some of the key questions in the organization and evolution of genomes, chromosomes and gene families.

The overarching theme of my thesis is the application of comparative genomics tools and techniques to understand the organization and evolution of genomes, chromosomes and gene families. My studies to this end have included characterization of the newly sequenced low coverage tammar wallaby genome to obtain a better assembly, comparative analysis of gene families in the gene dense regions of the human X chromosome to understand their origin and identification and annotation of the olfactory receptor gene family in vertebrates. Several key concepts and bioinformatics tools that were used throughout the thesis have been explained in Appendix.

I participated in a major effort to sequence and assemble the genome of the tammar wallaby (*Macropus eugenii*), the Australian model marsupial. The tammar wallaby genome was sequenced to 2× coverage by the Australian Genome Research Facility (AGRF) and the Baylor College of Medicine (BCM) (Graves *et al.* 2003). My role was to identify microsatellite markers for obtaining comprehensive genetic linkage map of the tammar wallaby chromosomes. It is essential to assign the chromosomal locations to the genomic sequence obtained by large scale sequencing projects to understand how genomes are organized and evolve (Lewin *et al.* 2009). The second chapter titled "Organization and evolution of the tammar wallaby genome" discusses the research work on the identification of microsatellite markers for the linkage map of the tammar wallaby genome. The tammar wallaby genome is used as the case study here showing how the *de novo* sequenced genome can be approached in targeted manner to maximize the efficiency in terms of time, economy and accuracy.

In the next part of my research, I used physical mapping of genes and multispecies comparative analysis to determine how the human X chromosome has evolved. Previous comparative analysis of the human X chromosome showed that the eutherian mammal X chromosome acquired an autosomal region in the last 100 million year since eutherian radiation. However, the comparative analysis of the human X chromosome with the chicken genome assembly also revealed another block on the human X chromosome with distinct evolutionary history. I present my research on the comparative analysis of these distinct blocks on the human X chromosome with multiple species in the chapter titled "Evolution of the human X chromosome". This chapter highlights the use of genetic maps to infer correct homology and suggests that careful selection of the ortholog identification methods is required to elucidate the evolutionary history of gene families.

In the last part of my research, I used modern genome analysis tools for identification of the olfactory receptor genes in vertebrates from publicly available genomic sequences. The olfactory receptor gene family is the largest gene family in mammals. In fish it comprises only 100 genes, but it has been expanded to approximately 1000 genes in mammals. The current phylogenetic classification is not informative for tracing the evolution of this large gene family in vertebrates. I will present my work on the identification of this gene family in vertebrates in the chapter titled "Evolution of the olfactory receptor gene family in vertebrates". I have also developed a classification system for this gene family to make future studies more consistent by eliminating the need to revisit the classification when more sequence data is available. This work emphasizes the use of pairwise similarity scores and partitioning based clustering to group closely related genes in large-scale annotations.

In the final section of this thesis, I will discuss the importance of comparative genomics in the post genome sequence era with emphasis on the identification and use of homologous syntenic blocks. I will present my research work in the light of recent advancements in DNA sequencing technology and provide a brief summary of the key findings.

2 Organization and evolution of the tammar wallaby genome

Marsupials were described as peculiar animals with resemblance to a greyhound with a long tail and jumping like a hare by Captain James Cook, the first European visitor to the Australian continent, in 1770 (c.f. Dixon 2008). Indeed marsupials are peculiar and they have not only been popular in children's stories, but have been of great importance for biomedical, cytological and evolutionary studies. Tammar wallaby particularly makes a better choice for a model organism than the opossum because it is more appropriate in size for repetitive blood and tissue sample collection and the pregnancies are highly manipulative for captive breeding. They have been extensively used in a variety of biomedical research fields including lactation, growth, olfaction, pathology, immunology, digestion, sexual differentiation, metabolism, embryology, pregnancy, endocrinology, placentation, biomechanics, contraception, hematology, parturition, chromosome evolution, epigenetics, imprinting, neurobiology, and behavior (reviewed in Hickford *et al.* 2009).

The Australian Research Council (ARC) Centre for Kangaroo Genomics in Australia (http://kangaroo.genomics.org.au) was developing vast numbers of resources for genetics and genomics studies for the tammar wallaby. One of the major tammar wallaby genomics resources was the low coverage (2×) whole genome shotgun sequencing obtained by a joint effort of the Baylor College of Medicine (BCM) and the Australian Genome Research Facility (AGRF) (Graves *et al.* 2003). It is essential to supplement any genome sequencing effort with a genetic map to identify evolutionary conserved elements (phylogenetic footprinting), identify small scale rearrangements and reconstruct ancestral karyotypes for elucidating the rate and direction of genome rearrangements (Lewin *et al.* 2009). To aid in the assembly of the low coverage tammar wallaby genome, a concerted effort was launched to physically map genes to tammar wallaby chromosomes and generate a genetic linkage map (Alsop *et al.* 2005, Deakin *et al.* 2008, Zenger *et al.* 2002). These genetic physical and linkage maps were then to be integrated to provide a framework for assembly, as well as to locate phenotypic markers.

I was involved in the generation of the genetic linkage map. The following chapter discusses various bioinformatics resources used to target regions for filling gaps in the current linkage map. A second-generation tammar wallaby linkage map was being generated by my collaborators at the University of Sydney (primary contact: Ms Chen Wei Wang, PhD Student, Faculty of Veterinary Science, University of Sydney, Australia). I identified microsatellite markers, which were used by Ms Chen Wei Wang in the linkage analysis using hybrid phase-known backcrosses from genetically distinct tammar wallabies from Kangaroo Island and Garden Island. I will co-author the publication on the second-generation tammar wallaby linkage map in the near future.

2.1 Introduction

Marsupials belong to the therian mammal infraclass called Metatheria, which diverged from placental mammals, the eutherian mammal infraclass called Eutheria, 148 million years ago (Bininda-Emonds *et al.* 2007). The common ancestor of both marsupials and eutherian mammals diverged from the mammalian subclass Prototheria (monotreme mammals) 168 million years ago (Bininda-Emonds *et al.* 2007). This divergence time places marsupials uniquely in the mammalian phylogeny as the link between monotreme mammals and eutherian mammals.

The study of marsupial genomes would shed light onto the organization of the therian mammal ancestral karyotype and help us predict the evolutionary paths of mammalian genomes. Although, full-length genome sequence of South American marsupial, opossum (*Monodelphis domestica*) has been available since 2007 (Mikkelsen *et al.* 2007), it is important to understand the organization of tammar wallaby genome because they have diverged some 80 million years ago, much the same as the divergence time between human and mouse (Bininda-Emonds *et al.* 2007). Moreover, very few resources apart from the genome sequence exist for opossum genomics compared to tammar wallaby. For example, in NCBI EST database, only 265 mRNA sequences are available for opossum compared to 14,878 mRNA sequences from the mammary gland of female tammar wallaby (Lefevre *et al.* 2007). This EST data can be used to predict more accurate gene models specific for marsupials.

The low coverage tammar wallaby genome sequences have recently been assembled to form 1,174,382 contigs with N50 size of 2.6 Kb (*i.e.* 50% of the total non-redundant bases sequenced from the tammar wallaby genome have been assembled into contigs larger than 2.6 Kb). These contigs have been organized into 277,711 larger scaffolds with N50 size of 36.6 Kb. However, this extremely large number of contigs and scaffolds are not yet assigned to chromosomes. The genomic sequences without a good genetic map provides no information about the genome rearrangements (Lewin *et al.* 2009). Therefore it is essential to construct genetic maps for tammar wallaby to infer karyotype rearrangements and aid the assembly of the large number of contigs generated after the low coverage genome sequencing.

2.1.1 Genetic maps

Gene/genetic maps refer to the arrangement of the genetic markers on the chromosomes. There are two types of maps; the genetic linkage map, in which the relative order of markers can be obtained by measuring the recombination frequency of markers in a genetic cross, and the physical map, in which absolute or relative location of markers can

be obtained by either fluorescence *in situ* hybridization (FISH) or measuring the co-occurrence frequency of two or more markers in hybrid cell panel.

2.1.1.1 Genetic linkage maps

Eukaryotic chromosomes are molecules of nuclear DNA harbouring genes, regulatory elements and other nucleotide sequences along with proteins like histone that aid in packaging of DNA in the nucleus. These chromosomes replicate, divide and are passed on to daughter cells to ensure the genetic diversity and survival of the progeny. The pair of homologous chromosomes undergoes recombination and exchange of material during meiosis. It is the frequency of this homologous recombination between two markers that is used to construct genetic linkage map. If two markers are far apart on the chromosome, they tend to have higher recombination frequency between them and hence their alleles are more likely to split apart during recombination. Conversely, if two markers are physically close on the chromosome, they tend to have a lower recombination frequency between them and hence their alleles will tend to stay together during recombination. Thus the frequency of recombination between two markers can be used as a measure to elucidate the relative distance between two or more markers for linkage mapping. The frequency of recombination is calculated by statistical analysis of the segregation of parental polymorphic markers in progeny, to generate a genetic linkage map. Single nucleotide polymorphism (SNP), microsatellite polymorphism, restriction fragment length polymorphism (RFLP), amplified fragment length polymorphism (AFLP), protein polymorphisms and phenotypes can be used as markers for construction of genetic linkage maps.

The first genetic linkage map was deduced for six sex-linked genes in *Drosophila* based on the segregation of eye colour and wing size phenotypes in progeny (Sturtevant 1913). Since then, genetic linkage maps have been useful in range of applications from the discovery and mapping of economically important traits in farm animals (*e.g.* Beraldi *et al.* 2007, Georges *et al.* 1995) and disease susceptibility loci in human (*e.g.* St George-Hyslop *et al.* 1992) to comparative analysis of the human and mouse genomes (Nadeau and Taylor 1984). A genetic linkage map indicates which markers are likely to be on the same chromosomes, and the outcome is the identification of "linkage groups" of genes that segregate together. These linkage groups represent physical chromosomes, chromosome arms or regions, but provide no information on the physical location or orientation of these groups.

2.1.1.2 Physical maps

Physical maps, in contrast to genetic linkage maps, are used to deduce the absolute or relative location of markers along the length of the chromosome. There are three major

approaches used for constructing physical maps, *viz.*, somatic cell hybrid analysis, radiation hybrid analysis and *in situ* hybridization analysis (reviewed in Chowdhary and Raudsepp 2005). In somatic cell hybrid analysis, somatic cells of one species (donor) are fused with somatic cells of the recipient species (usually rodent). This fusion is facilitated partly by transformation of recipient cells during which they incorporate donor chromosomes in the nucleus. In subsequent cell divisions of the fused cells, recipient cells will preferentially and randomly lose donor cell chromosomes, thus resulting in hybrid cell clones with one or more donor chromosomes. A panel of such hybrid cell clones is selected for analysis and characterized for the frequency of the co-occurrence of markers. The frequency of co-occurrence of two markers on the same chromosome will be higher than the frequency of co-occurrence of two markers on different chromosomes in a hybrid cell clone. Thus, the syntenic relationship of markers can be deduced using somatic cell hybrid analysis.

Radiation hybrid analysis is similar to somatic cell hybrid analysis except that before fusion, donor cells are irradiated with lethal doses of X-rays to fragment the donor chromosomes. Thus in radiation hybrid panel, different combinations of chromosomal *fragments* of donor cells (unlike different combinations of *whole* chromosomes of donor cells) are present in the hybrid cell clone. A radiation hybrid panel is superior to a somatic cell hybrid panel because not only it can resolve synteny of markers but also offers greater resolution to obtain the relative order of markers as well. Once the syntenic relationship of markers is obtained by somatic cell hybrid or radiation hybrid analysis, one of the markers within a syntenic block can be physically localized onto metaphase chromosomes to assign chromosome number to the syntenic markers.

In situ hybridization techniques are used for identifying absolute physical location of markers on the chromosomes. In this technique, marker-specific DNA is fluorescence-labeled and used as a probe. The hybridization of probe DNA to the homologous chromosomal segments can be visualized under microscope to identify chromosomal location of the probe (marker).

2.1.1.3 Advantages and limitations of different genetic maps

Physical mapping by *in situ* hybridization is the only technique for obtaining the absolute location of a marker on a chromosome. It is more practical to obtain the relative location of only a few markers by *in situ* hybridization owing to limited number of fluorescent dyes available for tagging the probe in contrast to theoretically unlimited number of markers that can be relatively ordered by hybrid cell analysis and genetic linkage mapping. Moreover, the relative order of thousands of markers can be deduced using genetic linkage mapping or hybrid cell analysis compared to *in situ* hybridization (Warrington and Bengtsson 1994). Except for the *in situ* hybridization technique, all other techniques also

offer the advantage of adding new markers to same map at a subsequent stage to increase the density of the map. Another disadvantage of *in situ* hybridization is that although probes as smalls as 1 kb can be used, routinely considerably larger (30-200 Kb) probes would be required for ease of detection and hybridization. In contrast, markers as small as a single nucleotide can be used in all other techniques. Nevertheless, physical mapping by *in situ* hybridization is the only technique that can reveal the location of a genetic linkage group or syntenic markers identified by hybrid cell analysis.

Genetic linkage mapping has its disadvantages as well. The recombination frequency varies along the length of the chromosome (Fullerton *et al.* 2001, Gorlov *et al.* 1994) and also between male and female of the species (Broman *et al.* 1998, Vanoorschot *et al.* 1992). These variations can lead to varying length of the genetic linkage map between the males and females of a species or a negatively correlated genetic linkage map length when compared to the physical size of the chromosomes. Even radiation hybrid panels and somatic cell hybrid panels have their own disadvantages, due to the instability of the hybrid clones during cell division. Moreover, the uptake of a chromosome or chromosomal fragment by recipient cells in hybrid cell clones may lead to incomplete representation of the genome, thus reducing the coverage of the linkage maps. Considering these advantages and disadvantages, it is therefore clear that the generation of both the physical and linkage map provides the optimum solution for assigning loci to chromosomes. Genetic maps have been valuable resource in comparative analysis of genomes and assigning chromosomes to contigs and scaffolds generated by whole genome sequencing efforts, which in turn have been useful in understanding the evolution of genomes.

2.1.2 Comparative genetic mapping

The targeted selection of markers for genetic mapping and subsequent comparative analysis can be made in one of several ways. The first is to use the same set of homologous markers for comparative mapping to reveal arrangement of these markers across divergent species. One of the early attempts to aid such comparative mapping was the development of degenerate Polymerase Chain Reaction (PCR) primers for amplification of known genes (Jiang *et al.* 1998, Lyons *et al.* 1997, Venta *et al.* 1996). The overall efficiency of these degenerate primer pairs to amplify unique products ranged from 16.7% to 52.4% for primers designed by Lyons *et al.* (1997) and 36.4% to 80.8% for primers designed by Jiang *et al.* (1998). The efficiency of amplification was 84% for the primers designed for 11 genes by Venta *et al.* (1996). The lower efficiency is mainly because of the fact that very limited gene sequence data was available at that stage to identify highly conserved regions for better primer design. However, with the availability of DNA sequence data from numerous species, universal primers can be designed more efficiently by either including more species data or including data from closely related species. U-probe is software that uses multiple species alignment to identify highly conserved

oligonucleotides for either screening BAC library or PCR reactions (Kellner *et al.* 2005). The success rate for identification of the correct BAC clone using oligonucleotides designed by the U-probe system was as high as 98%. Regardless of the tools used for probe design, the use of homologous markers for comparative mapping is more informative because *all of the markers used to create genetic maps* are useful in cross-species comparative analysis of the marker rearrangements.

The second way of creating more informative genetic maps is the use of markers that lie close to the breakpoints in synteny. The physical mapping in the tammar wallaby was approached by isolating BAC clones containing genes that were specifically selected to target the end points of conserved syntenic blocks between the human and opossum genomes (Deakin *et al.* 2008). It is known from comparative studies that vertebrate genomes have evolved in blocks of conserved synteny and there are specific sites in the genome that are more prone to breakpoints involving karyotype rearrangements (discussed in section 1.2). Therefore genetic maps targeting breakpoints of karyotype rearrangement are the most informative genetic maps for understanding genome organization since it captures many of the known breakpoints in vertebrate genomes. To construct a breakpoint-targeted genetic map in a species, homologous syntenic blocks between a closely related species (*e.g.* opossum is the most closely related species to tammar wallaby) and a distantly related species (*e.g.* human) are first identified. The genes that lie at the ends of these conserved syntenic blocks can be mapped either by linkage analysis or physical mapping. The resultant genetic map can be compared with other species to learn about the karyotype rearrangements specific to the species under investigation, or groups of closely related species, or even between different vertebrate classes. This breakpoint targeted genetic mapping forms the basis of this chapter in the thesis.

2.1.3 Assisted assembly of low coverage genomes

Genetic mapping is essential to infer major karyotype rearrangements in the vertebrate lineage (discussed in section 1.2). There are several vertebrate genomes sequenced at a low coverage (National Human Genome Research Institute). One of the biggest challenges of obtaining meaningful information out of low coverage genomes is the assembly of such low coverage genomes (Green 2007). For example, the cat genome was sequenced at a low coverage (1.9×) similar to that of the tammar wallaby, and both genomes are approximately 3 billion base pairs (Pontius *et al.* 2007). The assembly of the low coverage cat genome sequence, not surprisingly, generated a large number of contigs (hence more gaps in the final assembly) with the N50 size of the contigs equal to only 2,378 bp (*i.e.* 50% of the total non-redundant bases sequenced from the cat genome have been assembled into contigs larger than 2,378 bp). Only 65% of the euchromatic genome was represented in the cat genome assembly. Unordered contigs generated by the assembly of such low

coverage sequencing cannot be efficiently localized onto chromosomes since the total number of contigs is very large (817,956 contigs for the cat genome). Therefore chromosome scale or genome scale comparative analysis becomes rather difficult.

The availability of a radiation hybrid map consisting of 1,680 ordered marker loci, however, facilitated the organization of a large number of contigs on the chromosomes in the cat genome assembly. 54% of the cat euchromatic genome, although present in numerous small contigs and scaffolds, was eventually assigned a chromosome. The cat genome is one of the only examples showing that even the low coverage genome sequences can be ordered and oriented for comparative analysis. Although markers in radiation hybrid map were randomly chosen markers (Murphy *et al.* 2007), the density of markers was sufficient enough to aid in the assembly of this genome sequence. Generally, the higher the number of markers, the more expensive the studies become. Therefore it is crucial to limit the number of markers required to cover the entire genome of the organism. For this reason, it was essential to use targeted physical and linkage mapping to assist the assembly of the low coverage tammar wallaby genome assembly.

2.1.4 Marsupial genetic maps

The first linkage analysis in marsupials revealed that genes for adenosine deaminase, glucose phosphate isomerase, and protease inhibitor proteins form one linkage group in fat-tailed dunnart (*Sminthopsis crassicaudata*) (Bennett *et al.* 1986). Similarly, a second linkage group in fat-tailed dunnart consisted of genes for superoxide dismutase, transferrin and 6-phosphogluconate dehydrogenase proteins. This linkage analysis showed for the first time that the meiotic recombination frequency in female marsupials is much lower than in male marsupials. This contrasted with observations that in humans the male recombination frequency is lower than female (Broman *et al.* 1998). Like the fat-tailed dunnart, females of opossum also have lower recombination frequency compared to males (Vanoorschot *et al.* 1992) but, in tammar wallaby, it was initially shown that female recombination frequency was higher than males (McKenzie *et al.* 1995). These contradictory results for the tammar wallaby were more recently reconciled by using more markers and more informative meiosis (Zenger *et al.* 2002) and it was shown that in tammar wallaby, like opossum and fat-tailed dunnart, the recombination frequency in females is lower than males. It is interesting to note that this recombination frequency difference between male and female marsupials was a regional effect and not a general genome-wide phenomenon suggesting differences in rates of recombination along the length of the chromosome (Bennett *et al.* 1986, Samollow *et al.* 2004, Zenger *et al.* 2002). Genetic linkage analysis not only revealed that there were differences in male and female meiotic recombination frequency, but also the relative order of markers in the synteny.

The comprehensive linkage maps exists for both tammar wallaby and opossum. The first generation linkage map of opossum consisted of 83 loci representing nine linkage groups with a total map length of 633 cM (Samollow *et al.* 2004). Eight of the total nine linkage groups were autosomal and one was linked to the X chromosome. The continued effort to cover the entire genome and obtain physical location of linkage groups, enlarged the opossum linkage map to cover approximately 90% of the genome by using 150 marker loci (Samollow *et al.* 2007). All the linkage groups were subsequently assigned to chromosomes by using fluorescence *in situ* hybridization. To aid the assembly of the opossum genome, 381 BAC clones representing assembly scaffolds were also physically localized to the opossum metaphase chromosomes (Duke *et al.* 2007). The opossum genetic linkage and physical maps were integrated to inform the assembly of the opossum genome (Mikkelsen *et al.* 2007). Despite the significant coverage of the opossum genome by genetic maps, there are approximately 1,000 genes in the opossum genome sequence assembly that have not yet been assigned to the chromosomes.

The first generation linkage map of tammar wallaby consisted of 64 genetic markers representing nine linkage groups with a total map length of 824 cM (Zenger *et al.* 2002). The tammar wallaby karyotype consists of seven autosome pairs and one sex chromosome pair ($2n$ = 16). Eight out of nine tammar wallaby linkage groups were assigned to autosomes with two linkage groups representing chromosome 1 and one linkage group was assigned to the X chromosome.

The markers chosen for the tammar wallaby linkage analysis were randomly selected, but genetic maps derived from the random selection of markers are not ideally suited for comparative analysis because the inclusion of all the breakpoint regions in the synteny cannot be guaranteed. For example, if the randomly chosen markers are far apart on the chromosomes, the internal rearrangements between these two markers cannot be discovered since there is no data available. Similarly two closely located markers are also not very informative because they are more likely to represent the solid region of the genome (~95% of the genome) where the probability of the breakpoint is minimal according to the fragile breakage model discussed in section 1.2.1 (Pevzner and Tesler 2003b). Therefore it is essential to carefully select markers for genetic maps so that efficient cross species comparisons of breakpoints and karyotype rearrangement can be made as discussed in sections 2.1.2 and 2.1.3.

2.1.5 Tammar wallaby genome mapping: ongoing efforts

The marsupial karyotype is extremely conserved and the homology between chromosomes across the entire infraclass is known by chromosome painting studies (Rens *et al.* 2003, Rens *et al.* 1999). The number of chromosomes in marsupials varies from 2n = 10 in the swamp wallaby (2n = 11 in male swamp wallabies) to 2n = 32 in the rufous rat

kangaroo (Hayman 1990). Despite the large differences in the number of chromosomes between species, chromosome painting studies have shown that the marsupial karyotype can be divided into 19 blocks of conserved synteny (Glas *et al.* 1999, Rens *et al.* 2003). This means that genetic maps and assembly of **two** marsupial species can help identify all the 19 blocks of conserved synteny in marsupials. These blocks can thus be used for comparative analysis of all marsupials.

The opossum genome assembly is already available in the public domain (Mikkelsen *et al.* 2007). Recently, a low coverage tammar wallaby genome assembly has also been made available. However, tammar wallaby genome sequences have not yet been assigned to chromosomes owing to difficulty of assembling this low coverage genome (assembly version 1.1 can be viewed in NCBI nucleotide database with "macropus eugenii[Organism]" as the search term). Therefore, physical mapping of genes to tammar wallaby chromosomes was initiated in Professor Jennifer Graves' laboratory as part of the work of the ARC Centre of Excellence for Kangaroo Genomics. BAC clones were isolated that contained orthologs of mapped human genes, and these were physically mapped to tammar wallaby chromosomes by fluorescence *in situ* hybridization technique. This was initially a lengthy gene-by-gene effort; however, it was vastly accelerated by employing a strategy of identifying conserved blocks of synteny between human and opossum, and mapping only the endpoints of the conserved syntenic blocks (Deakin *et al.* 2008).

Additionally, tammar wallaby linkage mapping studies were also in progress by the efforts of collaborators at the University of Sydney (primary contact: Ms Chen Wei Wang, PhD Student, Faculty of Veterinary Science, University of Sydney, Australia). The tammar wallaby linkage analyses were performed using hybrid phase-known backcrosses from genetically distinct tammar wallaby populations from Kangaroo Island and Garden Island (described in McKenzie *et al.* 1993). Prior to my involvement in the project, a tedious process was followed to identify microsatellite markers. In this process, DNA was extracted from physically mapped BAC clones containing known genes. The BAC clone DNA was digested by restriction enzymes and 300-1000 bp fragments were size selected. These size selected fragments were hybridized against three types of dinucleotide repeats ($AC_{(15)}$, $AG_{(15)}$ and $AT_{(20)}$ fixed on nylon membrane. The hybridized DNA fragments were eluted and checked for the presence of microsatellite repeats by a PCR amplification step. Once the presence of microsatellite repeats was confirmed, the DNA fragments containing repeats were cloned into a plasmid vector and sequenced to obtain flanking sequences from which PCR primers were designed for polymorphism studies. *The important point to note here is that physically mapped BAC clones were characterized for the presence of the microsatellite repeats, which in turn could be used for linkage analysis. This step facilitates transferring of physical mapping information to linkage groups.* However, the process of characterizing physically mapped BAC clones for microsatellite repeats was extremely time consuming, costly and inefficient. Major limitations of this strategy were that the

physically mapped BAC clone might not have any microsatellite repeats, and low efficiency of cloning and hybridization would further reduce the chances of finding microsatellite repeats even if they were present.

2.1.6 Aims

The main aim of the following research was to make strategic use of the physical mapping data and tammar wallaby whole genome shotgun sequences to identify microsatellite repeats in the close vicinity of mapped genes. Physical mapping data were used to target gap regions in the linkage map and a novel bioinformatics based framework was designed to increase the efficiency of identifying microsatellite repeats for use in the linkage analysis. The framework presented in this chapter can be used to integrate the physical map and linkage map of any *de novo* sequenced mammalian genome to decrease the cost and time and increase the accuracy.

2.2 Methods

2.2.1 Identification of homologous syntenic blocks between human and opossum

Homologous syntenic blocks were identified between human and opossum using one-to-one ortholog information for these two species, which is available in the Ensembl database (Vilella *et al.* 2009). If two consecutive genes met the following criteria, they were added in a homologous syntenic block. Consider genes *A* and *B* as two *consecutive* genes on the human chromosome 1 separated by a 200 Kb (d^h) intergenic region. Their corresponding one-to-one opossum orthologs were genes *a* and *b* on the opossum chromosome 1, which were separated by a 500 Kb (d^o) intergenic region. The total amount of insertion/deletion (d') between genes *A* and *B* since the divergence of human and opossum was calculated as $|d^h - d^o|$ (= 300 Kb in this example). If any two consecutive genes on the chromosome have insertion/deletion (d') below the maximum allowed, they formed a homologous syntenic block. The maximum insertion/deletion limit was set to 1 Mb for this study therefore genes *A* and *B* formed a homologous syntenic block. Similarly if *C* and *c* were human and opossum one-to-one orthologous genes, and these genes were situated immediately downstream of genes *B* and *b* in respective species, they were compared next. If the insertion/deletion between genes *B* and *C* was less than the maximum insertion/deletion allowed, they will be added to the same homologous syntenic block as genes *A* and *B*, otherwise a new homologous syntenic block will be formed using genes *C* and *c*. Both human and opossum genomes were used as the reference genome in turn, and results of d' = 500 Kb, 1 Mb, 1.5 Mb, 2 Mb, 2.5 Mb and 3 Mb are shown for comparisons. The Perl script

used for identifying homologous syntenic blocks is provided in the supplementary information CD.

2.2.2 Pre-processing of the tammar wallaby trace sequences

The tammar wallaby whole genome shotgun sequences were obtained from the NCBI website <ftp://ftp.ncbi.nih.gov/pub/TraceDB/macropus_eugenii/> (National Centre for Biotechnology Information Trace Archives). A Perl script was developed to remove poor quality ends from all of the tammar wallaby whole genome shotgun (WGS) sequences such that the first 25 bases of both the 5' and 3' ends have quality values of more than 25 (quality value 20 = 99% accuracy and quality values are represented in on log scale). The Perl script used for processing tammar wallaby WGS sequences is provided in the supplementary information CD.

2.2.3 Annotation of tammar wallaby WGS sequences using opossum genes

Tammar wallaby WGS sequences were annotated for genes using opossum genes by reciprocal best hit criterion. cDNA sequences for all opossum genes were downloaded from the Ensembl v51 web server. Opossum cDNA sequences were used as queries to search the tammar wallaby WGS sequences using BLAST (parameters: -p megablast –v 100 –b 100 –e 1.0e-6 –W 12 –t 21 –N 0) (Zhang *et al.* 2000). All of the tammar wallaby hit sequences were reciprocally used as queries to search the cDNA sequence database of the opossum using BLAST (parameters: -p megablast –v 1 –b 1 –e 1.0e-6 –W 12 –t 21 –N 0) retaining only the best hits for each query sequence. The results were processed after reciprocal BLAST search to establish the reciprocal best hit relationship between opossum genes (represented by cDNA) and the tammar wallaby WGS sequences.

2.2.4 Identification of microsatellite repeats in target regions that fill gaps in the linkage map

Ms Chen Wei Wang provided approximate locations of gaps in the linkage map on the tammar wallaby ideogram (Alsop *et al.* 2005). Physically mapped genes closest to linkage map gaps were identified from the most current physical map of the tammar genome available at that stage (unpublished data, available through Dr. Janine Deakin).

The gap regions in the linkage map were targeted by the use of annotations of WGS sequences. First, the gene closest to the linkage map gap was used as an anchor gene and WGS sequences for the anchor gene were searched for dinucleotide microsatellites repeats containing at least ten repeat units. If no microsatellite repeat was found in WGS

sequences for the anchor gene, these sequences were extended to form larger contigs by recursive BLAST search. If the larger contigs contained no microsatellite repeat sequence, then the gene closest to the anchor gene in the *same* homologous syntenic block was used as a new anchor and the process of microsatellite search was repeated until at least one microsatellite was discovered for the gap region of the linkage map. The anchor genes were chosen to be in the same homologous syntenic block as physically mapped gene to avoid karyotype rearrangements. This ensures easy integration of the linkage map and the physical map.

WGS sequences were extended to form larger contigs by recursive MegaBLAST search to identify overlapping WGS sequences (parameters: -p 98 -v 10 -b 10 -e 1.0e-20 -F F) (Zhang *et al.* 2000). Recursive searches were performed until no more overlapping sequences were identified, up to maximum of five iterations. All overlapping sequences and their mate-pair sequences were assembled into larger contigs using the CAP3 program (Huang and Madan 1999). The larger contigs were reciprocally aligned to opossum cDNA sequences to confirm that the reciprocal best hit criterion was still fulfilled. This ensured the accuracy of the assembly process.

2.3 Results

2.3.1 Identification of homologous syntenic blocks between human and opossum

The probability of a breakpoint in the homologous syntenic block (HSB) is extremely low (Pevzner and Tesler 2003a). Therefore, these HSBs can be used to avoid major karyotype rearrangements, and fill the gaps in the linkage map. HSBs are usually identified by comparing the locations of orthologous genes (or any other orthologous DNA segment) between two species (Bourque and Pevzner 2002, Pan *et al.* 2005). In principle, by walking along the length of the chromosome in a reference species (human genome in this study), genes are added to the HSB if there is no disruption in the relative order of orthologs in the subject species (opossum genome in this study). Disruptions in the relative order of orthologs are indicators of karyotype rearrangements and hence the breakpoint in the synteny. Ensembl annotations for orthologous genes between human and opossum were used to identify HSBs between marsupial and eutherian lineage (Vilella *et al.* 2009). There were 13,279 one-to-one orthologs between human and opossum (Ensembl v55). The HSBs between human and opossum were identified by comparing the relative gene order and gene location in both genomes based on the maximum size of the insertion/deletion allowed as represented in Table 1.

Table 1 Number of homologous syntenic blocks between opossum and human genome with varying insertion/deletion allowed.

Maximum indel allowed in Mb	0.5	**1.0**	1.5	2.0	2.5	3.0
Number of HSBs	805	**628**	560	519	497	481
Human genome covered in Mb	1814	**2147**	2283	2384	2426	2469
% of Human genome covered	58	**69**	73	76	78	79
Opossum genome covered in Mb	2175	**2651**	2863	3033	3111	3179
% of Opossum genome covered	60	**74**	80	84	87	88
Human as reference genome						
Median HSB size in Mb	1.28	**1.64**	1.74	1.74	1.89	1.9
Largest HSB size in Mb	25.26	**32.5**	65.71	108.55	108.55	108.55
Opossum as reference genome						
Median HSB size in Mb	1.55	**1.96**	2.01	2.08	2.44	2.55
Largest HSB size in Mb	30.15	**43.06**	88.06	147.48	147.48	147.48

The number of homologous syntenic blocks decreased as the maximum insertion/deletion size allowed was increased. An increase in the coverage of the genome was correlated with an increase in the median size of the HSBs and an increase in the largest size of the HSBs. The method used in this study for HSB identification was not optimized to achieve mutually exclusive HSBs, that is, smaller HSBs may sometimes be embedded in larger HSBs in the gene dense regions where the distance between two genes is not large in either species. Further refinement of this method would be required to obtain mutually exclusive HSB. Nevertheless, the method is stringent and sufficient to obtain the HSB information required for targeted physical mapping or linkage mapping. The method could be used to compare any two genomes, and would be especially useful for species in which orthologous segments have already been identified and chromosomal locations were available.

2.3.2 Identification of tammar wallaby reciprocal best hit sequences for opossum genes

WGS sequences generated by Sanger technology often contain poor quality ends where the confidence of base calling is unreliable. It is essential to trim poor quality ends to avoid assembly errors and obtain reliable pairwise alignment scores during annotation step (Chou and Holmes 2001). The tammar wallaby trace archive repository contained 9,897,327 WGS sequences representing 9,056,336,520 bp. Of these, 9,391,912 sequences (~95%) were retained after quality trimming, and consisted of 5,273,763,918 bp (~58%). The quality cutoff threshold used was set to be very stringent to remove as many erroneous bases as possible. The stringent poor quality trimming does not affect further analysis since the aim to search for microsatellite repeats and not to assemble the whole genome. A database of trimmed tammar wallaby WGS sequences was locally set up to assign them to genes by reciprocal best hit searches.

Opossum cDNA sequences were downloaded from the Ensembl v55 web server. This cDNA set included all transcript sequences of known and predicted genes by the Ensembl pipeline (Hubbard *et al.* 2007). The dataset included 33,279 transcript sequences for 20,193 genes. Reciprocal best hit searches were performed by using these sequences as queries first. 323,332 tammar wallaby sequences were reciprocally related (by direct association, Figure 2, case D) to 18,252 opossum genes (~90% of all opossum genes annotated in Ensembl database).

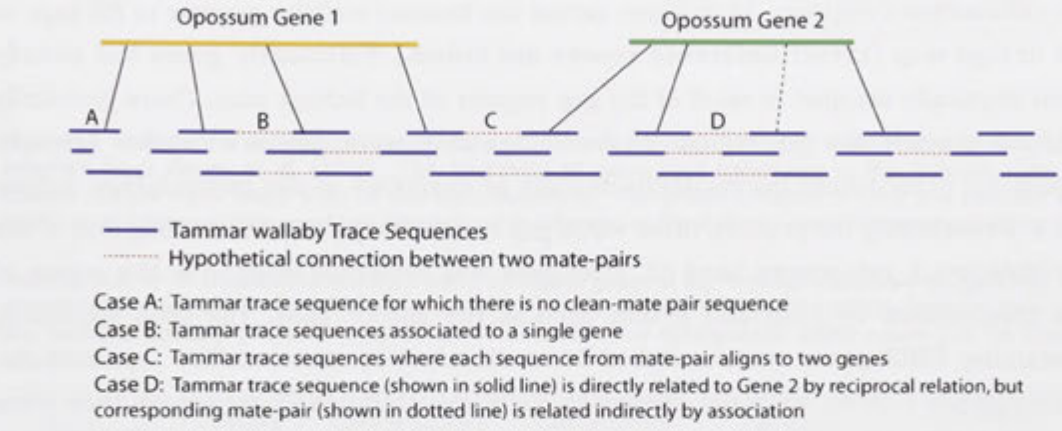


Figure 2 Tammar wallaby reciprocal best hits for opossum genes. cDNA sequences of opossum genes (orange and green bars) are used as queries to search tammar wallaby trace archive database (blue bars). Red dotted line shows the connection between two mate-pair sequences.

Once 323,332 tammar wallaby sequences were assigned to 18,252 opossum genes by the reciprocal best hit criterion, corresponding mate-pair sequences were also assigned to opossum genes (indirect association, Figure 2, Case D). Mate-pair sequences represent two ends of the same clone and therefore in principle, they are physically close together. The inclusion of mate-pair sequences resulted in a total of 576,212 WGS sequences associated with 18,252 opossum genes either by a direct reciprocal relationship or indirectly because of the mate-pair relationship. Theoretically, if 323,332 tammar wallaby sequences are associated with opossum genes, then 323,332 × 2 = 646,664 sequences should have been associated with opossum genes because two, and only two, sequences form a mate-pair.

The discrepancy between the actual number of WGS sequences related to opossum genes (576,212 sequences) and that of the theoretical number of WGS sequences related to opossum genes (646,664 sequences) was partly because not all the WGS sequences had good quality mate-pair sequence (Figure 2, Case A). The other reason for discrepancy was both mate-pair sequences may have been directly related to the same opossum gene, so they were counted in the directly related 323,332 WGS sequences (Figure 2, Case B). There were 6,285 WGS sequences that were associated with more than one opossum gene (Figure 2, Case C). Such cases were noted when two genes are very close together, or the

Ensembl pipeline has annotated a single gene as two distinct genes. WGS sequences were labeled as tammar wallaby orthologs for opossum genes if they were reciprocally related either directly or indirectly. WGS sequences containing genes were used to search for microsatellite repeats for the linkage map.

2.3.3 Identification of microsatellite repeats in target regions that fill gaps in the linkage map

My collaborators required 31 markers across the tammar wallaby genome to fill gaps in the linkage map (**Error! Reference source not found.**). Fortunately, genes had already been physically mapped in most of the gap regions of the linkage map. These physically mapped genes in the gap regions of the linkage map were chosen as anchor genes to search for dinucleotide microsatellite repeats as described in the methodology section 2.2.4. To exemplify the process, there was a gap in the linkage map on the long arm of the chromosome 7, cytogenetic band q1. *TSHR* gene was physically mapped in this region of the chromosome by FISH and it was used as the anchor gene. The WGS sequences containing *TSHR* genes were searched for a dinucleotide microsatellite repeat. If the microsatellite repeats were not present in *TSHR* containing WGS sequences, they were extended to form larger contigs as described in section 2.2.4 and searched for dinucleotide microsatellite repeat. If the dinucleotide microsatellite repeats were not present in the larger contigs of *TSHR* gene, then the gene closer to *TSHR* gene in the same HSB as *TSHR* gene was chosen as new anchor gene (*GTF2A1* in this example). The process was repeated until a dinucleotide microsatellite repeat was found. The use of anchor genes in the same homologous syntenic block that contained *TSHR* gene ensures reliable integration of the linkage map with the physical map since both the linkage map marker and physically mapped gene belong to the same HSB.

I identified 26 microsatellite repeats for linkage analysis (Table 2, Table 3). All markers were polymorphic in the hybrid phase-known backcrosses derived from genetically distinct Kangaroo Island and Garden Island tammar wallabies (McKenzie *et al.* 1993). Linkage analyses were conducted using 353 informative meiosis. All 26 markers were subsequently incorporated in the linkage map of tammar wallaby.

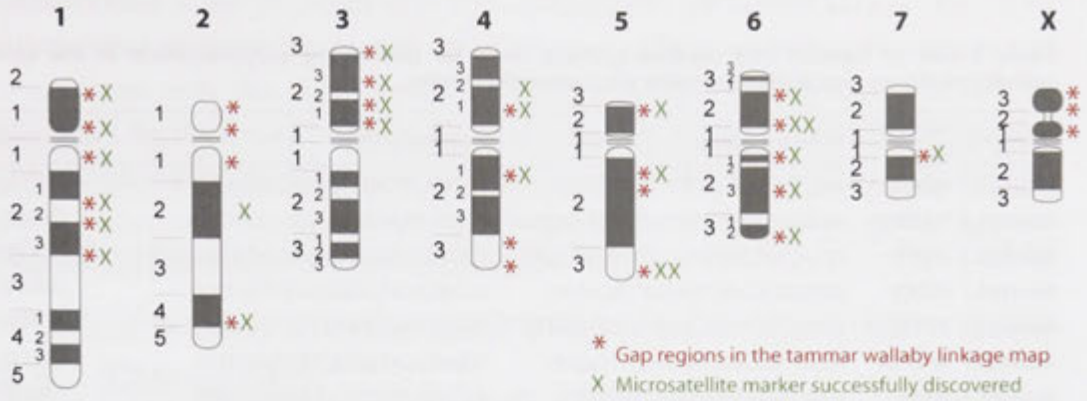


Figure 3 Ideogram showing chromosomes of female tammar wallaby when stained with DAPI adapted from Alsop *et al.* (2005). The locations of gaps (red astericks) in the linkage map are shown on the right hand side of the chromosomes. The green crosses shows the number and location of microsatellite markers discovered to fill gaps in the linkage map of tammar wallaby.

Table 2 List of microsatellite markers used to fill the gaps in the tammar wallaby linkage map. The chromosomal location of markers is obtained by integrating the tammar wallaby linkage map with the tammar wallaby physical map of chromosomes. ? = cytogenetic band could not be clearly identified.

Marker ID	Anchor gene	Tammar wallaby chromosome	Chromosome band	Opossum chromosome	Opossum chromosome location in Mb
Meu1p1.1	*AQP8*	1	p1	6	144.6
Meu1p1.2	*RAMP3*	1	p1	6	273.9
Meu1q1.1	*ASB7*	1	q1	1	139.1
Meu1q2.1	*ODZ2*	1	q2	1	366.6
Meu1q2.2	*TCERG1*	1	q2	1	1.7
Meu1q3.1	*MAT1A*	1	q3	5	103.8
Meu2q2.1	*USF1*	2	q2	2	456.4
Meu2q4.1	*TBX4*	2	q4	2	171.1
Meu3p2.1	*EEF2K*	3	p2	6	61.6
Meu3p2.2	*GABBR2*	3	p2	6	86.1
Meu3p2.3	*LFNG*	3	p2	6	77.9
Meu3p3.1	*DNAH12*	3	p3	6	208.4
Meu4p2.1	*CDH12*	4	p2	6	184.1
Meu4p2.2	*KDSR*	4	p2	3	142.1
Meu4q1.1	*KIAA1012*	4	q1	3	200.9
Meu5p2.1	*PTCHD1*	5	p2	4	43.6
Meu5q2.1	*CLDN18*	5	q2	4	100.2
Meu5q3.1	*BCL3*	5	q3	4	371.0
Meu5q3.2	*HPX*	5	q3	4	414.5
Meu6p2.1	*NOP14*	6	p2	2	501.8
Meu6p2.2	*TNIP2*	6	p2	5	231.4
Meu6p2.3	*ZNF143*	6	p2	5	249.3
Meu6q2.1	*COL4A2*	6	q2	7	82.2
Meu6q2.2	*SLAIN1*	6	q2	7	137.6
Meu6q3.1	*SNED*	6	q3	7	251.3
Meu7q2.1	*TSHR*	7	q?	1	467.7

Table 3 List of forward and reverse primers used for identifying polymorphism in the tammar wallaby pedigree population for each microsatellite marker.

Marker ID	Anchor gene	Forward primer	Reverse primer	Microsatellite
Meu1p1.1	*AQP8*	AACTTTGGTGTCTTGGTGGAA	TTTCAGTCACTGGGCTGAAGT	CA(35)
Meu1p1.2	*RAMP3*	ACACATAGTCACTCTCCTTTACCG	CAGAGAAGGGAGCCTGTTTAG	CA(22)
Meu1q1.1	*ASB7*	GGTCAGAGGACAACTAGGTTGAAG	CATACAGAGGCAAAAGCATAACTG	TG(30)GA(11)
Meu1q2.1	*ODZ2*	AGCCCATAGTCAGGCACATAC	GCACATAGAGGGAGTTGTCCA	CA(34)
Meu1q2.2	*TCERG1*	GACATATTAGCTGCTCTTCAGTGTTC	GAGCTTGCTATGTCTGAAGGCTAC	AC(19)
Meu1q3.1	*MAT1A*	ACATGGGGGTAAAACTTGGAC	TGAACCATGTCCTCTGACTCC	CA(17)
Meu2q2.1	*USF1*	ATAGGGAATGCAGCAGGTTG	ATCAGCTGTTCTAAGGCCACA	CA(28)
Meu2q4.1	*TBX4*	TCACTCTATATCGGTCAGAGGACA	GGTCTGGGACAGTAAATTCTTCAC	CA(29)
Meu3p2.1	*EEF2K*	AGGGCATCCCAAGATTCTTACT	GCAGTGAAAATGACTAGGAGGAG	TA(17)
Meu3p2.2	*GABBR2*	CTCCCAAGCTAGGAAACAACC	CAAGACCGTATCAGAGGCAAA	GT(27)
Meu3p2.3	*LFNG*	TGCACTCCATGAAGACACTTG	TCACTGGATTCAGATGGCTCT	GT(15)
Meu3p3.1	*DNAH12*	CTGTCAAGTCTGAAGTGGACAGA	GAGTTAATACTGGCGTCTTGGAG	TG(16)
Meu4p2.1	*CDH12*	TGCTACTACCCCATCTCTCTCTC	CTTTCCAAAAGAACCAGAGCA	CT(25)
Meu4p2.2	*KDSR*	TCTGTGTTCATTATCCGTGACA	CATTGTGAGAAAGAGCCATCTG	AC(26)
Meu4q1.1	*KIAA1012*	CTCTTTCATTCCTAGACACACTGG	GCAAGAAGAATGATGGACACAC	AC(24)
Meu5p2.1	*PTCHD1*	TTTTTCTTCTCCCCCGTACC	TGGCCTTGAAGCATACTTATTG	GT(26)
Meu5q2.1	*CLDN18*	GCAGAGCTGGCATTAGATGA	TTTGTTCAATGACCCCCAAT	AC(13)
Meu5q3.1	*BCL3*	AATGAGGGACAAGCAAGCTC	AGTTGACCTCAGGGCAGTGT	CA(18)
Meu5q3.2	*HPX*	GATCTCAGAAACATGGCCAGA	CTGTACCCTCAAACCTTGTGC	GT(22)
Meu6p2.1	*NOP14*	CCACCCCTCAGTGTTTCAGTAT	GGTTAATGGGGCTTAGGATAGG	AT(21)
Meu6p2.2	*TNIP2*	CATGTCACCTGGAACTTTTTCA	GTGTTGTATAGCTCAGTTTCAGATAGC	GT(13)
Meu6p2.3	*ZNF143*	GTTTATCACACCCAGGGACTGT	GGTTAAGGTGCCAAAAGAGGTA	AC(19)
Meu6q2.1	*COL4A2*	GAGAGGTCAGGGAAGGGTATCT	TAAACCAGGTACTCCTGGGAAA	TA(17)
Meu6q2.2	*SLAIN1*	CAGAGATTTTTGCCAGCAGAC	CCCAACCTTTCAAGTAGAATGC	TA(34)GA(13)
Meu6q3.1	*SNED*	TCCTCCAAATCCTCTTCCAGT	CACTGCAAGCACCACTGTCT	AC(21)
Meu7q2.1	*TSHR*	TCTATGAGCCAAGAACTCCAGA	GATGTTAGCAACAGAGATCATGGTA	AT(16)

There were only three gap regions in the tammar wallaby linkage map in which no polymorphic microsatellites were discovered. The short arm of the tammar wallaby chromosome X (Xp) is largely heterochromatic and contains the nucleolus organizer region (NOR), the site of repetitive ribosomal RNA sequences (Alsop *et al.* 2005). There are only two genes physically mapped in this Xp region, and no orthologous sequences could be found for these in the WGS sequences by the reciprocal best hit criterion. Therefore no microsatellite markers were discovered in this region. Likewise, the short arm of the tammar wallaby chromosome 2 is very small and only two genes have been physically mapped to this region. Again, no reciprocally related sequences were found for physically mapped genes on the chromosome 2p, so no microsatellite markers could be identified for this gap in the linkage map. In addition, there were no physically mapped genes in the tammar wallaby chromosome 2q1 and chromosome 4q3, so these were not included for a microsatellite search.

My collaborators at the University of Sydney subsequently performed linkage analysis for 26 polymorphic microsatellite markers. The locations of these markers in the linkage map were consistent with the physical mapping data for tammar wallaby chromosomes. My collaborators in University of Sydney carried out polymorphism studies and linkage analysis for generating the tammar wallaby linkage map. I am a co-author on papers being prepared describing the tammar wallaby linkage map and integrated map (linkage map overlap with physical map) by Chen Wei Wang.

2.4 Discussion

Robust physical and linkage maps are essential in comparative genomics to understand the evolution of genomes (Lewin *et al.* 2009). The number of markers included in these maps must be optimized, since both the physical and linkage mapping are laborious and expensive processes. My results indicate that the low coverage genome sequences can efficiently be used to target regions of interest for obtaining good quality physical map or linkage map.

The genomic sequencing era has greatly advanced the field of comparative genomics by providing ultimate details at the molecular level. Between the year 2006 and 2008, 24 mammalian species were selected for low coverage sequencing to aid the functional annotation of the human genome (National Human Genome Research Institute). It was noted early in the sequencing era that mammalian genomes are conserved not only for protein coding regions, but in the regions outside of genes as well. One of the example studies showed that more than half of the conserved elements between the human chromosome 21 and mouse/dog genome lie in regions outside of known exons on the chromosome (Frazer *et al.* 2001). The conserved non-coding elements often have regulatory role in gene expression (elegantly discussed in Hardison 2000). It was this aspect of the mammalian genome conservation that led to enormous amount of sequences available from variety of organisms for comparative studies. More recently, the drop in the cost of genome sequences by several orders of magnitude (discussed in (Mardis 2008) has led to a new initiative for obtaining sequences from more than 10,000 vertebrate species (Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species 2009). It is evident that large amount of sequences will be available in the future for comparative analysis.

However, it should be noted that low coverage genome sequences are difficult to assign to chromosomes (Green 2007). The tammar wallaby genome provides a good example because it has been sequenced to a low coverage of 2×. Assignment of 1,174,382 contigs to chromosomes for obtaining a picture of the genome organization is virtually impossible. Therefore it is necessary to combine the first pass assembly of tammar wallaby sequences with the physical map and linkage map to order contigs and assign them to the

chromosomes. This study and other mapping efforts for the tammar wallaby genome (Deakin *et al.* 2008), have shown that targeting physical or linkage maps for the breaks in synteny can efficiently generate a more comprehensive overview of the genome which can be directly compared for major karyotype rearrangements and also obtaining decent assembly of the genome. The systematic use of low coverage WGS sequences in this study for the identification of microsatellite markers in the gap regions of the tammar wallaby linkage map has allowed for the generation of a good quality contiguous linkage map for the tammar wallaby chromosomes. The linkage map has been integrated with the physical map to reveal the completeness of the linkage map. This contiguous map of the tammar wallaby chromosomes will now be used with the homologous syntenic block information derived by opossum/human comparisons to assign chromosomes to large number of contigs. This will aid in creating molecular map of the 19 conserved blocks of synteny between marsupials (Rens *et al.* 2003).

This example also demonstrates how low coverage genome sequence can be used systematically to obtain maximum data for comparative analysis. Firstly, as and when trace sequences are publicly available, they should be processed to trim poor quality regions and vectors so that more specific pairwise sequence similarity searches can be performed. Trimming of poor quality regions increases the sensitivity of the search methods employed (Malde 2008). Poor quality sequences at the beginning (5' end) and end (3' end) region cause problems in detecting overlaps for contig formation (Pop *et al.* 2002). Although stringent trimming of low quality sequences removes the larger proportion of sequence data as ambiguous ends, it improves confidence in downstream analysis. I designed a simple and robust Perl script to trim poor quality regions when the quality values for sequences are available. This pre-processing step allowed identification of overlaps between two or more sequences by pairwise alignments using BLAST searches. This is useful since the assembly of a large number of sequences generated by whole genome shotgun approach is not feasible on a desktop computer, but a BLAST search can be performed in real time on a desktop computer. A BLAST search for finding overlapping sequences reduces the search space required for downstream assembly programs and makes it amenable to carrying out such analysis for regions of interest on a desktop computer.

A large number of genomic sequences in publicly available databases are assembled into contigs that are not yet assigned to chromosomes. This is especially true for target species whose genomes have been sequenced at low coverage. Indeed, even more comprehensively sequenced genomes, such as the platypus genome that was sequenced at 6× coverage (Warren *et al.* 2008), suffer from a dearth of mapping information that could be used to assign the contigs to chromosomes. As a result more than 19,000 genes (total ~21,000 genes) are still not assigned to chromosomes. Targeting the regions of evolutionary breakpoints allows a reduction in the number of markers required for

physical mapping or linkage analysis and yet covers the greater proportion of the genome. For example, there are 628 homologous syntenic blocks between a eutherian genome (human genome) and a marsupial genome (opossum genome) covering approximately 70% of both genomes. If we were to obtain the physical map or the linkage map of marsupials, only 628×2 = 1,256 markers would be required (one from each end of the homologous syntenic block) to cover the genome comprehensively and obtain the relative gene order in marsupials. This accelerates the process of obtaining the genome architecture of an organism and also reduces the cost of overall project. The cat genome was initially sequenced at a lower coverage and then assembled to form larger contigs that were assigned to the cat chromosomes with the help of radiation hybrid map and homologous syntenic block between five vertebrate species (Pontius *et al.* 2007). I have made use of the homologous syntenic block information between human and opossum genomes to target regions for linkage mapping. This strategy can be applied to any mammalian genome so that rather than using homologous syntenic block information for just ordering contigs, it can be strategically used to create a physical map or linkage map of the breakpoint regions to achieve a more thorough picture of genome organization.

The tammar wallaby genome project has recently reached three of its major milestones; the physical map of the genome (Deakin *et al.* 2008) and personal communication), the linkage map of the genome (personal communication with Chen Wei Wang, University of Sydney) and the first draft assembly and annotation of the low coverage sequence (Hubbard *et al.* 2007). Future work requires organization of the tammar wallaby assembly into larger scaffolds using homologous syntenic blocks and then assigning those scaffolds to tammar wallaby chromosomes based on physical mapping data and linkage mapping data. Comparisons of the tammar wallaby genome with genomes of other marsupial and other mammals will reveal insights in to the evolution of mammalian genome in general and marsupial genomes for specific studies.

I have used these comparative methods in two studies involving marsupial sequence; firstly to resolve a mystery concerning the evolutionary origin of the human X, and secondly to characterize the marsupial olfactory gene family and investigate its evolution.

3 Evolution of the human X chromosome

Evolution of the sex chromosomes follows a different trajectory from that of the autosomes (Charlesworth *et al.* 1987). Some major driving forces that shape the evolution of differentiated sex chromosomes are male vs. female fitness, heterozygosity of one of the sex chromosomes, and the restriction of recombination on the sex-specific chromosome. Comparisons of the gene content of sex chromosomes of mammals and other vertebrates showed that the sex chromosomes of therian mammals are relatively young, having evolved in the last 166 million years since the divergence of therian mammals from the common ancestor with monotreme mammals.

Comparisons of the sex chromosomes of marsupials and eutherian mammals also show that the eutherian X and Y chromosomes have gained an autosomal segment even more recently, since their divergence from the common ancestor with marsupials in the last 148 million years. The conserved region of the X (XCR) and the added region (XAR) represent two evolutionary blocks, which are also separate in monotreme mammals and birds.

The availability of genomic sequence data from multiple species permitted a detailed examination of the origin of genes on the human X chromosome. These studies claimed that two regions of the human X (thus the X of all eutherian mammals) has an additional evolutionary layer that was added to both sex chromosomes before their split from marsupials (Kohn *et al.* 2004, Ross *et al.* 2005). Reciprocal best hit or best hit analysis provided the basis for this argument for three evolutionary layers on the eutherian sex chromosomes.

In this chapter I present my research on the evolution of the human X chromosome. I physically localized the human X chromosome genes on tammar wallaby metaphase chromosomes to show the conservation pattern. Moreover, I traced the evolutionary history of the genes from the cytogenetic band Xp11 and Xq28 by neighbour-joining phylogenetic analysis. The results refute the hypothesis that the human X is composed of three evolutionary layers and provide a much simpler model of the eutherian X chromosome evolution. My research work also emphasizes the need for good physical map data and choice of appropriate methodology or multiple lines of evidence to understand the evolution of chromosomes, chromosome regions or genes.

I have co-first-authored a research article based on this work with my supervisor, Dr. Margaret L. Delbridge, which was recently published in the Genome Research (Delbridge *et al.* 2009).

3.1 Introduction

Many forms of sex determination are found in the animal kingdom. The trigger that determines the difference between males and females of the species range from environmental cues (usually temperature) to genetic sex determination in animals without obvious sex chromosomes, to highly differentiated sex chromosomes. Those with genetic sex determination show either male heterogamety or female heterogamety; for species with differentiated sex chromosomes this results in either XY male: XX female (like humans), or ZZ male ZW female systems (like birds). Within amniotes (reptiles, birds and mammals), sex chromosomes have independently evolved from a pair of autosomes at least three times (Vallender and Lahn 2004, Veyrunes *et al.* 2008). In other words, the sex chromosomes in therian mammals are not homologous in their genetic composition to the sex chromosomes of monotreme mammals, birds, snakes or other reptiles. Likewise, the sex chromosomes of snakes are not homologous in their genetic composition to birds or other reptiles. It is important to understand the phylogenetic relationship of animals to gain better insights in to the evolutionary changes that have occurred millions of years ago.

3.1.1 Animal phylogeny

Studies of the evolution of chromosomes, genomes or genes are greatly aided by comparing similarities and differences between different species that are closely or distantly related. Fossil records, anatomical structures and molecular signatures have helped deduce the phylogenetic relationships between different species and enabled the construction of a phylogenetic tree of life that is generally accepted. A comprehensive analysis of protein sequence data and genome sequence data can now be used to address in detail the phylogenetic relationships between living species. In this chapter I concentrate mainly on vertebrate species, but some references are also made to worms, insects and flies.

The last common ancestor of bilaterians (animals with bilateral symmetry with a front end and a back end, and upside and downside) arose about 615 million years ago (MYA) (Figure 4) (Peterson *et al.* 2004). The last common ancestor of protostomes (including arthropods; *e.g.* flies, wasps, ants, and bees and nematodes like *C. elegans*) evolved about 550 MYA and vertebrates diverged from the last common ancestor with protostomes about 520 MYA (Adoutte *et al.* 2000, Peterson *et al.* 2004).

Phylum Vertebrata consists of seven classes. There are three classes of fish; the jawless fish, the cartilaginous fish and the bony fish. The fish classes of the vertebrate phylum shared common ancestor with terapods (four-limbed animals) 430 MYA (Blair Hedges and Kumar 2004). Amniotes (the reptile, aves and mammal classes) whose egg contains an

amniotic sac, chorion membrane and allantois, which helped them adapt to the drier environment of the land, radiated from the common ancestor with amphibians (Class Amphibia) around 370 MYA. Birds and turtles diverged from the common ancestor with lizards and snakes about 285 MYA: snakes and lizards shared the common ancestor approximately 220 MYA and birds and turtles last shared the common ancestor approximately 272 MYA (Rest *et al.* 2003). Class Mammalia last shared a common ancestor with birds 310 MYA (Blair Hedges and Kumar 2004). The egg-laying monotremes (mammalian Subclass Prototheria) diverged from the common ancestor with therians (mammalian Subclass Theria) about 166 MYA and marsupials (infraclass Metatheria) diverged from the common ancestor with placental mammals (infraclass Eutheria) about 148 MYA (Bininda-Emonds *et al.* 2007). I will use the above-mentioned phylogenetic relationships of animals to discuss the evolution of the human X chromosome in following sections of this chapter.

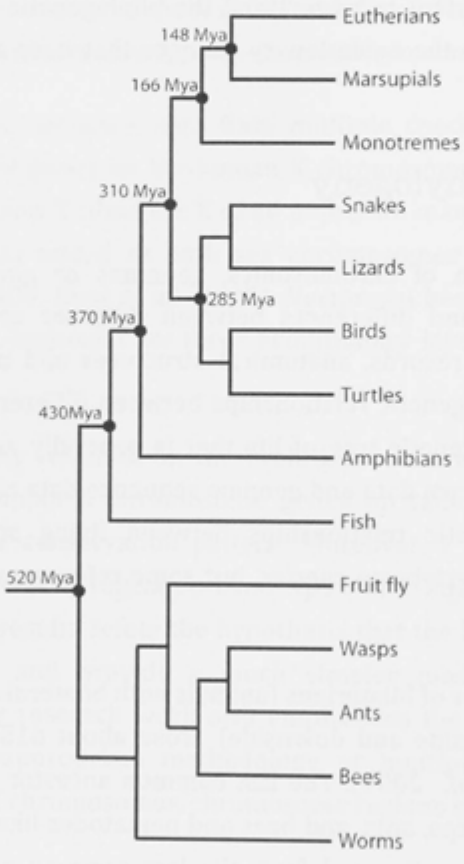


Figure 4 Phylogenetic relationship of vertebrates with invertebrates and divergence times of major vertebrate classes and subclasses. Mya = Million years ago

3.1.2 Vertebrate sex determination systems

Mostly all vertebrates reproduce sexually after the fusion of a female egg and male sperm. However, the mechanism that dictates the development of a vertebrate organism in to male or female is different in different groups. For example, environmental dependent sex determination (ESD) pathways are found in some fish and many reptile species (Devlin and Nagahama 2002, Ospina-Álvarez and Piferrer 2008, Sarre *et al.* 2004). The incubation temperature of the eggs plays a crucial role in most ESD systems, so they are referred to as temperature-dependent sex determination (TSD). However, other environmental factors such as the social surroundings (Francis and Barlow 1993), the dietary conditions of mothers (Warner *et al.* 2007), and pH of the rearing environment (Baroiller *et al.* 1999) have also been shown to influence sex determination pathways in some fish and many lizards.

In contrast to lizards and fish, mammals, birds, snakes and amphibians use a genetic sex determination (GSD) pathway, whereby a gene or genes that could be localized on the sex chromosomes trigger gonad differentiation. Many genes can influence sex determination pathway, but they are usually under the control of a single gene that triggers the male or female development. For example in therian mammals, the male-dominant *SRY* gene present on the Y chromosome is the male determining factor (Sinclair *et al.* 1990), however, *SOX9, FGF9, WNT4* and *RSPO1* genes among others also influence the sex determination pathway (reviewed in DiNapoli and Capel 2008). Similarly, the double dose of the *DMRT1* gene on the Z chromosome is the dominant male determining factor in birds (Smith *et al.* 2009) but the entire pathway of sex determination is also sensitive to expression of *SOX9, AMH, HINTW, FET1,* and *CYP19A1* genes (reviewed in Smith *et al.* 2007). The sex determining gene has been identified in only two other species; a transposed copy of *DMRT1* defines a novel Y chromosome in medaka species, and a W chromosome in the toad *Xenopus* (reviewed in Graves 2008). In some reptiles there appears to be a polygenic mode of sex determination.

How do these sex determination master switches end up being on the sex chromosomes? Or is it the presence of these sex determination master switches like *SRY* and *DMRT1* genes that defines the fate of a chromosome to become sex chromosome?

3.1.3 Evolution of sex chromosomes in vertebrates

The evolution of novel sex chromosomes starts with the acquisition of a sex determining locus/gene on one of the homologous chromosomes. Acquisition of sex determining locus may occur via a mutation in a gene already involved in the pathway, or in another pathway that changes its tissue specificity. The transition from TSD to GSD has been speculated to occur when polygenic influences on sex determination are usurped by a strong sex

determining factor that evolves in the population to dominate the process; however, the observation of a lizard that has GSD over a range of temperatures but TSD at the extreme (Quinn *et al.* 2007) suggests that there may be an underlying GSD mechanism.

Nevertheless, once sex determining locus is acquired by one member of an autosomal pair (proto Y or W), other genes that confer an advantage to that sex accumulate near this locus. For instance, many spermatogenesis genes were accumulated with *SRY* on the mammal Y. The presence of several sex-specific genes sets up selection for suppression of homologous recombination; for instance XY recombination is suppressed in male mammals and ZW recombination in female birds. Then more sex-specific mutations are accumulated around the site of the sex determination locus (Figure 5) (Rice 1987) and the non-recombining region is extended.

The absence of recombination in a region leads to an accumulation of recessive lethal or deleterious mutations on the Y (Charlesworth 1978, Muller 1918). As these mutations never become homozygous, such recurrent mutations from wild type to recessive alleles are unconstrained by selection. Thus, on an evolutionary time scale, one of the sex chromosomes becomes fixed for rare recessive, loss-of-function alleles at most loci, and deletion of the region can then occur. Lack of recombination and accumulation of recessive alleles (mutations) occurs due to stochastic loss of the least mutated Y chromosome ("Muller's ratchet"; Felsenstein 1974). In addition, inefficient selection occurs because the whole chromosome is the unit of selection, due to the absence of recombination; thus advantageous new alleles on otherwise degenerate Y chromosomes sweep through the population by a "hitchhiker effect" (Rice 1987), or conversely an advantageous new allele is eliminated from the population because of poor genetic background. This degradation of non-recombining loci leads to a degraded male specific Y chromosome or female specific W sex chromosomes. The Y chromosome has the added disadvantage that it is confined to the male lineage and therefore exposed to the high mutation and low DNA repair environment of the testis at each generation and therefore variation in Y chromosome is higher.

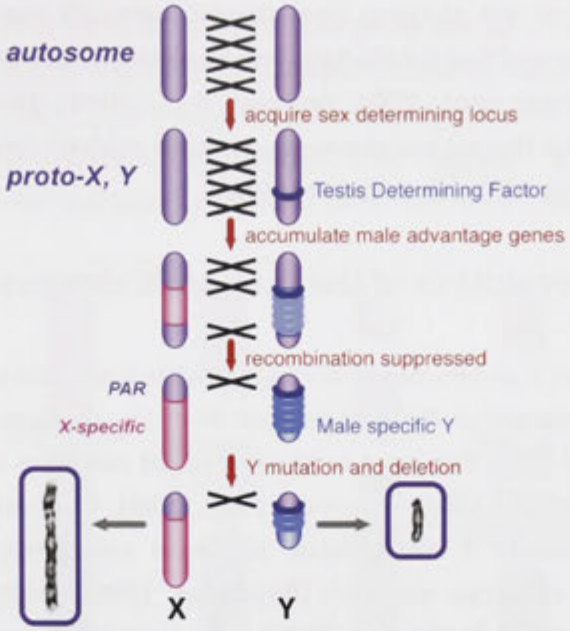


Figure 5 The sex chromosome evolution in the XX/XY system (adapted from Graves 2006). Once the sex determining locus is gained on the one of the autosomal pairing partners (purple), homologous recombination is suppressed around the sex determining locus followed by accumulation of male beneficial mutations and degradation around the sex determining locus. This results in the degradation of the male-specific Y chromosome.

Thus genetic factors that control the sex determination process can ultimately lead to either a male-specific Y chromosome or a female-specific W chromosome. In species with genetic sex determination, sex chromosomes are either heteromorphic (cytogenetically distinct from each other) as in mammals and birds, or homomorphic (cytogenetically similar to each other) as in fish and amphibians. Birds and snakes and many other reptiles have female heterogamety (ZZ males and ZW female, ZZ/ZW system) whereas therian mammals have male heterogamety (XX females and XY males, XX/XY system). Monotreme mammals form a unique mammalian group because they possess multiple sex chromosomes; male platypus have five X and five Y chromosomes and female platypus have five pairs of X chromosomes (Bick *et al.* 1973, Rens *et al.* 2004). Amphibians and fish show enormous plasticity in sex chromosome composition, with both the male heterogamety with XX/XY system, and the female heterogamety with ZZ/ZW system in related species and even within the same species (Ananias *et al.* 2007, Eggert 2004, Miura 2007, Volff *et al.* 2007).

Susumo Ohno proposed that if XX/XY in mammals and ZZ/ZW in birds evolved from a same homologous pair of autosomes, then gene content of the mammalian X chromosome and the bird's Z chromosome would be same (Ohno 1967). Ohno's hypothesis was tested by comparative gene mapping and genome comparisons, which showed that sex chromosomes of birds, snakes and therian mammals are not homologous (Matsuda *et al.*

2005, Nanda *et al.* 1999, Ross *et al.* 2005). Interestingly, the comparative mapping and sequence analysis of the platypus (monotreme mammal) sex chromosomes with the chicken genome showed that the chicken Z chromosome shares genes with the platypus X_5 chromosome (Grutzner *et al.* 2004, Veyrunes *et al.* 2008). These comparative analysis results highlight that therian sex chromosomes have evolved independently compared to monotreme mammal and bird sex chromosomes.

3.1.4 The evolution of the human X chromosome

The eutherian X and Y chromosomes are suggested to have arisen from an autosomal pair following the divergence of the monotremes from the common ancestor with therians approximately 166 MYA, but prior to the marsupial radiation approximately 148 MYA (Bininda-Emonds *et al.* 2007, Veyrunes *et al.* 2008). Comparative gene mapping has shown that the human X chromosome is almost completely homologous to the X chromosome of all eutherian mammals (Boyd *et al.* 1998, Murphy *et al.* 2005, Raudsepp *et al.* 2002, Rodriguez Delgado *et al.* 2009). This remarkable conservation, much greater than for any other mammal chromosome, was first pointed out by Susumo Ohno, and suggested to be due to constraints posed by the evolution of a chromosome-wide inactivation system that accomplishes dosage compensation of the X (Ohno 1967).

Comparative mapping between eutherian and marsupial mammals in the 1970s by somatic cell hybrid analysis revealed that X chromosome conservation was not complete in all mammals; marsupials broke "Ohno's Law" (Graves *et al.* 1979). By mapping the kangaroo orthologs of human X genes, it was discovered that the human X chromosome was made up of two distinct regions. The long arm of the human X chromosome and the pericentric region is conserved on the X chromosome in all therian mammals including marsupials and eutherians (X conserved region, or XCR; Figure 6) (Spencer *et al.* 1991b), and so must have been ancestral in therians.

However, the rest of the short arm of the human X is autosomal in marsupials, and it was proposed to have been added to the X chromosome of the eutherian mammals after they diverged from a common ancestor with marsupials (X added region, or XAR) (Spencer *et al.* 1991a). The XAR is homologous to the short arm of the tammar wallaby chromosome 5 (5p) (Spencer *et al.* 1991a) and to regions in the short arm of opossum chromosome 7 and 4 (Mikkelsen *et al.* 2007). The XCR and the XAR are therefore two blocks of the human X chromosome with a distinct evolutionary history. This is also consistent with the finding that XCR genes are conserved on the short arm of the chicken chromosome 4, whereas XAR genes are located, in virtually the same order, on chicken chromosome 1q (Kohn *et al.* 2004, Ross *et al.* 2005).

In the platypus, a monotreme mammal, human XAR genes are located on the platypus chromosomes 15 and 18 (Figure 6). Human XCR genes are all co-located on a single chromosome, but, surprisingly it is not an X chromosome, but chromosome 6 (Veyrunes *et al.* 2008, Waters *et al.* 2005). The X chromosomes in platypus have homology, not with the therian X, but to the chicken Z chromosome. This suggested that monotreme sex chromosomes may have evolved directly from a bird-like ancestral system (Graves 2008).

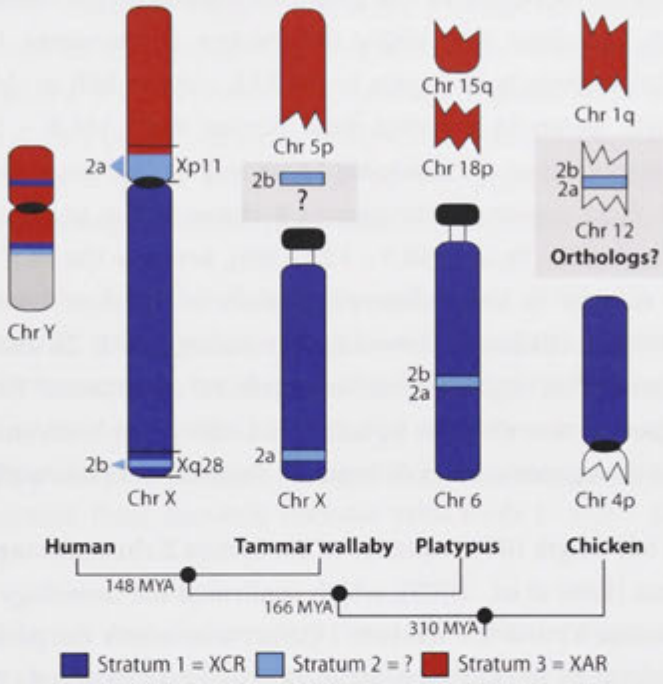


Figure 6 The evolution of the human X chromosome. Comparative analysis of the human X chromosome with marsupials, platypus and chicken shows two distinct evolutionary blocks; the X-added region (XAR, red) and the X-conserved region (XCR, navy blue). Since both the XAR and the XCR are autosomal in platypus and chicken, it suggests that the sex chromsomes have evolved independently in therian mammals compared to platypus and chicken lineages. One recent report (Kohn *et al.* 2004) has also suggested the presence of the third evolutionary block (light blue) on the human X chromosome. The origin of that third evolutionary block is under question in this study. Comparative analysis data between human and tammar wallaby were adapted from (Delbridge *et al.* 2009, Graves 1995), human and platypus from (Veyrunes *et al.* 2008) and human and chicken from (Kohn *et al.* 2004).

The availability of complete DNA sequence of the human X chromosome made it possible to perform fine scale comparisons with genomes of the chicken and fish. The comparisons were made, using a set of approximately 300 marker genes, between the human X chromosome and chicken, pufferfish and zebrafish genomes (Kohn *et al.* 2004). Most genes within the XCR mapped to chicken chromosome 4p, and most XAR genes to the chicken chromosome 1. More specifically, genes from human Xpter-p11.2 (0 – 46 Mb), including the pseudoautosomal region 1 (PAR1), were conserved on the chicken chromosome 1 (from 1q13-1q31, 104 – 122 Mb) in virtually the same order as found on the human X chromosome. This region is widely known as the X added region (XAR) (Graves 1995), and it was described as evolutionary strata 3 in that particular study

(Kohn *et al.* 2004). Nearly the whole of the region from proximal Xp11.2 to the centromere, and the entire long arm of the human X chromosome was completely conserved in a single segment on the chicken chromosome 4p. The gene order for this region was not conserved between human and chicken. This region is widely known as the X conserved region (XCR) (Graves 1995), and it was described as evolutionary strata 1 in that particular study (Figure 6) (Kohn *et al.* 2004).

A major puzzle was the discovery of two gene rich regions on the human X chromosome that did not share homology with either the chicken chromosome 1q or the chicken chromosome 4p; the regions homologous to the XAR and the XCR in chicken respectively (Kohn *et al.* 2004). Genes in a region from human Xp11 (46.0 – 54.1 Mb) showed homology with mainly chicken chromosome 12, as well as chicken chromosomes 1 and 4. They defined this Xp11 region on the human X chromosome as stratum 2a. Genes on another region from human Xq28 (150.9 – 152.3 Mb), between the *BGN* and *CTAG2* genes, also seemed not to map to the conserved clusters on chicken 1 and 4, but showed homology with multiple chicken chromosomes including 12, 1, 26 and other micro and macro chromosomes. This Xq28 region was defined as stratum 2b (Figure 6). The evolutionary stratum 2 was thought be added on the proto-X chromosome of therian mammals after the divergence mammals from the common ancestor with birds.

Subsequently, the full-length DNA sequence of the human X chromosome was compared to the chicken genome (Ross *et al.* 2005), which confirmed the homology of stratum 1 (the long arm of the human X) and the stratum 3 (short arm minus the pericentric region) of the human X chromosome to chicken genome as previously reported (Kohn *et al.* 2004). However, this comparison using full-length DNA sequence could not establish the origin of stratum 2a, which was shown to share homology with the chicken chromosomes 1, 12 and 4. The boundary for XAR and XCR however was identified between *RGN* and *NDUFB11* genes (46.83 – 46.88 Mb) (Ross *et al.* 2005). The position of the XAR-XCR boundary is supported by comparison of the human X chromosome with the opossum genome (Mikkelsen *et al.* 2007). As the position of the XCR-XAR boundary is well defined, therefore stratum 2a is located inside the XCR. It was commented by Ross *et al.* (2005) that stratum 2a should not be considered as XCR and no mention of stratum 2b genes was made in that report.

An independent method to assess the age of various regions of the human X was to compare the divergence of genes that have copies on the X and Y chromosome (Lahn and Page 1999). Nucleotide substitutions can either alter the amino acid sequences (non-synonymous substitutions) or not alter the amino acid sequences (synonymous substitutions) because of redundancy in the genetic code. Synonymous substitutions are nearly neutral and the rate of synonymous substitution at a synonymous site (*Ks*) can be used as a measure of divergence time. Therefore by comparison of the *Ks* values for genes

with copies on the X and the Y chromosome, a relative estimate of their divergence time can be obtained.

Based on the comparisons of *Ks* values for 19 genes with copies on the X and the Y chromosome, the human X chromosome was divided into four evolutionary strata, each representing different stages of the X chromosome evolution (Lahn and Page 1999). *These strata are different to strata defined by Kohn et al. (2004).* The Lahn and Page (1999) model of the X chromosome evolution suggested that the X conserved region was divided into two strata where stratum 1 represents the long arm of the X chromosome and stratum 2 represents Xp11 genes. This indicated that stratum 2 genes were evolving at a rate different to the stratum 1 genes, similar to observations made by Kohn *et al.* (2004), who indicated that stratum 2 genes were added on the proto-X chromosome of therian mammals and therefore might be expected to evolve at a different rate to stratum 1 genes.

The age of the stratum 2 predicted by Lahn and Page (1999) and more recently by (Skaletsky *et al.* 2003), was approximately 150 million years, suggesting stratum 2 has been part of the X chromosome for at least 150 million years. This approximate age for stratum 2 defined by Lahn and Page (1999) of 150 million years is in agreement with the theory of its independent addition to the proto-X chromosome some time after the divergence of mammals from common ancestor with birds (~310 – 168 million years) (Kohn *et al.* 2004).

The agreement between two independent lines of evidence suggested an additional evolutionary layer on the human X chromosome and it required further investigation in regard to the precise origin of the stratum 2 genes defined by Kohn *et al.* (2004).

3.1.5 Evolution and origin of the human stratum 2a (from Xp11) and stratum 2b (from Xq28) genes

More recently, the availability of genomic sequence data from lizard, frog, platypus, opossum and other mammals can provide us with insight into the origin of stratum 2 genes. The mapping of the human X chromosome stratum 2a and stratum 2b genes to the platypus chromosome 6, along with other genes from the long arm of the human X chromosome, suggests that these two regions are part of the conserved region of the human X chromosome (Veyrunes *et al.* 2008). *However, comparisons with the chicken genome suggest that these regions are not part of the conserved region of the X chromosome but are instead a separate evolutionary block on the human X chromosome.*

Comparative data suggest that stratum 2a and 2b were contiguous in an ancient vertebrate ancestor. Stratum 2a and 2b genes co-localize on pufferfish chromosome 9 and zebrafish linkage-group 8 (Grutzner *et al.* 2002, Kohn *et al.* 2004). Genomic sequence

data of frog (*Xenopus tropicalis*), lizard (*Anoles carolinensis*), and platypus (*Ornithorhynchus anatinus*) suggests that stratum 2a and 2b are co-localised in these species as well. In the frog, scaffold_154, scaffold_456 and scaffold_507 contain genes from both stratum 2a and stratum 2b (*Xenopus tropicalis* genome assembly, version 4.1, August 2005). Likewise, in the anole lizard, scaffold_32, scaffold_406 and scaffold_481 contain genes from stratum 2a and stratum 2b (*Anolis carolinensis* genome assembly, AnoCar1.0, February 2007). Similarly, in platypus, contigs Ultra403 and Ultra519 have genes from both these regions (*Ornithorhynchus anatinus* genome assembly, Ornithorhynchus_anatinus-5.0.1, March 2007). In platypus, contigs Ultra403 and Ultra519 have been physically localised on the chromosome 6 with other genes from the XCR by fluorescent *in-situ* hybridization (FISH) (Veyrunes *et al.* 2008).

However, the mapping of the genes in this single evolutionary block to different autosomes in chicken makes the origin of this region unclear. The data on which the conclusion of the independent origin of stratum 2a / 2b was based must therefore be scrutinized. The method used to compare approximately 300 genes from the human X chromosome with the chicken genome was not disclosed in the previous report (Kohn *et al.* 2004). However, another report (Ross *et al.* 2005) of comparisons of the human X chromosome to the chicken genome used BLASTZ global alignment method (Schwartz *et al.* 2000) and arrived at the similar conclusion to that of Kohn *et al.* (2004) for stratum 2a genes. The relationship of the stratum 2b with the chicken genome is not mentioned at all in the report by Ross *et al.* (2005). It is assumed that both reports (Kohn *et al.* 2004, Ross *et al.* 2005) used BLAST based methods for comparing the human X chromosome with the chicken genome, since both reports present similar results.

BLAST is a pairwise alignment tool in which the alignments between a query sequence and hit sequences are formed based on the predefined substitution matrix with penalties for mismatches and gaps (Altschul *et al.* 1990, Altschul *et al.* 1997). The resultant alignments are scored for each hit sequence by calculating the statistical significance of each hit, and all hits are ranked from highest similarity to lowest similarity. Consider a query sequence $Query^a$ from the human X chromosome aligned to three hit sequences, Hit^a, Hit^b and Hit^c, in the chicken genome with the true ortholog Hit^a scoring highest and the paralogs Hit^b and Hit^c scoring lowest. This process is very robust. However, there is always the problem of incorrectly annotating paralogs as if they were orthologs if the data analyzed is incomplete. For instance, if the Hit^a sequence is absent from chicken genome then Hit^b or Hit^c will score highest against $Query^a$ from the human X chromosome. This results in calling Hit^b or Hit^c as the ortholog of the $Query^a$ from the human X chromosome. As demonstrated, BLAST-like alignment methods for the identification of orthologs may detect false positives in the absence of true orthologs. For this reason, phylogenetic analyses are more robust than BLAST-like alignments when incomplete genomic data are analyzed.

In order to clarify whether or not stratum 2a and 2b of the human X chromosome have an independent origin and are truly orthologous to the genes on the chicken chromosomes 1, 12 and 26, I examined the phylogenetic relationship of gene families from human Xp11 and Xq28 including all homologs in the chicken genome. Phylogenetic analysis of gene families from these two strata will help us understand the relationship between the chicken genes and that of human stratum 2a and 2b genes. I have also included homologous sequences from the avian EST data in the phylogenetic analysis to account for the possibility that the chicken data are incomplete.

3.1.6 Aims

The following research is designed to answer three major questions. First, I aimed to determine the location of human Xp11 and Xq28 genes in marsupials to confirm that these genes have been part of the X conserved region since the divergence of therian mammals from the common ancestor with monotremes. Secondly, I aimed to compare the locations of all homologs (orthologs and paralogs) of the human Xp11 and Xq28 genes in the human, rat, opossum and chicken genomes to infer syntenic relationships between human stratum 2a and 2b genes and genes found in the chicken genome. My third aim was to infer the phylogenetic relationship between the human Xp11 and Xq28 genes with the homologs in the chicken genome.

3.2 Methods

3.2.1 Physical location of human Xq28 orthologs in tammar wallaby

A tammar wallaby BAC clone library was screened with radioactively labeled overgo probes. Overgo probes are 40 bp oligonucleotide probes constructed by annealing together two 24-mer oligonucleotide sequences that overlap over 8 bp. Complementary strands are end-filled with radioactive ^{32}P-dCTP and ^{32}P-dATP and the resulting radioactively labeled 40 bp oligonucleotides are used as probes.

For the genes of interest, opossum (*Monodelphis domestica*) cDNA sequences were obtained from Ensembl (http://www.ensembl.org/index.html). Opossum cDNA sequences were then used to search the tammar wallaby whole-genome shotgun (WGS) trace archive database using BLAST at the NCBI website (http://www.ncbi.nlm.nih.gov/). Significantly good, unique sequences from the tammar wallaby WGS trace archives were then selected and used to search the opossum genome assembly to establish the reciprocal best hit relationship, to reduce the false-positive discovery. Reciprocally related tammar wallaby sequences for the genes of interest were selected to design overgos. Overgos were

designed using the local copy of overgo-maker program available from Washington University Genome Sequencing Centre website (http://genome.wustl.edu/tools/software/overgo.cgi). The overgo probe sequences were checked to make sure that they were non-repetitive (Table 4).

Table 4 List of genes from human Xq28 and corresponding 40 bp overgo probes selected for BAC clone library screening and physical mapping in tammar wallaby.

Gene Name	40 bp probe sequence
TMEM185A	TGCCATCCTCTTGTGTCCTTATACACACACACACACACAC
ARD1A	GTTGTGAGGGAAAGAGAACAGGACCTCTTTCCCTCACAAC
ATP2B3	TTGTGGGAGAACTCACTGCATTGCTGTTGTGGGCATCGA
PLXNB3	AGGAATTGGCCAATGGGATCAGAGAGTTGGGAGGCATTCA
HMGB3	AGAGGAAAGAGAAGGGGTTGGACTAGGTAGATGACCCCTA
MTMR1	CGTGAATACCTGAGGTCTCATTCGTGGATTCCATCGGACA
G6PD	CAAAGAAGGAGGAAGTATGTCTGGTAGCTGCTGAATGGCA
ARHGAP4	ACCAGCTAGTTCAGACGCTCATTGTGCAGCCCGAATTGAT
SLC6A8	GCACGGTAGATTTGCTGGTGGAAATGAGGGATGGGAAATG
L1CAM	CAAATCCCTCTCAGCACCTTTCAGTCTGTCTCATCCTCCT
VAMP7	GGCTATCCTTTTCGCAGTTGTTGCCAGAGGCACAACCATC
STK23F	CTGGTGGGGATAATTGGAGCTAACTCAGATGGTCTGTGTG
F8	CTCACCCTCTTCCCATTTTCTGGGGATACTGTCTACATGG
CNGA2	TCCCTCAACCTGCAATGTCTGAAGCCCTCAATCTCACACA
DUSP9	ATGGGGATCTGCTTGTAGTGGATGTCGCCGTCCTTTTCAA
PDZD4	TGGCATGACCACAGATGATGATGCCGTGAGTGAGATGAAG
IDH3G	ACAAGACAGCTTGGGTCAGAGAGAAATGTATCCAGTACCC
RAB39B	GGAAATCCAGACCTTGTCCGAAAGGCACAATGGCAGATGT

3.2.2 BAC library screening

The tammar wallaby Bacterial Artificial Chromosome clone (BAC clone) library (Me_Kba) was obtained from Arizona Genomic Institute (AGI) (Tucson, AZ, USA). The library contains BAC clones with average insert size of 166 Kb. The Me_Kba library covers 11.36 genome equivalents on 11 nylon filters. In a single library screening experiment, at least 3 library filters were used (~3× genome coverage).

The Me_Kba library was screened with overgo probes using the BACPAC resources overgo hybridization protocol available at http://bacpac.chori.org/overgohyb.htm (Ross *et al.* 1999). Briefly, radioactively labeled overgo probes were hybridized to BAC clone library filters overnight at 60°C. Filters were then washed and exposed to autoradiography films (Amersham Hyperfilm MP, GE Healthcare UK Ltd, Little Chalfont, Buckinghamshire, UK) in cassettes at -80°C overnight. Autoradiography films were then developed, fixed and washed as per manufacturer's recommendations. Positive signals were identified using the grid supplied by AGI. If the positive signals were not visible after the overnight exposure, filters were exposed to new autoradiography films for an additional 14 days. The resulting positive BAC clones were then isolated from the in-house tammar wallaby

Me_Kba library glycerol stock. Multiple probes (more than 5 probes) were used simultaneously in any library screening experiment.

Positive clones were grown on a Luria-Bertani (LB) agar plate with 15 μg/ml chloramphenicol. A single colony for each positive clone was subcultured in 15-20 ml LB medium supplemented with 15 μg/ml chloramphenicol for enrichment. Positive BAC clones were tested for false positive signals and individual BAC clones were assigned to the gene of interest by Dot-Blot method (because multiple probes were used to screen the library) (Deakin *et al.* 2008). For the Dot-Blot, 0.5 μl of culture was spotted on Hybond N+ nylon filters (GE Healthcare UK Ltd, Little Chalfont, Buckinghamshire, UK) placed on LB agar plate with 15 μg/ml chloramphenicol. Plates were incubated at 37°C overnight. Nylon filters were removed from the plates next day and colonies on the nylon filters were lysed, denatured and neutralized as per manufacturer's recommendations. The resulting BAC clone DNA was fixed to the nylon filters by soaking in 0.4 M NaOH for 20 minutes on Whatman 3MM chromatography paper (Whatman International Ltd, Maidstone, Kent, UK). Dot-Blots thus obtained were screened using one probe at a time using the same protocol as BAC library screening described earlier. Glycerol stocks were maintained at –80°C for positive BAC clones with glycerol/culture ratio of 4/1.

3.2.3 DNA isolation from BAC clones

DNA was isolated from positive BAC clone using Wizard® Plus SV Minipreps DNA Purification System (Promega Corporation, WI, USA) as per the manufacturer's protocol, scaled up to accommodate 15 ml of BAC clone culture. The isolated DNA was stored at –20°C until needed.

3.2.4 BAC clone confirmation by sequencing

Each of the positive BAC clones identified by the Dot-Blot method were confirmed again by sequencing of a gene specific PCR product from the BAC clone (Table 5). For this, fragments of DNA for genes of interest were amplified from positive BAC clones by the polymerase chain reaction (PCR) with the following parameters. The template concentration of 1-20 ng per reaction and primer concentration of 0.4 μM to 1.0 μM was used for a 10 – 25 μl reaction volume. Reactions were carried out using GoTaq® Green Master Mix (Promega Corporation, WI, USA) as per the manufacturer's recommendations. PCR cycling conditions were as follows: an initial denaturing temperature of 94°C for 10 minutes followed by 35 cycles of denaturation at 94°C for 30 seconds, annealing at 60°C for 30 seconds, and extension at 72°C for 1 minute. A final extension at 72°C for 10 minutes was carried out and the PCR was incubated at 4°C until further processing. PCR products were analyzed on a 2% TAE buffer agarose gel. Electrophoresis was carried out

at 100 volts for 45 – 60 minutes. Genomic DNA was used as a positive control, and plain MilliQ water was used as a negative control for the reactions. Positive BAC clones for *VAMP7* and *F8* genes were not confirmed by sequencing.

Table 5 List of genes and corresponding PCR primers used for sequence gene fragments from the positive BAC clones. Expected product size is also listed against each gene.

Gene Name	Forward primer	Reverse primer	~Product size
TMEM185A	ACTTAGTTGTGCCACCCTG	CAGTATCTGTTGCAGCCTGT	471 bp
ARD1A	CTGTACTCACAGCCAACTTTC	TTGAGGCCTGGTCCATAA	312 bp
ATP2B3	AGATTGTGGGAGAACTCACC	ACTTGGCATCTGAAGACATG	254 bp
PLXNB3	GCATTGCTGTATCTCTGGG	AAGGTCTAAGGACTTCGGAA	625 bp
HMGB3	GGGACTGTGACTAGAGTGCTT	CTCCTGACTTCCAAATCCA	877 bp
MTMR1	CACCGTGAATACCTGAGGT	CTAGCTCCCATCTGCATTT	676 bp
G6PD	CAGGAGGGCCTCATATGTA	GTTCTACTCACCCAGAAACTCT	423 bp
ARHGAP4	GAGGGACTTTAGGGTCATCT	TCCTGCTTCACCCTCTTTA	804 bp
SLC6A8	ACAATGGCCAGCTACCAT	CAAGGGAGTAGGAGTGACTTAC	993 bp
L1CAM	GAGTAACTCCAGACCCAGTCT	ATGAGGAGGAGTGACCACAC	582 bp
STK23F	CCCTGGAGCTGTATCTTGA	CAGTGACCACTCGTATTTCTC	288 bp
CNGA2	GCAGGGTGTGGATATAGTTG	TCCTGAATGGGAGCTCAT	317 bp
DUSP9	TTCCACACTCTATGCCCAC	CCTTACCACTTGTAGCTTCAG	718 bp
PDZD4	TGCTAGAGGAGGAGAGTCTCTA	TGATCCAGTTGTCCAGGAT	937 bp
IDH3G	CATAATTTGCTCCAGCCAC	AGGGTTCTTCTAGTATGTCGG	186 bp
RAB39B	CCTCAGAAGAGGTTGTCAAA	ACCTCACTGTAGGAGAAGCTAC	455 bp

Gel bands corresponding to the PCR products were excised, and the DNA was isolated from the gel using the QIAquick Gel Extraction Kit (QIAGEN GmbH, Hilden, Germany) as per manufacturer's recommendations. 7.5 µl of PCR product was mixed with 0.5 µl of one 20 µM forward strand PCR primers for each gene of interest. The mixture was sent to AGRF (Brisbane, Australia) for direct sequencing.

3.2.5 Fluorescent *in-situ* hybridization (FISH)

Approximately 1 µg of DNA isolated from the positive BAC clones was used as a probe for FISH. Briefly, the BAC clone DNA was labeled with SpectrumOrange dUTP or SpectrumGreen dUTP (Abbott Molecular Inc., Des Plaines, IL, USA) by nick translation. The fluorescently labeled BAC clone probes were hybridized to tammar wallaby male metaphase chromosomes on slides (Alsop *et al.* 2005) overnight at 37°C. The unbound probe was removed by washing slides with 0.4× SSC / 0.3% (v/v) Tween20 for 2 minutes at 60°C followed by a second wash with 2× SSC / 0.1% (v/v) Tween20 for 1 minute at room temperature. Metaphase chromosomes were counterstained with DAPI (4', 6'-diamino-2-phenylindole) in VectaShield (Vector Laboratories Inc., Burlingame, CA, USA). The fluorescent signals were observed with a Zeiss Axioplan2 epifluorescence microscope. A SPOT RT Monochrome CCD (charge-coupled device) camera (Diagnostic Instruments Inc., Sterling Heights, MI, USA) was used to capture DAPI stained chromosome images and

fluorescent signals. The images were merged using IP Lab imaging software (Scanalytics Inc, Fairfax, VA, USA).

3.2.6 Use of the Ensembl database

The Ensembl v53 (March 2009) data was used for comparative analysis. Using the Ensembl BioMart tool, orthologs of 243 genes from human Xp11 and Xq28 were identified in the platypus, chicken, anole lizard, and the frog genomes (Vilella *et al.* 2009). Ortholog data was analyzed using Perl scripts.

Autosomal paralogs of the human Xp11 and Xq28 genes in human were identified from Ensembl (Vilella *et al.* 2009). Local gene order and gene content in the 500 kb flanking regions (upstream and downstream) of each human paralog was identified from the Ensembl database. The human paralogs of the human Xp11 and Xq28 genes and their 500 kb flanking regions, are referred to here as "*paralogous regions*". One-to-one orthologs were identified in the rat, opossum and chicken genomes for the genes in paralogous regions of the human genome. Orthologous genes are the genes derived from a common ancestral gene as a result of the speciation. Paralogous genes are derived as a result of duplication before or after the speciation event (outparalogs and inparalogs respectively). Comparative analysis was performed for local gene order and gene content in rat, opossum and chicken for each human paralogous region to trace the evolutionary history of the gene families of human Xp11 and Xq28 genes.

To exemplify the above process briefly, the *IRAK1* gene lies in human Xq28 region. The paralogous genes of *IRAK1* are *IRAK2* (human chromosome 3: 10,181,563 – 10,260,427 bp) *IRAK3* (human chromosome 12: 64,869,270-64,928,684 bp), and *IRAK4* (human chromosome 12: 42,439,047-42,468,166 bp). For all three genes, *IRAK2*, *IRAK3*, and *IRAK4*, the paralogous regions were identified by subtracting 500 kb from their start positions (upstream, 5' flanking) and adding 500 kb to their end positions (downstream, 3' flanking). Protein-coding genes in the 5' and 3' flanking regions were then identified (Hubbard *et al.* 2007). This gave us the local gene content and gene order in the paralogous regions. One-to-one orthologs for the genes in the paralogous regions were identified in the rat, the opossum and the chicken genome.

3.2.7 TreeFam database

The TreeFam database is a collection of phylogenetic trees for all animal genes (Li *et al.* 2006, Ruan *et al.* 2007). Plant and yeast genes were used as outgroups for the construction of phylogenetic trees in this database. The phylogenetic trees of genes were helpful in understanding the orthologous and paralogous relationship within a gene

family. The TreeFam phylogenetic trees for Xp11 and Xq28 genes were examined for chicken orthologs.

3.2.8 Reciprocal best-hit search

The reciprocal best hit criterion was used to identify chicken/zebrafinch EST/cDNA sequences that were reciprocally related to the human Xp11 and Xq28 genes. Chicken EST/cDNA sequences were downloaded from the BBSRC ChickEST database at <http://www.chick.manchester.ac.uk/> (Boardman *et al.* 2002, Hubbard *et al.* 2005). Zebrafinch EST/cDNA sequences were downloaded from the Songbird Neurogenomics (SoNG) Initiative website http://titan.biotec.uiuc.edu/songbird/ (Replogle *et al.* 2008). Human cDNA sequences (including all alternatively spliced transcripts) for Xp11 and Xq28 genes were downloaded from Ensembl v53.

Human cDNA sequences were used as query sequences to search the chicken/zebrafinch EST/cDNA database using nucleotide BLAST search tool (Altschul *et al.* 1990, Altschul *et al.* 1997). Default search parameters were used except the e-value cut-off was set to 1.0. Chicken/zebrafinch EST/cDNA sequences that aligned significantly (e-value < 1.0) to the human cDNA sequences for Xp11 and Xq28 genes were then isolated. Reciprocally, chicken/zebrafinch EST/cDNA sequences were used as queries to search the human cDNA sequence database using nucleotide BLAST search tool. Only the first hit was retained for all queries in the reciprocal search. The results of both forward (human→chicken/zebrafinch) and reciprocal (chicken/zebrafinch→human) searches were parsed to establish a one-to-one reciprocal relationship. All reciprocally identified chicken/zebrafinch EST/cDNA sequences were then mapped to the chicken genome using the MegaBlast tool of the BLAST suite (Zhang *et al.* 2000). To determine the location of the EST/cDNA sequences, the percent identity of the alignment was set to be 95% or higher for the chicken EST/cDNA sequences, and 80% or higher for zebrafinch EST/cDNA sequences. 95% sequence identity was chosen to allow for single nucleotide polymorphisms (since EST/cDNA sequences were obtained from more than one individual) and sequencing errors.

3.2.9 Building phylogenetic trees including the chicken/zebrafinch EST/cDNA sequences

The chicken/zebrafinch EST/cDNA sequences were translated in all six reading frames to make a translated sequence database. All protein sequences that make up a gene family in the TreeFam database were extracted. These protein sequences were used to search for homologous chicken/zebrafinch EST/cDNA sequences using BLASTP for protein sequences with default parameters. A sub-dataset of the homologous chicken/zebrafinch

EST/cDNA translated sequences was created for each gene family after parsing the BLAST search results. This sub-dataset was then filtered using the Hidden-Markov model (HMM) search using the HMMER program with an e-value cut-off of 0.1. Gene family HMMs obtained from the TreeFam database were used as queries. Gene family sequences obtained from the TreeFam database and chicken/zebrafinch EST/cDNA sequences specific to the gene family were aligned using ClustalW (Ma *et al.* 2007, Thompson *et al.* 1997). ClustalW parameters for pair-wise alignments for multiple sequence alignments were set to PWGAPOPEN = 50, PWGAPEXT = 2, GAPOPEN = 3, GAPEXT = 1, NUMITER=25, MAXDIV=30 and GONNET matrix. Gap opening and extension penalty were set higher than default values to allow partial sequences to match with gaps at the ends rather than in between residues. TreeBeST (Li 2006) was then used to filter and retain columns with a score greater than 15 (Thompson *et al.* 1997), and construct neighbour-joining trees based on the *p*-distance and Kimura's corrections for transition and transversion rates. The resulting neighbour-joining phylogenetic trees were read for topology using the "ortho" module of the TreeBeST program. The chicken/zebrafinch EST/cDNA sequences identified as orthologs of the human Xp11 and Xq28 genes were mapped to the chicken genome using MegaBlast tool of the BLAST suite as described earlier.

3.3 Results

3.3.1 Location of stratum 2a and stratum 2b genes in tammar wallaby, opossum and platypus

To test whether the human Xq28 region is a part of the conserved region of the X chromosome by the criterion of its position in marsupials, I physically localized 18 genes from this region on the tammar wallaby metaphase chromosomes by FISH (section 3.2.1 to 3.2.5). Gene specific BAC clones were used as probes to hybridize to metaphase chromosomes of the tammar wallaby. These clones were obtained from the tammar wallaby BAC clone library by screening for the selected 18 genes using radioactively labeled overgo probes.

A total of 23 BAC clones were isolated that were positive for one or more of the 18 human Xq28 genes (Table 6) by the Dot-blot method and/or sequencing of a gene specific PCR amplification product from the BAC clone using gene specific primers. The sequences of the PCR amplification products from the BAC clones were identical to the tammar wallaby gene sequence, indicating that BAC clones isolated for each gene were true positives. All of these BAC clones map to the tammar wallaby X chromosome (Figure 7, Figure 8, Figure 9, Figure 10, Figure 11).

Table 6 List of human Xq28 genes and corresponding positive tammar wallaby Me_Kba BAC clone ID. The mapping location of each clone on the tammar wallaby chromosomes is also shown. All these genes are conserved on the X chromosome of all eutherian mammals.

Gene Name	BAC-clone ID	Tammar Chromosome Location
TMEM185A	12H21	Xq
ARD1A	31M16, 57G1	Xq
ATP2B3	44B18, 48L7, 48N8	Xq
PLXNB3	328I20	Xq
HMGB3	70E7	Xq
MTMR1	129P4	Xq
G6PD	381I4, 409J12, 120M12, 45A17	Xq
ARHGAP4	31M16	Xq
SLC6A8	57G1, 44B18, 48L7, 48N8	Xq
L1CAM	57G1	Xq
VAMP7	87J20, 378A5, 12E16	Xq
STK23F	328I20	Xq
F8	398A5	Xq
CNGA2	428L19	Xq
DUSP9	113H6, 277A14, 148G5	Xq
PDZD4	328I20	Xq
IDH3G	328I20	Xq
RAB39B	142J18, 3O7	Xq

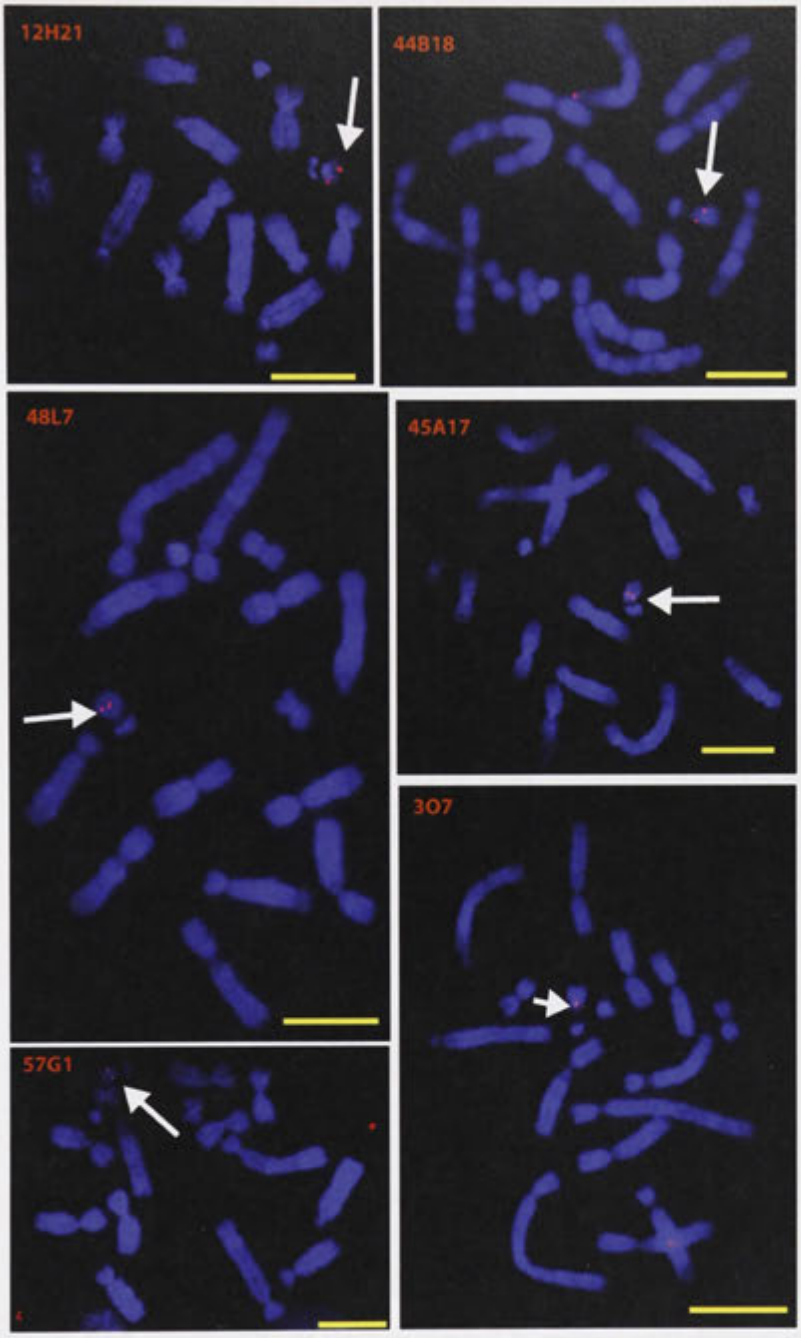


Figure 7 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.

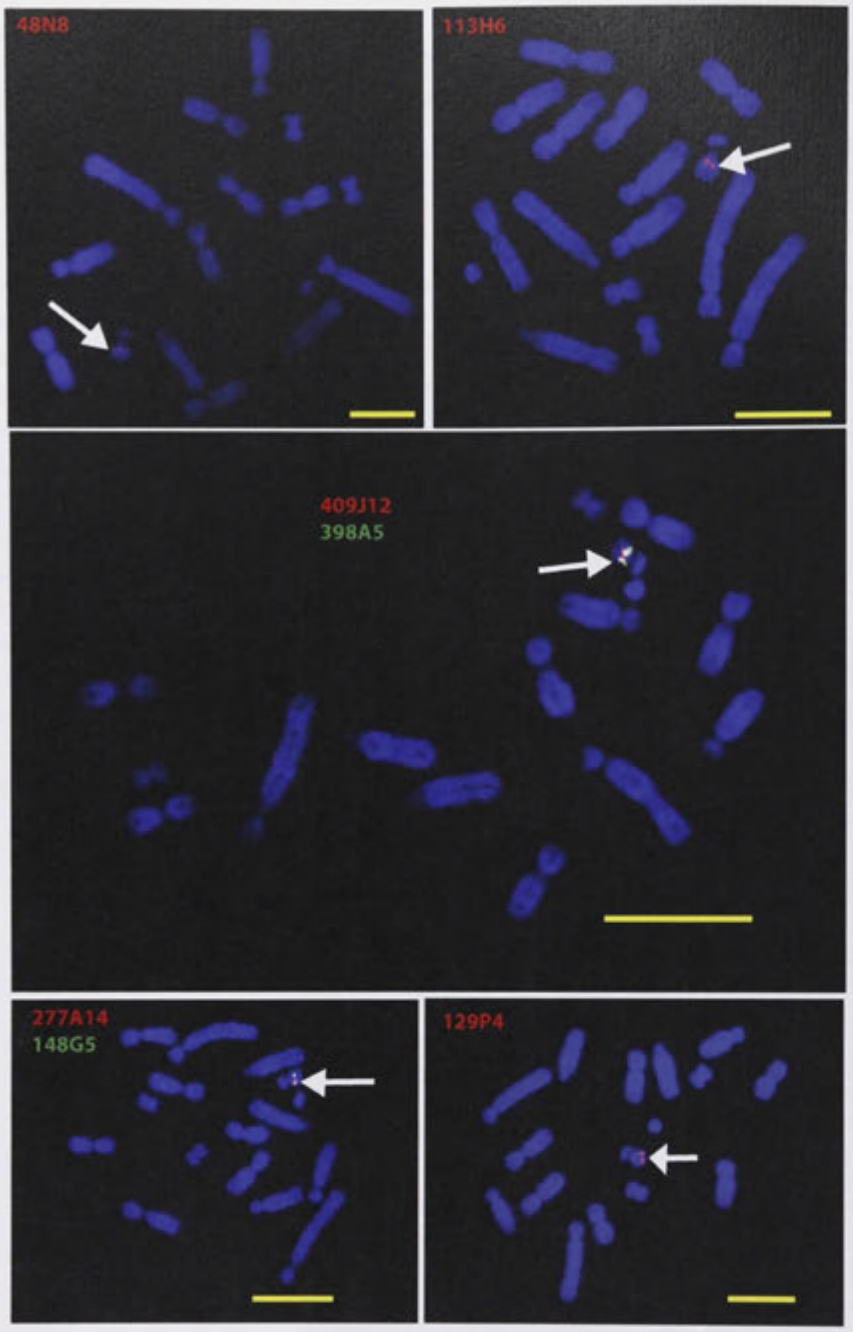
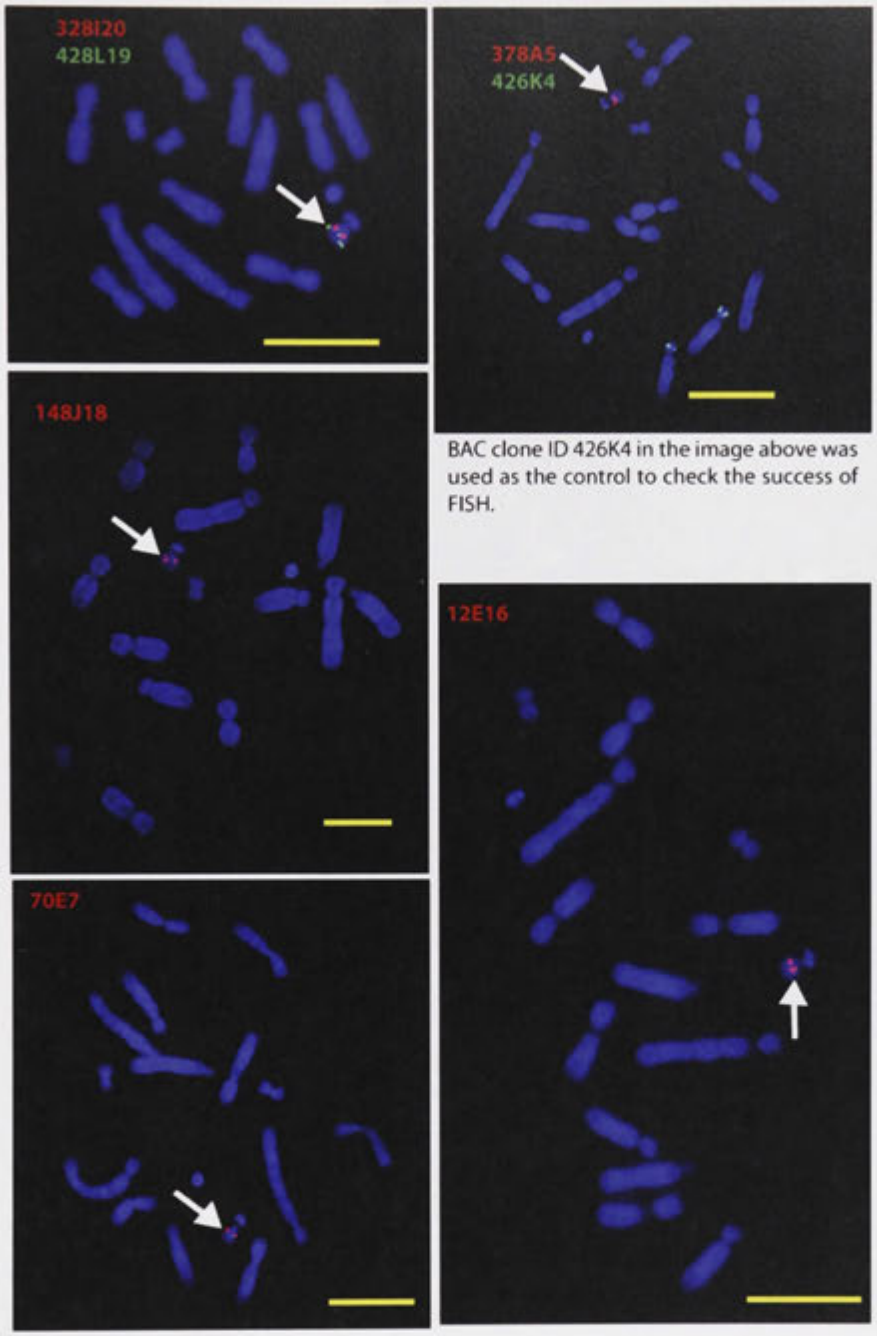


Figure 8 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.



BAC clone ID 426K4 in the image above was used as the control to check the success of FISH.

Figure 9 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.

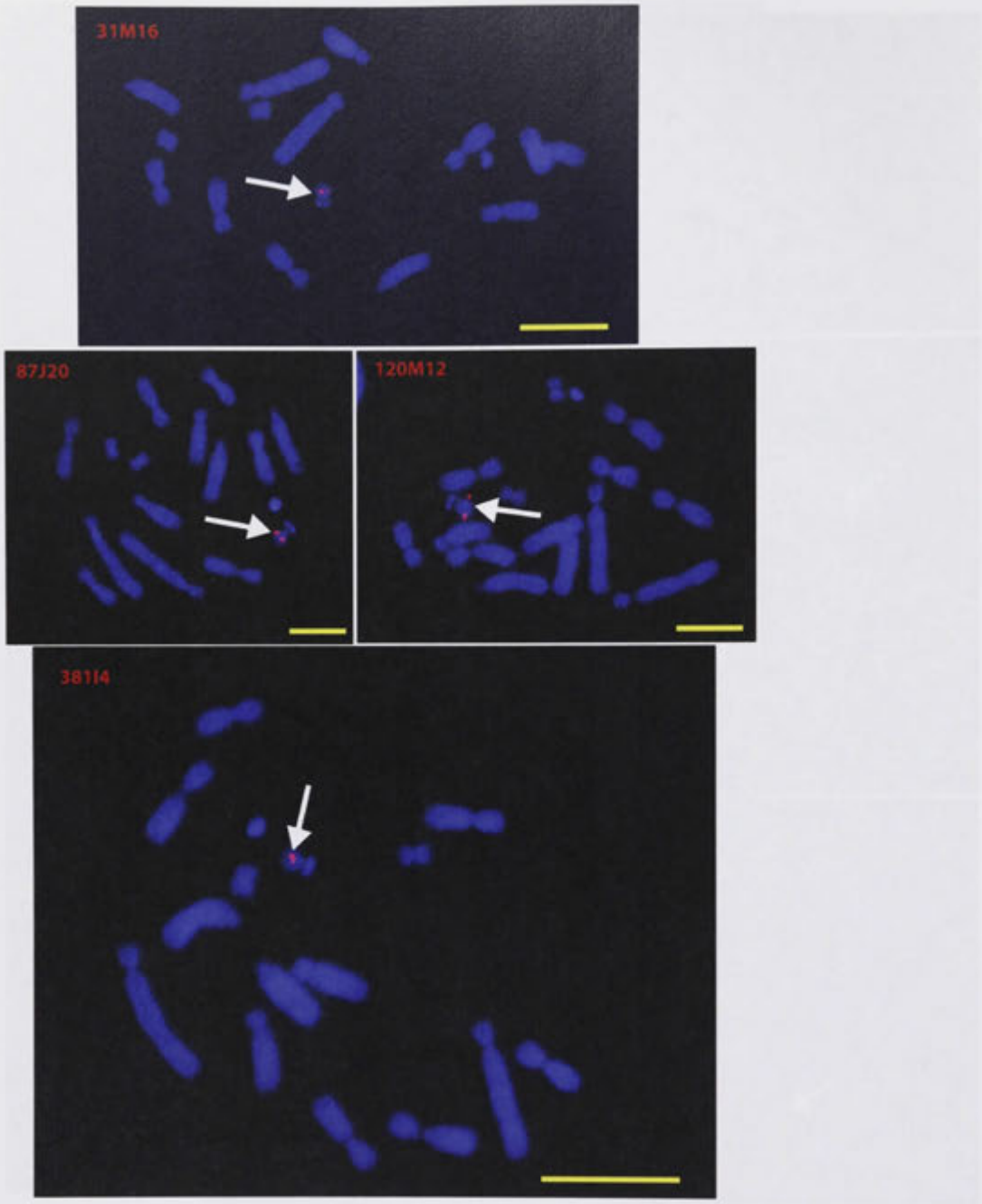


Figure 10 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.

My results were consistent with early mapping of some of the Stratum 2a and 2b genes. Four genes from stratum 2a (*GPR173, JARID1C, RIBC1* and *HUWE1*) were already known to map to the tammar wallaby X chromosome (Delbridge *et al.* 2009). Stratum 2a and 2b genes are also conserved on the X chromosome of the opossum (Ensembl v55 and Delbridge *et al.* 2009).

Therian mammals have diverged from the common ancestor with platypus 166 MYA. Physical localization of stratum 2a and 2b genes on the X chromosome of distantly related eutherians and marsupials suggests that these strata have been conserved on the X chromosome in all therian mammals. My mapping results show that if stratum 2a and 2b were added to the therian X chromosome as an independent evolutionary block, this must have occurred prior to the radiation of therian mammals, approximately 148 MYA.

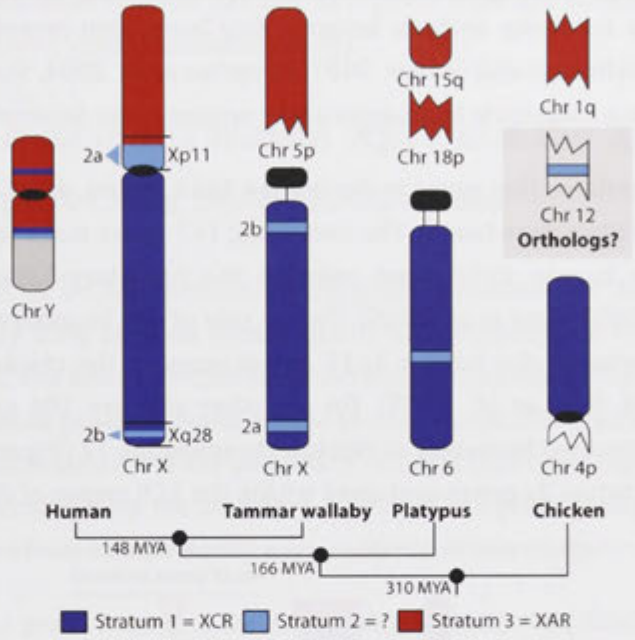


Figure 11 The schematic representation of the human Xq28 gene mapping in tammar wallaby. Stratum 2b (light blue) is present on the X chromosome of the tammar wallaby with other genes from the human XCR (dark blue) indicating that stratum 2b is part of the X-conserved region. Comparative analysis data between human and tammar wallaby were adapted from (Delbridge *et al.* 2009, Graves 1995), human and platypus from (Veyrunes *et al.* 2008) and human and chicken from (Kohn *et al.* 2004).

The hypothesis of stratum 2a and 2b comprising an independent evolutionary block on the human X chromosome, was then also tested by examining the location of these genes in the monotremes. The conserved region of the human X chromosome is conserved entirely on platypus chromosome 6 (Veyrunes *et al.* 2008). I found that Stratum 2a and 2b genes co-localize on at least two platypus contigs, including Ultra403 (0.9 Mb) and Ultra519 (9.9 Mb) that have been localised to platypus chromosome 6 by FISH. This data is consistent with at least 10 other contigs that are homologous to the conserved region of the X chromosome that have been localised on the platypus chromosome 6 (Veyrunes *et al.* 2008, Waters *et al.* 2005).

This result also suggests that stratum 2a and 2b have been part of the therian XCR and the proto-X chromosome in platypus, indicating that stratum 2a and 2b have had a similar origin to other therian XCR genes since mammals diverged from birds 310 MYA.

3.3.2 Identification of the human genes within Stratum 2a and 2b

The gene content of human cytogenetic bands Xp11 and Xq28 were then compared with the chicken, the opossum, and rat genomes. There are 99 cancer/testis antigen genes on the human X chromosome (Ross *et al.* 2005). These genes were concentrated in, but not limited to, the Xp11 and Xq28 regions. The cancer/testis antigen gene families were excluded from the following analysis because they have been recently expanded in the primate lineage (Delbridge and Graves 2007, Kouprina *et al.* 2004, Stevenson *et al.* 2007) and therefore do not contribute to analysis of the origins of the Stratum 2a and 2b genes.

There are 186 protein-coding genes in the human Xp11 region, 43 of which are members of cancer/testis antigen gene family. The remaining 143 genes make up the dataset for the Xp11 region. The human Xp11 band contains the boundary between XCR and XAR (Mikkelsen *et al.* 2007, Ross *et al.* 2005). On one side of this boundary were 37 genes that lie in the XAR region of the human Xp11 homologous to the chicken chromosome 1q (Kohn *et al.* 2004, Ross *et al.* 2005). On the other side are 106 genes that belong to stratum 2a, with reported homology to chicken chromosome 12 (Figure 12). I investigated the origin of the stratum 2a genes contained within the XCR region of the human Xp11.

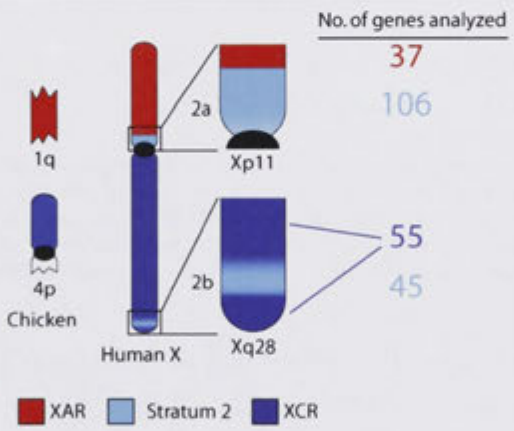


Figure 12 The schematic representation of the human X chromosome. The X-added region (XAR, red) is homologous to the chicken chromosome 1q region and the X-conserved region (XCR, navy blue) is homologous to the chicken chromosome 4p. The origin of 151 genes from the stratum 2a and the stratum 2b (light blue) is investigated by multi-spcecies comparative analysis.

Similarly, I found that human Xq28 included 121 protein-coding genes, of which 21 genes belong to the cancer/testis antigen gene family. The 100 remaining genes made up the dataset of human Xq28. The long arm of the human X chromosome (including cytogenetic band Xq28) is conserved on the X chromosome across all therian mammals (Graves 2006) and this region is largely homologous to the chicken chromosome 4p. The exception is the stratum 2b within the cytogenetic band Xq28 (Kohn *et al.* 2004). Stratum 2b consists of 45

protein coding genes, which are reported to share homology with multiple chicken chromosomes including chicken chromosomes 1, 12, 26 and other macro and micro chromosomes (Kohn *et al.* 2004). The two regions flanking stratum 2b in human Xq28 are homologous to the chicken chromosome 4p.

The human Xp11 and Xq28 dataset analyzed in this research therefore consisted of 243 genes (supplementary table 1); 143 genes from the human Xp11 region (37 genes from the XAR + 106 genes from stratum 2a) and 100 genes from the human Xq28 region (55 genes from the XCR + 45 genes from stratum 2b).

3.3.3 Orthologs of the human Xp11 and the Xq28 genes in tetrapods using Ensembl database

There were a total of 21,343 protein-coding genes annotated in the human genome (Ensembl v53). Only 63% of these human genes have orthologs in the chicken genome (Vilella *et al.* 2009). The low correspondence between the number of orthologous genes in the chicken and human genomes could be due to species-specific deletion and duplication of genes and gene divergence, but the possibility remains that some regions of the chicken genome might be missing from the assembly. A total of 243 genes (143 genes from human Xp11, and 100 genes from human Xq28) were analyzed in this study.

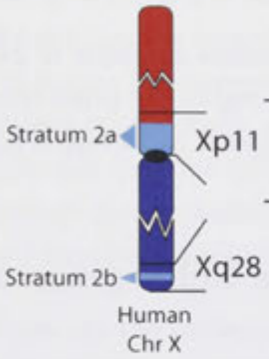
Of the 37 of the 143 genes belong to the XAR in human Xp11, 28 (75%) have orthologs in the chicken genome (Figure 13). 27 of these chicken orthologs map to chicken chromosome 1q (Ensembl annotations), consistent with previous reports (Kohn *et al.* 2004, Ross *et al.* 2005).

Only one chicken ortholog, for human gene *DUSP21,* maps to another chicken chromosome (chromosome 15). This particular chicken ortholog is related to the human *DUSP21* gene by a one-to-many relationship. A one-to-many orthologous relationship is defined for a gene when a single gene in human is orthologous to several genes in the chicken genome. This is likely to be the result of a chicken lineage specific duplication event after the divergence of the chicken lineage from the common ancestor with the human lineage.

Within Xq28 there are 55 genes in the known conserved region of the X chromosome, flanking the stratum 2b genes. Of these, 39 human genes have 38 orthologs in the chicken genome (71% of the human genes). Examining the locations of chicken orthologs using Ensembl annotations, it was revealed that 31 of the 38 chicken orthologs from the conserved region of the human Xq28 map to chicken chromosome 4p, consistent with previous reports (Kohn *et al.* 2004, Ross *et al.* 2005).

Of the seven genes that map elsewhere, one of the chicken orthologs lies on a genomic contig that has not been assigned to any chicken chromosome (chrUn). The remaining

orthologs mapped to chicken chromosomes 1 (outside the known XAR homologous region), chicken chromosome 4 (outside the known XCR homologous region) and chicken chromosome 14.



	No. of Genes Analysed	No. of Chicken Orthologs	No. of Anole Lizard Orthologs	No. of Frog Orthologs
Stratum 2a	37	28 (75%)	28 (75%)	27 (72%)
Xp11	106	16 (15%)	73 (68%)	50 (47%)
Stratum 2b	55	39 (71%)	32 (58%)	28 (50%)
Xq28	45	4 (9%)	39 (86%)	31 (68%)

Stratum 1 = XCR Stratum 2 = ? Stratum 3 = XAR

Figure 13 Schematic representation of the human Xp11 and Xq28 orthologs in chicken, lizard and frog. The data is retrieved from Ensembl database using BioMart tool.

Thus at least 70% of the genes within the XAR and XCR regions of human Xp11 and Xq28 shared orthology with the chicken genome, and most of these genes map to the respective XCR and XAR homologous regions in the chicken on chromosome 4p and 1q respectively. In contrast, only 13% of the stratum 2a and 2b genes had chicken orthologs (Figure 13, supplementary table 2). Four chicken orthologs of stratum 2a genes mapped to chicken chromosome 4p, showing that these genes are likely to be part of the homologous region of the XCR. Four other orthologs have not yet been assigned to any chicken chromosome (chrUn), so could also be located on chicken chromosome 4p. The remaining chicken homologs of stratum 2a genes map to the chicken chromosomes 1q (outside the known XAR homologous region), Z, 12, 14, and 18. However, the homologs mapping to chromosomes other than 4p or unassigned contig (chrUn) had a one-to-many correspondence with the human genes, so they may represent chicken lineage specific duplication of genes like *DUSP21* orthologs described earlier.

Similarly only four genes out of the 45 stratum 2b genes have chicken orthologs (Figure 13, supplementary table 2). Which are localized on chicken chromosomes 1, 12 and an unassigned contig. Therefore, the Ensembl database has identified no chicken orthologs for many human genes from the stratum 2a and 2b. Few genes for which the Ensembl database suggests the presence of orthologs in the chicken genome, share one-to-many orthologous relationship and hence orthologs by descent could not established unambiguously.

Both the frog and the anole lizard have more orthologs to human stratum 2a and 2b genes than does the chicken genome (Figure 13, supplementary table 2). This indicated that stratum 2a and 2b regions have been well conserved in vertebrates at least since

tetrapod/fish split 430 MYA (Blair and Hedges 2005), and therefore their absence/loss was confined to the chicken lineage.

Stratum 2a and 2b genes have been stably co-localised in distantly related species including frog, lizard and platypus (section 3.1.5). This strongly suggests that the two regions were originally located together, but became separated into two regions in the therian lineage. *My search of chicken orthologs for stratum 2a and 2b returned with at least 4 genes in the region mapping to the chicken chromosome 4p*, suggesting that stratum 2a and 2b genes, after all, were part of the conserved region of the therian X chromosome that is represented within the XCR homologous region of the chicken chromosome 4p. The absence of many stratum 2a and 2b genes from the chicken genome suggests that either these genes were lost from the chicken lineage, or are missing from the current chicken genome assembly.

The inconsistency of these results with previous reports raises two questions. First, what evidence was there to support the claims (Kohn *et al.* 2004, Ross *et al.* 2005) that these two regions form a separate evolutionary block on the human X chromosome? Second, was the region deleted from the chicken genome as a single event because stratum 2a and 2b are co-localised in frog, lizard and platypus? An analysis of the paralogous regions of the stratum 2a and 2b genes extended the analysis of the chicken homologs of these regions. This analysis was complemented by a search for orthologous genes of stratum 2a and 2b in an independent sequence data source comprised of chicken and zebrafinch EST/cDNA sequences.

3.3.4 Comparative analysis of the human Xp11 and Xq28 gene families with the rat, opossum, and the chicken genome

The Ensembl database contains multi-species comparative analysis that is mainly focused on evolution of gene families (Vilella *et al.* 2009). I used Ensembl homology assignments to analyze 171 genes from the Xp11 and Xq28 regions (including stratum 2a and 2b) with autosomal paralogs in the human genome. The 171 X-borne genes were found to have 832 autosomal paralogs scattered across all the autosomes (supplementary table 3). 92 out of the 171 genes (≈54%) have four paralogs or fewer in the human genome. This proportion is consistent with genome wide data indicating that 61% of the protein-coding genes in the human genome have four or fewer paralogs (Vilella *et al.* 2009).

Because previous reports (Kohn *et al.* 2004, Ross *et al.* 2005) claimed that the human stratum 2a and 2b regions have homology to chicken chromosomes 1, 12 and 26, these chicken chromosomes were closely examined for their homology with the human

chromosomes. The local gene content and gene order in human, rat, opossum and chicken is extremely well conserved for the human Xp11 and Xq28 paralogous regions (Figure 14).

Gene Name	Human		Rat		Opossum		Chicken	
	Gene ID	Chr 1 (Mb)	Gene ID	Chr 13 (Mb)	Gene ID	Chr 2 (Mb)	Gene ID	Chr 26 (Mb)
SOX13	ENSG00000143842	202.3	ENSRNOG00000028353	46.3	ENSMODG00000001192	108.7	ENSGALG00000000583	1.5
ETNK2	ENSG00000143845	202.4	ENSRNOG00000028368	46.3	ENSMODG00000001210	108.8	ENSGALG00000000557	1.5
KISS1	ENSG00000170498	202.4			ENSMODG00000025661	109.5		
GOLT1A	ENSG00000174567	202.4	ENSRNOG00000002936	46.2	ENSMODG00000001474	109.5		
PIK3C2B	ENSG00000133056	202.7	ENSRNOG00000029938	46.0	ENSMODG00000001583	109.9	ENSGALG00000000623	1.6
MDM4	ENSG00000198625	202.8	ENSRNOG00000009696	45.9	ENSMODG00000001605	110.0	ENSGALG00000000636	1.6
LRRN2	ENSG00000170382	202.9	ENSRNOG00000009691	45.8	ENSMODG00000024800	110.1	ENSGALG00000000639	1.6
NFASC	ENSG00000163531	203.1	ENSRNOG00000030515	45.5	ENSMODG00000001662	110.7	ENSGALG00000000644	1.7
CNTN2	ENSG00000184144	203.3	ENSRNOG00000009033	45.4	ENSMODG00000001692	110.9	ENSGALG00000000653	1.8
RBBP5	ENSG00000117222	203.3	ENSRNOG00000021289	45.4	ENSMODG00000001708	110.9	ENSGALG00000000668	1.8
TMEM81	ENSG00000174529	203.3	ENSRNOG00000028752	45.4	ENSMODG00000024792	110.9	ENSGALG00000000656	1.8
TMCC2	ENSG00000133069	203.5	ENSRNOG00000000033	45.2	ENSMODG00000001757	111.0	ENSGALG00000000673	1.9
NUAK2	ENSG00000163545	203.5	ENSRNOG00000000034	45.2	ENSMODG00000001769	111.1	ENSGALG00000000680	1.9
KLHDC8A	ENSG00000162873	203.6	ENSRNOG00000000036	45.2	ENSMODG00000001778	111.2	ENSGALG00000000684	1.9
LEMD1	ENSG00000186007	203.6	ENSRNOG00000039735	45.1	ENSMODG00000025662	111.3		
PCTK3	ENSG00000117266	203.7	ENSRNOG00000001137	45.0	ENSMODG00000001796	111.5	ENSGALG00000000692	2.0
MFSD4	ENSG00000174514	203.8	ENSRNOG00000024657	44.9	ENSMODG00000001804	111.5	ENSGALG00000000695	2.0
SLC45A3	ENSG00000158715	203.9	ENSRNOG00000007591	44.8	ENSMODG00000001824	111.7	ENSGALG00000000703	2.0
RAB7L1	ENSG00000117280	204.0			ENSMODG00000001843	111.9	ENSGALG00000000712	2.1
SLC41A1	ENSG00000133065	204.0			ENSMODG00000001859	111.9	ENSGALG00000000721	2.1
PM20D1	ENSG00000162877	204.1	ENSRNOG00000039745	44.7	ENSMODG00000001870	112.0	ENSGALG00000000724	2.1
SLC26A9	ENSG00000174502	204.1	ENSRNOG00000029514	44.7	ENSMODG00000001887	112.0	ENSGALG00000000745	2.1
AVPR1B	ENSG00000198049	204.4	ENSRNOG00000006891	44.5	ENSMODG00000001960	112.3	ENSGALG00000000788	2.2
CTSE	ENSG00000196188	204.5	ENSRNOG00000006963	44.6	ENSMODG00000001939	112.2	ENSGALG00000000786	2.2
IKBKE	ENSG00000143466	204.7	ENSRNOG00000025100	44.2	ENSMODG00000002016	112.7	ENSGALG00000013356	2.3
DYRK3	ENSG00000143479	204.9	ENSRNOG00000004870	44.1	ENSMODG00000002080	112.9	ENSGALG00000000863	2.3
MAPKAPK2	ENSG00000162889	204.9	ENSRNOG00000004726	44.0	ENSMODG00000002087	113.0	ENSGALG00000000883	2.3
IL10	ENSG00000136634	205.0	ENSRNOG00000004647	44.0	ENSMODG00000002097	113.1	ENSGALG00000000892	2.4
IL24	ENSG00000162892	205.1	ENSRNOG00000004470	43.8	ENSMODG00000025663	113.4		
C1orf116	ENSG00000182795	205.3	ENSRNOG00000004341	43.7	ENSMODG00000002153	113.6	ENSGALG00000001091	2.4
YOD1	ENSG00000180667	205.3	ENSRNOG00000025704	43.7	ENSMODG00000002159	113.6	ENSGALG00000023958	2.4
PRKCE1	ENSG00000139836	205.5	ENSRNOG00000004162	43.6	ENSMODG00000002191	113.7	ENSGALG00000001117	2.4
C4BPB	ENSG00000123843	205.3	ENSRNOG00000004125	43.6	ENSMODG00000002211	113.7		
CD55	ENSG00000196352	205.6	ENSRNOG00000003927	43.3	ENSMODG00000002282	113.9	ENSGALG00000023951	2.5

Figure 14 An example of comparison of human Xp11 and Xq28 paralogous regions in rat, opossum and chicken. The flanking regions corresponding to Xp11 paralogs (red rows) and Xq28 paralogs (blue rows) in human show extremely well conserved gene content and gene order between human (yellow cloumn), rat (yellow-green column), opossum (blue-green column) and chicken (gray column). A region on human chromosome 1 (202.3 – 205.6 Mb) contains six paralogs of the human Xp11 and Xq28 genes. This region is conserved on the rat chromosome 13 (43.3 – 46.3 Mb) in reverse orientation compared to the human gene order. The same region is conserved in the same orientation on the opossum chromosome 2 (108.7 – 113.9 Mb) and the chicken chromosome 26 (1.5 – 2.5 Mb).

When the gene order and gene content of all paralogous regions were compared the rat, opossum and chicken genomes, the blocks of conserved synteny were clearly evident. The region of chicken chromosome 1 that showed homology to the human Xp11 and Xq28 regions is more closely related in gene order and gene content to human chromosomes 7, 12 and 10 (Figure 15, supplementary table 4). Similarly, the genes on the chicken chromosome 12 are conserved on the human chromosome 3 and genes on the chicken chromosome 26 are conserved on the human chromosomes 1 and 10. This indicated that the genes from the chicken chromosome 1, 12 and 26 reported as orthologs (Kohn *et al.* 2004, Ross *et al.* 2005) are in fact paralogous to the human Xp11 and Xq28 regions. None of the chicken chromosome 1, 12 or 26 has any genes that are orthologous to the human X chromosome as per Ensembl annotations.

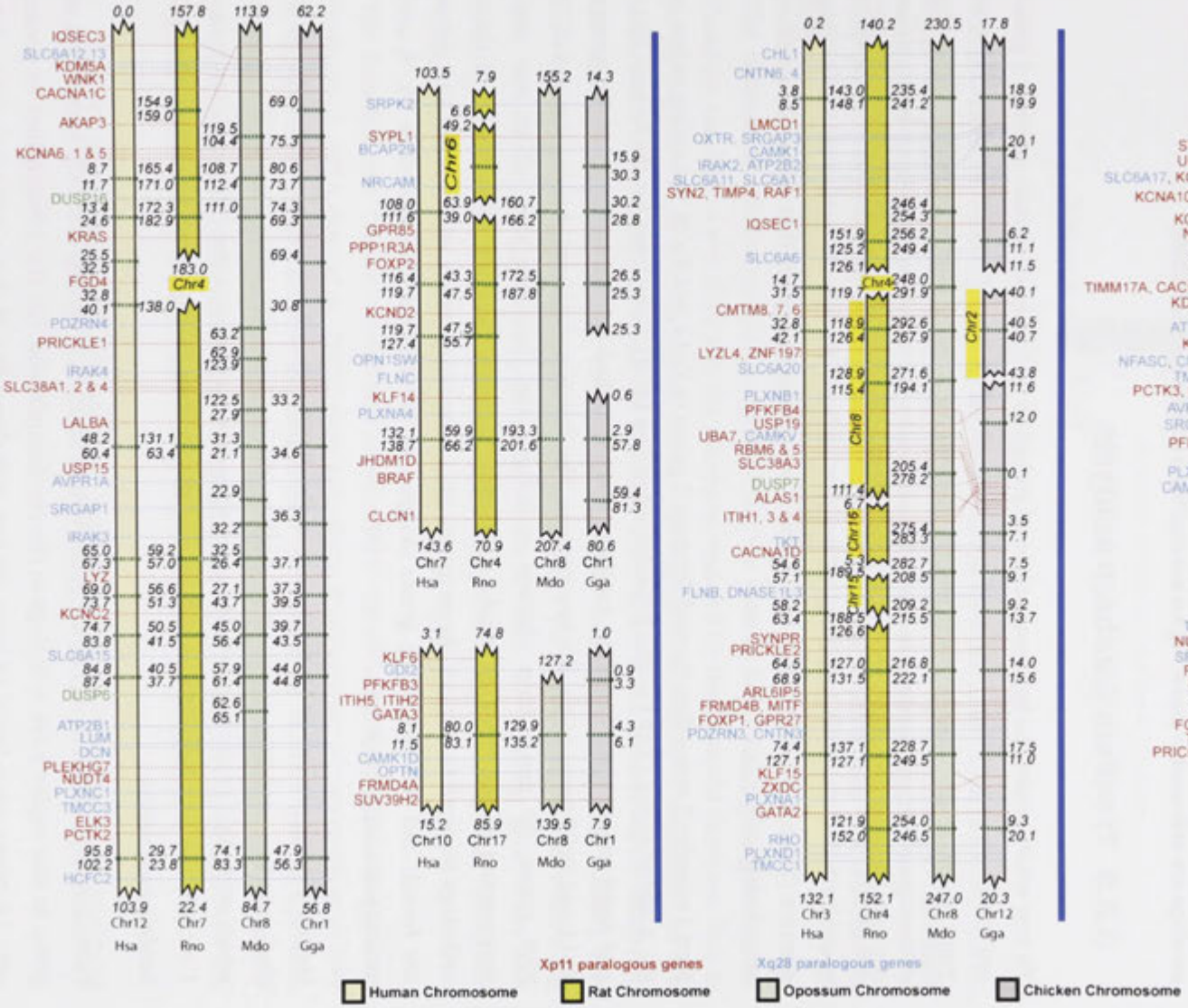

Xp11 paralogous genes
Xq28 paralogous genes
Human Chromosome
Rat Chromosome
Opossum Chromosome
Chicken Chromosome

Figure 15 Conservation of Xp11 and Xq28 paralogs and their genomic contexts in different species. Schematic representation of the location of Xp11 paralogs (red) and Xq28 paralogs (blue), including 1 Mb of genomic context surrounding each, on chicken (Gga, gray) chromosomes 1, 12, and 26. Conservation of the positions of these genes is indicated by the red and blue dotted lines across human (Hsa, yellow), rat (Rno, yellow-green), and opossum (Mdo, blue-green) chromosomes. Chromosome (Chr) numbers are indicated either below or at the side of each conserved segment, and the start and end points of the conserved sections along the chromosomes are indicated in megabases from the terminus of the short arm. Small intervals between the conserved blocks on a single chromosome are indicated by green horizontal dotted lines, and the start and end points of the intervals are also indicated in megabases from the tip of the short arm. Members of the *DUSP* gene family are found in both Xp11 and Xq28, and their paralogs are indicated in green. Figure adapted from (Delbridge *et al.* 2009).

3.3.5 TreeFam database analysis

To test whether the chicken genes claimed to be the orthologs of human stratum 2 genes are orthologs or paralogs, I analyzed gene trees obtained from TreeFam database (Li *et al.* 2006, Ruan *et al.* 2007). Phylogenetic analysis of a gene family is better suited for deducing orthology/paralogy relationships when reciprocal best hit or best hit analyses are inconclusive or ambiguous. The results from multi-species comparisons of paralogous genes revealed that the homologs of human stratum 2a and 2b genes in the chicken genome (Kohn *et al.* 2004, Ross *et al.* 2005) are paralogous genes that were misidentified as orthologous genes.

First I examined genes that flanked Stratum 2 genes in Xp11 and Xq28, belonging either to the X added region (XAR) or the X conserved region (XCR). Gene trees for the human Xp11 and Xq28 genes in the TreeFam database were searched for the orthologs of the human Xp11 and Xq28 genes in the chicken genome. The TreeFam database showed that out of 37 XAR genes, 30 genes have chicken orthologs, and therefore this region was well represented in the TreeFam database (Figure 16, supplementary table 5). 26 of these orthologs mapped to chicken chromosome 1q, in the same region where other XAR genes are found, and the remaining genes mapped to the chicken chromosomes 15, 9, and unmapped contigs (chrUn).

Similarly 34 genes out of 55 XCR genes in the Xq28 region have chicken orthologs. 31 of these chicken orthologs mapped to chicken chromosome 4p, within the XCR region to which other XCR genes mapped. The remaining orthologs map to the chicken chromosome 1 outside the XAR homologous region, and the chicken chromosome 4q outside the XCR homologous region.

I obtained very different results for genes within Stratum 2, finding that only about 10% of genes in this region have orthologs in the chicken genome. For the 106 stratum 2a genes, only 14 genes were found to have chicken orthologs; five of these orthologs mapped to chicken chromosome 4p in the XCR homologous region and three orthologs were on the unmapped contig (chrUn). The remaining homologs mapped to chicken chromosomes Z, 14, 12, and 18. The TreeFam database therefore contained no orthologs for 92 Stratum 2a

genes. Likewise, the TreeFam database contained no chicken orthologs for the 45 stratum 2b genes.

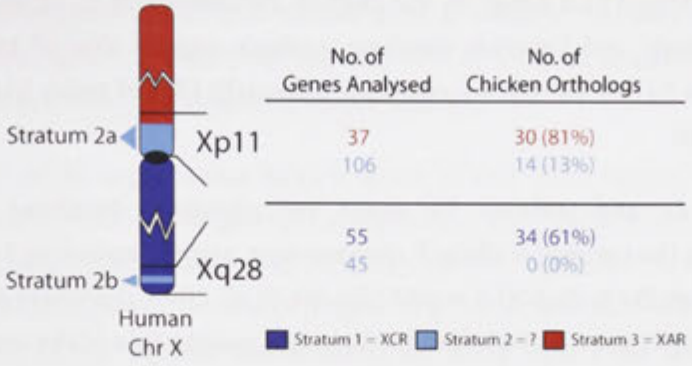


Figure 16 Schematic representation of the TreeFam database search for the human Xp11 and Xq28 orthologs in the chicken genome. 243 genes were analyzed from the Xp11 and Xq28 regions. The regions flanking stratum 2a and 2b in Xp11 (XAR, red) and in Xq28 (XCR, dark blue) are shown to have greater (>61%) number of orthologs in the chicken genome. However, only 13% of stratum 2a genes (light blue) have orthologs in the chicken genome and stratum 2b genes (light blue) do not have any orthologs identified by the TreeFam database.

The number of chicken orthologs reported for the human Xp11 and Xq28 genes in the TreeFam database were similar to the numbers reported by the Ensembl database. This is partly expected since both databases use the same genomic assembly of organisms for the annotation of orthologs. However, in the year 2008 when these datasets were analyzed, the gene tree building method used by the Ensembl database (Clamp *et al.* 2003) was different to the method used by the TreeFam database (Li *et al.* 2006). It was reassuring that, despite significant differences in the method used for building gene trees and ortholog annotations, both the Ensembl and TreeFam database show similar results for the numbers of chicken orthologs of human Xp11 and Xq28 genes.

The Ensembl and TreeFam databases results differ significantly from previous analyses (Kohn *et al.* 2004, Ross *et al.* 2005) since both these databases use phylogenetic gene tree based methods rather than the reciprocal best hit or best hit method used in previous reports. Ensembl and TreeFam database did not find as many orthologs for stratum 2 genes. In contrast, previous studies were based on BLAST searches, and therefore the reported number of orthologs may actually be paralogs and not orthologs.

3.3.6 Exploring chicken and zebrafinch EST/cDNA sequence data

The above analysis suggested the following key conclusions. Comparative analysis of paralogous regions of the human Xp11 and Xq28 genes showed that genes on chicken chromosomes 1, 12 and 26 were in fact homologous to the human chromosomes 1, 3, 7, 10

and 12 and not the human X chromosome. Conversely, the human X chromosome genes do not have orthologs on the chicken chromosomes 1, 12 and 26 as suggested by previous reports (Kohn *et al.* 2004, Ross *et al.* 2005). Also, analysis of the Ensembl and TreeFam databases confirmed that genes on the chicken chromosomes 1, 12 and 26 are paralogs. However, Ensembl and TreeFam database analysis suggest that of 151 genes from the human stratum 2a and 2b regions, only approximately 13% of genes have orthologs in the chicken genome.

The stratum 2a and stratum 2b genes are physically localized on the platypus chromosome 6, the tammar wallaby X chromosome, and the opossum X chromosome with other genes from the human XCR region (Deakin *et al.* 2008, Delbridge *et al.* 2009, Hore *et al.* 2007, Veyrunes *et al.* 2008). Also the Ensembl annotations of the anole lizard and frog assemblies suggest that stratum 2a and 2b are co-localized (discussed in section 3.1.5). What happened to stratum 2a and 2b in the chicken then? Were these regions deleted from the chicken genome or is there a possibility that these regions are absent from the current chicken genomic assembly?

Lineage specific deletions are extremely hard to figure out since genomic assembly of the local regions of interest must be accurate and complete. Therefore, first I explored the chicken/zebrafinch EST/cDNA sequence database to check whether the chicken assembly in these two regions is complete.

I used an independent data source to test the hypothesis that absence of the chicken orthologs of at least 87% genes from stratum 2a and 2b in both the TreeFam database and the Ensembl v53 database merely reflects incomplete chicken genomic assembly. Independent data sources were the chicken and zebrafinch EST/cDNA sequence databases. The chicken EST/cDNA sequence database was created from 64 cDNA libraries from 21 different adult and embryonic tissues, and therefore represented a more or less complete transcriptome profile (gene architecture) of the chicken genome (Hubbard *et al.* 2005). The chicken EST/cDNA database also represents an independent source of sequences from the genomic sequences used in the chicken assembly. This increased the chance of finding these genes if they are present in the chicken but simply missing from the assembly. Zebrafinch EST/cDNA data were also used because it would reveal whether regions corresponding to the human Xp11 and Xq28 are deleted from the whole avian lineage or just the chicken genome.

3.3.7 Reciprocal best hit search

The chicken/zebrafinch EST/cDNA database was searched using cDNA sequences for the human Xp11 and Xq28 genes (Ensembl v53) to isolate homologous chicken/zebrafinch sequences. A reciprocal search was performed against all human cDNA sequences to

confirm the reciprocal best hit relationship between sequences. Human cDNA sequences were therefore paired with their reciprocal best-hit chicken/zebrafinch EST/cDNA sequences (supplementary table 6). More than one EST/cDNA sequence was identified as the reciprocal best-hit match for each human gene because of alternative splicing of a transcript of a gene captured as distinct sequences in EST/cDNA sequencing. Reciprocal best hits are called orthologs for simplicity in this section.

As for the analysis of the regions that flank stratum 2a and 2b in human Xp11 and Xq28 respectively, the number of orthologs represented in the EST/cDNA data sets was similar to that of Ensembl and TreeFam database. 26 genes out of 37 genes in the XAR were found to have chicken/zebrafinch EST orthologs (Figure 17), and nearly all of these (24 of the 26 orthologs) mapped to the chicken chromosome 1q along with other XAR homologous genes. The other two orthologs mapped to the chicken chromosomes 5, and unmapped contig. Similarly, out of 55 XCR genes, 26 genes have chicken orthologs (Figure 17), and 24 of the 26 orthologs mapped to the chicken chromosome 4p along with other XCR homologous genes in the chicken genome. The other two orthologs mapped to chicken chromosomes 20, and unmapped contig. Chicken genes are qualified as orthologs based on the reciprocal best hit relationship in this analysis. However, the genes that map to chicken chromosomes other than 4p and 1q may not be true orthologs. This does not affect the conclusions of this analysis since the number of genes that map to chicken chromosomes 4p and 1q is significantly higher than the number of genes that do not map to chicken chromosomes 4p and 1q.

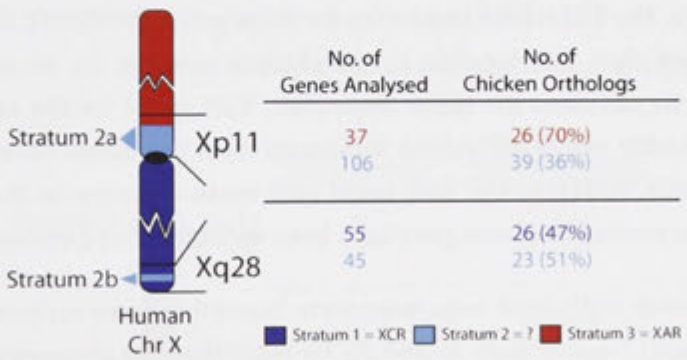


Figure 17 Schematic representation of the reciprocally related chicken/zebrafinch EST/cDNA sequences to the human Xp11 and Xq28 genes. The number of genes with reciprocal best hits from the EST/cDNA sequence database in the XAR (red) and the XCR (dark blue) is similar to the previous reports, the Ensembl database and the TreeFam database. However, the number of genes with reciprocal best hits in EST/cDNA sequence databse in the human stratum 2a and 2b (light blue) is significantly higher than the previous reports, the Ensembl database and the TreeFam database.

I found that 39 genes from stratum 2a have reciprocal best hits in the EST/cDNA sequence database (Figure 17). This number was significantly higher than the number of orthologs reported by Ensembl v53 database (14 chicken orthologs) and the TreeFam database (14

chicken orthologs). Six of these chicken orthologs mapped to chicken chromosome 4p amongst other XCR homologous genes, and six mapped to an unmapped contig (chrUn) in the chicken. EST/cDNA sequences were also found for at least 13 stratum 2a genes that did not map to any parts of the chicken genome using the specified criterion of 95% or higher pairwise sequence identity with the chicken genome.

Similarly for stratum 2b genes, whereas Ensembl v53 reports only four chicken orthologs and TreeFam database (version 7.0) reports no orthologs in the chicken genome, there were reciprocal best hits for at least 23 genes in the EST/cDNA sequence data (Figure 17). Reciprocal best hits for 17 stratum 2b genes did not map to any chromosomes in the current chicken genome assembly.

Previous reports (Kohn *et al.* 2004, Ross *et al.* 2005) suggested that chicken orthologs of most human stratum 2a and 2b genes mapped to chicken chromosomes other than chromosome 1q (the XAR homologous region) and chromosome 4p (the XCR homologous region). However, I found that only 11 genes from stratum 2a and 6 genes from stratum 2b had reciprocal best hits in the EST/cDNA sequences that mapped to regions outside the XAR and XCR homologous regions (chicken chromosomes 1, 5, 6, 10, 12, 13 and 14). These exceptional genes were not necessarily true orthologs, since reciprocal best hit searches are not an absolute measure of a one-to-one relationship, since, in the absence of the true best hit in the sequence data, the second best hit will be promoted as the best hit (discussed in section 3.1.5).

Other inconsistencies were also found which probably resulted from poor assembly of the chicken genome. The EST/cDNA sequences for three genes (*SUV39H1*, *CCNB3*, and *SMC1A*) mapped to more than one location in the chicken genome, so no conclusive genomic location could be obtained for these sequences. This could be the result of inaccurate EST/cDNA assembly where EST/cDNA sequences from two genes have been inaccurately assembled as one. Alternatively, this could also mean an error in the chicken genome assembly where contigs from one gene have been wrongly placed on two chromosomes.

Chicken/zebrafinch EST/cDNA sequences were found that were reciprocally related to 62 genes (41% genes) from stratum 2a and 2b, far more than the 20 reported by the Ensembl v53 database (13% genes) and 14 reported by the TreeFam database (9% genes). The presence of a large number of EST/cDNA sequences that were reciprocally related to the human stratum 2a and 2b genes and did ***not*** map to any part of the chicken genome, suggests that stratum 2a and 2b genes had orthologs in the chicken, but they were under-represented in the current chicken genome assembly. The reciprocally related EST/cDNA sequences that did not map to the XCR homologous region on the chicken chromosome 4p or the XAR homologous region on the chicken chromosome 1q did not cluster on chicken chromosomes 1, 12 or 26 as previously reported.

The reciprocal best hit strategy has its own demerits whereby paralogous genes can often be misidentified as orthologs, due to species-specific duplications, and deletions of genes after speciation. Neighbour-joining phylogenetic gene trees are better suited to recover orthologous/paralogous relationships in a gene family.

3.3.8 Phylogenetic analysis of the human Xp11 and Xq28 genes including chicken/zebrafinch EST/cDNA sequences

Neighbour-joining gene trees were constructed for gene families of the human Xp11 and Xq28 genes including chicken/zebrafinch EST/cDNA sequences. Branch lengths were estimated by calculating the *p*-distance (number of substitutions / total number of sites) followed by Kimura's correction (for the rate of transitions *vs* transversions); both calculated in the TreeBeST program (Li 2006). The resultant gene trees were reconciled with the known species tree to infer orthologs and paralogs using the "ortho" module of the TreeBeST program. Chicken/zebrafinch orthologs were mapped to chicken chromosomes to find their locations in the chicken genome (supplementary table 7).

Again, I first tested flanking genes from the XAR and XCR. I found that 21 genes out of 37 genes (56%) from the XAR have chicken/zebrafinch orthologs as identified by neighbour-joining phylogenetic trees (Figure 18): 19 of these genes map to chicken chromosome 1q with other XAR homologous genes and the remaining two genes map to chicken chromosome 9.

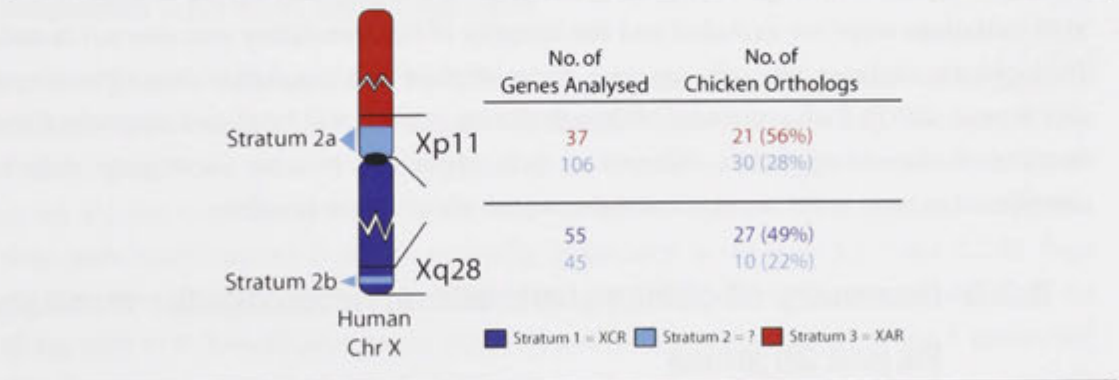


Figure 18 Schematic representation of the neighbour-joining phylogenetic tree analysis for the human Xp11 and Xq28 genes which included chicken/zebrafinch EST/cDNA sequences. Approximately 50% of the XAR (red) and the XCR (navy blue) genes were found to have chicken orthologs in the EST/cDNA database. Likewise approximately 25% of genes from stratum 2a and 2b (light blue) are also found to have chicken orthologs in the EST/cDNA database.

Similarly, for 55 XCR genes, 27 genes (49% of 55 genes) were found to have chicken/zebrafinch orthologs (Figure 18): 21 of these orthologs are on chicken

chromosome 4p with other XCR homologous genes, and three orthologs do not map to the chicken genome. The remaining orthologs map to chicken chromosomes 1, 2, 8, 9, 11 and chrUn. Since orthologs of the *ZNF275* gene (ENSG00000063587) mapped to multiple chicken chromosomes, it is likely that the *ZNF275* gene family has been expanded in the chicken lineage so a one-to-one relationship could not be established. The neighbour-joining phylogenetic gene tree analysis showed that approximately 50% of the genes from the XAR and the XCR had orthologs in the chicken/zebrafinch EST/cDNA sequences, comparable to the Ensembl and TreeFam databases. Therefore, this increased my confidence that the results obtained for stratum 2a and 2b are reliable.

Of 106 genes analyzed from stratum 2a, at least 30 genes (28%) had orthologs represented in the chicken/zebrafinch EST/cDNA sequences (Figure 18). Of these, 4 chicken orthologs mapped to the chicken chromosome 4p in the XCR homologous region and 15 chicken orthologs did not map to the chicken genome. 7 orthologs were located on the unmapped contigs. The remaining 4 orthologs mapped to chicken chromosomes 1, 4 (outside the XAR and XCR), 9 and 14. Similarly for stratum 2b, of 45 genes analyzed, at least 10 genes (22%) have orthologs in the chicken/zebrafinch EST/cDNA database (Figure 18): 7 genes did not map to any chicken chromosomes, and the remaining 3 genes mapped to chicken chromosomes 1, 8 and an unmapped contig.

Fewer chicken orthologs for Xp11 and Xq28 genes were found by the neighbour-joining phylogenetic tree analysis than the number of reciprocally related genes (section 3.3.7). Chicken/zebrafinch EST/cDNA sequences were not tested for their completeness in terms of length. Moreover, these sequences were translated in all six reading frames and the translated sequences that passed through the HMMER search filter step (section 3.2.9) were directly used in neighbour-joining analysis. The sequencing errors causing frame-shift mutations were not excluded and the integrity of the translation was also not tested. Phylogenetic analyses are very sensitive to incomplete and inaccurate data (Rosenberg and Kumar 2001). Two sequences, although closely related, will be placed distantly if the number of sites compared is different in both sequences because incomplete data is considered as gaps in the sequence and these gaps attract more penalties.

3.3.9 Summary of chicken/zebrafinch orthologs for stratum 2a and 2b genes

Two independent analyses were performed using the chicken/zebrafinch EST/cDNA sequences (which are from an independent source of sequence to the chicken genomic sequences). The results show that 47 genes from stratum 2a and 2b (of a total of 151 genes) have putative orthologs in the chicken/zebrafinch EST/cDNA sequences (Figure 19).

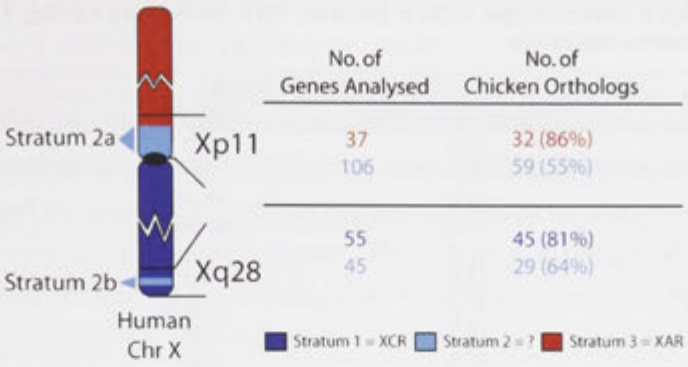


Figure 19 Schematic representation of the number of chicken orthologs for the human Xp11 and Xq28 genes in EST/cDNA sequence data. More than 80% of the XAR (red) and the XCR (dark blue) genes have orthologs represented in the EST/cDNA sequence data, which is similar to the number of orthologs suggested by the TreeFam and the Ensembl databases. The stratum 2a and 2b (light blue) genes are under-represented in the genomic sequence data and hence the TreeFam and Ensembl databases suggest significantly lower number of orthologs for these regions in the chicken. Chicken/zebrafinch EST/cDNA sequences have significantly larger proportion (approximately 58%) of orthologs for the human stratum 2a and 2b genes.

37 genes were analyzed from the XAR in human cytogenetic band Xp11. Of these, 32 genes have orthologs in the chicken/zebrafinch genome. The five genes that did not have avian orthologs were *CXorf27*, *CXorf31*, *CXorf38*, *GPR82*, and *CHST7*. Likewise, of the 55 genes from the XCR in the human Xq28, 45 genes had avian orthologs. The primary focus of the current analysis was the human stratum 2a and 2b genes. However, flanking regions (the XAR and XCR) were included in the analysis to compare the results of this analysis against the ortholog assignments by the TreeFam and Ensembl database. The results for the XAR and XCR genes are comparable to that of the ortholog assignments by the TreeFam and the Ensembl database suggesting the methods used in this work are reliable for ortholog assignments using the EST/cDNA sequences.

For 106 genes of stratum 2a, at least 59 genes have avian orthologs and for 45 genes of stratum 2b, at least 29 genes have avian orthologs (Table 7, Figure 19). This means that 58% of genes have avian orthologs. Most of the chicken orthologs of stratum 2a and 2b genes did not map to any region in the current chicken genome assembly, indicating that they were truly missing from the assembly (discussed in sections 3.3.7 and 3.3.8). Four chicken orthologs for the human stratum 2a genes mapped to chicken chromosome 4p along with XCR homologous genes, suggesting that this region is part of the X conserved region.

Table 7 Summary of the human Xp11 and Xq28 stratum 2 genes that have novel chicken orthologs. HSAX = the human X chromosome, GGA = chicken, Un = Unmapped contig, Absent = Not present in the chicken genome assembly.

Xp11 Stratum 2a genes			
Gene Name	**HSAX-Location (Mb)**	**GGA Chr**	**GGA Chr Location**
UBA1	46.9	Un	27.9
NDUFB11	46.9	Un	7.9
RBM10	46.9	Absent	*
ZNF41	47.2	Absent	*
UXT	47.4	Absent	*
WASF4	47.5	Un	35.6
ZNF81	47.6	Absent	*
SLC38A5	48.2	Un	17.2
WDR13	48.3	Un	15.8
EBP	48.3	Absent	*
PORCN	48.3	Absent	*
SUV39H1	48.4	Un	21.4
WAS	48.4	Absent	*
GLOD5	48.5	4	11.4
SLC35A2	48.6	Un	46.9
PQBP1	48.6	Absent	*
GRIPAP1	48.7	Absent	*
OTUD5	48.7	Absent	*
WDR45	48.8	Un	18.8
TFE3	48.8	Absent	*
PRICKLE3	48.9	Absent	*
SYP	48.9	Absent	*
MAGIX	48.9	Absent	*
PLP2	48.9	Absent	*
CCDC22	49.0	Absent	*
CLCN5	49.6	4	9.6
CCNB3	49.9	Absent	*
SHROOM4	50.4	4	1.8
BMP15	50.7	4	1.8
JARID1C	53.2	Un	61.7
SMC1A	53.4	Un	45.1
HUWE1	53.6	Absent	*
WNK3	54.2	Absent	*
TSR2	54.5	Absent	*
GNL3L	54.6	Absent	*
PFKFB1	55.0	Un	38.6
RRAGB	55.8	4	11.5
KLF8	56.3	Un	53.4
FAAH2	57.3	4	1.7
Xq28 Stratum 2b genes			
FAM58A	152.5	Absent	*
BCAP31	152.6	Absent	*
PDZD4	152.7	Absent	*
IDH3G	152.7	Absent	*
SSR4	152.7	Absent	*
L1CAM	152.8	Absent	*
ARD1A	152.8	Absent	*
HCFC1	152.9	Absent	*
RENBP	152.9	Absent	*
FLNA	153.2	Un	2
TAZ	153.3	Un	15.8
FAM50A	153.3	Absent	*
UBL4A	153.4	Un	52.4
FAM3A	153.4	Absent	*
G6PD	153.4	Absent	*

The XCR is well conserved on the X chromosome of all therian mammals, and its homolog, chromosome 6, in platypus. Also this arrangement of XCR genes is preserved in the lizard

and frog. This makes it unlikely that stratum 2a and 2b represent a separate evolutionary block on the human X chromosome (Kohn *et al.* 2004).

A simpler explanation for the evolution of the human X chromosome is that the human X chromosome consists simply of the X conserved region (XCR) homologous to the chicken chromosome 4p and the X added region (XAR) homologous to the chicken chromosome 1q (Figure 20), as previously proposed (Graves 1995). It is more likely that most chicken orthologs of the stratum 2 genes are missing from the chicken genome assembly.

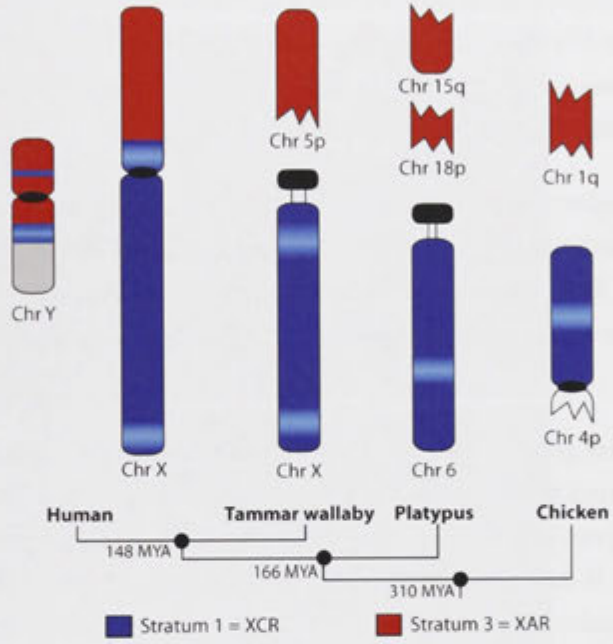


Figure 20 Schematic representation of the evolution of the human X chromosome. The X conserved region (XCR, navy blue) is conserved on the X chromosome of all therian mammals and it is autosomal in platypus and chicken. However, the X added region (XAR, red) is autosomal in marsupials, platypus and the chicken. The stratum 2a and 2b (light blue) are shown in this research to be part of the XCR and should not considered as a separate evolutionary block. The human X chromosome is composed of only two evolutionary layers, the XCR and the XAR.

3.4 Discussion: the evolution of the human X chromosome

In this chapter I report the analysis of the human Xp11 and Xq28 regions to investigate the origins of stratum 2 on the human X chromosome. Stratum 2 was claimed to have evolved independently of the XAR and XCR, and was proposed to be a separate evolutionary block on the human X chromosome (Kohn *et al.* 2004, Ross *et al.* 2005). The results presented in this chapter challenge this hypothesis and show that the human X chromosome is made up of only two evolutionary blocks, that is, the X added region (XAR) and the X conserved region (XCR).

The XAR is homologous to a region between 104 Mb to 122 Mb on the long arm of the chicken chromosome 1 (1q) (Kohn *et al.* 2004, Ross *et al.* 2005). The local gene order of homologous regions on both the chicken chromosome 1 and on the human XAR is largely maintained, with only few rearrangements. In marsupial lineages, the XAR remains intact in the potoroo and marsupial species with a 2n=14 karyotype. It has, however, undergone fission/fusion events in some marsupial lineages (Deakin *et al.* 2008, Rens *et al.* 2003). The XAR is homologous to the short arm and the pericentric regions of the long arm of the tammar wallaby chromosome 5 (Deakin *et al.* 2008). In the opossum, the XAR is homologous to two chromosomes; the pericentric region of the short arm of chromosome 4 (4p) and the pericentric region of the short and the long arm of chromosome 7 (Mikkelsen *et al.* 2007). The XAR is homologous to the telomeric region of the long arm of the platypus chromosomes 15 (15q) and a region in the short arm of the platypus chromosome 18 (18p) (Figure 20) (Veyrunes *et al.* 2008).

The X conserved region (XCR) is conserved on the X chromosome of all therian mammals (marsupials and eutherians). The proposed stratum 2 is embedded within the regions homologous to the XCR in marsupials and the platypus and hence is most likely to be the part of the XCR on the human X chromosome. The XCR is homologous to the chicken chromosome 4p.

However, genes from gene dense regions in the human cytogenetic bands Xp11 and Xq28 are under-represented in the current chicken genomic assembly. I found that many Xp11 and Xq28 orthologous genes were present in the EST/cDNA sequences but not in the genomic assembly. It has been shown that the human XCR region is highly rearranged compared to the chicken genome (Ross *et al.* 2005), and the two marsupial genomes. The gene order in the common ancestor of amniotes for the XCR remains to be elucidated.

The human cytogenetic bands Xp11 and Xq28 are the most GC rich regions of the human X chromosome and they are considerably higher in GC content than rest of the human genome as well (Costantini *et al.* 2006, Saccone *et al.* 1996). These two regions are also considerably higher in their short interspersed nuclear elements (SINE) content, lower in long interspersed nuclear elements (LINE) content and highly transcribed, from genes with shorter introns (Versteeg *et al.* 2003). The unusually high GC content and high gene density for Xp11 and Xq28 homologous regions is conserved across tetrapods (Costantini *et al.* 2009, Costantini *et al.* 2006). It is known that higher GC content of the DNA strand usually forms structures that cause polymerase synthesis to terminate prematurely (Hanvey *et al.* 1988, Samadashwily *et al.* 1993). The relatively high G+C content of the human Xp11 and the human Xq28 homologous regions could be the reason that these regions were not successfully sequenced, and could explain why they are poorly represented in the chicken genome.

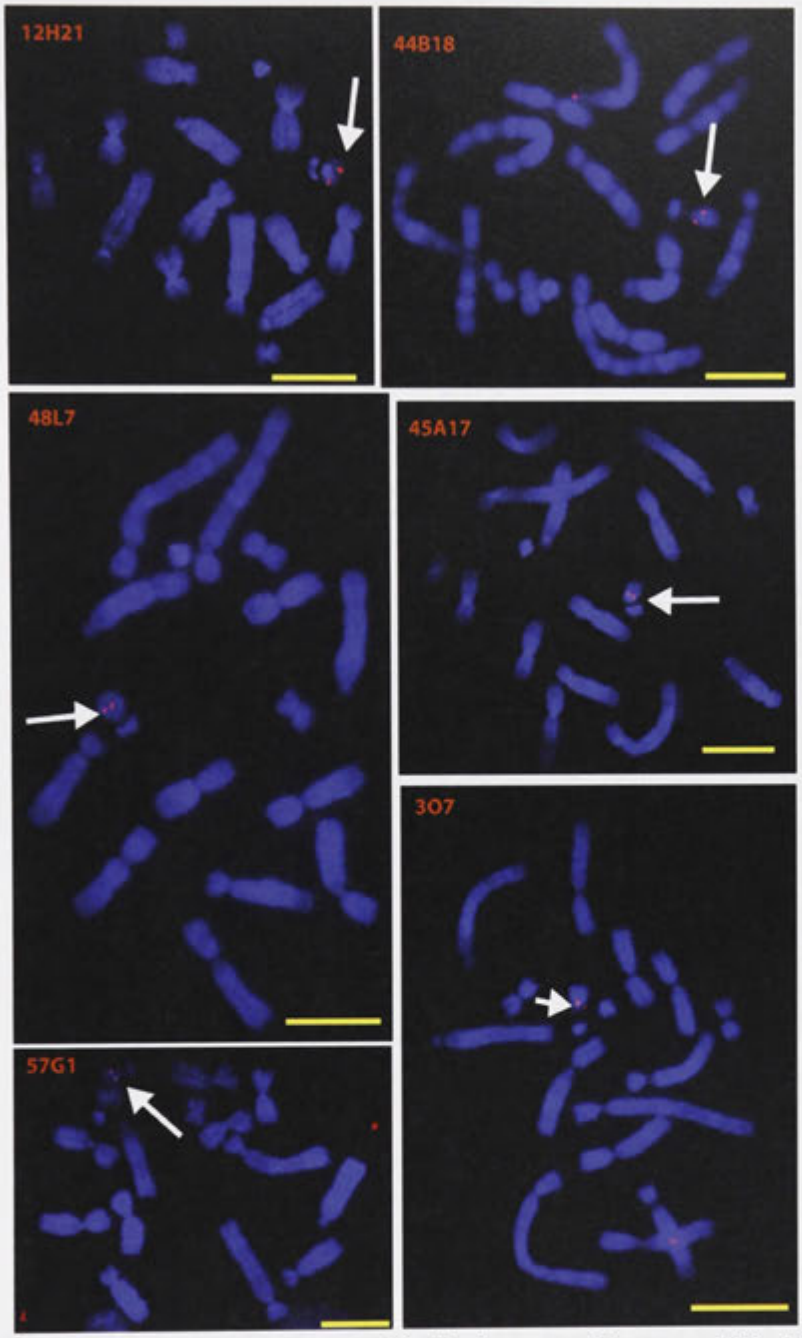


Figure 7 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.

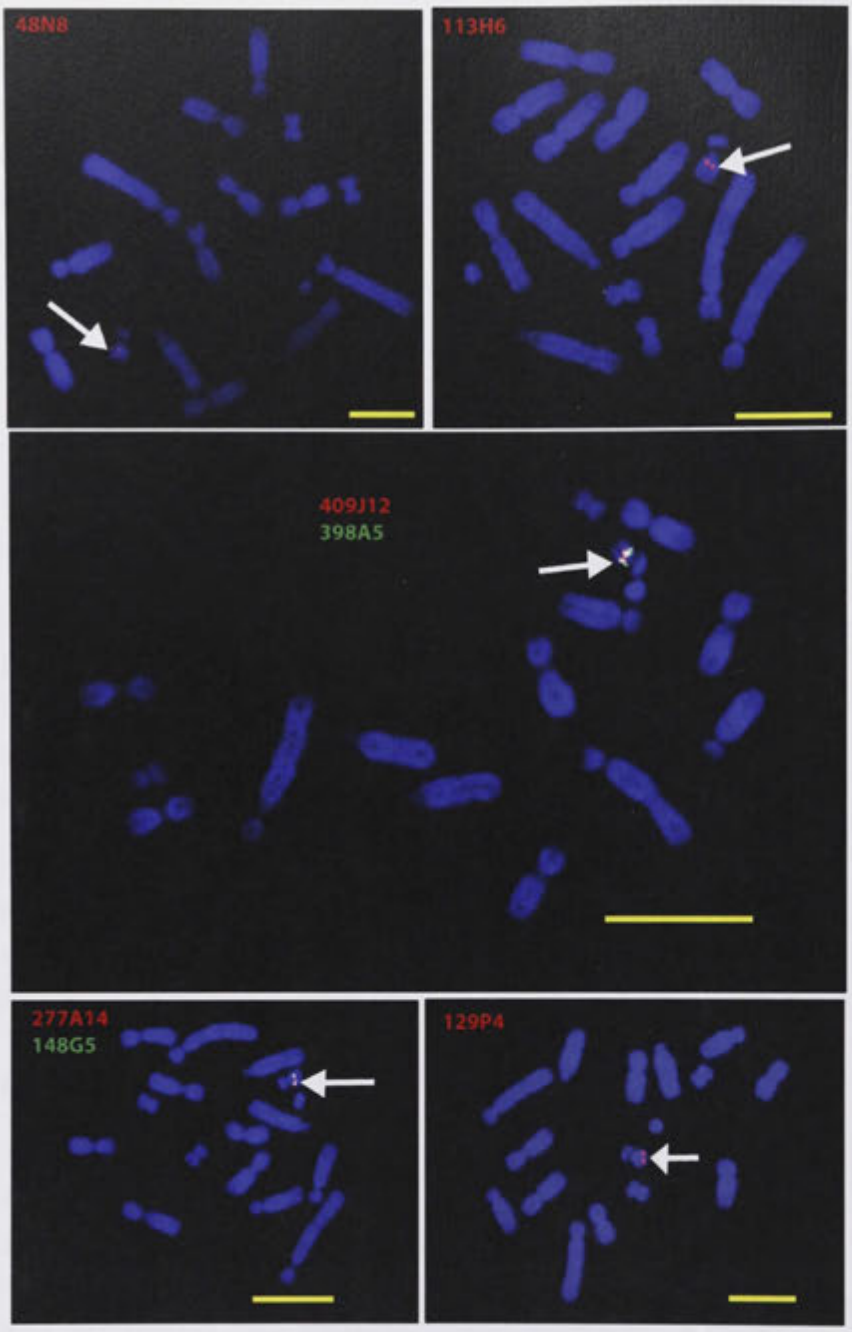
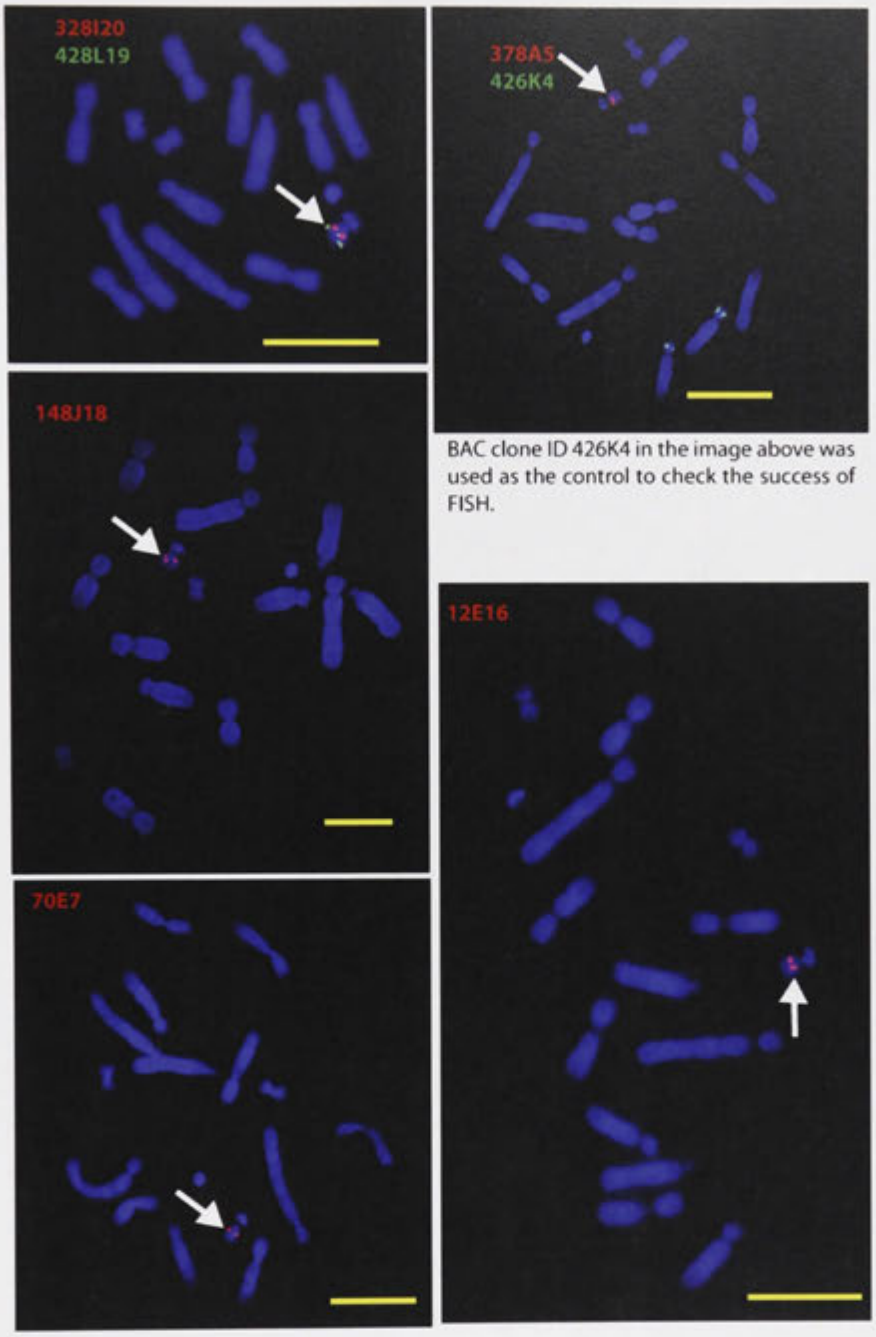


Figure 8 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.



BAC clone ID 426K4 in the image above was used as the control to check the success of FISH.

Figure 9 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.

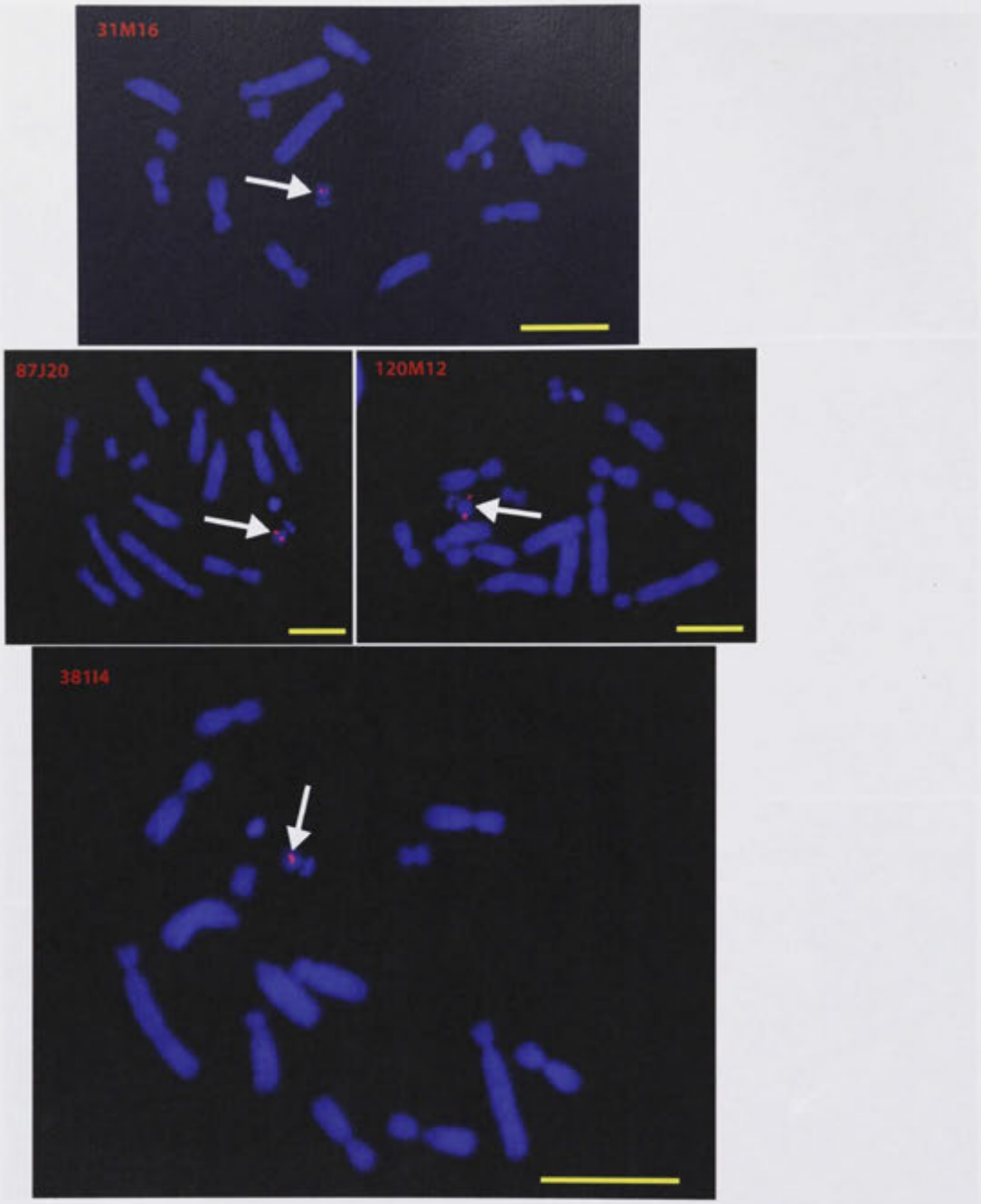


Figure 10 Fluorescent *in-situ* hybridization images of tammar wallaby Me_KBa library BAC clones identified as positives for the Xq28 genes. The scale bar in each image represents 10 μm. Male metaphase chromosomes spreads were chosen for FISH to see if any of the genes map to the Y chromosome as well.

My results were consistent with early mapping of some of the Stratum 2a and 2b genes. Four genes from stratum 2a (*GPR173, JARID1C, RIBC1* and *HUWE1*) were already known to map to the tammar wallaby X chromosome (Delbridge *et al.* 2009). Stratum 2a and 2b genes are also conserved on the X chromosome of the opossum (Ensembl v55 and Delbridge *et al.* 2009).

Therian mammals have diverged from the common ancestor with platypus 166 MYA. Physical localization of stratum 2a and 2b genes on the X chromosome of distantly related eutherians and marsupials suggests that these strata have been conserved on the X chromosome in all therian mammals. My mapping results show that if stratum 2a and 2b were added to the therian X chromosome as an independent evolutionary block, this must have occurred prior to the radiation of therian mammals, approximately 148 MYA.

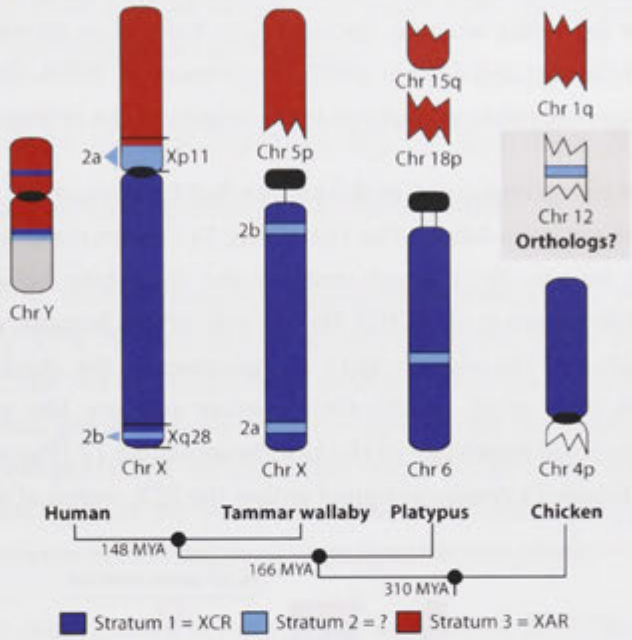


Figure 11 The schematic representation of the human Xq28 gene mapping in tammar wallaby. Stratum 2b (light blue) is present on the X chromosome of the tammar wallaby with other genes from the human XCR (dark blue) indicating that stratum 2b is part of the X-conserved region. Comparative analysis data between human and tammar wallaby were adapted from (Delbridge *et al.* 2009, Graves 1995), human and platypus from (Veyrunes *et al.* 2008) and human and chicken from (Kohn *et al.* 2004).

The hypothesis of stratum 2a and 2b comprising an independent evolutionary block on the human X chromosome, was then also tested by examining the location of these genes in the monotremes. The conserved region of the human X chromosome is conserved entirely on platypus chromosome 6 (Veyrunes *et al.* 2008). I found that Stratum 2a and 2b genes co-localize on at least two platypus contigs, including Ultra403 (0.9 Mb) and Ultra519 (9.9 Mb) that have been localised to platypus chromosome 6 by FISH. This data is consistent with at least 10 other contigs that are homologous to the conserved region of the X chromosome that have been localised on the platypus chromosome 6 (Veyrunes *et al.* 2008, Waters *et al.* 2005).

This result also suggests that stratum 2a and 2b have been part of the therian XCR and the proto-X chromosome in platypus, indicating that stratum 2a and 2b have had a similar origin to other therian XCR genes since mammals diverged from birds 310 MYA.

3.3.2 Identification of the human genes within Stratum 2a and 2b

The gene content of human cytogenetic bands Xp11 and Xq28 were then compared with the chicken, the opossum, and rat genomes. There are 99 cancer/testis antigen genes on the human X chromosome (Ross *et al.* 2005). These genes were concentrated in, but not limited to, the Xp11 and Xq28 regions. The cancer/testis antigen gene families were excluded from the following analysis because they have been recently expanded in the primate lineage (Delbridge and Graves 2007, Kouprina *et al.* 2004, Stevenson *et al.* 2007) and therefore do not contribute to analysis of the origins of the Stratum 2a and 2b genes.

There are 186 protein-coding genes in the human Xp11 region, 43 of which are members of cancer/testis antigen gene family. The remaining 143 genes make up the dataset for the Xp11 region. The human Xp11 band contains the boundary between XCR and XAR (Mikkelsen *et al.* 2007, Ross *et al.* 2005). On one side of this boundary were 37 genes that lie in the XAR region of the human Xp11 homologous to the chicken chromosome 1q (Kohn *et al.* 2004, Ross *et al.* 2005). On the other side are 106 genes that belong to stratum 2a, with reported homology to chicken chromosome 12 (Figure 12). I investigated the origin of the stratum 2a genes contained within the XCR region of the human Xp11.

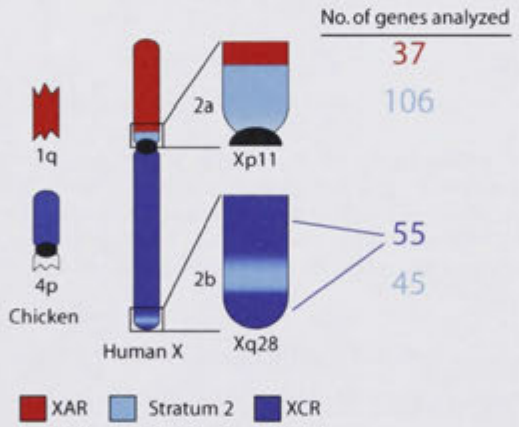


Figure 12 The schematic representation of the human X chromosome. The X-added region (XAR, red) is homologous to the chicken chromosome 1q region and the X-conserved region (XCR, navy blue) is homologous to the chicken chromosome 4p. The origin of 151 genes from the stratum 2a and the stratum 2b (light blue) is investigated by multi-spcecies comparative analysis.

Similarly, I found that human Xq28 included 121 protein-coding genes, of which 21 genes belong to the cancer/testis antigen gene family. The 100 remaining genes made up the dataset of human Xq28. The long arm of the human X chromosome (including cytogenetic band Xq28) is conserved on the X chromosome across all therian mammals (Graves 2006) and this region is largely homologous to the chicken chromosome 4p. The exception is the stratum 2b within the cytogenetic band Xq28 (Kohn *et al.* 2004). Stratum 2b consists of 45

protein coding genes, which are reported to share homology with multiple chicken chromosomes including chicken chromosomes 1, 12, 26 and other macro and micro chromosomes (Kohn *et al.* 2004). The two regions flanking stratum 2b in human Xq28 are homologous to the chicken chromosome 4p.

The human Xp11 and Xq28 dataset analyzed in this research therefore consisted of 243 genes (supplementary table 1); 143 genes from the human Xp11 region (37 genes from the XAR + 106 genes from stratum 2a) and 100 genes from the human Xq28 region (55 genes from the XCR + 45 genes from stratum 2b).

3.3.3 Orthologs of the human Xp11 and the Xq28 genes in tetrapods using Ensembl database

There were a total of 21,343 protein-coding genes annotated in the human genome (Ensembl v53). Only 63% of these human genes have orthologs in the chicken genome (Vilella *et al.* 2009). The low correspondence between the number of orthologous genes in the chicken and human genomes could be due to species-specific deletion and duplication of genes and gene divergence, but the possibility remains that some regions of the chicken genome might be missing from the assembly. A total of 243 genes (143 genes from human Xp11, and 100 genes from human Xq28) were analyzed in this study.

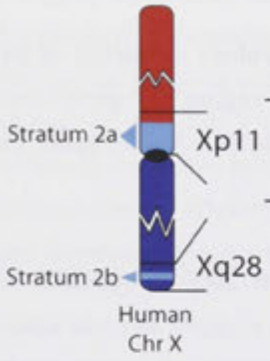
Of the 37 of the 143 genes belong to the XAR in human Xp11, 28 (75%) have orthologs in the chicken genome (Figure 13). 27 of these chicken orthologs map to chicken chromosome 1q (Ensembl annotations), consistent with previous reports (Kohn *et al.* 2004, Ross *et al.* 2005).

Only one chicken ortholog, for human gene *DUSP21*, maps to another chicken chromosome (chromosome 15). This particular chicken ortholog is related to the human *DUSP21* gene by a one-to-many relationship. A one-to-many orthologous relationship is defined for a gene when a single gene in human is orthologous to several genes in the chicken genome. This is likely to be the result of a chicken lineage specific duplication event after the divergence of the chicken lineage from the common ancestor with the human lineage.

Within Xq28 there are 55 genes in the known conserved region of the X chromosome, flanking the stratum 2b genes. Of these, 39 human genes have 38 orthologs in the chicken genome (71% of the human genes). Examining the locations of chicken orthologs using Ensembl annotations, it was revealed that 31 of the 38 chicken orthologs from the conserved region of the human Xq28 map to chicken chromosome 4p, consistent with previous reports (Kohn *et al.* 2004, Ross *et al.* 2005).

Of the seven genes that map elsewhere, one of the chicken orthologs lies on a genomic contig that has not been assigned to any chicken chromosome (chrUn). The remaining

orthologs mapped to chicken chromosomes 1 (outside the known XAR homologous region), chicken chromosome 4 (outside the known XCR homologous region) and chicken chromosome 14.



	No. of Genes Analysed	No. of Chicken Orthologs	No. of Anole Lizard Orthologs	No. of Frog Orthologs
Stratum 2a	37	28 (75%)	28 (75%)	27 (72%)
Xp11	106	16 (15%)	73 (68%)	50 (47%)
Stratum 2b	55	39 (71%)	32 (58%)	28 (50%)
Xq28	45	4 (9%)	39 (86%)	31 (68%)

Stratum 1 = XCR Stratum 2 = ? Stratum 3 = XAR

Figure 13 Schematic representation of the human Xp11 and Xq28 orthologs in chicken, lizard and frog. The data is retrieved from Ensembl database using BioMart tool.

Thus at least 70% of the genes within the XAR and XCR regions of human Xp11 and Xq28 shared orthology with the chicken genome, and most of these genes map to the respective XCR and XAR homologous regions in the chicken on chromosome 4p and 1q respectively. In contrast, only 13% of the stratum 2a and 2b genes had chicken orthologs (Figure 13, supplementary table 2). Four chicken orthologs of stratum 2a genes mapped to chicken chromosome 4p, showing that these genes are likely to be part of the homologous region of the XCR. Four other orthologs have not yet been assigned to any chicken chromosome (chrUn), so could also be located on chicken chromosome 4p. The remaining chicken homologs of stratum 2a genes map to the chicken chromosomes 1q (outside the known XAR homologous region), Z, 12, 14, and 18. However, the homologs mapping to chromosomes other than 4p or unassigned contig (chrUn) had a one-to-many correspondence with the human genes, so they may represent chicken lineage specific duplication of genes like *DUSP21* orthologs described earlier.

Similarly only four genes out of the 45 stratum 2b genes have chicken orthologs (Figure 13, supplementary table 2). Which are localized on chicken chromosomes 1, 12 and an unassigned contig. Therefore, the Ensembl database has identified no chicken orthologs for many human genes from the stratum 2a and 2b. Few genes for which the Ensembl database suggests the presence of orthologs in the chicken genome, share one-to-many orthologous relationship and hence orthologs by descent could not established unambiguously.

Both the frog and the anole lizard have more orthologs to human stratum 2a and 2b genes than does the chicken genome (Figure 13, supplementary table 2). This indicated that stratum 2a and 2b regions have been well conserved in vertebrates at least since

tetrapod/fish split 430 MYA (Blair and Hedges 2005), and therefore their absence/loss was confined to the chicken lineage.

Stratum 2a and 2b genes have been stably co-localised in distantly related species including frog, lizard and platypus (section 3.1.5). This strongly suggests that the two regions were originally located together, but became separated into two regions in the therian lineage. *My search of chicken orthologs for stratum 2a and 2b returned with at least 4 genes in the region mapping to the chicken chromosome 4p*, suggesting that stratum 2a and 2b genes, after all, were part of the conserved region of the therian X chromosome that is represented within the XCR homologous region of the chicken chromosome 4p. The absence of many stratum 2a and 2b genes from the chicken genome suggests that either these genes were lost from the chicken lineage, or are missing from the current chicken genome assembly.

The inconsistency of these results with previous reports raises two questions. First, what evidence was there to support the claims (Kohn *et al.* 2004, Ross *et al.* 2005) that these two regions form a separate evolutionary block on the human X chromosome? Second, was the region deleted from the chicken genome as a single event because stratum 2a and 2b are co-localised in frog, lizard and platypus? An analysis of the paralogous regions of the stratum 2a and 2b genes extended the analysis of the chicken homologs of these regions. This analysis was complemented by a search for orthologous genes of stratum 2a and 2b in an independent sequence data source comprised of chicken and zebrafinch EST/cDNA sequences.

3.3.4 Comparative analysis of the human Xp11 and Xq28 gene families with the rat, opossum, and the chicken genome

The Ensembl database contains multi-species comparative analysis that is mainly focused on evolution of gene families (Vilella *et al.* 2009). I used Ensembl homology assignments to analyze 171 genes from the Xp11 and Xq28 regions (including stratum 2a and 2b) with autosomal paralogs in the human genome. The 171 X-borne genes were found to have 832 autosomal paralogs scattered across all the autosomes (supplementary table 3). 92 out of the 171 genes (≈54%) have four paralogs or fewer in the human genome. This proportion is consistent with genome wide data indicating that 61% of the protein-coding genes in the human genome have four or fewer paralogs (Vilella *et al.* 2009).

Because previous reports (Kohn *et al.* 2004, Ross *et al.* 2005) claimed that the human stratum 2a and 2b regions have homology to chicken chromosomes 1, 12 and 26, these chicken chromosomes were closely examined for their homology with the human

chromosomes. The local gene content and gene order in human, rat, opossum and chicken is extremely well conserved for the human Xp11 and Xq28 paralogous regions (Figure 14).

Gene Name	Human		Rat		Opossum		Chicken	
	Gene ID	Chr 1 (Mb)	Gene ID	Chr 13 (Mb)	Gene ID	Chr 2 (Mb)	Gene ID	Chr 26 (Mb)
SOX13	ENSG00000143842	202.3	ENSRNOG00000028353	46.3	ENSMODG00000001192	108.7	ENSGALG00000000583	1.5
ETNK2	ENSG00000143845	202.4	ENSRNOG00000028368	46.3	ENSMODG00000001210	108.8	ENSGALG00000000557	1.5
KISS1	ENSG00000170498	202.4			ENSMODG00000025661	109.5		
GOLT1A	ENSG00000174567	202.4	ENSRNOG00000002936	46.2	ENSMODG00000001474	109.5		
PIK3C2B	ENSG00000133056	202.7	ENSRNOG00000029938	46.0	ENSMODG00000001583	109.9	ENSGALG00000000623	1.6
MDM4	ENSG00000198625	202.8	ENSRNOG00000009696	45.9	ENSMODG00000001605	110.0	ENSGALG00000000636	1.6
LRRN2	ENSG00000170382	202.9	ENSRNOG00000009691	45.8	ENSMODG00000024800	110.1	ENSGALG00000000639	1.6
NFASC	ENSG00000163531	203.1	ENSRNOG00000030515	45.5	ENSMODG00000001662	110.7	ENSGALG00000000644	1.7
CNTN2	ENSG00000184144	203.3	ENSRNOG00000009033	45.4	ENSMODG00000001692	110.9	ENSGALG00000000653	1.8
RBBP5	ENSG00000117222	203.3	ENSRNOG00000021289	45.4	ENSMODG00000001708	110.9	ENSGALG00000000668	1.8
TMEM81	ENSG00000174529	203.3	ENSRNOG00000028752	45.4	ENSMODG00000024792	110.9	ENSGALG00000000656	1.8
TMCC2	ENSG00000133069	203.5	ENSRNOG00000000033	45.2	ENSMODG00000001757	111.0	ENSGALG00000000673	1.9
NUAK2	ENSG00000163545	203.5	ENSRNOG00000000034	45.2	ENSMODG00000001769	111.1	ENSGALG00000000680	1.9
KLHDC8A	ENSG00000162873	203.6	ENSRNOG00000000036	45.2	ENSMODG00000001778	111.2	ENSGALG00000000684	1.9
LEMD1	ENSG00000186007	203.6	ENSRNOG00000039735	45.1	ENSMODG00000025662	111.3		
FOXP3	ENSG00000117586	203.7	ENSRNOG00000008137	45.0	ENSMODG00000001795	111.5	ENSGALG00000000692	2.0
MFSD4	ENSG00000174514	203.8	ENSRNOG00000024657	44.9	ENSMODG00000001804	111.5	ENSGALG00000000695	2.0
SLC45A3	ENSG00000158715	203.9	ENSRNOG00000007591	44.8	ENSMODG00000001824	111.7	ENSGALG00000000703	2.0
RABTL1	ENSG00000117280	204.0			ENSMODG00000001843	111.9	ENSGALG00000000712	2.1
SLC41A1	ENSG00000133065	204.0			ENSMODG00000001859	111.9	ENSGALG00000000721	2.1
PM20D1	ENSG00000162877	204.1	ENSRNOG00000039745	44.7	ENSMODG00000001870	112.0	ENSGALG00000000724	2.1
SLC26A9	ENSG00000174502	204.1	ENSRNOG00000029514	44.7	ENSMODG00000001887	112.0	ENSGALG00000000745	2.1
AVPR1B	ENSG00000198049	204.4	ENSRNOG00000006891	44.5	ENSMODG00000001960	112.3	ENSGALG00000000768	2.2
CTSE	ENSG00000196188	204.5	ENSRNOG00000006963	44.6	ENSMODG00000001939	112.2	ENSGALG00000000786	2.2
IKBKE	ENSG00000143466	204.7	ENSRNOG00000025100	44.2	ENSMODG00000002016	112.7	ENSGALG00000013356	2.3
DYRK3	ENSG00000143479	204.9	ENSRNOG00000004870	44.1	ENSMODG00000002080	112.9	ENSGALG00000000863	2.3
MAPKAPK2	ENSG00000162889	204.9	ENSRNOG00000004726	44.0	ENSMODG00000002087	113.0	ENSGALG00000000883	2.3
IL10	ENSG00000136634	205.0	ENSRNOG00000004647	44.0	ENSMODG00000002097	113.1	ENSGALG00000000892	2.4
IL24	ENSG00000162892	205.1	ENSRNOG00000004470	43.8	ENSMODG00000025663	113.4		
C1orf116	ENSG00000182795	205.3	ENSRNOG00000004341	43.7	ENSMODG00000002153	113.6	ENSGALG00000001091	2.4
YOD1	ENSG00000180667	205.3	ENSRNOG00000025704	43.7	ENSMODG00000002159	113.6	ENSGALG00000023958	2.4
PRKAG2	ENSG00000123838	205.9	ENSRNOG00000004160	43.6	ENSMODG00000002191	113.7	ENSGALG00000001117	2.4
C4BPB	ENSG00000123843	205.3	ENSRNOG00000004125	43.6	ENSMODG00000002211	113.7		
CD55	ENSG00000196352	205.6	ENSRNOG00000003927	43.3	ENSMODG00000002282	113.9	ENSGALG00000023951	2.5

Figure 14 An example of comparison of human Xp11 and Xq28 paralogous regions in rat, opossum and chicken. The flanking regions corresponding to Xp11 paralogs (red rows) and Xq28 paralogs (blue rows) in human show extremely well conserved gene content and gene order between human (yellow cloumn), rat (yellow-green column), opossum (blue-green column) and chicken (gray column). A region on human chromosome 1 (202.3 – 205.6 Mb) contains six paralogs of the human Xp11 and Xq28 genes. This region is conserved on the rat chromosome 13 (43.3 – 46.3 Mb) in reverse orientation compared to the human gene order. The same region is conserved in the same orientation on the opossum chromosome 2 (108.7 – 113.9 Mb) and the chicken chromosome 26 (1.5 – 2.5 Mb).

When the gene order and gene content of all paralogous regions were compared the rat, opossum and chicken genomes, the blocks of conserved synteny were clearly evident. The region of chicken chromosome 1 that showed homology to the human Xp11 and Xq28 regions is more closely related in gene order and gene content to human chromosomes 7, 12 and 10 (Figure 15, supplementary table 4). Similarly, the genes on the chicken chromosome 12 are conserved on the human chromosome 3 and genes on the chicken chromosome 26 are conserved on the human chromosomes 1 and 10. This indicated that the genes from the chicken chromosome 1, 12 and 26 reported as orthologs (Kohn *et al.* 2004, Ross *et al.* 2005) are in fact paralogous to the human Xp11 and Xq28 regions. None of the chicken chromosome 1, 12 or 26 has any genes that are orthologous to the human X chromosome as per Ensembl annotations.

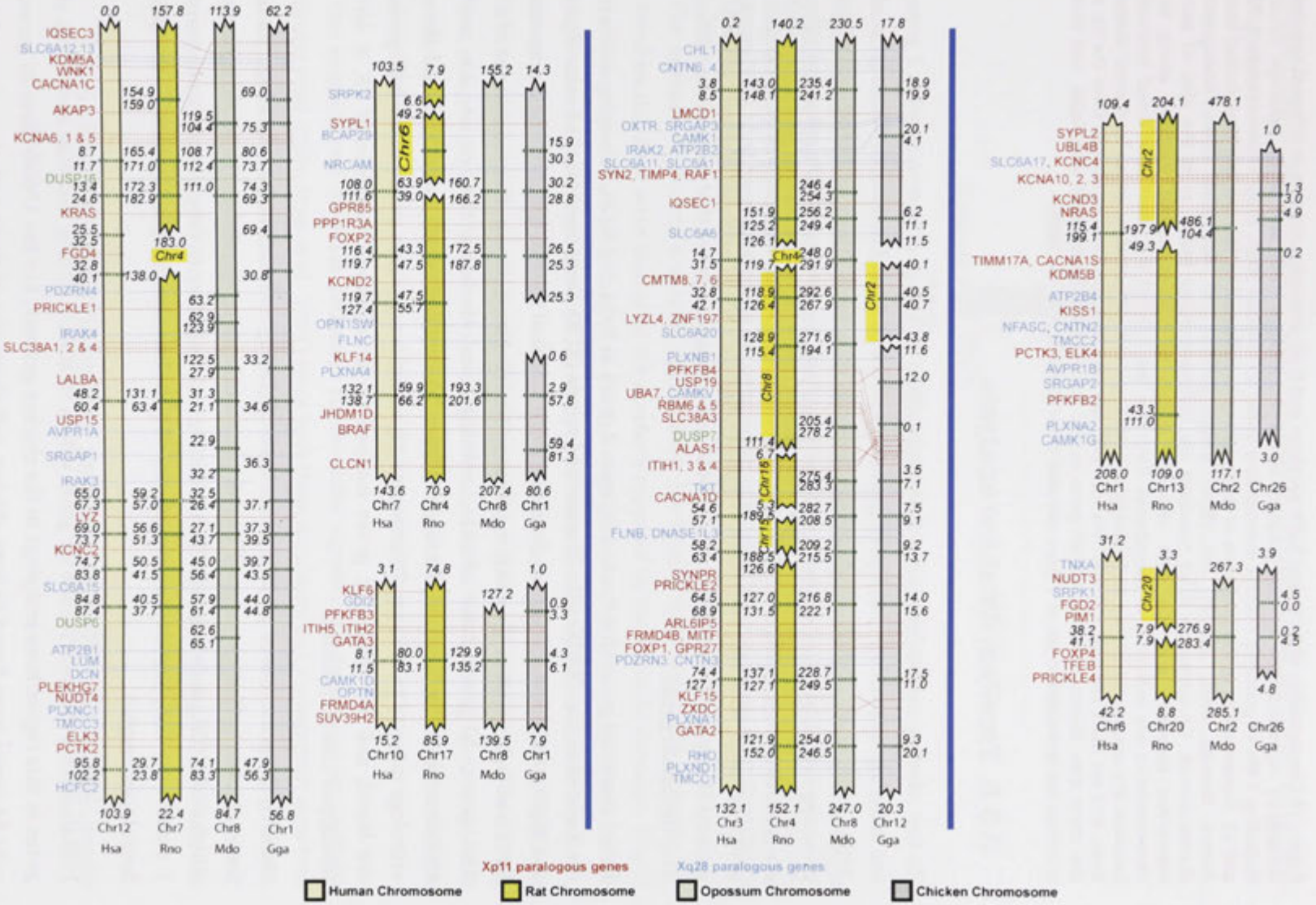

Xp11 paralogous genes
Xq28 paralogous genes
Human Chromosome
Rat Chromosome
Opossum Chromosome
Chicken Chromosome

Figure 15 Conservation of Xp11 and Xq28 paralogs and their genomic contexts in different species. Schematic representation of the location of Xp11 paralogs (red) and Xq28 paralogs (blue), including 1 Mb of genomic context surrounding each, on chicken (Gga, gray) chromosomes 1, 12, and 26. Conservation of the positions of these genes is indicated by the red and blue dotted lines across human (Hsa, yellow), rat (Rno, yellow-green), and opossum (Mdo, blue-green) chromosomes. Chromosome (Chr) numbers are indicated either below or at the side of each conserved segment, and the start and end points of the conserved sections along the chromosomes are indicated in megabases from the terminus of the short arm. Small intervals between the conserved blocks on a single chromosome are indicated by green horizontal dotted lines, and the start and end points of the intervals are also indicated in megabases from the tip of the short arm. Members of the *DUSP* gene family are found in both Xp11 and Xq28, and their paralogs are indicated in green. Figure adapted from (Delbridge *et al.* 2009).

3.3.5 TreeFam database analysis

To test whether the chicken genes claimed to be the orthologs of human stratum 2 genes are orthologs or paralogs, I analyzed gene trees obtained from TreeFam database (Li *et al.* 2006, Ruan *et al.* 2007). Phylogenetic analysis of a gene family is better suited for deducing orthology/paralogy relationships when reciprocal best hit or best hit analyses are inconclusive or ambiguous. The results from multi-species comparisons of paralogous genes revealed that the homologs of human stratum 2a and 2b genes in the chicken genome (Kohn *et al.* 2004, Ross *et al.* 2005) are paralogous genes that were misidentified as orthologous genes.

First I examined genes that flanked Stratum 2 genes in Xp11 and Xq28, belonging either to the X added region (XAR) or the X conserved region (XCR). Gene trees for the human Xp11 and Xq28 genes in the TreeFam database were searched for the orthologs of the human Xp11 and Xq28 genes in the chicken genome. The TreeFam database showed that out of 37 XAR genes, 30 genes have chicken orthologs, and therefore this region was well represented in the TreeFam database (Figure 16, supplementary table 5). 26 of these orthologs mapped to chicken chromosome 1q, in the same region where other XAR genes are found, and the remaining genes mapped to the chicken chromosomes 15, 9, and unmapped contigs (chrUn).

Similarly 34 genes out of 55 XCR genes in the Xq28 region have chicken orthologs. 31 of these chicken orthologs mapped to chicken chromosome 4p, within the XCR region to which other XCR genes mapped. The remaining orthologs map to the chicken chromosome 1 outside the XAR homologous region, and the chicken chromosome 4q outside the XCR homologous region.

I obtained very different results for genes within Stratum 2, finding that only about 10% of genes in this region have orthologs in the chicken genome. For the 106 stratum 2a genes, only 14 genes were found to have chicken orthologs; five of these orthologs mapped to chicken chromosome 4p in the XCR homologous region and three orthologs were on the unmapped contig (chrUn). The remaining homologs mapped to chicken chromosomes Z, 14, 12, and 18. The TreeFam database therefore contained no orthologs for 92 Stratum 2a

genes. Likewise, the TreeFam database contained no chicken orthologs for the 45 stratum 2b genes.

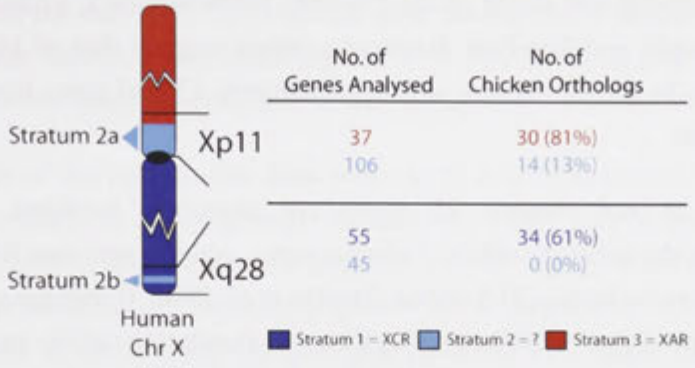


Figure 16 Schematic representation of the TreeFam database search for the human Xp11 and Xq28 orthologs in the chicken genome. 243 genes were analyzed from the Xp11 and Xq28 regions. The regions flanking stratum 2a and 2b in Xp11 (XAR, red) and in Xq28 (XCR, dark blue) are shown to have greater (>61%) number of orthologs in the chicken genome. However, only 13% of stratum 2a genes (light blue) have orthologs in the chicken genome and stratum 2b genes (light blue) do not have any orthologs identified by the TreeFam database.

The number of chicken orthologs reported for the human Xp11 and Xq28 genes in the TreeFam database were similar to the numbers reported by the Ensembl database. This is partly expected since both databases use the same genomic assembly of organisms for the annotation of orthologs. However, in the year 2008 when these datasets were analyzed, the gene tree building method used by the Ensembl database (Clamp *et al.* 2003) was different to the method used by the TreeFam database (Li *et al.* 2006). It was reassuring that, despite significant differences in the method used for building gene trees and ortholog annotations, both the Ensembl and TreeFam database show similar results for the numbers of chicken orthologs of human Xp11 and Xq28 genes.

The Ensembl and TreeFam databases results differ significantly from previous analyses (Kohn *et al.* 2004, Ross *et al.* 2005) since both these databases use phylogenetic gene tree based methods rather than the reciprocal best hit or best hit method used in previous reports. Ensembl and TreeFam database did not find as many orthologs for stratum 2 genes. In contrast, previous studies were based on BLAST searches, and therefore the reported number of orthologs may actually be paralogs and not orthologs.

3.3.6 Exploring chicken and zebrafinch EST/cDNA sequence data

The above analysis suggested the following key conclusions. Comparative analysis of paralogous regions of the human Xp11 and Xq28 genes showed that genes on chicken chromosomes 1, 12 and 26 were in fact homologous to the human chromosomes 1, 3, 7, 10

and 12 and not the human X chromosome. Conversely, the human X chromosome genes do not have orthologs on the chicken chromosomes 1, 12 and 26 as suggested by previous reports (Kohn *et al.* 2004, Ross *et al.* 2005). Also, analysis of the Ensembl and TreeFam databases confirmed that genes on the chicken chromosomes 1, 12 and 26 are paralogs. However, Ensembl and TreeFam database analysis suggest that of 151 genes from the human stratum 2a and 2b regions, only approximately 13% of genes have orthologs in the chicken genome.

The stratum 2a and stratum 2b genes are physically localized on the platypus chromosome 6, the tammar wallaby X chromosome, and the opossum X chromosome with other genes from the human XCR region (Deakin *et al.* 2008, Delbridge *et al.* 2009, Hore *et al.* 2007, Veyrunes *et al.* 2008). Also the Ensembl annotations of the anole lizard and frog assemblies suggest that stratum 2a and 2b are co-localized (discussed in section 3.1.5). What happened to stratum 2a and 2b in the chicken then? Were these regions deleted from the chicken genome or is there a possibility that these regions are absent from the current chicken genomic assembly?

Lineage specific deletions are extremely hard to figure out since genomic assembly of the local regions of interest must be accurate and complete. Therefore, first I explored the chicken/zebrafinch EST/cDNA sequence database to check whether the chicken assembly in these two regions is complete.

I used an independent data source to test the hypothesis that absence of the chicken orthologs of at least 87% genes from stratum 2a and 2b in both the TreeFam database and the Ensembl v53 database merely reflects incomplete chicken genomic assembly. Independent data sources were the chicken and zebrafinch EST/cDNA sequence databases. The chicken EST/cDNA sequence database was created from 64 cDNA libraries from 21 different adult and embryonic tissues, and therefore represented a more or less complete transcriptome profile (gene architecture) of the chicken genome (Hubbard *et al.* 2005). The chicken EST/cDNA database also represents an independent source of sequences from the genomic sequences used in the chicken assembly. This increased the chance of finding these genes if they are present in the chicken but simply missing from the assembly. Zebrafinch EST/cDNA data were also used because it would reveal whether regions corresponding to the human Xp11 and Xq28 are deleted from the whole avian lineage or just the chicken genome.

3.3.7 Reciprocal best hit search

The chicken/zebrafinch EST/cDNA database was searched using cDNA sequences for the human Xp11 and Xq28 genes (Ensembl v53) to isolate homologous chicken/zebrafinch sequences. A reciprocal search was performed against all human cDNA sequences to

confirm the reciprocal best hit relationship between sequences. Human cDNA sequences were therefore paired with their reciprocal best-hit chicken/zebrafinch EST/cDNA sequences (supplementary table 6). More than one EST/cDNA sequence was identified as the reciprocal best-hit match for each human gene because of alternative splicing of a transcript of a gene captured as distinct sequences in EST/cDNA sequencing. Reciprocal best hits are called orthologs for simplicity in this section.

As for the analysis of the regions that flank stratum 2a and 2b in human Xp11 and Xq28 respectively, the number of orthologs represented in the EST/cDNA data sets was similar to that of Ensembl and TreeFam database. 26 genes out of 37 genes in the XAR were found to have chicken/zebrafinch EST orthologs (Figure 17), and nearly all of these (24 of the 26 orthologs) mapped to the chicken chromosome 1q along with other XAR homologous genes. The other two orthologs mapped to the chicken chromosomes 5, and unmapped contig. Similarly, out of 55 XCR genes, 26 genes have chicken orthologs (Figure 17), and 24 of the 26 orthologs mapped to the chicken chromosome 4p along with other XCR homologous genes in the chicken genome. The other two orthologs mapped to chicken chromosomes 20, and unmapped contig. Chicken genes are qualified as orthologs based on the reciprocal best hit relationship in this analysis. However, the genes that map to chicken chromosomes other than 4p and 1q may not be true orthologs. This does not affect the conclusions of this analysis since the number of genes that map to chicken chromosomes 4p and 1q is significantly higher than the number of genes that do not map to chicken chromosomes 4p and 1q.

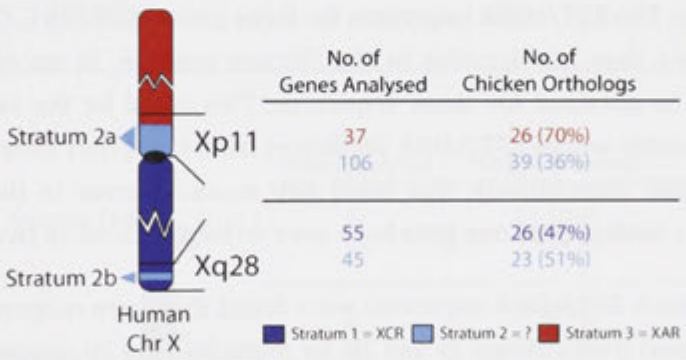


Figure 17 Schematic representation of the reciprocally related chicken/zebrafinch EST/cDNA sequences to the human Xp11 and Xq28 genes. The number of genes with reciprocal best hits from the EST/cDNA sequence database in the XAR (red) and the XCR (dark blue) is similar to the previous reports, the Ensembl database and the TreeFam database. However, the number of genes with reciprocal best hits in EST/cDNA sequence databse in the human stratum 2a and 2b (light blue) is significantly higher than the previous reports, the Ensembl database and the TreeFam database.

I found that 39 genes from stratum 2a have reciprocal best hits in the EST/cDNA sequence database (Figure 17). This number was significantly higher than the number of orthologs reported by Ensembl v53 database (14 chicken orthologs) and the TreeFam database (14

chicken orthologs). Six of these chicken orthologs mapped to chicken chromosome 4p amongst other XCR homologous genes, and six mapped to an unmapped contig (chrUn) in the chicken. EST/cDNA sequences were also found for at least 13 stratum 2a genes that did not map to any parts of the chicken genome using the specified criterion of 95% or higher pairwise sequence identity with the chicken genome.

Similarly for stratum 2b genes, whereas Ensembl v53 reports only four chicken orthologs and TreeFam database (version 7.0) reports no orthologs in the chicken genome, there were reciprocal best hits for at least 23 genes in the EST/cDNA sequence data (Figure 17). Reciprocal best hits for 17 stratum 2b genes did not map to any chromosomes in the current chicken genome assembly.

Previous reports (Kohn *et al.* 2004, Ross *et al.* 2005) suggested that chicken orthologs of most human stratum 2a and 2b genes mapped to chicken chromosomes other than chromosome 1q (the XAR homologous region) and chromosome 4p (the XCR homologous region). However, I found that only 11 genes from stratum 2a and 6 genes from stratum 2b had reciprocal best hits in the EST/cDNA sequences that mapped to regions outside the XAR and XCR homologous regions (chicken chromosomes 1, 5, 6, 10, 12, 13 and 14). These exceptional genes were not necessarily true orthologs, since reciprocal best hit searches are not an absolute measure of a one-to-one relationship, since, in the absence of the true best hit in the sequence data, the second best hit will be promoted as the best hit (discussed in section 3.1.5).

Other inconsistencies were also found which probably resulted from poor assembly of the chicken genome. The EST/cDNA sequences for three genes (*SUV39H1*, *CCNB3*, and *SMC1A*) mapped to more than one location in the chicken genome, so no conclusive genomic location could be obtained for these sequences. This could be the result of inaccurate EST/cDNA assembly where EST/cDNA sequences from two genes have been inaccurately assembled as one. Alternatively, this could also mean an error in the chicken genome assembly where contigs from one gene have been wrongly placed on two chromosomes.

Chicken/zebrafinch EST/cDNA sequences were found that were reciprocally related to 62 genes (41% genes) from stratum 2a and 2b, far more than the 20 reported by the Ensembl v53 database (13% genes) and 14 reported by the TreeFam database (9% genes). The presence of a large number of EST/cDNA sequences that were reciprocally related to the human stratum 2a and 2b genes and did ***not*** map to any part of the chicken genome, suggests that stratum 2a and 2b genes had orthologs in the chicken, but they were under-represented in the current chicken genome assembly. The reciprocally related EST/cDNA sequences that did not map to the XCR homologous region on the chicken chromosome 4p or the XAR homologous region on the chicken chromosome 1q did not cluster on chicken chromosomes 1, 12 or 26 as previously reported.

The reciprocal best hit strategy has its own demerits whereby paralogous genes can often be misidentified as orthologs, due to species-specific duplications, and deletions of genes after speciation. Neighbour-joining phylogenetic gene trees are better suited to recover orthologous/paralogous relationships in a gene family.

3.3.8 Phylogenetic analysis of the human Xp11 and Xq28 genes including chicken/zebrafinch EST/cDNA sequences

Neighbour-joining gene trees were constructed for gene families of the human Xp11 and Xq28 genes including chicken/zebrafinch EST/cDNA sequences. Branch lengths were estimated by calculating the *p*-distance (number of substitutions / total number of sites) followed by Kimura's correction (for the rate of transitions *vs* transversions); both calculated in the TreeBeST program (Li 2006). The resultant gene trees were reconciled with the known species tree to infer orthologs and paralogs using the "ortho" module of the TreeBeST program. Chicken/zebrafinch orthologs were mapped to chicken chromosomes to find their locations in the chicken genome (supplementary table 7).

Again, I first tested flanking genes from the XAR and XCR. I found that 21 genes out of 37 genes (56%) from the XAR have chicken/zebrafinch orthologs as identified by neighbour-joining phylogenetic trees (Figure 18): 19 of these genes map to chicken chromosome 1q with other XAR homologous genes and the remaining two genes map to chicken chromosome 9.

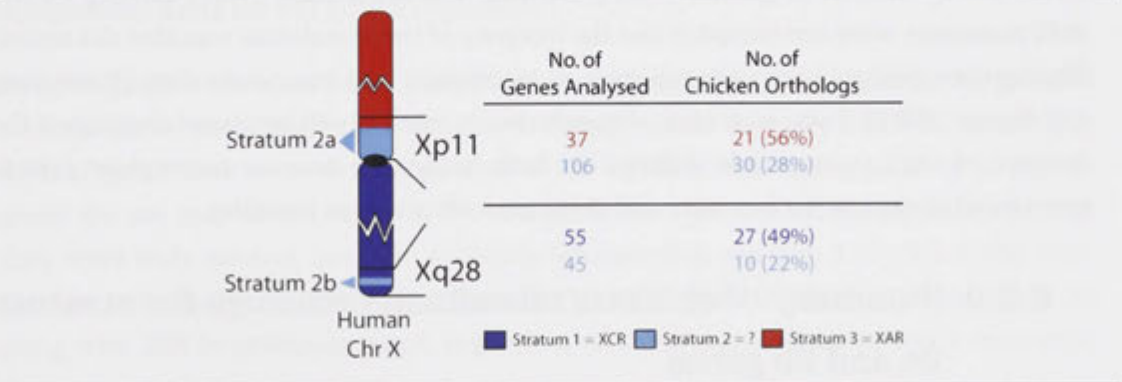


Figure 18 Schematic representation of the neighbour-joining phylogenetic tree analysis for the human Xp11 and Xq28 genes which included chicken/zebrafinch EST/cDNA sequences. Approximately 50% of the XAR (red) and the XCR (navy blue) genes were found to have chicken orthologs in the EST/cDNA database. Likewise approximately 25% of genes from stratum 2a and 2b (light blue) are also found to have chicken orthologs in the EST/cDNA database.

Similarly, for 55 XCR genes, 27 genes (49% of 55 genes) were found to have chicken/zebrafinch orthologs (Figure 18): 21 of these orthologs are on chicken

chromosome 4p with other XCR homologous genes, and three orthologs do not map to the chicken genome. The remaining orthologs map to chicken chromosomes 1, 2, 8, 9, 11 and chrUn. Since orthologs of the *ZNF275* gene (ENSG00000063587) mapped to multiple chicken chromosomes, it is likely that the *ZNF275* gene family has been expanded in the chicken lineage so a one-to-one relationship could not be established. The neighbour-joining phylogenetic gene tree analysis showed that approximately 50% of the genes from the XAR and the XCR had orthologs in the chicken/zebrafinch EST/cDNA sequences, comparable to the Ensembl and TreeFam databases. Therefore, this increased my confidence that the results obtained for stratum 2a and 2b are reliable.

Of 106 genes analyzed from stratum 2a, at least 30 genes (28%) had orthologs represented in the chicken/zebrafinch EST/cDNA sequences (Figure 18). Of these, 4 chicken orthologs mapped to the chicken chromosome 4p in the XCR homologous region and 15 chicken orthologs did not map to the chicken genome. 7 orthologs were located on the unmapped contigs. The remaining 4 orthologs mapped to chicken chromosomes 1, 4 (outside the XAR and XCR), 9 and 14. Similarly for stratum 2b, of 45 genes analyzed, at least 10 genes (22%) have orthologs in the chicken/zebrafinch EST/cDNA database (Figure 18): 7 genes did not map to any chicken chromosomes, and the remaining 3 genes mapped to chicken chromosomes 1, 8 and an unmapped contig.

Fewer chicken orthologs for Xp11 and Xq28 genes were found by the neighbour-joining phylogenetic tree analysis than the number of reciprocally related genes (section 3.3.7). Chicken/zebrafinch EST/cDNA sequences were not tested for their completeness in terms of length. Moreover, these sequences were translated in all six reading frames and the translated sequences that passed through the HMMER search filter step (section 3.2.9) were directly used in neighbour-joining analysis. The sequencing errors causing frame-shift mutations were not excluded and the integrity of the translation was also not tested. Phylogenetic analyses are very sensitive to incomplete and inaccurate data (Rosenberg and Kumar 2001). Two sequences, although closely related, will be placed distantly if the number of sites compared is different in both sequences because incomplete data is considered as gaps in the sequence and these gaps attract more penalties.

3.3.9 Summary of chicken/zebrafinch orthologs for stratum 2a and 2b genes

Two independent analyses were performed using the chicken/zebrafinch EST/cDNA sequences (which are from an independent source of sequence to the chicken genomic sequences). The results show that 47 genes from stratum 2a and 2b (of a total of 151 genes) have putative orthologs in the chicken/zebrafinch EST/cDNA sequences (Figure 19).

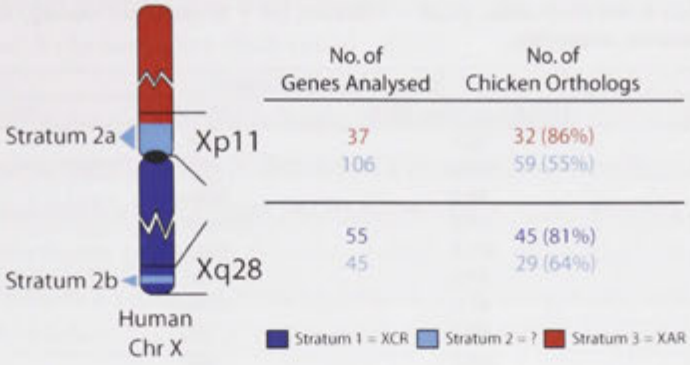


Figure 19 Schematic representation of the number of chicken orthologs for the human Xp11 and Xq28 genes in EST/cDNA sequence data. More than 80% of the XAR (red) and the XCR (dark blue) genes have orthologs represented in the EST/cDNA sequence data, which is similar to the number of orthologs suggested by the TreeFam and the Ensembl databases. The stratum 2a and 2b (light blue) genes are under-represented in the genomic sequence data and hence the TreeFam and Ensembl databases suggest significantly lower number of orthologs for these regions in the chicken. Chicken/zebrafinch EST/cDNA sequences have significantly larger proportion (approximately 58%) of orthologs for the human stratum 2a and 2b genes.

37 genes were analyzed from the XAR in human cytogenetic band Xp11. Of these, 32 genes have orthologs in the chicken/zebrafinch genome. The five genes that did not have avian orthologs were *CXorf27*, *CXorf31*, *CXorf38*, *GPR82*, and *CHST7*. Likewise, of the 55 genes from the XCR in the human Xq28, 45 genes had avian orthologs. The primary focus of the current analysis was the human stratum 2a and 2b genes. However, flanking regions (the XAR and XCR) were included in the analysis to compare the results of this analysis against the ortholog assignments by the TreeFam and Ensembl database. The results for the XAR and XCR genes are comparable to that of the ortholog assignments by the TreeFam and the Ensembl database suggesting the methods used in this work are reliable for ortholog assignments using the EST/cDNA sequences.

For 106 genes of stratum 2a, at least 59 genes have avian orthologs and for 45 genes of stratum 2b, at least 29 genes have avian orthologs (Table 7, Figure 19). This means that 58% of genes have avian orthologs. Most of the chicken orthologs of stratum 2a and 2b genes did not map to any region in the current chicken genome assembly, indicating that they were truly missing from the assembly (discussed in sections 3.3.7 and 3.3.8). Four chicken orthologs for the human stratum 2a genes mapped to chicken chromosome 4p along with XCR homologous genes, suggesting that this region is part of the X conserved region.

Table 7 Summary of the human Xp11 and Xq28 stratum 2 genes that have novel chicken orthologs. HSAX = the human X chromosome, GGA = chicken, Un = Unmapped contig, Absent = Not present in the chicken genome assembly.

Xp11 Stratum 2a genes			
Gene Name	**HSAX-Location (Mb)**	**GGA Chr**	**GGA Chr Location**
UBA1	46.9	Un	27.9
NDUFB11	46.9	Un	7.9
RBM10	46.9	Absent	*
ZNF41	47.2	Absent	*
UXT	47.4	Absent	*
WASF4	47.5	Un	35.6
ZNF81	47.6	Absent	*
SLC38A5	48.2	Un	17.2
WDR13	48.3	Un	15.8
EBP	48.3	Absent	*
PORCN	48.3	Absent	*
SUV39H1	48.4	Un	21.4
WAS	48.4	Absent	*
GLOD5	48.5	4	11.4
SLC35A2	48.6	Un	46.9
PQBP1	48.6	Absent	*
GRIPAP1	48.7	Absent	*
OTUD5	48.7	Absent	*
WDR45	48.8	Un	18.8
TFE3	48.8	Absent	*
PRICKLE3	48.9	Absent	*
SYP	48.9	Absent	*
MAGIX	48.9	Absent	*
PLP2	48.9	Absent	*
CCDC22	49.0	Absent	*
CLCN5	49.6	4	9.6
CCNB3	49.9	Absent	*
SHROOM4	50.4	4	1.8
BMP15	50.7	4	1.8
JARID1C	53.2	Un	61.7
SMC1A	53.4	Un	45.1
HUWE1	53.6	Absent	*
WNK3	54.2	Absent	*
TSR2	54.5	Absent	*
GNL3L	54.6	Absent	*
PFKFB1	55.0	Un	38.6
RRAGB	55.8	4	11.5
KLF8	56.3	Un	53.4
FAAH2	57.3	4	1.7
Xq28 Stratum 2b genes			
FAM58A	152.5	Absent	*
BCAP31	152.6	Absent	*
PDZD4	152.7	Absent	*
IDH3G	152.7	Absent	*
SSR4	152.7	Absent	*
L1CAM	152.8	Absent	*
ARD1A	152.8	Absent	*
HCFC1	152.9	Absent	*
RENBP	152.9	Absent	*
FLNA	153.2	Un	2
TAZ	153.3	Un	15.8
FAM50A	153.3	Absent	*
UBL4A	153.4	Un	52.4
FAM3A	153.4	Absent	*
G6PD	153.4	Absent	*

The XCR is well conserved on the X chromosome of all therian mammals, and its homolog, chromosome 6, in platypus. Also this arrangement of XCR genes is preserved in the lizard

and frog. This makes it unlikely that stratum 2a and 2b represent a separate evolutionary block on the human X chromosome (Kohn *et al.* 2004).

A simpler explanation for the evolution of the human X chromosome is that the human X chromosome consists simply of the X conserved region (XCR) homologous to the chicken chromosome 4p and the X added region (XAR) homologous to the chicken chromosome 1q (Figure 20), as previously proposed (Graves 1995). It is more likely that most chicken orthologs of the stratum 2 genes are missing from the chicken genome assembly.

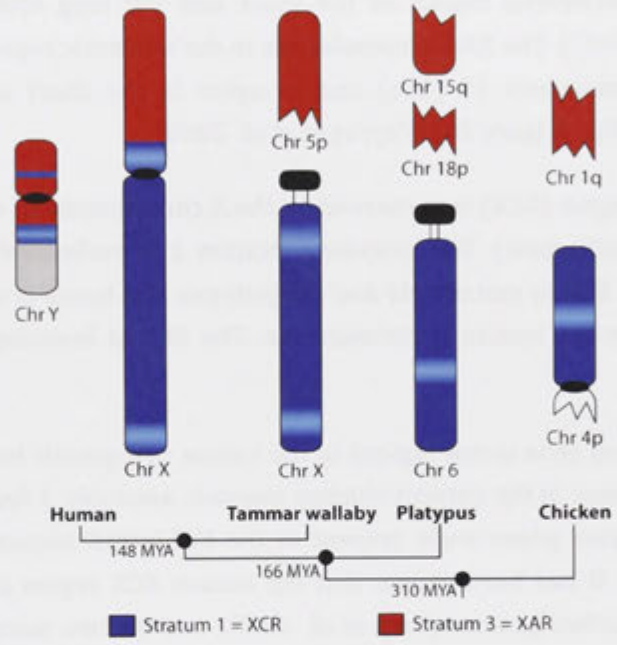


Figure 20 Schematic representation of the evolution of the human X chromosome. The X conserved region (XCR, navy blue) is conserved on the X chromosome of all therian mammals and it is autosomal in platypus and chicken. However, the X added region (XAR, red) is autosomal in marsupials, platypus and the chicken. The stratum 2a and 2b (light blue) are shown in this research to be part of the XCR and should not considered as a separate evolutionary block. The human X chromosome is composed of only two evolutionary layers, the XCR and the XAR.

3.4 Discussion: the evolution of the human X chromosome

In this chapter I report the analysis of the human Xp11 and Xq28 regions to investigate the origins of stratum 2 on the human X chromosome. Stratum 2 was claimed to have evolved independently of the XAR and XCR, and was proposed to be a separate evolutionary block on the human X chromosome (Kohn *et al.* 2004, Ross *et al.* 2005). The results presented in this chapter challenge this hypothesis and show that the human X chromosome is made up of only two evolutionary blocks, that is, the X added region (XAR) and the X conserved region (XCR).

The XAR is homologous to a region between 104 Mb to 122 Mb on the long arm of the chicken chromosome 1 (1q) (Kohn *et al.* 2004, Ross *et al.* 2005). The local gene order of homologous regions on both the chicken chromosome 1 and on the human XAR is largely maintained, with only few rearrangements. In marsupial lineages, the XAR remains intact in the potoroo and marsupial species with a 2n=14 karyotype. It has, however, undergone fission/fusion events in some marsupial lineages (Deakin *et al.* 2008, Rens *et al.* 2003). The XAR is homologous to the short arm and the pericentric regions of the long arm of the tammar wallaby chromosome 5 (Deakin *et al.* 2008). In the opossum, the XAR is homologous to two chromosomes; the pericentric region of the short arm of chromosome 4 (4p) and the pericentric region of the short and the long arm of chromosome 7 (Mikkelsen *et al.* 2007). The XAR is homologous to the telomeric region of the long arm of the platypus chromosomes 15 (15q) and a region in the short arm of the platypus chromosome 18 (18p) (Figure 20) (Veyrunes *et al.* 2008).

The X conserved region (XCR) is conserved on the X chromosome of all therian mammals (marsupials and eutherians). The proposed stratum 2 is embedded within the regions homologous to the XCR in marsupials and the platypus and hence is most likely to be the part of the XCR on the human X chromosome. The XCR is homologous to the chicken chromosome 4p.

However, genes from gene dense regions in the human cytogenetic bands Xp11 and Xq28 are under-represented in the current chicken genomic assembly. I found that many Xp11 and Xq28 orthologous genes were present in the EST/cDNA sequences but not in the genomic assembly. It has been shown that the human XCR region is highly rearranged compared to the chicken genome (Ross *et al.* 2005), and the two marsupial genomes. The gene order in the common ancestor of amniotes for the XCR remains to be elucidated.

The human cytogenetic bands Xp11 and Xq28 are the most GC rich regions of the human X chromosome and they are considerably higher in GC content than rest of the human genome as well (Costantini *et al.* 2006, Saccone *et al.* 1996). These two regions are also considerably higher in their short interspersed nuclear elements (SINE) content, lower in long interspersed nuclear elements (LINE) content and highly transcribed, from genes with shorter introns (Versteeg *et al.* 2003). The unusually high GC content and high gene density for Xp11 and Xq28 homologous regions is conserved across tetrapods (Costantini *et al.* 2009, Costantini *et al.* 2006). It is known that higher GC content of the DNA strand usually forms structures that cause polymerase synthesis to terminate prematurely (Hanvey *et al.* 1988, Samadashwily *et al.* 1993). The relatively high G+C content of the human Xp11 and the human Xq28 homologous regions could be the reason that these regions were not successfully sequenced, and could explain why they are poorly represented in the chicken genome.

The results presented here indicate the value of using multiple strategies for tracing the evolutionary history of gene families. Phylogenetic analysis combined with the syntenic conservation of the genes provides the most reliable estimates of the evolution of gene families (Vilella *et al.* 2009). Any methods involving only pair-wise comparisons should be viewed carefully, as incomplete data and lineage specific gene loss often leads results away from the true evolutionary trajectory.

4 Evolution of the olfactory receptor gene family in vertebrates

The olfactory receptor gene family is widely studied in vertebrates and invertebrates but its large size has hampered holistic analysis, at least in vertebrates. Therefore, I sought to develop a computational pipeline that automatically identifies members of this gene family in vertebrates. The olfactory gene family identification pipeline is synchronized with the Ensembl database to reflect changes as a result of newer version of the assembly of an organism or newly obtained assembly of an organism. This will aid in circumventing the procedural steps required to identify this gene family in vertebrates by all researchers seeking to use the most recent annotations of the assembly.

My second aim in this section was to develop a classification and nomenclature system for the olfactory receptor gene family. Neighbour-joining phylogenetic trees are routinely used to classify this large gene family into groups of closely related sequences. However, the presence of thousands of these olfactory receptor genes in vertebrates prohibits the usage of phylogenetic analysis, both from the computation and theoretical points of view. Therefore, I have used a clustering based algorithm that is capable of handling large amounts of data as a proxy for the neighbour-joining algorithm.

I have used the newly developed classification to understand the dynamics of the olfactory receptor gene family evolution in vertebrates, which is discussed in the following chapter.

4.1 Introduction

The sense of smell (olfaction) is remarkable in animals. Olfaction enables animals to find food, safeguard themselves from predators, find mates for reproducing, and delineate territory. The olfactory pathway in vertebrates can be categorized broadly in four major compartments (Shepherd 1972); olfactory receptors, olfactory neurons, olfactory bulb, and the lateral olfactory tract. Odor particles are recognized by olfactory receptors in the nasal epithelium, which then send signals along the bipolar neurons to the olfactory bulb. The signals are processed in the olfactory bulb and relayed by the lateral olfactory tract to higher regions of the brain for processing.

Olfactory receptor genes are members of the gene superfamily coding for G-protein coupled receptors (GPCR) (Fredriksson *et al.* 2003). GPCRs are cell membrane proteins with seven hydrophobic alpha-helical transmembrane domains connected by alternating intracellular and extracellular loops. GPCRs undergo conformational changes upon binding to extracellular stimuli (ligands) to activate heterotrimeric G-proteins (guanine nucleotide-binding proteins) (Gilman 1987, Oldham and Hamm 2008). The active G-protein undergoes conformational changes in the G_{α} and $G_{\beta\gamma}$ subunits to activate effector proteins inside the cell (Oldham and Hamm 2006). Numerous types of ligands including ions, odorants, amines, peptides, proteins, lipids, nucleotides, and photons stimulate GPCRs. The GPCR superfamily not only serves the essential function of connecting the organism with the environment through the sense of olfaction, taste and vision, but it also plays a vital role in important physiological pathways by interacting with hormones, neurotransmitters, biogenic amines and drug molecules.

GPCRs can be broadly classified into five major families in the human genome: the glutamate, rhodopsin, adhesion, frizzled/taste2 and secretin families (Fredriksson *et al.* 2003). This classification system is based on the analysis of phylogenetic trees constructed by neighbour joining and maximum parsimony methods. The rhodopsin family is the largest GPCR family with approximately 710 members; other families have fewer than 25 members in the human genome. Olfactory receptors belong to the rhodopsin family of GPCRs along with cannabinoid receptors, hormone receptors, protein receptors, nucleotide receptors and others (Horn *et al.* 2003).

4.1.1 The olfactory receptor gene family in animals

Olfaction is common to most animal species, and the olfactory receptor (OR) genes are found in both the invertebrates and the vertebrates. Amongst invertebrates, *Drosophila spp.* contain approximately 60 olfactory receptor genes (Robertson *et al.* 2003), the mosquito genome contains approximately 79 olfactory receptor genes (Hill *et al.* 2002), and the honeybee genome contains approximately 163 OR genes (Robertson and Wanner

2006). In contrast, the nematode (*C. elegans*) genome hosts approximately 1300 chemosensory receptor genes that may be involved in the sense of olfaction (Robertson and Thomas 2006). Insect olfactory receptor genes contains introns, which contrasts with the vertebrate OR genes that contain a single exon (Hildebrand and Shepherd 1997).

The vertebrate OR gene family was first discovered in the rat in 1991 (Buck and Axel 1991). Because vertebrate OR receptor genes are intronless, exon boundaries are easy to identify (the gene starts with the start codon and ends with the stop codon), hence complex intron modeling is not required (Henderson *et al.* 1997). This enables easy identification of the OR genes in vertebrate species for the comparative analysis and evolutionary studies. Novel olfactory genes have been identified in variety of vertebrates using genomic sequence data and homology searches (Glusman *et al.* 2000, Hayden *et al.* 2009, Niimura 2009, Quignon *et al.* 2005, Steiger *et al.* 2009b, Steiger *et al.* 2008, Warren *et al.* 2008). The number of OR genes in vertebrates varies from as few as 34 genes in the spotted green pufferfish, to as many as 1638 OR genes in the frog. There are approximately 1000 OR genes in most mammals (Niimura and Nei 2005b). OR genes are located in clusters on almost all chromosomes in vertebrates (Aloni *et al.* 2006). Despite the presence of nearly 1000 OR genes, the expression of OR genes is highly regulated such that only one OR gene is expressed in one neuron of the olfactory epithelium (Lomvardas *et al.* 2006).

4.1.2 Evolution of the OR gene repertoire

Comparative analysis between the human and mouse OR genes shows that the human genome has lost a large number of functional genes, and the mouse genome has gained a large number of OR genes (Niimura and Nei 2005a). The expansion of the OR gene family seems to be the result of *genomic drift* due to random duplications and deletion of genes (Niimura and Nei 2006, Nozawa *et al.* 2007). Current evidence strongly favours the birth-and-death model of evolution (Nei and Rooney 2005) that hypothesizes that OR genes arise as a result of tandem duplication events (birth process). Different copies of the duplicated genes are subsequently lost or become pseudogenes in separate lineages during evolution (death process). For example, opossums have a significantly large number of class II, clade AD genes and rodents have large number of class II, clade G genes compared to other mammals (Niimura and Nei 2007). This expansion and deletion are found only in specific sub-families of OR genes rather than the whole OR repertoire suggesting that during evolution certain sub-families are gained in a species and certain subfamilies are lost from the species. Since the effect of loss of OR genes on fitness cannot be directly tested, it is difficult to enumerate the exact evolutionary mechanisms that may have played a role in the evolution of this gene family.

However, analysis of OR genes in humans and other primates with trichromatic vision has shown that the loss of OR genes in primates correlates with the acquisition of trichromatic vision (Dong *et al.* 2009, Gilad *et al.* 2004). Similarly, the platypus leads a semi-aquatic lifestyle,has electroreception and the largest vertebrate vomeronasal type I gene repertoire that may all enable it to detect prey (Grus *et al.* 2007, Pettigrew 1999), which may explain why it has fewer OR genes. Other aquatic mammals also have considerably smaller OR repertoire compared to the land mammals (Kishida *et al.* 2007) since the sense of smell may have been replaced by other senses like echolocation in the sperm whales (Oelschlager and Kemp 1998).

Similar to the loss of OR genes, retention of functional OR genes is also driven by adaptation of the species to its environment. For example, nocturnal birds have a larger OR repertoire than their close diurnal relatives, as they rely on the sense of smell more than their day-active relatives (Steiger *et al.* 2009a). OR gene family 2/13 have been shown to be selectively retained in aquatic mammals compared to their land dwelling cousins (Hayden *et al.* 2009).

When the nucleotide diversity in the OR pseudogenes was compared with the functional OR genes, it was revealed that pseudogenes have similar nucleotide diversity as observed in intergenic regions (neutral regions) whereas functional OR genes are relatively less diverse. This suggests that there is some constraint that is preserving the integrity of the functional OR genes. It is evident that a combination of mechanisms exists for the evolution of OR gene family: one that expands this gene family (genomic drift), the second that facilitates loss of OR genes from the genome (neutral evolution) and last but not the least, one that retains certain sub-families of OR genes (adaptation).

4.1.3 Classification of vertebrate OR genes

It is essential to classify and annotate OR genes from vertebrates to understand fine scale evolutionary signatures that may be shaping the OR repertoire in a species. Neighbour-joining phylogenetic trees have been routinely used to identify monophyletic clades and hence groups of closely related OR genes. One of the early studies on the amphibian OR genes revealed that vertebrate OR genes can be classified into two major classes: class I and II (Freitag *et al.* 1995). Class I genes in amphibians were more similar to the fish OR genes and class II genes were more similar to mammalian OR genes. The functional analysis of class I and II genes in *Xenopus* showed that class I genes were specialized in detecting water soluble odorants and class II genes were activated by air-borne odorants (Mezler *et al.* 2001). Therefore, this classification not only revealed evolutionary relationship, but was also supported by functional analysis of OR genes. This classification system was further developed and formally proposed to encompass OR genes from variety of vertebrate species (Glusman *et al.* 2000). Briefly, the neighbour-joining phylogenetic

tree of OR genes was analyzed to group OR genes into families and sub-families based on monophyletic clades. The largest group of OR genes forming a monophyletic clade were grouped as a family such that all members of the family share at least 40% identity in amino acid sequence. Similarly, the largest group of OR genes forming a monophyletic clade inside the family clade are grouped together as sub-family such that all members of the sub-family share at least 60% identity in amino acid sequences. The OR sequences that did not fit the above criteria were left for manual annotation into families and sub-families.

768 OR genes from 25 species were divided into two major classes by this classification: class I (64 OR genes) and class II (704). Class I receptor genes were found in all vertebrates whereas Class II receptors were confined to tetrapods, and therefore were thought to have evolved in tetrapods since their divergence from the common ancestor with the fish (Freitag *et al.* 1995). 64 OR genes from class I were classified into 17 families and 704 class II OR genes were classified into 14 families. A systematic nomenclature system was proposed for the vertebrate OR genes based on the family and sub-family identification from the neighbour-joining phylogenetic trees (Glusman *et al.* 2000). According to their nomenclature system, a functional OR gene name would appear as OR5AB1 and an OR pseudogene name would appear as OR5AB2P, where both genes belong to family 5 and sub-family AB. All OR genes in the family number 5 will share at least 40% identity in amino acid sequence and all genes in the subfamily AB will share at least 60% identity in amino acid sequence between them.

Although class I genes were predominantly present in the fish and frog, significant number of class I OR genes were later discovered in human and mouse genomes as well (Fuchs *et al.* 2001, Malnic *et al.* 2004, Zhang and Firestein 2002). Not only class I genes were present in mammals, but a single class II gene was also discovered in fish when genomic sequences for zebrafish became available (Niimura and Nei 2005b). This finding questioned the evolutionary relationship and functional relevance of the classification by Glusman *et al.* (2000). It was also shown by detailed analysis of OR genes from 50 different mammals that individual families annotated by Glusman *et al.* (2000) were not recovered as monophyletic clades and therefore their classification system was not suitable for the annotation of the OR genes in vertebrates.

Alternatively, a new classification system was developed to address evolutionary relationship of OR genes (Niimura and Nei 2005b). According to this classification system, OR genes can be classified into two major groups: type I and type II. The type I OR genes were found in all vertebrates including the jawless vertebrate sea lamprey, which is basal to the fish (Libants *et al.* 2009, Niimura 2009). These type I OR genes were further divided into six groups (α, β, γ, δ, ε, and ζ). Groups α and γ were found predominantly in tetrapods and are represented by only a single member in zebrafish and few pseudogenes in medaka and stickleback. This suggested that α and γ genes are involved in detecting air-

borne odorant molecules, and were elaborated in air-breathing tetrapods. In contrast, groups δ, ε, and η were found only in fish and not tetrapods, which suggested that these groups of genes are involved in detecting water-soluble odorant molecules. Group β genes are found in fish and tetrapods, which suggested that these genes detect both air-borne and water-soluble odorant molecules (Niimura 2009). The sea lamprey type I genes formed a separate monophyletic clade that could not be grouped in the above six groups.

The classification by Niimura and Nei (2005a) is robust since it captures evolutionary relationship of OR genes from diverse vertebrate classes. However, one of the major limitations is that it does not offer sufficient resolution to understand the evolutionary dynamics of this gene family in vertebrates. For example, most of the tetrapod OR genes were classified as the group γ genes. This is essentially saying that all tetrapods have OR genes without any indication of expansion or contraction of groups of OR genes that may have undergone different selective pressures. It is shown that avian OR gene family have undergone large scale adaptive evolution, whereby majority of OR genes from chicken are distinct to the OR genes from the zebrafinch (Steiger *et al.* 2009b). This example strongly suggest that a comprehensive classification framework is required that can encompass OR genes from all vertebrates and simultaneously provide insight into the evolution of this large gene family at a sufficient resolution.

4.1.4 Construction of an olfactory receptor gene family database

The olfactory receptor gene family is the largest gene family in mammals, and significant numbers of genes are present also in fish, amphibians and birds. Therefore it is essential to have a centralized repository of this gene family so that researchers have easy access to all data relating to olfactory receptor genes. The need for the database is evident from the existence of at least two methods for identification of OR genes (Niimura and Nei 2007, Quignon *et al.* 2005), as well as two separate and competing methods for classifying this gene family in vertebrates. Genome sequences from a large number of vertebrates are available in the public databases, and OR genes sequences they contain need to be included into specialist OR gene databases. Few databases exist for olfactory receptor gene family, but they are either not specific to olfactory receptors (Horn *et al.* 2003, Papasaikas *et al.* 2004) or they are limited to only few species and not updated since their release (Crasto *et al.* 2002, Olender *et al.* 2004). It is also important to update the OR gene family repertoire to reflect the changes in the assembly of the genome of a species and new experimental evidence since even most recently published study used the genomic sequence data from the Ensembl release in December 2008 (Hayden *et al.* 2009).

4.1.5 Aims

There are therefore two major aims of this research. My first aim is to develop a reliable and robust data mining system that is capable of updating OR repertoires from vertebrates as and when the genomic sequence assembly of a species is first made available or is updated. My second aim is to classify olfactory receptor gene family into evolutionary related meaningful groups such that they can be used to test for evolutionary processes that shape this rather dynamic gene family.

4.2 Methods

Genomic sequence data for 45 vertebrate species were obtained from the web-server of the Ensembl database (v56) (Table 8). I used Perl scripting language to automate the data-mining process and downstream classification of the OR repertoire. I also used Bioperl modules for processing the BLAST search outputs and the FASTA search outputs (Stajich *et al.* 2002).

Table 8 List of species investigated for identifying OR repertoire.

Species name	Common Name	Abbreviation	Assembly build
Homo sapiens	Human	HUMAN	GRCh37
Pan troglodytes	Chimpanzee	PANTR	CHIMP2.1
Gorilla gorilla	Gorilla	GORGO	gorGor1
Pongo pygmaeus	Orangutan	PONPY	PPYG2
Macaca mulatta	Macaque	MACMU	MMUL 1
Callithrix jacchus	Marmoset	CALJA	calJac3
Tarsius syrichta	Tarsier	TARSY	tarSyr1
Microcebus murinus	Mouse Lemur	MICMU	micMur1
Otolemur garnettii	Bushbaby	OTOGA	BUSHBABY1
Tupaia belangeri	Tree Shrew	TUPGB	TREESHREW
Mus musculus	Mouse	MOUSE	NCBIM37
Rattus norvegicus	Rat	RAT	RGSC3.4
Dipodomys ordii	Kangaroo rat	DIPOR	dipOrd1
Spermophilus tridecemlineatus	Squirrel	SPETR	SQUIRREL
Cavia porcellus	Guinea Pig	CAVPO	cavPor3
Oryctolagus cuniculus	Rabbit	RABIT	RABBIT
Ochotona princeps	Pika	OCHPR	pika
Vicugna pacos	Alpaca	LAMVI	vicPac1
Sus scrofa	Pig	PIG	Sscrofa9
Tursiops truncatus	Dolphin	TURTR	turTru1
Bos taurus	Cow	BOVIN	Btau 4.0
Equus caballus	Horse	HORSE	EquCab2
Felis catus	Cat	FELCA	CAT
Canis familiaris	Dog	CANFA	BROADD2
Myotis lucifugus	Microbat	MYOLU	MICROBAT1
Pteropus vampyrus	Megabat	PTEVA	pteVam1
Erinaceus europaeus	Hedgehog	ERIEU	HEDGEHOG
Sorex araneus	Shrew	SORAR	COMMON SHREW1
Loxodonta africana	Elephant	LOXAF	loxAfr2
Procavia capensis	Hyrax	PROCA	proCap1
Echinops telfairi	Lesser hedgehog tenrec	ECHTE	TENREC
Dasypus novemcinctus	Armadillo	DASNO	dasNov2
Choloepus hoffmanni	Sloth	CHOHO	choHof1
Monodelphis domestica	Opossum	MONDO	BROADO5
Macropus eugenii	Tammar wallaby	MACEU	Meug 1.0
Ornithorhynchus anatinus	Platypus	ORNAN	OANA5
Gallus gallus	Chicken	CHICK	WASHUC2
Taeniopygia guttata	Zebra Finch	TAEGU	taeGut3.2.4
Anolis carolinensis	Anole Lizard	ANOCA	AnoCar1.0
Xenopus tropicalis	Xenopus tropicalis	XENTR	JGI4.1
Tetraodon nigroviridis	Tetraodon	TETNG	TETRAODON8
Takifugu rubripes	Fugu	TAKRU	FUGU4
Gasterosteus aculeatus	Stickleback	GASAC	BROADS1
Oryzias latipes	Medaka	ORYLA	MEDAKA1
Danio rerio	Zebrafish	DANRE	Zv8

4.2.1 Identification of the olfactory receptor gene family

The data mining process is similar to that previously reported with few modifications (Niimura and Nei 2007). The data mining process is summarized in the flowchart (Figure 21). Briefly, the genomic sequences of each target species were searched using functional OR gene sequences of human, dog, opossum and platypus obtained from the HORDE database (Olender *et al.* 2004) using TBLASTN program (protein queries *vs.* translated nucleotide database) (Altschul *et al.* 1990, Altschul *et al.* 1997). Target database sequence matches were filtered to retain alignments longer than 20 residues. For each sequence retained after the filtering step, up to 2,000 bp of DNA sequence was obtained from the 5'

and 3' flanking regions. The flanking regions were obtained to identify the start and stop codon. 2,000 bp flanking region was chosen to traverse through hypothetical assembly gaps that are placed in a contig or between contigs during the gap closure stages of the assembly (Jaffe *et al.* 2003).

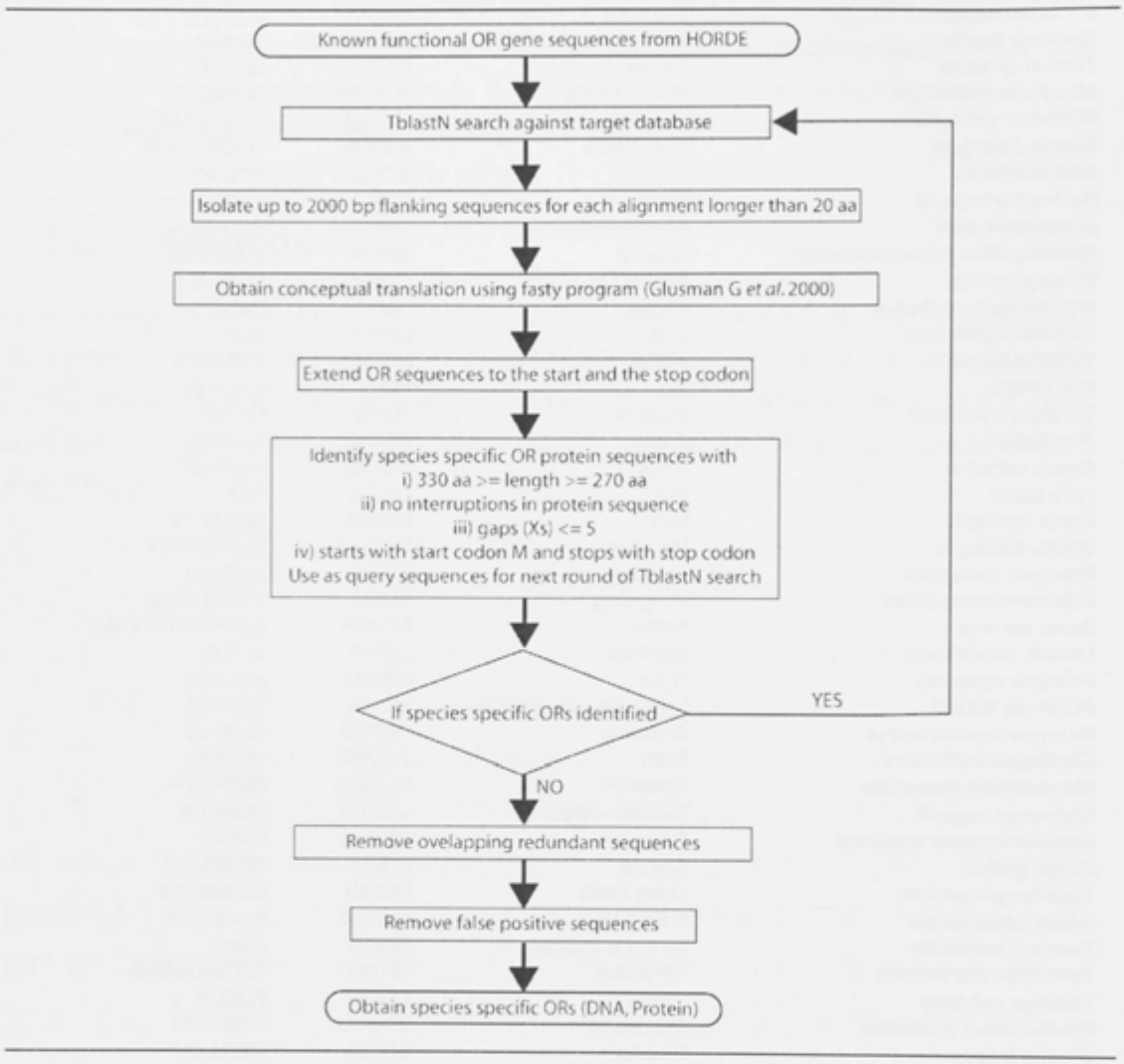


Figure 21 Flowchart depicting the data mining process used for the identification of olfactory receptor gene family in vertebrates.

Resultant OR-like sequences, which aligned with the known OR sequences for at least 20 residues, were conceptually translated to obtain OR-like protein sequences using FASTY program (-Z) (Glusman *et al.* 2000, Pearson 1994, Pearson and Lipman 1988). FASTY alignment program is used for searching protein sequence database (human, dog, opossum and platypus functional OR sequences in this case) using nucleotide sequences (OR-like sequences obtained after TBLASTN search). The FASTY program appropriately penalizes frameshift mutations and generates a single contiguous alignment unlike TBLASTN alignment in which the resultant alignments are broken into two high scoring

pairs (HSPs) if there is a frameshift mutation (Figure 22). Species-specific OR-like sequences were identified after the FASTY alignment step.

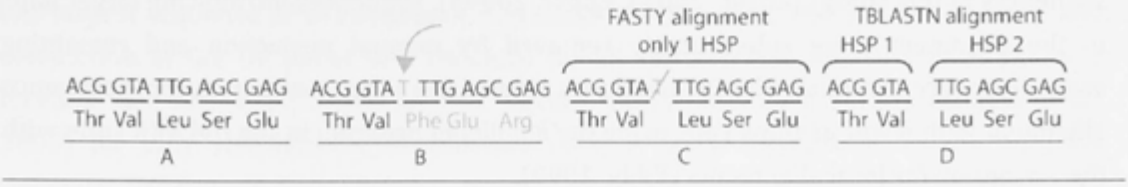


Figure 22 Diagramatic representation of the difference between FASTY and TBLASTN alignments. An OR sequence (A) when incorporates an insertional mutation (B) it causes shifting of the reading frame. (C) The FASTY program will appropriately penalize the mutation and produce single contiguous alignments, whereas, (D) the TBLASTN program will break the alignment into two high scoring pairs (HSPs).

The species-specific OR-like sequences were processed to obtain protein sequences which met the following criteria: (1) the sequence started with a start codon and ended with a stop codon, (2) the sequence contained no frame-shift mutations or in-frame stop codons (except the terminal stop codon) (3) the sequence length was greater than 270 amino acids (4) the sequence had fewer than 5 gaps represented by 'X' in the protein sequences. These gaps represented the hypothetical assembly gaps and not the alignment gaps. This step was performed to identify putative species-specific functional OR genes. Remaining sequences that did not fulfill the above four criteria were labeled as pseudogene/truncated sequences.

The genomic sequences of the target species were then searched recursively using the species-specific functional OR genes as queries for the TBLASTN program (Altschul *et al.* 1990, Altschul *et al.* 1997). Recursive TBLASTN searches were performed to identify all distantly related homologs of OR genes, until the process converged and no more OR like sequences could be identified.

All of the above steps were repeated for all 45 vertebrate species with genomic sequence data available from the Ensembl database.

4.2.2 Removal of false positive sequences

The OR sequences obtained after the data mining were processed to remove false positive sequences using hidden-Markov model (HMM) search (Eddy 1998) against a database containing HMMs for OR gene family and non-OR outgroup gene families. HMM offers greater sensitivity and specificity in distinguishing between OR protein sequences and non-OR GPCR sequences (Karchin *et al.* 2002). OR gene family HMM was obtained as follows. Protein sequences translated from functional OR genes from anole lizard, zebra finch, tammar wallaby, megabat and dog were used to construct hidden-Markov model for the OR gene family (OR-HMM). 1,221 sequences from the above five species were first

used to construct a multiple sequence alignment using the MUSCLE program (Edgar 2004a, Edgar 2004b). MUSCLE program was chosen for multiple sequence alignment because of its speed and accuracy in handling large number of sequences (1,221 sequences in this study) (Edgar 2004a, Edgar 2004b). Sequences introducing larger gaps in the alignment were subsequently removed by manual inspection and remaining sequences were realigned. The OR-HMM was created from the resultant multiple sequence alignment with fewer genuine gaps using the *hmmbuild* program in the HMMER suite with the '–s' option for local alignments (Eddy 1998).

Similarly HMMs were also obtained for the non-OR gene families. Class A rhodopsin-like G-protein coupled receptor genes were chosen as outgroup gene families because they are the closest homologs of olfactory receptors (Horn *et al.* 2003, Horn *et al.* 1998). OR genes also belong to the Class A rhodopsin like G-protein coupled receptor gene family. Protein sequences for ten non-OR gene families were obtained from GPCRDB website (Table 9) (Horn *et al.* 2003, Horn *et al.* 1998). Multiple sequence alignments and HMMs for outgroup families (outgroup-HMMs) were derived as described above for the OR-HMM.

Subsequently, OR-HMM and outgroup-HMMs were calibrated using *hmmcalibrate* program to increase search sensitivity (Eddy 1998). The database containing OR-HMM and outgroup-HMMs was searched to identify and remove false-positive OR-like sequences. Multiple sequence alignments of each outgroup gene family and olfactory receptor gene family that were used to construct HMM database are provided in the supplementary information.

Table 9 List of the Class A rhodopsin like GPCR gene families that were used as outgroup sequences for discriminating between true positive and false positive OR like sequences.

Gene family name	Average protein sequence length	Number of sequences used
Cannabinoid receptor family	438 aa	30
Gonadotropin-releasing hormone receptor family	395 aa	94
Hormone protein receptor family	707 aa	85
Leukotriene B4 receptor family	356 aa	12
Melatonin receptor family	395 aa	37
Nucleotide-like receptor family	350 aa	94
Peptide receptor family	360 aa	100
Prostanoid receptor family	398 aa	93
Thyrotropin-releasing hormone & Secretagogue receptor family	383 aa	38
Rhodopsin Vertebrate receptor family	378 aa	100

Briefly, species-specific OR-like sequences obtained after data mining were used as queries to search HMM database using the *hmmpfam* program in the HMMER suite. If a query sequence identified the OR-HMM as the best match, it was considered to be a true positive, but if any sequence identified the any of the outgroup-HMM as the best match it was discarded as a false positive sequence.

The true positive OR sequences were further processed to remove redundant sequences by comparing genomic locations of each OR gene. Two or more sequences were identified as redundant if they overlapped at the same genomic location and subsequently only the one largest sequence of all redundant sequences was retained. The genomic location and orientation of the OR genes was obtained for all species. This information of genomic location was formatted in the generic feature format version 3 (GFF3) (Eilbeck *et al.* 2005) for easy visualization as olfactory receptor gene family tract in the Ensembl database (Hammond and Birney 2004). GFF3 files are provided in supplementary information.

4.2.3 Classification of OR genes in clusters of closely related sequences

The vertebrate OR gene family was classified into the clusters of closely related sequences using pairwise similarity score as the distance measure and Markov clustering (MCL) algorithm (van Dongen 2000). The MCL algorithm performs partition clustering of sparse matrix graphs based on flow simulations (Enright *et al.* 2002, Van Dongen 2008). The result of the MCL algorithm is the natural partition clusters of closely related sequences controlled by the value of *inflation* parameter. Therefore, MCL algorithm can in effect group all the sequences with higher pairwise similarity score (for example, orthologs and inparalogs), and partition groups of sequences with low pairwise similarity scores (relatively diverged sequences).

The sparse matrix graph used as input for the MCL algorithm was created as follows. An all-*vs*-all pairwise sequence similarity search was carried out using the BLASTP program with length of database size parameter set to -z = 1,000,000,000 (Altschul *et al.* 1990, Altschul *et al.* 1997). The database length was set to 1,000,000,000 to obtain consistent e-values in the future when new sequences are added to the vertebrate OR repertoire. The output of all-*vs*-all alignments was processed using Perl script to generate a sparse matrix graph representing pairwise relationship of OR sequences from all species. Pairwise alignments longer than 50% of either the query or hit length were included in sparse matrix graph. The sparse matrix graph was obtained as follows. Consider *Sequence A* and *Sequence B* as two OR genes. When *Sequence A* was used as the query in a BLASTP search, it identified *Sequence B* as one of the matches with an e-value = E^1. Similarly when *Sequence B* was used as the query in a BLASTP search, it identified *Sequence A* as one of the matches with an e-value = E^2. The similarity score for *Sequence A* and *Sequence B* is then calculated as follows:

$$S^1 = -\log_e(E^1) \quad \text{……(1)}$$

$$S^2 = -\log_e(E^2) \quad \text{……(2)}$$

$$Average(AB) = (S^1 + S^2) / 2 \quad (3)$$

S^1 and S^2 were individual sequence similarity scores and *Average(AB)* represented the pairwise similarity score for a pair of sequences. The average score was taken into account because BLASTP searches uses heuristics to perform the search and therefore the e-value returned from a search using *Sequence A* as the query and *Sequence B* as the hit is not always the same as the e-value returned from a search using *Sequence B* as the query and *Sequence A* as the hit (Altschul *et al.* 1990, Altschul *et al.* 1997). Thus, the average of the two similarity scores gave a symmetric score for any two sequences. The similarity score obtained from the above equation for any two sequences was used to construct a sparse matrix graph. In the graph, *Sequence A* and *Sequence B* are nodes of the graph and *Average(AB)* is the weighted edge between nodes *Sequence A* and *Sequence B*.

The *inflation* parameter in the MCL algorithm dictated granularity of the clusters. At lower inflation values, the clusters were coarser and at higher inflation values, the clusters were tighter (Figure 23). Inflation values between 2.0 and 3.4 in increments of 0.2 (values between 1.2 and 5 are recommended) were used. Pruning parameters were set to *-P* 70000, *-R* 8000 and *-S* 8000 such that final "jury-synopsis" (measure of cluster significance) results were more than 95 out of a total of 100 (jury-synopsis reflects the mathematical bootstrapping applied in the MCL algorithm).

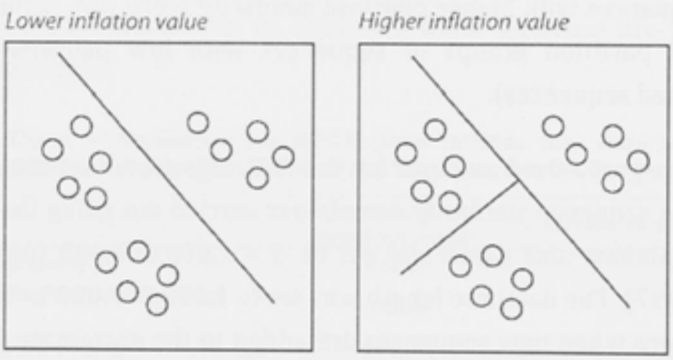


Figure 23 Schematic diagram showing the influence of inflation value on the resultant clustering in MCL algorithm. The sequences are represented as circles (nodes), whcih are connected to all the other nodes in this graph. Nodes that are closer to each other have higher similarity scores than nodes that are further. Fine grained clusters can be obtained by increasing the inflation value.

I have also used Niimura (2009) dataset in all-*vs*-all BLASTP searches and subsequent classification by the MCL algorithm to associate existing classification with the classification proposed in the present study.

4.2.4 Phylogenetic analysis of representative ORs from each cluster to estimate cluster validity

Neighbour-joining phylogenetic gene trees were constructed to investigate the relationship between the OR sequences in different clusters. Up to four functional OR genes were randomly chosen from each for phylogenetic analysis (taxon sampling; Hedtke *et al.* 2006). Multiple sequence alignments of the randomly sampled OR protein sequences were constructed using the MUSCLE program. Alignment columns with a score less than 15 were removed (maximum score per column = 100 when the residue is conserved in all sequences) (Thompson *et al.* 1997). Neighbour-joining phylogenetic trees were constructed using *p*-distance and Kimura's correction for the transition/transversion ratio using TreeBeST program with 100 bootstraps (Li 2006). The random sampling of ORs, multiple sequence alignment and neighbour-joining tree construction were repeated 100 times. A majority rule consensus tree was obtained from the resultant 100 neighbour-joining phylogenetic trees using the CONSENSE program in the PHYLIP package (Felsenstein 2005). The resultant phylogenetic trees were visualized using the Dendroscope program, which reads nexus format files to display phylogenetic trees (Huson *et al.* 2007). The entire process of random sampling, neighbour-joining and consensus tree construction was performed for clusters obtained by using varying inflation values.

4.2.5 Repetition of the data mining, classification and phylogenetic analysis steps

The data mining process, subsequent classification and phylogenetic analysis described in sections 4.2.1, 4.2.2, 4.2.3 and 4.2.4 was repeated because the FASTY program with default parameters introduced systematic errors in aligning OR-like sequences to known OR protein sequences (section 4.2.1). The FASTA program assumes that the database sequences are distant relatives of the query sequence when calculating the e-value (Pearson 1994, Pearson and Lipman 1988). OR protein sequences were highly similar to each other, so the e-value calculations were confounded when conceptual translations were sought. This resulted in no significant hits to the database sequences below the default e-value of 10, subsequently resulting in lower number of OR genes found for some species compared to previous reports. I tried setting the length of the database to 9,999,999,999 and –z 11 option for achieving accurate e-values as recommended by the author of the program (personal communication with Prof. William R. Pearson). The –z = 11 option uses pairwise alignment scores weighted by the regression of mean scores *vs* length of the library sequence for calculating e-value. However, this did not solve the e-value calculation problem. Therefore, I tested the FASTY program again by adding 1,129 sequences from the solute carrier gene family (TreeFam database ID: TF313792) to add

heterogeneity in the database used during FASTY searches. Sequence heterogeneity in the database adds contrast required to calculate e-value by reducing the proportion of number of hits identified in the database compared to the total number of sequences in the database. The addition of the solute carrier gene family sequences provided the FASTY program sufficient heterogeneity in the database and subsequently more reliable e-value calculations were obtained.

Once the problem of heterogeneous database was solved, repetition of the data mining, classification and phylogenetic analysis steps were performed. The classification of OR gene family in the first round resulted in 95 clusters that had functional OR genes from 42 species. Three sequences were randomly chosen from each of these clusters to represent diverse OR genes. 285 OR sequences (95 clusters × 3 sequences) were used as queries for searching genomic sequences of each of the 45 vertebrate species as described in section 4.2.1. These OR sequences were also used as database sequences along with the solute carrier gene family sequences for obtaining the conceptual translation as described in section 4.2.1. The OR sequences used as queries and database sequences are provided in the supplementary information.

4.2.6 Phylogenetic analysis of fish and reptile OR genes by maximum likelihood

All functional OR genes from all five species of fish were identified and corresponding protein sequences were used to construct multiple sequence alignments using the MUSCLE program. This multiple sequence alignment was used for phylogenetic analysis by maximum likelihood (Guindon *et al.* 2005). Multiple sequence alignment was submitted to PhyML website (http://www.atgc-montpellier.fr/phyml/) and phylogenetic tree for fish OR genes was obtained by using the LG substitution model (Le and Gascuel 2008), and estimating the proportion of invariable sites from the data. The branch support in the phylogenetic tree was obtained by implementing the approximate likelihood ratio test (Anisimova and Gascuel 2006).

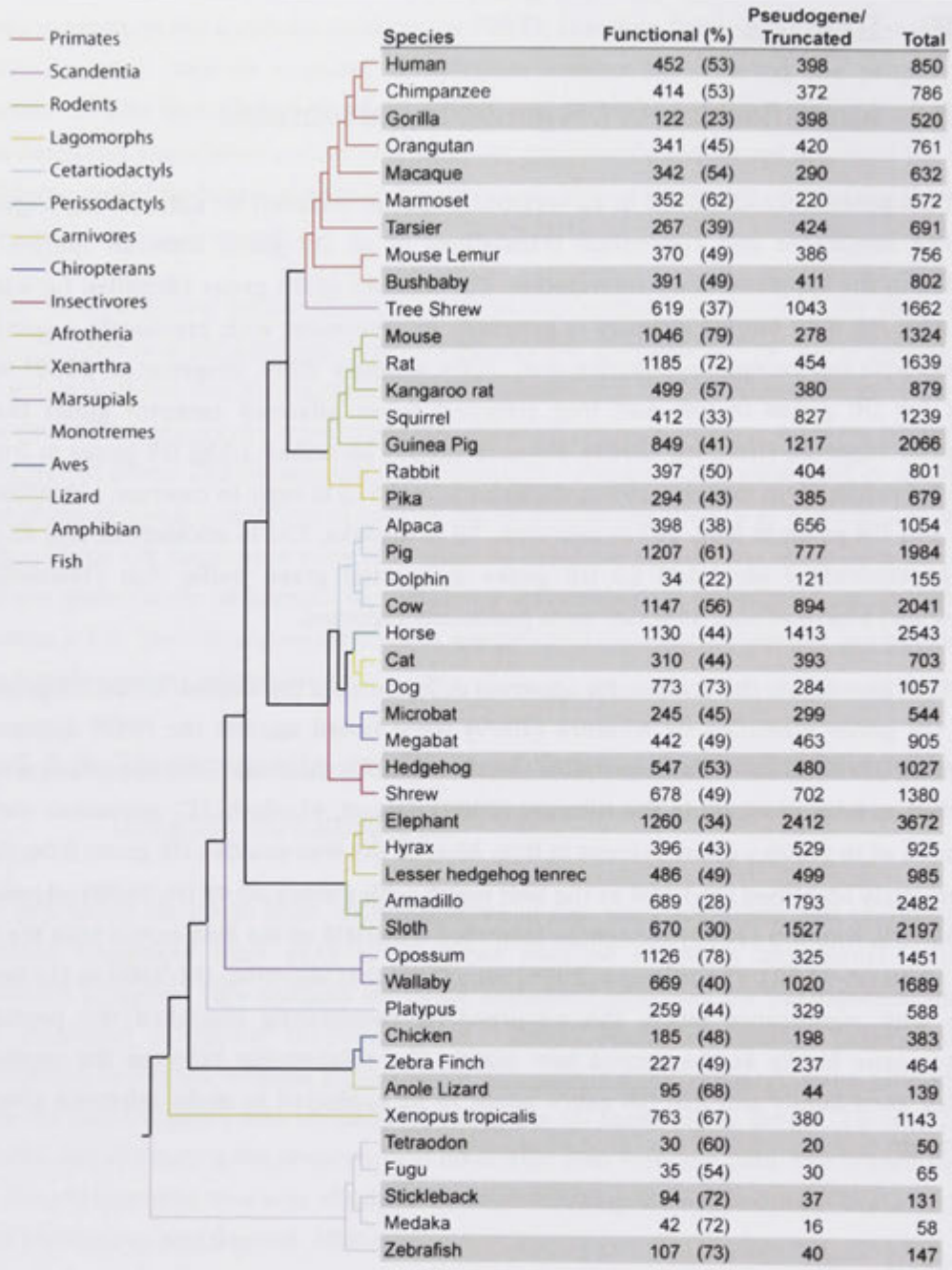
Similarly all functional OR gene sequences from chicken, zebrafinch and anole lizard were also examined by phylogenetic analysis. The same process, as used for fish OR genes, was used to construct phylogenetic tree for chicken, zebrafinch and anole lizard OR genes.

4.3 Results

4.3.1 Identification of OR genes in vertebrates

46,621 OR genes were identified in 45 vertebrate species included for data mining (Figure 24). DNA sequences and conceptual translations of all OR genes from all species is provided in the supplementary information. The number of OR genes identified for each species by my data mining strategy is generally in agreement with previously reported numbers with few exceptions (Hayden *et al.* 2009, Niimura 2009, Steiger *et al.* 2008). My search for OR genes in fish and frog returned fewer olfactory receptor genes than previously reported (Niimura 2009). Niimura (2009) identified 1,638 OR genes in frog, 176 in zebrafish, 98 in medaka, 159 in stickleback, and 125 in fugu. In contrast, I identified only 1,141 OR genes in frog, 147 in zebrafish, 58 in medaka, 131 in stickleback, and 65 in fugu. In contrast, I identified 50 OR genes in spotted green puffer fish (*Tetraodon nigroviridis*) where as only 34 genes were previously reported.

To further investigate the reasons for apparent differences in the number of the OR genes, 2,921 OR genes identified by Niimura (2009) were tested against the HMM database containing OR-HMM and outgroup-HMMs (section 4.2.2 for methods). 366 sequences were identified as false positives in the Niimura (2009) dataset, of which 317 sequences were annotated as to group γ genes present in frog. Also, all the true positive OR genes from the present study identified OR-HMM as the best match with z-score >3.48 (P>99.98) whereas only 95.40% Niimura (2009) sequences identified OR-HMM as the best match with the z-score >3.48 (P>99.98). The Niimura (2009) sequences that identified OR-HMM as the best match with significance below the z-score=3.48, consistently identified the peptide receptor gene family as the second best match. The relationship between the peptide receptor gene family and fish OR genes needs to be evaluated to make inference about their origin.



Species	Functional (%)	Pseudogene/ Truncated	Total
Human	452 (53)	398	850
Chimpanzee	414 (53)	372	786
Gorilla	122 (23)	398	520
Orangutan	341 (45)	420	761
Macaque	342 (54)	290	632
Marmoset	352 (62)	220	572
Tarsier	267 (39)	424	691
Mouse Lemur	370 (49)	386	756
Bushbaby	391 (49)	411	802
Tree Shrew	619 (37)	1043	1662
Mouse	1046 (79)	278	1324
Rat	1185 (72)	454	1639
Kangaroo rat	499 (57)	380	879
Squirrel	412 (33)	827	1239
Guinea Pig	849 (41)	1217	2066
Rabbit	397 (50)	404	801
Pika	294 (43)	385	679
Alpaca	398 (38)	656	1054
Pig	1207 (61)	777	1984
Dolphin	34 (22)	121	155
Cow	1147 (56)	894	2041
Horse	1130 (44)	1413	2543
Cat	310 (44)	393	703
Dog	773 (73)	284	1057
Microbat	245 (45)	299	544
Megabat	442 (49)	463	905
Hedgehog	547 (53)	480	1027
Shrew	678 (49)	702	1380
Elephant	1260 (34)	2412	3672
Hyrax	396 (43)	529	925
Lesser hedgehog tenrec	486 (49)	499	985
Armadillo	689 (28)	1793	2482
Sloth	670 (30)	1527	2197
Opossum	1126 (78)	325	1451
Wallaby	669 (40)	1020	1689
Platypus	259 (44)	329	588
Chicken	185 (48)	198	383
Zebra Finch	227 (49)	237	464
Anole Lizard	95 (68)	44	139
Xenopus tropicalis	763 (67)	380	1143
Tetraodon	30 (60)	20	50
Fugu	35 (54)	30	65
Stickleback	94 (72)	37	131
Medaka	42 (72)	16	58
Zebrafish	107 (73)	40	147

Figure 24 Phylogenetic species tree showing the relationship between vertebrate species used in this analysis. The number of functional OR genes (% of total), the number of pseudogenes/truncated OR genes and the number of total OR genes in each species examined is also listed next to the name of the species.

The results presented here suggest that OR-like sequences identified during data mining should be tested for false positives. HMM profile search is more effective in filtering out false positives compared to phylogenetic tree based method since more outgroup gene families can be used without compromising sensitivity. Niimura (2009) has shown that genes from groups θ1, θ2, κ, and λ are atypical to other OR genes and may have been

misidentified as ORs during their search for OR genes. At least 21 genes from groups θ1, θ2, κ, and λ were identified as false positives when searched against HMM database used in this study.

4.3.2 Classification of OR genes in clusters of closely related sequences

The classification of OR genes into groups of closely related sequences enables elucidation of the evolutionary history of the gene family in vertebrates and identification of mechanisms that shape the OR repertoire in individual species. I have used pairwise alignment scores as a distance measure and MCL algorithm to identify clusters of closely related OR genes. The resultant clusters represents groups of sequences that are most closely related to each other in a group than they are to any other OR gene in any other group. The granularity of clusters in the MCL algorithm is controlled by the *inflation* parameter. I have tested *inflation* values ranging from 2.0 to 3.4 in the increments of 0.2 (Table 10). Higher inflation value results in fine-grained clusters and relatively more number of clusters.

Table 10 Number of clusters obtained as a result of the inflation value used.

Inflation value	2.0	2.2	2.4	2.6	2.8	3.0	3.2	3.4
Number of clusters	97	101	105	109	112	123	130	140

Inflation values higher than 2.8 resulted in some clusters with only one OR gene in them. The formation of a cluster with only one OR gene in it suggests that the OR gene is not related to any other OR gene from any other cluster. However, it seems more reasonable to assume that each OR gene is related to at least one other OR gene. Therefore, the clusters obtained using inflation values 3.0, 3.2 and 3.4 were not favourable. All sets of clusters obtained as a result of varying inflation values were tested to see if each individual cluster is recovered as a monophyletic clade in the neighbour-joining phylogenetic trees.

4.3.3 Phylogenetic analysis of representative ORs from each cluster to estimate cluster validity

There were 22,396 functional OR genes identified in 45 vertebrate species. Traditionally OR genes were classified by phylogenetic analysis. However, phylogenetic analysis including all 22,396 functional OR genes is computationally impractical. Therefore, I used the MCL algorithm for classification of OR genes. The validity of the classification by the MCL algorithm was tested by the neighbour-joining phylogenetic analysis of randomly sampled functional OR genes from each cluster. If the MCL algorithm is comparable to

neighbour-joining phylogenetic trees in identifying closely related OR genes, then each cluster should be recovered as a monophyletic clade.

Different sets of clusters were obtained for different inflation values (Table 10). A set of clusters obtained for a particular inflation value was chosen for phylogenetic analysis. For example, 97 clusters obtained as a result of the inflation value equal to 2.0 were tested first as described in the method section 4.2.4. Up to four functional OR genes (some clusters may have less than four functional genes) were randomly chosen (100 times) and neighbour-joining phylogenetic trees were constructed. A majority rule consensus tree was then constructed using 100 neighbour-joining phylogenetic trees. The consensus tree was manually inspected to identify the number of clusters that were recovered as monophyletic clade. The whole process was then repeated for 101 clusters obtained as a result of inflation value equal to 2.2 and so on.

If OR genes from the same cluster forms monophyletic clade in all 100 neighbour-joining trees, then they will be recovered as monophyletic clade in the consensus tree as well with the node support for the monophyly equal to 100. The consensus trees for each inflation value were manually inspected. It was revealed in the consensus trees that not all clusters could be recovered as monophyletic clades (Table 11). Based on this phylogenetic analysis, clustering obtained by using 2.2 as the inflation value, results in classification of OR genes such that only 15.84% of the clusters are not recovered as monophyletic clades. This was lower than for all the other inflation values and therefore results of inflation value of 2.2 were considered significant and used for further analysis.

Table 11 Phylogenetic analysis *vs* varying inflation values used in the MCL algorithm.

Inflation value	2.0	**2.2**	2.4	2.6	2.8	3.0	3.2	3.4
Number of clusters	97	**101**	105	109	112	123	130	140
Non-monophyletic clusters (%)	17 (17.52)	**16 (15.84)**	18 (17.14)	18 (16.51)	23 (20.53)	26 (21.13)	25 (19.23)	28 (20.00)

Closer inspection of the majority rule consensus tree for the inflation value 2.2 reveals that of all the clusters recovered as monophyletic clades, were monophyletic in more than 50% randomly sampled neighbour-joining phylogenetic trees: 82% of were monophyletic in more than 80 trees and 18% were monophyletic in 50-80% trees (Figure 25). The distance measure used in the neighbour-joining algorithm was the *p*-distance with Kimura's correction (section 4.2.4) whereas the distance measure used for the clustering was the –log(e-value). Inherent differences in the *p*-distance and the e-value calculations could be the cause for certain clusters being not recovered as monophyletic clades. However, it is comforting to note that large proportion (85%) of clusters are recovered as monophyletic clades which has formed the basis of OR gene family classification in the past (Glusman *et al.* 2000, Niimura 2009). It would be most appropriate to classify OR genes using the MCL algorithm and –log(e-value) as the distance measure to cope up with the ever increasing number of genomic sequences available.

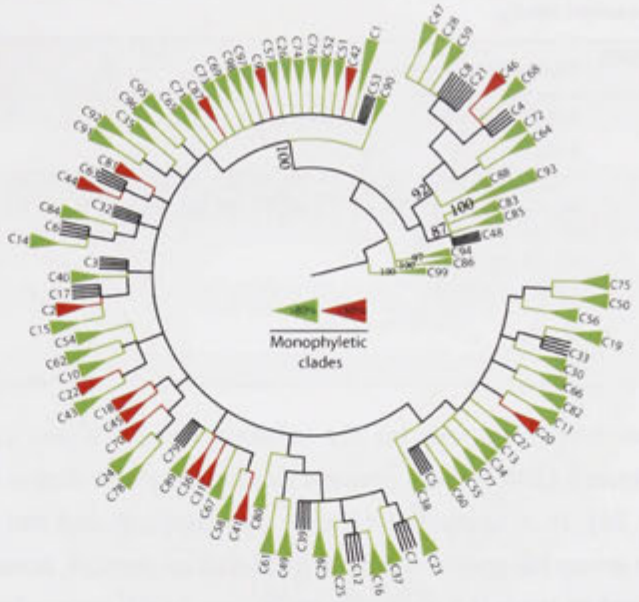


Figure 25 Majority rule consensus tree derived from 100 randomly sampled neighbour-joining phylogenetic trees representing vertebrate OR genes. OR gene clusters obtained by using 2.2 inflation value are represented with a prefix 'C' followed by identification number. The OR gene clusters recovered as monophyletic clades are collapsed and represented as green triangle if they were recovered as monophyletic clades in 80 or more neighbour-joining trees and red triangles if they were recovered in less than 80 trees. Black branches represent clusters that were not recovered as monophyletic clades.

4.3.4 OR gene family annotation transfer

There are two existing annotation systems for the OR gene family (Glusman *et al.* 2000, Niimura and Nei 2005b). I have used 2,555 OR gene sequences annotated according to the Niimura and Nei (2005) system and 1,956 sequences annotated according to the Glusman *et al.* (2000) system for the transfer of annotations from the old systems to the classification and annotation system presented here. I used previously annotated OR sequences with the OR sequences identified by me in all-*vs*-all BLASTP searches and subsequent classification by the MCL algorithm as described in section 4.2.3.

Niimura and Nei (2005) classified vertebrate OR genes into two major groups; type I and type II. Type I OR genes were further classified into six subgroups (α, β, γ, δ, ε, and ζ) and type II OR genes into five subgroups (η, θ1, θ2, κ, and λ) (Niimura 2009). The identity of subgroups θ1, θ2, κ, and λ as OR genes is questionable (Niimura 2009) and therefore not included for the annotation transfer. My classification system can divide the subgroups annotated by Niimura and Nei (2005) into one or more clusters of closely related to OR genes (Table 12).

Table 12 Relationship between Niimura and Nei (2005) classification and the classification presented in the present study.

Niimura and Nei (2005) classification	Cluster IDs associated with each group
Group α	4, 8, 21, 46, 59
Group β	8, 88
Group γ	1, 3, 5, 6, 7, 9, 11, 12, 14, 18, 20, 22, 24, 25, 26, 27, 30, 32, 33, 35, 38, 39, 41, 42, 43, 44, 45, 49, 51, 52, 53, 55, 56, 57, 60, 61, 63, 66, 69, 65, 70, 71, 74, 80, 81, 87, 90, 91, 92, 95, 96
Group δ	48, 83, 85
Group ε	94
Group ζ	86, 99
Group η	93

The majority rule consensus tree for the inflation value 2.2 also confirmed that all the subgroups of Niimura (2009) were present as monophyletic clades with significant node support (Figure 26). It is important to note that Niimura and Nei (2005) classification system is able to group OR genes into closely related sequences, however, it does not offer resolution required to trace the evolutionary history of OR genes. For example, according to Niimura (2009) almost all mammalian OR genes (more than 44,000) will be classified in the group γ. This is informative as far as concluding that mammalian OR genes are significantly different than fish OR genes. But it does not tell anything about species specific traits of evolutionary trajectories for OR genes. I have designed a classification system in which species-specific expansions and contractions of OR gene family can be studied to explore the evolutionary paths that might be shaping this gene family in vertebrates.

Apart from the classification used by Niimura and Nei (2005), another classification and annotation system exists for the OR gene family in vertebrates (Glusman *et al.* 2000). This classification system divides OR genes into 17 families: four class I families thought to be involved in detecting water-soluble odorant molecules and remaining class II families to be involved in detecting air-borne molecules. These families have been categorized into more than 200 sub-families. The basis of this classification system was also the monophyly of the closely related OR genes. However, recent analysis of the OR gene family encompassing 50 mammalian species revealed that not all families form monophyletic clades (Hayden *et al.* 2009), hence undermining the use of such classification system. Therefore I would recommend that this classification system is not suitable and should not be used for the OR gene family.

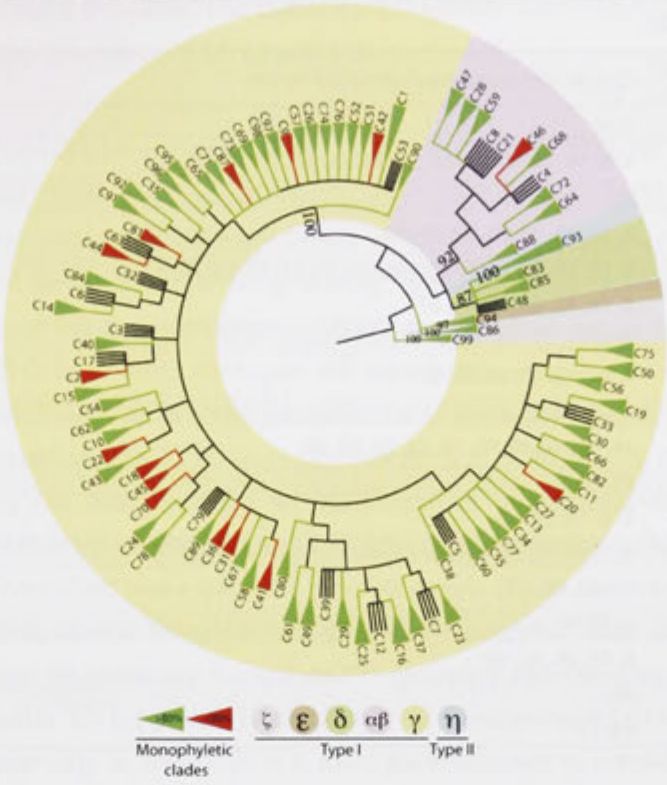


Figure 26 Majority rule consensus tree derived from 100 randomly sampled neighbour-joining phylogenetic trees representing vertebrate OR genes. OR gene clusters obtained by using 2.2 inflation value are represented with a prefix 'C' followed by identification number. The OR gene clusters recovered as monophyletic clades are collapsed and represented as green triangle if they were recovered as monophyletic clades in 80 or more neighbour-joining trees and red triangles if they were recovered in less than 80 trees. Black branches represent clusters that were not recovered as monophyletic clades.The highlighted regions show that my classification system is able to recapitulate the previous classification system (Niimura and Nei 2005a). For example, all group γ OR genes (yellow highlight) were present as monophyletic clade in all 100 random trees. Similarly other groups were also recovered with significant node support.

The analyses of 1,956 sequences, which were annotated according to Glusman et al. (2000) system, were assigned a cluster ID (Table 13). It is noted that each family corresponds to one or more clusters. The monophyly of families annotated according to Glusman *et al.* (2000) system could not be recovered in the phylogenetic analysis and hence it is not presented here. As an example, cluster 1 is recovered as monophyletic clade in the phylogenetic analysis. This cluster contains OR genes from both family 2 and 13 indicating that family 2 and 13 are actually not independently monophyletic and there is some overlap between family 2 and 13. It has been shown by independent analysis that family 2 and 13 can not be recovered as monophyletic clades individually since they overlap in the phylogenetic tree (Hayden *et al.* 2009). Cluster 1 is one of many examples undermining Glusman *et al.* (2000) classification where monophyly of each family is highly questionable and therefore it is strongly recommended that this classification system should be avoided to make inferences about the evolution of the OR gene family in vertebrates.

Table 13 Relationship between Glusman *et al.* (2000) classification and the classification presented in the present study.

Glusman et al. (2000) family classification	Cluster IDs associated with each family
1	3, 40
2	1, 9, 18, 31, 36, 41, 52, 53, 58, 65, 67, 73, 79, 89, 90
3	76
4	7, 12, 16, 23, 25, 29, 37, 39,
5	5, 10, 13, 27, 30, 33, 34, 38, 50, 53, 54, 56, 62, 75, 77
6	6, 14, 32, 42, 69, 84, 87
7	2, 3, 15, 17, 70, 74
8	5, 11, 19, 20, 33, 51, 55
9	5, 32, 60, 66, 82
10	18, 24, 45, 49, 61, 70, 74, 78, 96, 98
11	26, 81, 87
12	3, 5, 18, 57, 80
13	1, 9, 14, 18, 53, 57, 65, 73, 74,
14	10, 54, 62
51	4, 46, 68
52	8, 21, 28, 47, 59,
55	88
56	64, 72

4.3.5 A novel classification system and nomenclature for OR genes

The results described in sections 4.3.2, 4.3.3 and 4.3.4 strongly suggests that classification of the OR gene family by using the e-value as the distance measure and the MCL algorithm is valid and useful for elucidating the evolutionary history of OR genes in vertebrates. Therefore a novel classification system based on the analysis presented here is recommended for future use.

The OR gene name should begin with species code derived from UniProt's NEWT database, which is a taxonomy database and it is updated daily (Phan *et al.* 2003). For example the species code for *Homo sapiens* is HUMAN, for *Gallus gallus* it is CHICK and for *Bos taurus* it is BOVIN. The cluster ID should also be represented in the name as for example C1 for Cluster1 and C2 for Cluster2 and so on. Individual OR genes should be labeled with ascending numbers to identify them individually. A trailing 'P" would be denoted if the OR gene is a pseudogene or truncated gene.

To demonstrate the nomenclature system, consider five OR genes in Cluster1: one functional human OR gene, one human OR pseudogene, two functional chicken OR genes and one functional zebrafish OR gene. The corresponding nomenclature would be *HUMAN-C1OR1, HUMAN-C1OR2P, CHICK-C1OR3, CHICK-C1OR4, DANRE-C1OR5.* This nomenclature system is flexible to accommodate more OR genes in a cluster or within species as and

when new data becomes available. The classification system and nomenclature will greatly facilitate comparative genomics of OR gene family in the future.

4.3.6 Evolution of OR gene family: a new point of view

I have classified and annotated OR gene family in 45 vertebrate species for comparative analysis. OR genes have diverged early during the vertebrate evolution resulting in specific subgroups of OR genes (Niimura 2009). It was observed in one of the earlier comparisons that fish have more divergent OR repertoire compared to mammals or frog or birds (Niimura and Nei 2005a). The classification in the present study informs that fish OR genes can be classified into 11 clusters (Cluster Ids 5, 8, 9, 48, 83, 85, 86, 88, 93, 93 and 99) whereas using the same criteria for classification, mammalian OR genes can be classified in to 88 different clusters (Figure 27). This suggests that it is the mammalian OR genes that have diversified more than the fish OR genes. Other important finding about fish OR genes is that cluster 99 genes are present in zebrafish only and no other fish species. All of cluster 99 genes are located on the zebrafish chromosome 10 (34.76-34.82 Mb) flanked by cluster 86 (upstream) and cluster 93 (downstream) genes. Since, cluster 99 genes are present only in zebrafish, it is most parsimonious to consider that these OR genes have evolved only in the zebrafish lineage. Similarly, significantly large number of cluster 85 genes are present in medaka and stickleback (21 and 40 respectively) compared to only seven genes in fugu, four in tetraodon and one in zebrafish. This suggests that cluster 85 genes have expanded in medaka and stickleback and not any other fish species. The expansion of cluster 99 in zebrafish and cluster 85 in medaka and stickleback provides evidence in support of genomic drift whereby random duplication events have led to such expansions.

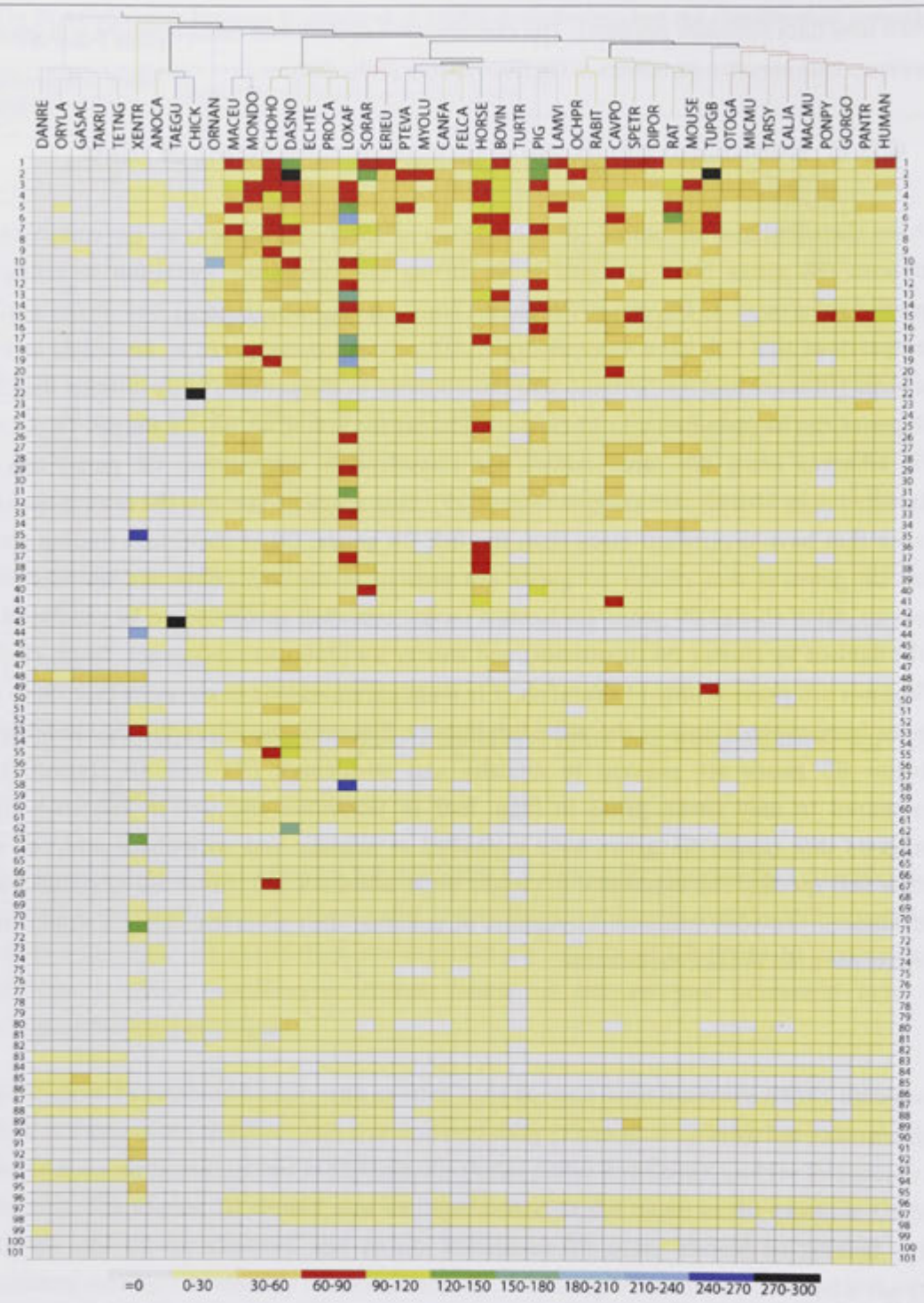


Figure 27 Heat map showing the number of OR genes in each cluster for each species in the intervals of 30. Four-alphabet species codes are represented at the top of the heat map which can be referred to Table 8 for full name and scientific name. The X-axis represents 45 species included for classification and the Y-axis represents the clusters identified by the MCL algorithm.

Functional OR genes from all five species of fish were also examined by phylogenetic analysis (section 4.2.6) to confirm the classification by the MCL algorithm (Figure 28). Only eight clusters were tested by phylogenetic analysis since there were no functional genes present in cluster 5, 8 and 9 from fish. Seven of the eight clusters analyzed (except cluster 99, 12 genes) were monophyletic in the phylogenetic tree. Notable exception were two OR genes, one from cluster 85 and one from cluster 48, that did not group with their respective cluster. Nevertheless, of 308 OR genes analyzed, 294 (95%) genes formed monophyletic clades that represented individual clusters identified in the present study. This suggests that the MCL algorithm is able to significantly capture the evolutionary relationship of OR genes.

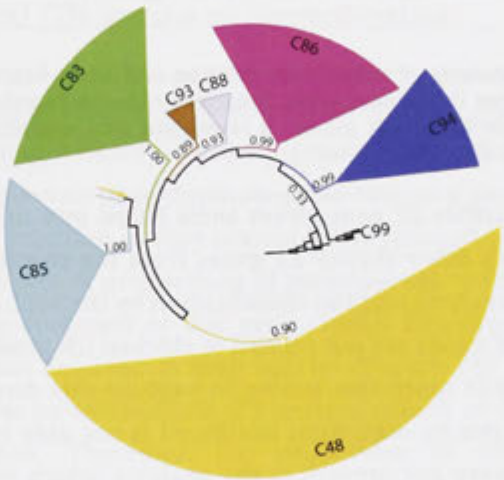


Figure 28 Phylogenetic analysis by maximum likelihood for functional OR genes from fish. All monophyletic clades with significant likelihood probability were collapsed. It is revealed that the classification by the MCL algorithm is consistent with the phylogenetic analysis since all monophyletic clades represent genes from a single cluster. Only one gene from cluster 85 and one from cluster 48 were not grouping with their respective clusters.

Similar to fish OR genes, avian OR genes are also of particular interest. I identified 383 OR genes in the chicken and 464 OR genes in the zebrafinch. 319 (83%) chicken OR genes were grouped together into cluster 22 along with 22 genes from anole lizard and one gene from hedgehog tenrec. Similarly, 452 (97%) of zebrafinch genes were grouped together into cluster 43 along with one gene from anole lizard, two genes from chicken and one gene from platypus. This suggests that birds have highly species-specific OR repertoire. The phylogenetic analysis of functional OR genes from birds and anole lizard (section 4.2.6 for methods) also reveals that significant proportion of bird OR repertoire is species-specific (Figure 29). This has been independently verified as well (Steiger *et al.* 2009b).

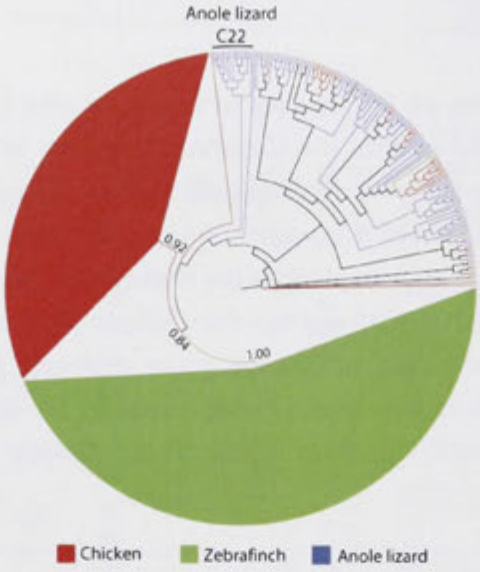


Figure 29 Phylogenetic analysis of zebrafinch, chicken and anole lizard functional OR genes. Red branches represent chicken OR genes, green branches represents zebra finch OR genes and blue branches represents anole lizard OR genes. It is evident that zebrafinch and chicken OR genes have evolved in a species-specific manner.

The MCL algorithm classifies 22 genes from anole lizard into the cluster 22. These 22 OR genes do not group with other cluster 22 genes from the chicken indicating discrepancy between phylogenetic analysis and the classification by the MCL algorithm. It is likely that these 22 anole lizard OR genes are not related to chicken OR genes from cluster 22 and the inflation parameter needs more fine tuning to capture this discrepancy. It is also likely that, phylogenetic analysis by maximum likelihood is not able to resolve the monophyly because pseudogenes were not present in the analysis, which may be the needed link to group chicken and anole lizard cluster 22 genes. Further investigations in the future would clarify this ambiguity.

Like fish and birds, frog OR repertoire has also evolved by species-specific expansion of OR genes. 1,143 OR genes were identified in the frog of which 771 (67%) of genes are present in four clusters: cluster 35 (257 genes), cluster 44 (219 genes), cluster 63 (138 genes), cluster 71 (122 genes) and cluster 95 (35 genes). Only cluster 63 contains one gene from armadillo, otherwise remaining all clusters are frog specific clusters. Only cluster 48 is shared between frog and fish only and no other mammals or birds. It is likely that genes from this cluster are essential for detecting water-soluble odorants. However, further studies would be required to confirm that hypothesis. Apart from frog specific clusters and cluster shared with fish, remaining all clusters with frog genes have gene from other mammalian species.

It has been recently shown that platypus has greater number of family 14 genes (corresponding to clusters 10, 54 and 62, Table 13) compared to other therian mammals (Hayden *et al.* 2009). There is no platypus OR gene in cluster 54 and 62. However, of the

total 588 OR genes in platypus, 195 (33%) OR genes are grouped into cluster 10. The number of platypus OR genes in cluster 10 is the highest of any mammalian species, second highest being 99 genes from shrew. This clearly indicates the expansion of cluster 10 genes in the platypus.

The number of OR genes between mammals varies greatly from species to species, for example, 520 to 850 in primates, 879 to 2,066 in rodents, 155 to 2,041 in cetartiodactyls, 703 to 1057 in carnivores, 925 to 3,672 in afrotherians, and 1,451 to 1689 in marsupials. The lowest number of OR genes were present in dolphin (155 genes) suggesting that these genes were not essential for dolphins in the aquatic environment and hence they have been lost. Although the range of the OR gene family in therian mammals varies from species to species, there is no clear evidence of species-specific expansion of any clusters.

4.3.7 Functional OR genes in vertebrates

Olfactory receptor genes are thought to have evolved through genomic drift where large genomic regions containing olfactory receptor genes gets duplicated within population (Nozawa *et al.* 2007). However, once duplicated, the olfactory genes are subject to either removal through natural selection or they remain active through adaptive evolution (Gilad *et al.* 2003). The comparison of proportions of pseudogenes *vs* functional genes helps us understand the adaptive evolution of this gene family. Significant number of olfactory receptor pseudogenes was detected in each species (Figure 24). However, species level comparisons do not allow us to elucidate if a certain subset of OR genes are being lost through natural selection. Therefore, the proportion of functional (or inversely pseudogenes) were counted for each cluster of OR genes (Figure 30). The definition of functional genes is strictly theoretical definition without any experimental evidence as explained in section 4.2.1. However, this does not affect the following conclusions about the proportion of functional *vs.* pseudogenes.

Cluster 15 and 98 has significantly higher proportion of pseudogenes (lower proportion of functional genes). Cluster 15 is present only in therian mammals (Figure 27) with only one gene in tarsier compared to significantly higher number in human (91 genes) and chimpanzee (83 genes). None of these human or chimpanzee cluster 15 genes could be identified as functional genes based on the criteria defined in section 4.2.1. Similarly, cluster 98 is also present in almost all mammals and not in any other species, and most of the cluster 98 genes pseudogenes. In contrast to clusters 15 and 98, there were 19 clusters with more than 60% of genes identified as functional genes.

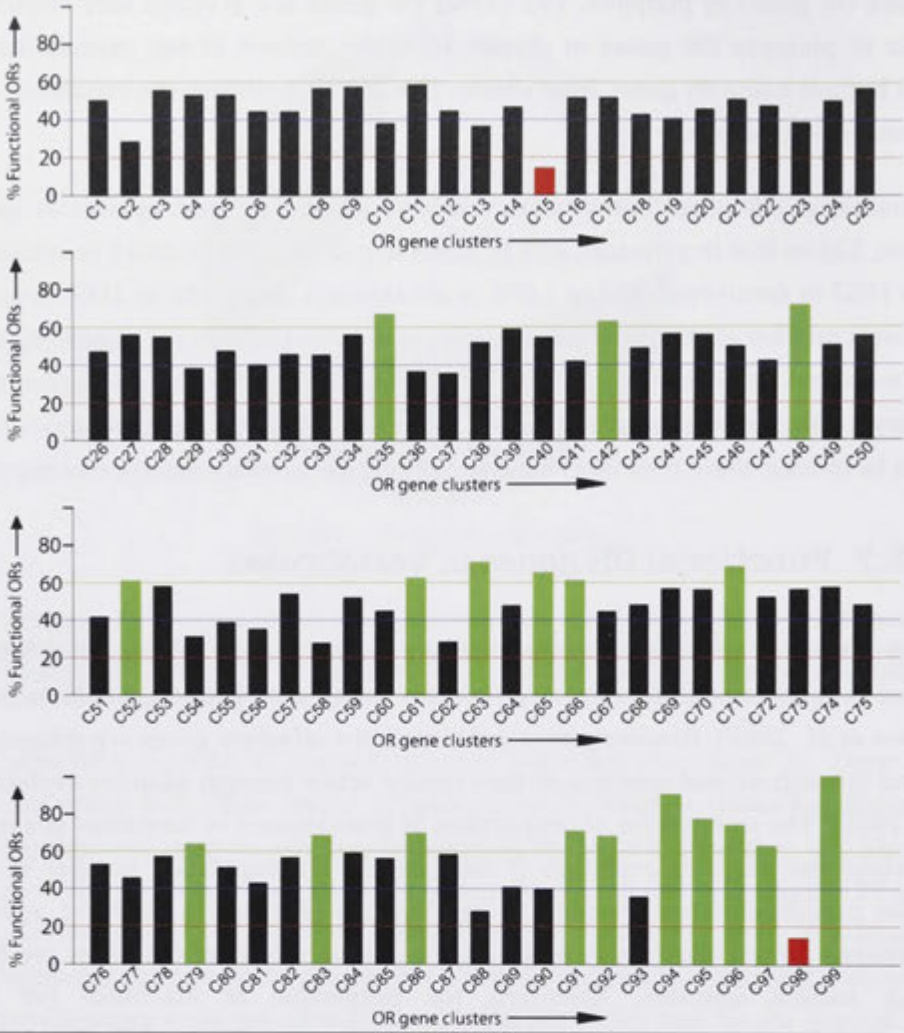


Figure 30 A histogram representing percentage of functional genes in each cluster. Green bars represent the presence of more than 60% functional genes in a cluster and red bar indicates less than or equal to 20% functional genes. Black bars indicate a range between 20 and 60% of functional OR genes.

The definite cause of removal of cluster 15 and 98 genes from mammals or retention of larger proportion of functional OR genes from certain cluster remain to be investigated. However, the classification by the MCL algorithm offers us opportunity to closely inspect closely related genes in multiple species for the signatures of natural selection or adaptive evolution; the distinctive advantage is that both functional and pseudogenes can be classified by the same framework.

4.4 Discussion

4.4.1 Data mining to establish vertebrate OR repertoires

I have identified and annotated OR genes in 45 vertebrate species. This offers unique opportunity to understand the dynamics of the evolution of OR gene family in vertebrates. I have developed a framework in Perl scripting language to identify this large gene family in any vertebrate genome for which genomic sequence data is currently available, and to accommodate more organisms later when their genomic sequence data becomes available in the future. The data mining process was streamlined with the Ensembl release of any new vertebrate genome sequences as well as updates of assemblies for any species. Various databases that contain members of the annotated olfactory receptor gene family do exists, but they are either not specific to olfactory receptors and contain entire GPCR family (Horn *et al.* 2003, Papasaikas *et al.* 2004), or are limited to only few species and not updated since their release (Crasto *et al.* 2002, Olender *et al.* 2004). Even in more recently published study, the genomic sequence data was used from the Ensembl release in December 2008 (Hayden *et al.* 2009). It is imperative to streamline identification and annotation of olfactory receptor gene family in vertebrates since more and more genomic sequence data is being gathered for comparative analysis (Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species 2009).

A program to identify olfactory receptor gene in a sequence has recently been published (Hayden *et al.* 2009), but it suffers from one of major drawbacks in the identification pipeline, which is they haven't addressed the issue of removal of false positive sequences. I have shown in the present study that previously non-OR genes have been misidentified as OR genes. Previously, non-OR genes were identified based on the manual inspection of phylogenetic trees (Glusman *et al.* 2000, Niimura 2009). However, the choice of outgroups used for such discrimination greatly affects the outcome in phylogenetic analysis (Piller and Bart 2009). In light of this evidence, HMM searches employed for the removal of false positive sequences is most appropriate and would increase the confidence in subsequent comparative analysis. Moreover, HMM based removal of false positive sequences have a distinctive advantage that it can easily incorporate more outgroup sequences without affecting the accuracy and speed of such implementation.

Another issue addressed in the present data-mining strategy is the use of FASTY searches for obtaining conceptual translation of OR genes (Glusman *et al.* 2000). Previously, the coding region of the OR genes were identified by walking out in either direction from the initial TBLASTN hits until a start or stop codon is identified in the given range (Niimura and Nei 2007). This extension in either direction is not based on the alignments of the sequences and therefore it will either underestimate or overestimate the length of the OR sequences (Figure 31). I used the FASTY program (Pearson 1994, Pearson and Lipman

1988) to extend the TBLASTN search matches in both direction, and obtained conceptual translations of the OR sequences based on pairwise alignments as previously described (Glusman et al. 2000). The use of FASTY searches should not affect the number of OR genes identified; however, the number of functional OR genes or pseudogenes genes will be different.

A

```
>212-Gga:Un_random:27237433:27238464:-1
          Length = 338

 Score =  653 bits (1685), Expect = 0.0,   Method: Compositional matrix adjust.
 Identities = 328/340 (96%), Positives = 329/340 (96%), Gaps = 2/340 (0%)

Query: 1   MGQGNETVITEFILEGFSGLNQRLQLFLSLVFLLIYLTTVMGNATIIFLVCVDNRLQIPM 60
           MGQGNETV+TEFILEGFSGLNQRLQLFLSLVFLLIYLTTVMGNATIIFLVCVDNRLQIPM
Sbjct: 1   MGQGNETVVTEFILEGFSGLNQRLQLFLSLVFLLIYLTTVMGNATIIFLVCVDNRLQIPM 60

//

Query: 301 KDAIAHATAQLFPQSRCISGQ*F*LSVTVSNAARAVGHTM 340
           KDAIAHATAQLFPQSRCISGQ F  SV V NAARAVGHTM
Sbjct: 301 KDAIAHATAQLFPQSRCISGQQF--SVIVPNAARAVGHTM 338
```

B

```
>212-Gga:Un_random:27237433:27238464:-1
          Length = 338

 Score =  631 bits (1627), Expect = 0.0,   Method: Compositional matrix adjust.
 Identities = 314/321 (97%), Positives = 315/321 (98%)

Query: 1   MGQGNETVITEFILEGFSGLNQRLQLFLSLVFLLIYLTTVMGNATIIFLVCVDNRLQIPM 60
           MGQGNETV+TEFILEGFSGLNQRLQLFLSLVFLLIYLTTVMGNATIIFLVCVDNRLQIPM
Sbjct: 1   MGQGNETVVTEFILEGFSGLNQRLQLFLSLVFLLIYLTTVMGNATIIFLVCVDNRLQIPM 60

//

Query: 301 KDAIAHATAQLFPQSRCISGQ 321
           KDAIAHATAQLFPQSRCISGQ
Sbjct: 301 KDAIAHATAQLFPQSRCISGQ 321
```

Figure 31 An example demonstrating the difference in the length of the predicted OR sequences. (A) A chicken OR sequence identified by the method used here (B) The same OR sequence identified by Niimura (2009). The alignment in the shaded area shows that Niimura (2009) did not extend the OR sequence beyond the stop codon and consequently identified the OR gene as a functional gene. The method used here identified the same OR sequence but labeled it as pseudogene since the stop codon lies within the putative coding region.

The data-mining pipeline presented here therefore can be used for establishing a database that is dedicated to identification and annotation of OR gene family in vertebrates.

4.4.2 Classification of OR genes

OR genes were classified into two major classes based on the evidence gathered from *Xenopus laevis*; class I and class II (Freitag *et al.* 1995). Class I genes were more closely related to fish OR genes and class II genes were more closely related to mammalian OR genes in *Xenopus laevis*. Based on the identification of OR genes from multiple vertebrate species, this classification system was put to test. It was shown for the first time that class I genes not only exists in fish but in some mammalian species as well (Glusman *et al.* 2000). The issue of class I and II was not fixed, but circumvented, and a systematic

nomenclature system was proposed. Class I genes were divided into 17 families and class II genes into 14 families, and these families were subdivided into several subfamilies as well. One of most recent analysis of OR genes from 50 mammalian species has clearly shown that not all families form monophyletic clades in the phylogenetic analysis (Hayden *et al.* 2009), the basis of the classification system, and therefore it would be inappropriate to use this classification system.

One of the other competing classification systems divides OR genes into two major groups; type I and II (Niimura and Nei 2005a). Type I genes were further divided into six groups and type II genes were divided into five subgroups. According to this classification system, the whole mammalian OR repertoire was essentially classified as group γ genes belonging to type I. This classification system was robust and tested using multiple chordate species. However, the major drawback was that it did not offer resolution to study species-specific gains or losses of OR genes.

Both these classification system used neighbour-joining phylogenetic trees infer groups of closely related OR genes. The complexity of the neighbour-joining algorithm increases both in time and space with an increase in the number of amino acid sequences used (Guo *et al.* 2007). Consider that number '*n*' of OR sequences need to be classified using the neighbour-joining phylogenetic tree analysis. Time and space complexity are functions of the number of sequences and time complexity will increase at the rate of n^3 and space complexity will increase at the rate of n^2 in the neighbour-joining algorithm. If the neighbour-joining algorithm requires 5 seconds to process n = 100 sequences, it will require 5000 seconds to process 1,000 sequences and 5 million seconds (approximately 58 days) to process 10,000 sequences. Similarly, in layman's term, if the neighbour-joining algorithm requires 5 units of memory on a computer for 100 sequences, it will require 500 units for 1,000 sequences and 50,000 units for 10,000 sequences. Multiple sequence alignments used for phylogenetic analysis will also add in to the time and space complexity for phylogenetic analysis. Therefore it becomes prohibitively difficult to perform neighbour-joining phylogenetic gene tree analysis to classify OR genes as more and more genomic sequence data become available from variety of organisms.

To overcome all the challenges that are existing for the classification of OR genes into groups of closely related sequences, I have used –log(e-value) as the distance measure and the MCL algorithm to process pairwise distances to identify groups of closely related sequences. The MCL algorithm was successfully used to identify other gene family clusters in whole vertebrate genomes (Enright *et al.* 2002, Vilella *et al.* 2009). However, no previous attempts had been made to obtain sub-family level classification. I have shown that by carefully adjusting the granularity parameter in the MCL algorithm, partition clusters of the OR gene family can be achieved. The MCL algorithm was able to process pairwise similarity graph in matter of hours and therefore it is more suitable to handle such a large gene family compared to the neighbour-joining algorithm. The classification

of OR genes into clusters of closely related sequences was tested by phylogenetic analysis and it was shown that significant number of clusters were recovered as monophyletic clades. This suggested that e-value from the BLAST searches can be used as a distance measure to classify OR genes.

Additionally the classification system presented here provides resolution necessary to understand the evolutionary dynamics of the OR gene family in vertebrates. It was shown using this classification that there are several OR clusters that are species specific. The classification presented here was also able to recapitulate the diversity of OR genes in all vertebrate species with sufficient resolution to elucidate selective pressures that may be acting on this gene family in the future. Specifically, analysis of each cluster for birth and death of OR genes using tools like CAFÉ (De Bie *et al.* 2006) can reveal the dynamics of this gene family at a finer level. These clusters of OR genes can also be used identify functional residues in the OR genes as previously have been shown (Man *et al.* 2007). These results and potential for deeper understanding are strongly favouring the classification of OR genes by using the MCL algorithm and BLAST e-values.

4.4.3 Alternative classification methods

Profile-hidden Markov models (Brown *et al.* 2005, Eddy 1998, Finn *et al.* 2008, Srivastava *et al.* 2007) and support vector machines (SVM) (Bhasin and Raghava 2004, Karchin *et al.* 2002) are algorithms that can be efficiently applied to family and sub-family classification. Significant success has been observed in the general classification of GPCR proteins into families and subfamilies based on HMM and SVM. The SVM algorithm is generally more sensitive but computationally more expensive than the HMM based methods (Karchin *et al.* 2002). In addition, the relationship between two sequences as orthologs should be known before using SVM or HMM based methods, to prevent the inclusion of false positives. Functional evidence and chromosomal location information are necessary to obtain from more number of species to isolate true orthologs so that SVM and HMM based methods can be used with confidence. My study provides basis for future studies that may employ SVM or HMM based classification systems.

4.4.4 Novel nomenclature of OR gene family

The aim of the proposed nomenclature is to allow easy comparative analysis of OR gene family in vertebrates. This nomenclature effectively represents the OR sequence similarity at a specified granularity. The name of the OR gene will show the cluster it belongs to and hence if one is interested in estimating sequence divergence or selection pressure within a cluster, all OR sequences of the same cluster can easily be isolated and analyzed. Clear advantage of a systematic nomenclature presented here is that researchers from various disciplines can have a unique and common name for each one of the OR gene for easy

reference. Despite advantages, one of the major limitations of this nomenclature would be that clear distinction between orthologs or paralogs would not be possible. However, this is intentionally avoided at this stage because we do not have sufficient mapping information to satisfactorily establish orthology/paralogy in all species. Since ortholog assignments were not attempted in this nomenclature, it was deemed as necessary to include species name and trailing unique number in the OR gene name.

5 Discussion

Comparative genomics has come a long way in the era of DNA sequencing technology. The reduced cost of obtaining DNA sequences from organisms has revolutionized the field. However, the power of sequencing has also increased the complexity of deducing meaningful relationship between homologous DNA in two species or within a species.

In this thesis I have addressed three key elements in comparative genomics: obtaining and integrating physical maps and genetic linkage maps for the tammar wallaby genome, distinguishing orthologs and paralogs in order to understand the evolution of the human X chromosome, and identifying and classifying members of the olfactory receptor gene family in vertebrates. These three topics cover three distinct aspects in the comparative genomics: genome evolution, chromosome evolution, and gene family evolution.

5.1 Tammar wallaby whole genome shotgun sequences: wealth of information

The DNA sequences from an organism are much less tractable and informative without a good map to guide its assembly and interpretation (Lewin *et al.* 2009). The NHGRI undertook the gala challenge of obtaining low coverage genomic sequences from 24 vertebrate species for comparative analysis (National Human Genome Research Institute).

The initial low coverage sequencing and the comparative analysis of the dog genome (Kirkness *et al.* 2003) and the cat genome (Pontius *et al.* 2007) established the importance of two strategies for making maximum use of low coverage genome sequences. First was to obtain a high-density physical map or genetic linkage map to anchor contigs to chromosomes. Second was to make use of the cross species comparative data to form scaffolds of contigs representing the homologous syntenic blocks between two or more species. The output of both these strategies was merged to create the final assembly of the low coverage cat genome and dog genome (Kirkness *et al.* 2003, Murphy *et al.* 2005, Pontius *et al.* 2007).

One of the first mammals to be the subject of twofold coverage was the tammar wallaby genome sequence, obtained at the low coverage through the joint effort of the Baylor College of Medicine and the Australian Genome Research Facility (Graves *et al.* 2003). It was therefore important to obtain a physical and linkage maps of this species in order to aid the assembly, and this was done by targeting regions that were syntenic in other marsupials and eutherians, by identifying breakpoints in the physical map (Deakin *et al.* 2008) and integrating the genetic linkage map using markers mapped on both.

In section 2, I presented my research work identifying microsatellite markers for the linkage map of tammar wallaby within or near to genes that have been (or could be) mapped physically. This enabled the integration of the two types of maps for this species, which was important for the map to be useful for localizing phenotypic traits. I identified microsatellite markers by exploring the low coverage whole genome shotgun sequences of tammar wallaby, a novel approach that I pioneered.

The syntenic blocks that were identified to be homologous between two species included orthologous genes arranged contiguously in the same order in two species that are compared. Traditionally, such homologous syntenic blocks were obtained by analyzing chromosome scale global alignments of the DNA (Kohn *et al.* 2006, Murphy *et al.* 2005, Pan *et al.* 2005). More recently, with the automatic identification of the orthologs in variety of organisms (Berglund *et al.* 2008, Li *et al.* 2006, Li *et al.* 2003, Remm *et al.* 2001, Vilella *et al.* 2009), the computation of chromosomal scale global alignments can be avoided to identify homologous syntenic blocks between two species (Ng *et al.* 2009).

Here I developed a simpler algorithm, which takes the orthologous gene sets in two species and their corresponding locations in each species to identify homologous syntenic blocks between any two species. 628 homologous syntenic blocks between marsupial (opossum) and eutherian (human) mammals were identified using this algorithm by allowing for 1 Mb of insertion/deletion between any two consecutive syntenic genes.

These homologous syntenic blocks allowed for a targeting of the gap regions in the linkage map. Successful identification of microsatellite markers in the linkage map gap regions and subsequent localization of these markers in the predicted gap regions of the linkage map demonstrated that this strategy of using homologous syntenic block information enhances throughput for the physical map and the linkage map. In retrospect, if the homologous syntenic block information is used for the physical map of the tammar wallaby, it would require approximately $628 \times 2 = 1256$ markers (one for each end of the homologous syntenic block) to cover approximately 70% of the genome.

My work presented in section 2 shows that any *de novo* genome with low coverage sequences can efficiently be processed to obtain maximum information for comparative analysis to reduce the cost and time required. This should be particularly valuable because more resources can be invested in understanding the evolution of the genome rather than organization of the genome.

More recently, new technologies have become available to obtain DNA sequence at a reasonably lower cost and in less time than traditional Sanger sequencing technology (reviewed in Shendure and Ji 2008). These have aided in obtaining better assembly of lower coverage mammalian genome. For example, 2× gorilla genome assembly was supplemented with 35× coverage Illumina short read assembly to obtain virtual gorilla

genome (http://www.ensembl.org/Gorilla_gorilla/Info/Index). Similarly, next generation sequencing can be used to obtain mate-pair or paired-end reads for tammar wallaby, and these reads can be used to obtain larger contigs and virtual tammar wallaby genome. Classical genetic tools like linkage analysis or cytogenetic techniques like FISH mapping, may be replaced by next-generation sequencing technologies in the future to obtain chromosomal maps for comparative analysis.

5.2 Evolution of the human X chromosome: a mystery resolved

It is essential to obtain a good physical map of the genomes for understanding genome evolution. The chromosomal location of the genes (markers) provides crucial information in understanding of their homology relationship. For example, orthologous genes in two species are separated through a speciation event only, and they are more likely to be present in the same homologous syntenic block between two species. One of the criteria for establishing orthology was recognized to be their presence in homologous synteny groups (Andersson *et al.* 1996).

Distinguishing orthologs and paralogs by their genome context turned out to be critical in resolving a long-running debate about the origins of the human X chromosome. In section 3, I present my work on the analysis of the human X chromosome evolution by comparing orthologous and paralogous regions of the genes on the human X chromosome in the cytogenetic bands Xp11 and Xq28. Previous reports (Kohn *et al.* 2004, Ross *et al.* 2005) based on the best hit pairwise comparisons suggested that two regions, Xp11 and Xq28, formed a separate evolutionary block on the human X chromosome because chicken homologs were located in a third genomic regions.

To investigate the origin of this putative separate evolutionary block, homologous syntenic blocks containing paralogous genes were compared in human, rat, opossum, and the chicken genome. Using this strategy I was able to show by comparative analysis that the best hit method of Kohn *et al.* (2004) and Ross *et al.* (2005) misidentified paralogous genes as orthologous genes. My analysis showed that the human Xp11 and Xq28 orthologous genes are not present in the current chicken genome assembly, because of that paralogs in other locations were identified as orthologs (Delbridge *et al.* 2009).

My work raised the question of whether these chicken orthologs were deleted from the chicken lineage, or whether they were merely missing from the current chicken genome assembly. To distinguish these possibilities, I analyzed the chicken EST/cDNA sequence data by phylogenetic gene tree analysis and reciprocal best hit method to confirm that the chicken orthologs of the human Xp11 and Xq28 genes are present in these libraries, so are merely missing from the current chicken genome assembly rather than being deleted from

the chicken lineage. Many chicken EST/cDNA sequences that I identified as orthologous to the human Xp11 and Xq28 genes did not map to any regions in the chicken genome, indicating that these regions were excluded from the chicken genome sequencing. The more detailed phylogenetic gene tree analysis that I used, and the reciprocal best hit method, revealed many Xp11 and Xq28 orthologous genes in the chicken EST/cDNA sequences. This showed that the best hit method used previously (Kohn *et al.* 2004, Ross *et al.* 2005) is not sufficient for establishing the orthologous relationship between genes and regions using incomplete data sets.

My research on the evolution of the human X chromosome leads to a simpler model of the evolution of the human X chromosome with only two evolutionary blocks made up of the X-conserved region and the X-added region (Graves 1995). It would be interesting to map the Xp11 and Xq28 orthologous genes to chicken metaphase chromosomes to compare their locations. However, physical mapping of smaller probes (EST/cDNA sequences) is not efficient or reliable, so it was not attempted as part of this research work.

False identification of paralogous genes as orthologous can have lead to serious confusions in comparative analysis. For example, the human X chromosome is enriched for sex and reproduction related genes (Saifi and Chandra 1999) and mental retardation genes (Zechner *et al.* 2001), and it was suggested that such genes accumulated on the X because it is unpaired in males; either because of direct selection of male advantageous genes, or sexual selection for male "ornament" genes that are expressed in XY hemizygous males but not XX females (Graves *et al.* 2002).

The evolutionary history of this biased gene content pattern was tested by functional analysis of homologous genes in the chicken genome (Kemkemer *et al.* 2009b). Their analysis assumed the three evolutionary blocks on the human X chromosome, and suggested that chicken chromosome 12 was also enriched for the brain function related genes, implying that the regions was enriched in brain genes even before it was co-opted into being a sex chromosome. This is an important conclusion, contradicting the theory that male-advantage or sexually selected genes accumulate on the X. However, I have shown here that the chicken chromosome 12 is a paralogous region and not the orthologous region. This implies that the biased expression of the chicken chromosome 12 genes in the chicken brain could be an effect of the subfunctionalization of genes, instead of the orthology of genes with an ancient function in the brain (Rastogi and Liberles 2005). Thus it is important to establish that putative orthologs in two different species have the same genome context so that comparisons of their function and evolution are meaningful.

5.3 Holistic analysis of the olfactory receptor gene family

The olfactory receptor gene family is the largest gene family in mammals, composed of hundreds or thousands of members with very similar sequences. Because of its size and the sequence similarity of its members, large scale analysis for genome evolution and gene family evolution often omits the olfactory receptor gene because it poses difficulties in automation and increases computational costs (Li *et al.* 2006).

Therefore it is essential to develop a suitable method to understand the evolution of this large gene family in vertebrates. I designed a pipeline specifically to automate the identification of the OR genes in vertebrates. Then a clustering based strategy was used to classify OR genes. The benefit of these research outcomes is the availability of data on the olfactory receptor gene family to the scientific community for future analysis. This platform will enable further analyses that will enhance our understanding of the evolution of the OR repertoire in vertebrates. The data-mining pipeline is synchronized with the Ensembl database so that newer versions of the assembly, or a newly available assembly of an organism, can be processed as and when it becomes available.

This platform for the olfactory receptor gene family will greatly enhance the capabilities to investigate the function of individual olfactory receptor genes, evaluate the evolutionary trajectory of this gene family in vertebrates and perform comparative analysis with a robust, continually updated dataset that has never been achieved for the olfactory receptor gene family.

6 Conclusions

I have shown in this research work that comparative genomics is a useful tool to understand the evolutionary mechanisms that shape the genomes and hence the phenomes. It is important to first understand the organization of the genome for which I have shown that appropriate use modern bioinformatics tools and techniques can significantly increase the throughput for construction of blocks of conserved synteny: the pieces of "the genome jigsaw puzzle". Tammar wallaby genomic sequence assembly pipeline will be greatly benefited from this analysis. Once the pieces of jigsaw puzzle are identified, comparative analyses can be performed to understand the organization of these pieces and it can illuminate major karyotype rearrangements. I have addressed the key issue regarding the evolution of the X chromosome by comparative analysis of the conserved blocks of synteny. This will greatly improve inferences regarding the evolution of the human X chromosome. Large block of conserved synteny can help identify major differences between species, like apples from oranges. However, it is the genes within these large blocks of conserved synteny that dictate finer differences between various apples. The analysis of large OR gene family presented here provides many avenues to identify evolutionary paths, selective pressures, peculiarities of genomes, expansion and contraction of gene families and functional properties for this gene family. It would be fair to say that comparative genomics is essential for understanding the evolution, be it at genome level, chromosome level or gene level.

7 References

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B. and de Rosa, R. (2000) The new animal phylogeny: Reliability and implications. *Proc. Natl. Acad. Sci. U. S. A.*, 97, p4453.

Aloni, R., Olender, T. and Lancet, D. (2006) Ancient genomic architecture for mammalian olfactory receptor clusters. *Genome Biology*, 7, pR88.

Alsop, A., Miethke, P., Rofe, R., Koina, E., Sankovic, N., Deakin, J., *et al.* (2005) Characterizing the chromosomes of the Australian model marsupial *Macropus eugenii* (tammar wallaby). *Chromosome Res.*, 13, p627.

Altschul, S., F., Gish, W., Miller, W., Myers, E., W. and Lipman, D., J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, 215, p403.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25, p3389.

Ananias, F., Modesto, A. D. S., Mendes, S. C. and Napoli, M. F. (2007) Unusual primitive heteromorphic ZZ/ZW sex chromosomes in *Proceratophrys boiei* (Anura, Cycloramphidae, Alsodinae), with description of C-Band interpopulational polymorphism. *Hereditas*, 144, p206.

Andersson, L., Archibald, A., Ashburner, M., Audun, S., Barendse, W., Bitgood, J., *et al.* (1996) Comparative genome organization of vertebrates. The first international workshop on comparative genome organization. *Mamm. Genome*, 7, p717.

Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol*, 55, p539.

Bailey, J. A., Baertsch, R., Kent, W. J., Haussler, D. and Eichler, E. E. (2004) Hotspots of mammalian chromosomal evolution. *Genome Biology*, 5.

Barbazuk, W. B., Korf, I., Kadavi, C., Heyen, J., Tate, S., Wun, E., *et al.* (2000) The syntenic relationship of the zebrafish and human genomes. *Genome Res.*, 10, p1351.

Baroiller, J. F., Guiguen, Y. and Fostier, A. (1999) Endocrine and environmental aspects of sex differentiation in fish. *Cellular and Molecular Life Sciences (CMLS)*, 55, p910.

Bennett, J. H., Hayman, D. L. and Hope, R. M. (1986) Novel sex differences in linkage values and meiotic chromosome behaviour in a marsupial. *Nature*, 323, p59.

Beraldi, D., McRae, A. F., Gratten, J., Pilkington, J. G., Slate, J., Visscher, P. M., *et al.* (2007) Quantitative trait loci (QTL) mapping of resistance to strongyles and coccidia in the free-living Soay sheep (Ovis aries). *Int. J. Parasitol.*, 37, p121.

Berglund, A. C., Sjolund, E., Ostlund, G. and Sonnhammer, E. L. (2008) InParanoid 6: eukaryotic ortholog clusters with inparalogs. *Nucleic Acids Res*, 36, pD263.

Bhasin, M. and Raghava, G. P. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res*, 32, pW383.

Bick, Y. A., Murtagh, C. and Sharman, G. B. (1973) The chromosomes of an egg-laying mammal Tachyglossus aculeatus (the echidna). *Cytobios*, 7, p233.

Bininda-Emonds, O. R. P., Cardillo, M., Jones, K. E., MacPhee, R. D. E., Beck, R. M. D., Grenyer, R., *et al.* (2007) The delayed rise of present-day mammals. *Nature*, 446, p507.

Blair Hedges, S. and Kumar, S. (2004) Precision of molecular time estimates. *Trends Genet.*, 20, p242.

Blair, J. E. and Hedges, S. B. (2005) Molecular phylogeny and divergence times of Deuterostome animals. *Mol. Biol. Evol.*, 22, p2275.

Boardman, P. E., Sanz-Ezquerro, J., Overton, I. M., Burt, D. W., Bosch, E., Fong, W. T., *et al.* (2002) A comparative collection of chicken cDNAs. *Curr. Biol.*, 12, p1965.

Bourque, G. and Pevzner, P. A. (2002) Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Res.*, 12, p26.

Boyd, Y., Blair, H. J., Cunliffe, P., Denny, P., Gormally, E. and Herman, G. E. (1998) Mouse chromosome X. *Mamm. Genome*, 8, pS361.

Breen, M., Thomas, R., Binns, M. M., Carter, N. P. and Langford, C. F. (1999) Reciprocal chromosome painting reveals detailed regions of conserved synteny between the karyotypes of the domestic dog (*Canis familiaris*) and human. *Genomics*, 61, p145.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L. and Weber, J. L. (1998) Comprehensive human genetic maps: Individual and sex-specific variation in recombination. *Am. J. Hum. Genet.*, 63, p861.

Brown, D., Krishnamurthy, N., Dale, J. M., Christopher, W. and Sjolander, K. (2005) Subfamily HMMs in functional genomics. *Pac. Symp. Biocomput.*, p322.

Buck, L. and Axel, R. (1991) A novel multigene family may encode odorant receptors: A molecular basis for odor recognition. *Cell*, 65, p175.

Carbone, L., Harris, R. A., Vessere, G. M., Mootnick, A. R., Humphray, S., Rogers, J., *et al.* (2009) Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. *Plos Genetics*, 5.

Charlesworth, B. (1978) Model for evolution of Y chromosomes and dosage compensation. *Proc. Natl. Acad. Sci. U. S. A.*, 75, p5618.

Charlesworth, B., Coyne, J. A. and Barton, N. H. (1987) The relative rates of evolution of sex-chromosomes and autosomes. *American Naturalist*, 130, p113.

Chou, H. H. and Holmes, M. H. (2001) DNA sequence quality trimming and vector removal. *Bioinformatics*, 17, p1093.

Chowdhary, B. P. and Raudsepp, T. (2005) Mapping genomes at the chromosome level. IN RUVINSKY, A. and GRAVES, J. A. M. (Eds.) *Mammalian Genomics.* Cambridge, CABI Publishing.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucl. Acids Res.*, 31, p38.

Clark, M. S. (1999) Comparative genomics: the key to understanding the Human Genome Project. *Bioessays*, 21, p121.

Costantini, M., Cammarano, R. and Bernardi, G. (2009) The evolution of isochore patterns in vertebrate genomes. *Bmc Genomics*, 10.

Costantini, M., Clay, O., Auletta, F. and Bernardi, G. (2006) An isochore map of human chromosomes. *Genome Res.*, 16, p536.

Crasto, C., Marenco, L., Miller, P. and Shepherd, G. (2002) Olfactory Receptor Database: a metadata-driven automated population from sources of gene and protein sequences. *Nucleic Acids Res*, 30, p354.

De Bie, T., Cristianini, N., Demuth, J. P. and Hahn, M. W. (2006) CAFE: a computational tool for the study of gene family evolution. *Bioinformatics,* 22, p1269.

Deakin, J., Koina, E., Waters, P., Doherty, R., Patel, V., Delbridge, M., *et al.* (2008) Physical map of two tammar wallaby chromosomes: A strategy for mapping in non-model mammals. *Chromosome Res.,* 16, p1159.

Dehal, P. and Boore, J. (2006) A phylogenomic gene cluster resource: the Phylogenetically Inferred Groups (PhIGs) database. *BMC Bioinformatics,* 7, p201.

Dehal, P. and Boore, J. L. (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *Plos Biology,* 3, p1700.

Delbridge, M. L. and Graves, J. A. (2007) Origin and evolution of spermatogenesis genes on the human sex chromosomes. *Soc Reprod Fertil Suppl,* 65, p1.

Delbridge, M. L., Patel, H. R., Waters, P. D., McMillan, D. A. and Graves, J. A. M. (2009) Does the human X contain a third evolutionary block? Origin of genes on human Xp11 and Xq28. *Genome Res.,* 19, p1350.

Devlin, R. H. and Nagahama, Y. (2002) Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. *Aquaculture,* 208, p191.

DiNapoli, L. and Capel, B. (2008) SRY and the standoff in sex determination. *Mol. Endocrinol.,* 22, p1.

Dixon, R. M. W. (2008) Australian aboriginal words in dictionaries: A history. *International Journal of Lexicography,* 21, p129.

Dong, D., He, G., Zhang, S. and Zhang, Z. (2009) Evolution of olfactory receptor genes in primates dominated by birth-and-death process. *Genome Biol Evol,* pevp026.

Duke, S. E., Samollow, P. B., Mauceli, E., Lindblad-Toh, K. and Breen, M. (2007) Integrated cytogenetic BAC map of the genome of the gray, short-tailed opossum, Monodelphis domestica. *Chromosome Res.,* 15, p361.

Eddy, S. R. (1998) Profile hidden Markov models. *Bioinformatics,* 14, p755.

Edgar, R. (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics,* 5, p113.

Edgar, R. C. (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.,* 32, p1792.

Eggert, C. (2004) Sex determination: the amphibian models. *Reproduction Nutrition Development,* 44, p539.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., *et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biology,* 6.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.,* 30, p1575.

Entrez Genome Database, http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome, National Centre of Biotechnology Information, U.S. National Library of Medicine, Bethesda MD, USA, Last accessed in 2009

Everts-van der Wind, A., Larkin, D. M., Green, C. A., Elliott, J. S., Olmstead, C. A., Chiu, R., *et al.* (2005) A high-resolution whole-genome cattle-human comparative map reveals details of mammalian chromosome evolution. *Proc. Natl. Acad. Sci. U. S. A.,* 102, p18526.

Felsenstein, J. (1974) The evolutionary advantage of recombination. *Genetics,* 78, p737.

Felsenstein, J. (2005), PHYLIP (Phylogeny Inference Package),

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res,* 36, pD281.

Fitch, W. M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.,* 19, p99.

Francis, R. C. and Barlow, G. W. (1993) Social control of primary sex differentiation in the *Midas Cichlid. Proc. Natl. Acad. Sci. U. S. A.,* 90, p10673.

Frazer, K. A., Sheehan, J. B., Stokowski, R. P., Chen, X., Hosseini, R., Cheng, J. F., *et al.* (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res.,* 11, p1651.

Fredriksson, R., Lagerstrom, M. C., Lundin, L. G. and Schioth, H. B. (2003) The G-protein-coupled receptors in the human genome form five main families. Phylogenetic analysis, paralogon groups, and fingerprints. *Mol. Pharmacol.,* 63, p1256.

Freitag, J., Krieger, J. r., Strotmann, J. and Breer, H. (1995) Two classes of olfactory receptors in xenopus laevis. *Neuron,* 15, p1383.

Fuchs, T., Glusman, G., Horn-Saban, S., Lancet, D. and Pilpel, Y. (2001) The human olfactory subgenome: from sequence to structure and evolution. *Hum. Genet.,* 108, p1.

Fullerton, S. M., Carvalho, A. B. and Clark, A. G. (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.,* 18, p1139.

Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. (2009) *J. Hered.,* 100, p659.

Georges, M., Nielsen, D., Mackinnon, M., Mishra, A., Okimoto, R., Pasquino, A. T., *et al.* (1995) Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics,* 139, p907.

Gilad, Y., Bustamante, C. D., Lancet, D. and Paabo, S. (2003) Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am. J. Hum. Genet.,* 73, p489.

Gilad, Y., Przeworski, M. and Lancet, D. (2004) Loss of olfactory receptor genes coincides with the acquisition of full trichromatic vision in primates. *PLoS Biol,* 2, pE5.

Gilman, A. G. (1987) G proteins: transducers of receptor-generated signals. *Annu. Rev. Biochem.,* 56, p615.

Glas, R., Leo, A., Delbridge, M., Reid, K., Ferguson-smith, M., O'Brien, P., *et al.* (1999) Chromosome painting in marsupials: Genome conservation in the Kangaroo family. *Chromosome Res.,* 7, p167.

Glusman, G., Bahar, A., Sharon, D., Pilpel, Y., White, J. and Lancet, D. (2000) The olfactory receptor gene superfamily: data mining, classification, and nomenclature. *Mamm. Genome,* 11, p1016.

Goodstadt, L., Heger, A., Webber, C. and Ponting, C. P. (2007) An analysis of the gene complement of a marsupial, Monodelphis domestica: evolution of lineage-specific genes and giant chromosomes. *Genome Res.,* 17, p969.

Gordon, L., Yang, S., Tran-Gyamfi, M., Baggott, D., Christensen, M., Hamilton, A., *et al.* (2007) Comparative analysis of chicken chromosome 28 provides new clues to the evolutionary fragility of gene-rich vertebrate regions. *Genome Res.,* 17, p1603.

Gorlov, I. P., Zhelezova, A. I. and Gorlova, O. (1994) Sex differences in chiasma distribution along two marked mouse chromosomes: differences in chiasma distribution as a reason for sex differences in recombination frequency. *Genet. Res.,* 64, p161.

Goureau, A., Yerle, M., Schmitz, A., Riquet, J., Milan, D., Pinton, P., *et al.* (1996) Human and porcine correspondence of chromosome segments using bidirectional chromosome painting. *Genomics,* 36, p252.

Graves, J. A., Chew, G. K., Cooper, D. W. and Johnston, P. G. (1979) Marsupial--mouse cell hybrids containing fragments of the marsupial X chromosome. *Somatic Cell Genet.,* 5, p481.

Graves, J. A. M. (1995) The origin and function of the mammalian Y chromosome and Y-borne genes - an evolving understanding. *Bioessays,* 17, p311.

Graves, J. A. M. (2006) Sex chromosome specialization and degeneration in mammals. *Cell,* 124, p901.

Graves, J. A. M. (2008) Weird animal genomes and the evolution of vertebrate sex and sex chromosomes. *Annu. Rev. Genet.,* 42, p565.

Graves, J. A. M., Gécz, J. and Hameister, H. (2002) Evolution of the human X - a smart and sexy chromosome that controls speciation and development. *Cytogenetic and Genome Research,* 99, p141.

Graves, J. A. M., Wakefield, M. J., Renfree, M. B., Cooper, D. W., Speed, T., Lindblad-Toh, K., *et al.* (2003), Proposal to sequence the genome of the model marsupial *Macropus eugenii* (tammar wallaby), Australian Genome Research Facility, Australia and Baylor College of Medicine, Washington University, St Louis, MO, USA Funding organizations: NHGRI, USA and Victoria State Government, Australia

Green, P. (2007) 2X genomes - Does depth matter? *Genome Res.,* 17, p1547.

Grus, W. E., Shi, P. and Zhang, J. (2007) Largest vertebrate vomeronasal type 1 receptor gene repertoire in the semiaquatic platypus. *Mol. Biol. Evol.,* 24, p2153.

Grutzner, F., Crollius, H. R., Lutjens, G., Jaillon, O., Weissenbach, J., Ropers, H.-H., *et al.* (2002) Four-hundred million years of conserved synteny of human Xp and Xq genes on three tetraodon chromosomes. *Genome Res.,* 12, p1316.

Grutzner, F., Rens, W., Tsend-Ayush, E., El-Mogharbel, N., O'Brien, P. C., Jones, R. C., *et al.* (2004) In the platypus a meiotic chain of ten sex chromosomes shares genes with the bird Z and mammal X chromosomes. *Nature,* 432, p913.

Guindon, S., Lethiec, F., Duroux, P. and Gascuel, O. (2005) PHYML Online - a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res,* 33, pW557.

Guo, M. Z., Li, J. F. and Liu, Y. (2007) A topological transformation in evolutionary tree search methods based on maximum likelihood combining p-ECR and neighbor joining. *Bmc Bioinformatics,* 9.

Hammond, M. P. and Birney, E. (2004) Genome information resources - developments at Ensembl. *Trends Genet.,* 20, p268.

Hanvey, J. C., Klysik, J. and Wells, R. D. (1988) Influence of DNA-sequence on the formation of non-B right-handed helices in oligopurine.oligopyrimidine inserts in plasmids. *J. Biol. Chem.,* 263, p7386.

Hardison, R. C. (2000) Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.,* 16, p369.

Hayden, S., Bekaert, M., Crider, T. A., Mariani, S., Murphy, W. J. and Teeling, E. C. (2009) Ecological adaptation determines functional mammalian olfactory subgenomes. *Genome Res.,* 20, p1.

Hayman, D. L. (1990) Marsupial cytogenetics. *Aust. J. Zool.,* 37, p331.

Hedtke, S. M., Townsend, T. M. and Hillis, D. M. (2006) Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol,* 55, p522.

Henderson, J., Salzberg, S. and Fasman, K. H. (1997) Finding genes in DNA with a Hidden Markov Model. *J. Comput. Biol.,* 4, p127.

Hickford, D., Frankenberg, S. and Renfree, M. B. (2009) The tammar wallaby, *Macropus eugenii*: A model kangaroo for the study of developmental and reproductive biology. *Cold Spring Harbor Protocols,* 2009, ppdb.emo137.

Hildebrand, J. G. and Shepherd, G. M. (1997) Mechanisms of olfactory discrimination: converging evidence for common principles across phyla. *Annu. Rev. Neurosci.,* 20, p595.

Hill, C. A., Fox, A. N., Pitts, R. J., Kent, L. B., Tan, P. L., Chrystal, M. A., *et al.* (2002) G Protein-Coupled Receptors in *Anopheles gambiae*. *Science,* 298, p176.

Hore, T. A., Koina, E., Wakefield, M. J. and Marshall Graves, J. A. (2007) The region homologous to the X-chromosome inactivation centre has been disrupted in marsupial and monotreme mammals. *Chromosome Res.,* 15, p147.

Horn, F., Bettler, E., Oliveira, L., Campagne, F., Cohen, F. E. and Vriend, G. (2003) GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res,* 31, p294.

Horn, F., Weare, J., Beukers, M. W., Horsch, S., Bairoch, A., Chen, W., *et al.* (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Res,* 26, p275.

Huang, X. Q. and Madan, A. (1999) Cap3: A DNA sequence assembly program. *Genome Res.,* 9, p868.

Hubbard, S. J., Grafham, D. V., Beattie, K. J., Overton, I. M., McLaren, S. R., Croning, M. D. R., *et al.* (2005) Transcriptome analysis for the chicken based on 19,626 finished cDNA sequences and 485,337 expressed sequence tags. *Genome Res.,* 15, p174.

Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.,* 35, pD610.

Hunkapiller, T., Kaiser, R. J., Koop, B. F. and Hood, L. (1991) Large-scale and automated DNA sequence determination. *Science,* 254, p59.

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M. and Rupp, R. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics,* 8, p460.

Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J. P., *et al.* (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.,* 13, p91.

Jiang, Z., Priat, C. and Galibert, F. (1998) Traced orthologous amplified sequence tags (TOASTs) and mammalian comparative maps. *Mamm. Genome,* 9, p577.

Karchin, R., Karplus, K. and Haussler, D. (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics,* 18, p147.

Kellner, W. A., Sullivan, R. T., Carlson, B. H. and Thomas, J. W. (2005) Uprobe: A genome-wide universal probe resource for comparative physical mapping in vertebrates. *Genome Res.,* 15, p166.

Kemkemer, C., Kohn, M., Cooper, D. N., Froenicke, L., Hogel, J., Hameister, H., *et al.* (2009a) Gene synteny comparisons between different vertebrates provide new insights into breakage and fusion events during mammalian karyotype evolution. *Bmc Evolutionary Biology,* 9.

Kemkemer, C., Kohn, M., Kehrer-Sawatzki, H., Fundele, R. and Hameister, H. (2009b) Enrichment of brain-related genes on the mammalian X chromosome is ancient and predates the divergence of synapsid and sauropsid lineages. *Chromosome Res.*, 17, p811.

Kemkemer, C., Kohn, M., Kehrer-Sawatzki, H., Minich, P., Hogel, J., Froenicke, L., *et al.* (2006) Reconstruction of the ancestral ferungulate karyotype by electronic chromosome painting (E-painting). *Chromosome Res.*, 14, p899.

Kent, W. J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res.*, 12, p656.

Kirkness, E. F., Bafna, V., Halpern, A. L., Levy, S., Remington, K., Rusch, D. B., *et al.* (2003) The dog genome: Survey sequencing and comparative analysis. *Science*, 301, p1898.

Kishida, T., Kubota, S., Shirayama, Y. and 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, p428.

Kohn, M., Hogel, J., Vogel, W., Minich, P., Kehrer-Sawatzki, H., Graves, J. A. M., *et al.* (2006) Reconstruction of a 450-My-old ancestral vertebrate protokaryotype. *Trends Genet.*, 22, p203.

Kohn, M., Kehrer-Sawatzki, H., Vogel, W., Graves, J. A. M. and Hameister, H. (2004) Wide genome comparisons reveal the origins of the human X chromosome. *Trends Genet.*, 20, p598.

Kosiol, C., Vinar, T., da Fonseca, R. R., Hubisz, M. J., Bustamante, C. D., Nielsen, R., *et al.* (2008) Patterns of positive selection in six mammalian genomes. *Plos Genetics*, 4.

Kouprina, N., Mullokandov, M., Rogozin, I. B., Collins, N. K., Solomon, G., Otstot, J., *et al.* (2004) The SPANX gene family of cancer/testis-specific antigens: Rapid evolution and amplification in African great apes and hominids. *Proc. Natl. Acad. Sci. U. S. A.*, 101, p3077.

Lahn, B. T. and Page, D. C. (1999) Four evolutionary strata on the human X chromosome. *Science*, 286, p964.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, 409, p860.

Larkin, D. M., Everts-van der Wind, A., Rebeiz, M., Schweitzer, P. A., Bachman, S., Green, C., *et al.* (2003) A cattle-human comparative map built with cattle BAC-ends and human genome sequence. *Genome Res.*, 13, p1966.

Larkin, D. M., Pape, G., Donthu, R., Auvil, L., Welge, M. and Lewin, H. A. (2009) Breakpoint regions and homologous synteny blocks in chromosomes have different evolutionary histories. *Genome Res.*, 19, p770.

Le, S. Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, 25, p1307.

Lefevre, C. M., Digby, M. R., Whitley, J. C., Strahm, Y. and Nicholas, K. R. (2007) Lactation transcriptomics in the Australian marsupial, Macropus eugenii: transcript sequencing and quantification. *Bmc Genomics*, 8.

Lemaitre, C., Zaghloul, L., Sagot, M. F., Gautier, C., Arneodo, A., Tannier, E., *et al.* (2009) Analysis of fine-scale mammalian evolutionary breakpoints provides new insight into their relation to genome organisation. *Bmc Genomics*, 10.

Lewin, H. A., Larkin, D. M., Pontius, J. and O'Brien, S. J. (2009) Every genome sequence needs a good map. *Genome Res.*

Li, H. (2006), TreeBeST: Tree building guided by species tree, http://treesoft.sourceforge.net/treebest.shtml

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J.-K., Osmotherly, L., *et al.* (2006) TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucl. Acids Res.*, 34, pD572.

Li, L., Stoeckert, C. J. and Roos, D. S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13, p2178.

Libants, S., Carr, K., Wu, H., Teeter, J., Chung-Davidson, Y.-W., Zhang, Z., *et al.* (2009) The sea lamprey Petromyzon marinus genome reveals the early origin of several chemosensory receptor families in the vertebrate lineage. *BMC Evolutionary Biology*, 9, p180.

Lomvardas, S., Barnea, G., Pisapia, D. J., Mendelsohn, M., Kirkland, J. and Axel, R. (2006) Interchromosomal interactions and olfactory receptor choice. *Cell*, 126, p403.

Longo, M. S., Carone, D. M., Green, E. D., O'Neill, M. J., O'Neill, R. J., Progra, N. C. S., *et al.* (2009) Distinct retroelement classes define evolutionary breakpoints demarcating sites of evolutionary novelty. *Bmc Genomics*, 10.

Lyons, L. A., Laughlin, T. F., Copeland, N. G., Jenkins, N. A., Womack, J. E. and O'Brien, S. J. (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nat. Genet.*, 15, p47.

Ma, L., G, B., Np, B., R, C., Pa, M., H, M., *et al.* (2007) ClustalW and ClustalX version 2.0. *Bioinformatics*, pbtm404.

Malde, K. (2008) The effect of sequence quality on sequence alignment. *Bioinformatics*, 24, p897.

Malnic, B., Godfrey, P. A. and Buck, L. B. (2004) The human olfactory receptor gene family. *Proc. Natl. Acad. Sci. U. S. A.*, 101, p2584.

Man, O., Willhite, D. C., Crasto, C. J., Shepherd, G. M. and Gilad, Y. (2007) A framework for exploring functional variability in olfactory receptor genes. *PLoS One*, 2, pe682.

Mardis, E. R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, 24, p133.

Matsuda, Y., Nishida-Umehara, C., Tarui, H., Kuroiwa, A., Yamada, K., Isobe, T., *et al.* (2005) Highly conserved linkage homology between birds and turtles: Bird and turtle chromosomes are precise counterparts of each other. *Chromosome Res.*, 13, p601.

Maxam, A. M. and Gilbert, W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U. S. A.*, 74, p560.

McKenzie, L. M., Collet, C. and Cooper, D. W. (1993) Use of a subspecies cross for efficient development of a linkage map for a marsupial mammal, the tammar wallaby (*Macropus eugenii*). *Cytogenet. Cell Genet.*, 64, p264.

McKenzie, L. M., Poole, W. E., Collet, C. and Cooper, D. W. (1995) Higher female than male recombination rates in a marsupial mammal, the tammar wallaby (*Macropus eugenii*). *Cytogenet. Cell Genet.*, 68, p64.

Meyer, A. and Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, 27, p937.

Mezler, M., Fleischer, J. and Breer, H. (2001) Characteristic features and ligand specificity of the two olfactory receptor classes from Xenopus laevis. *J. Exp. Biol.*, 204, p2987.

Mikkelsen, T. S., Wakefield, M. J., Aken, B., Amemiya, C. T., Chang, J. L., Duke, S., *et al.* (2007) Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature,* 447, p167.

Miura, I. (2007) An evolutionary witness: the frog *Rana rugosa* underwent change of heterogametic sex from XY male to ZW female. *Sexual Development,* 1, p323.

Moran, C. and James, J. W. (2005) Linkage mapping. IN RUVINSKY, A. and GRAVES, J. A. M. (Eds.) *Mammalian Genomics.* Cambridge, CABI Publishing.

Muller, H. J. (1918) Genetic variability, twin hybrids and constant hybrids, in a case of balanced lethal factors. *Genetics,* 3, p422.

Murphy, W. J., Davis, B., David, V. A., Agarwala, R., Schaffer, A. A., Wilkerson, A. J. P., *et al.* (2007) A 1.5-Mb-resolution radiation hybrid map of the cat genome and comparative analysis with the canine and human genomes. *Genomics,* 89, p189.

Murphy, W. J., Larkin, D. M., der Wind, A. E.-v., Bourque, G., Tesler, G., Auvil, L., *et al.* (2005) Dynamics of mammalina chromosome evolution inferred from multispecies comparative maps. *Science,* 309, p613.

Nadeau, J. H. and Taylor, B. A. (1984) Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. U. S. A.,* 81, p814.

Nanda, I., Shan, Z., Schartl, M., Burt, D. W., Koehler, M., Nothwang, H.-G., *et al.* (1999) 300 million years of conserved synteny between chicken Z and human chromosome 9. *Nat. Genet.,* 21, p258.

National Centre for Biotechnology Information Trace Archives, ftp://ftp.ncbi.nih.gov/pub/TraceDB/, National Library of Medicine and National Institutes of Health

National Human Genome Research Institute, http://www.genome.gov, Last accessed in 2009

Nei, M. and Rooney, A. P. (2005) Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.,* 39, p121.

Ng, M. P., Vergara, I. A., Frech, C., Chen, Q. K., Zeng, X. H., Pei, J., *et al.* (2009) OrthoClusterDB: an online platform for synteny blocks. *Bmc Bioinformatics,* 10.

Niimura, Y. (2009) On the origin and evolution of vertebrate olfactory receptor genes: comparative analysis among 23 chordate species. *Genome Biol Evol,* 2009, p34.

Niimura, Y. and Nei, M. (2005a) Evolutionary changes of the number of olfactory receptor genes in the human and mouse lineages. *Gene,* 346, p23.

Niimura, Y. and Nei, M. (2005b) Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc. Natl. Acad. Sci. U. S. A.,* 102, p6039.

Niimura, Y. and Nei, M. (2006) Evolutionary dynamics of olfactory and other chemosensory receptor genes in vertebrates. *J. Hum. Genet.,* 51, p505.

Niimura, Y. and Nei, M. (2007) Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLoS ONE,* 2, pe708.

Nozawa, M., Kawahara, Y. and Nei, M. (2007) Genomic drift and copy number variation of sensory receptor genes in humans. *Proc. Natl. Acad. Sci. U. S. A.,* 104, p20421.

O'Brien, K. P., Remm, M. and Sonnhammer, E. L. L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucl. Acids Res.,* 33, pD476.

O'Brien, S. J., Menotti-Raymond, M., Murphy, W. J., Nash, W. G., Wienberg, J., Stanyon, R., *et al.* (1999) The promise of comparative genomics in mammals. *Science,* 286, p458.

O'Brien, S. J. and Nash, W. G. (1982) Genetic mapping in mammals: chromosome map of domestic cat. *Science,* 216, p257.

O'Brien, S. J., Wienberg, J. and Lyons, L. A. (1997) Comparative genomics: lessons from cats. *Trends Genet.,* 13, p393.

Oelschlager, H. H. and Kemp, B. (1998) Ontogenesis of the sperm whale brain. *J. Comp. Neurol.,* 399, p210.

Ohno, S. (1967) *Sex chromosome and sex-linked genes* edited by Springer-Verlag, Berlin,

Oldham, W. M. and Hamm, H. E. (2006) Structural basis of function in heterotrimeric G proteins. *Q. Rev. Biophys.,* 39, p117.

Oldham, W. M. and Hamm, H. E. (2008) Heterotrimeric G protein activation by G-protein-coupled receptors. *Nat Rev Mol Cell Biol,* 9, p60.

Olender, T., Feldmesser, E., Atarot, T., Eisenstein, M. and Lancet, D. (2004) The olfactory receptor universe--from whole genome analysis to structure and evolution. *Genet Mol Res,* 3, p545.

Ospina-Ålvarez, N. and Piferrer, F. (2008) Temperature-dependent sex determination in fish revisited: Prevalence of a single sex ratio response pattern, and possible effects of climate change. *PLoS ONE,* 3, pe2837.

Pan, X. K., Stein, L. and Brendel, V. (2005) SynBrowse: a synteny browser for comparative sequence analysis. *Bioinformatics,* 21, p3461.

Papasaikas, P. K., Bagos, P. G., Litou, Z. I., Promponas, V. J. and Hamodrakas, S. J. (2004) PRED-GPCR: GPCR recognition and family classification server. *Nucleic Acids Res,* 32, pW380.

Pearson, W. R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.,* 24, p307.

Pearson, W. R. and Lipman, D. J. (1988) Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. U. S. A.,* 85, p2444.

Peterson, K. J., Lyons, J. B., Nowak, K. S., Takacs, C. M., Wargo, M. J. and McPeek, M. A. (2004) Estimating metazoan divergence times with a molecular clock. *Proc. Natl. Acad. Sci. U. S. A.,* 101, p6536.

Pettigrew, J. D. (1999) Electroreception in monotremes. *J. Exp. Biol.,* 202, p1447.

Pevzner, P. and Tesler, G. (2003a) Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. *Genome Res.,* 13, p37.

Pevzner, P. and Tesler, G. (2003b) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proc. Natl. Acad. Sci. U. S. A.,* 100, p7672.

Phan, I. Q., Pilbout, S. F., Fleischmann, W. and Bairoch, A. (2003) NEWT, a new taxonomy portal. *Nucleic Acids Res,* 31, p3822.

Piller, K. R. and Bart, H. L., Jr. (2009) Incomplete sampling, outgroups, and phylogenetic inaccuracy: a case study of the Greenside Darter complex (Percidae: Etheostomablennioides). *Mol. Phylogenet. Evol.,* 53, p340.

Pontius, J. U., Mullikin, J. C., Smith, D. R., Lindblad-Toh, K., Gnerre, S., Clamp, M., *et al.* (2007) Initial sequence and comparative analysis of the cat genome. *Genome Res.,* 17, p1675.

Pop, M., Salzberg, S. L. and Shumway, M. (2002) Genome sequence assembly: Algorithms and issues. *Computer,* 35, p47.

Prober, J. M., Trainor, G. L., Dam, R. J., Hobbs, F. W., Robertson, C. W., Zagursky, R. J., *et al.* (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science,* 238, p336.

Quignon, P., Giraud, M., Rimbault, M., Lavigne, P., Tacher, S., Morin, E., *et al.* (2005) The dog and rat olfactory receptor repertoires. *Genome Biology,* 6, pR83.

Quinn, A. E., Georges, A., Sarre, S. D., Guarino, F., Ezaz, T. and Graves, J. A. (2007) Temperature sex reversal implies sex gene dosage in a reptile. *Science,* 316, p411.

Rastogi, S. and Liberles, D. A. (2005) Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *Bmc Evolutionary Biology,* 5.

Raudsepp, T., Kata, S. R., Piumi, F., Swinburne, J., Womack, J. E., Skow, L. C., *et al.* (2002) Conservation of gene order between horse and human X chromosomes as evidenced through radiation hybrid mapping. *Genomics,* 79, p451.

Remm, M., Storm, C. E. and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.,* 314, p1041.

Rens, W., Grutzner, F., O'Brien P, C., Fairclough, H., Graves, J. A. and Ferguson-Smith, M. A. (2004) Resolution and evolution of the duck-billed platypus karyotype with an X1Y1X2Y2X3Y3X4Y4X5Y5 male sex chromosome constitution. *Proc. Natl. Acad. Sci. U. S. A.,* 101, p16257.

Rens, W., O'Brien, P. C. M., Fairclough, H., Harman, L., Graves, J. A. M. and Ferguson-Smith, M. A. (2003) Reversal and convergence in marsupial chromosome evolution. *Cytogenetic and Genome Research,* 102, p282.

Rens, W., O'Brien, P. C. M., Yang, F., Graves, J. A. M. and Ferguson-Smith, M. A. (1999) Karyotype relationships between four distantly related marsupials revealed by reciprocal chromosome painting. *Chromosome Res.,* 7, p461.

Replogle, K., Arnold, A., Ball, G., Band, M., Bensch, S., Brenowitz, E., *et al.* (2008) The Songbird Neurogenomics (SoNG) Initiative: Community-based tools and strategies for study of brain gene function and evolution. *BMC Genomics,* 9, p131.

Rest, J. S., Ast, J. C., Austin, C. C., Waddell, P. J., Tibbetts, E. A., Hay, J. M., *et al.* (2003) Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Mol. Phylogenet. Evol.,* 29, p289.

Rice, W. R. (1987) Genetic hitchhiking and the evolution of reduced activity of the Y sex chromosome. *Genetics,* 116, p161.

Robertson, H. M. and Thomas, J. H. (2006) The putative chemoreceptor families of C. elegans. *WormBook,* p1.

Robertson, H. M. and Wanner, K. W. (2006) The chemoreceptor superfamily in the honey bee, Apis mellifera: Expansion of the odorant, but not gustatory, receptor family. *Genome Res.,* 16, p1395.

Robertson, H. M., Warr, C. G. and Carlson, J. R. (2003) Molecular evolution of the insect chemoreceptor gene superfamily in Drosophila melanogaster. *Proc. Natl. Acad. Sci. U. S. A.,* 100, p14537.

Rodriguez Delgado, C. L., Waters, P. D., Gilbert, C., Robinson, T. J. and Graves, J. A. (2009) Physical mapping of the elephant X chromosome: conservation of gene order over 105 million years. *Chromosome Res.*

Rosenberg, M. S. and Kumar, S. (2001) Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc. Natl. Acad. Sci. U. S. A.,* 98, p10751.

Ross, M. T., Grafham, D. V., Coffey, A. J., Scherer, S., McLay, K., Muzny, D., *et al.* (2005) The DNA sequence of the human X chromosome. *Nature,* 434, p325.

Ross, M. T., LaBrie, S., McPherson, J. and Stanton, V. P. (1999) Screening large-insert libraries by hybridization. *Current Protocols in Human Genetics.* John Wiley & Sons, Inc.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J. M., Guo, Y., *et al.* (2007) TreeFam: 2008 Update. *Nucl. Acids Res.,* pgkm1005.

Ruiz-Herrera, A., Castresana, J. and Robinson, T. J. (2006) Is mammalian chromosomal evolution driven by regions of genome fragility? *Genome Biology,* 7.

Saccone, S., Caccio, S., Kusuda, J., Andreozzi, L. and Bernardi, G. (1996) Identification of the gene-richest bands in human chromosomes. *Gene,* 174, p85.

Saifi, G. M. and Chandra, H. S. (1999) An apparent excess of sex- and reproduction-related genes on the human X chromosome. *Proceedings of the Royal Society of London Series B-Biological Sciences,* 266, p203.

Samadashwily, G. M., Dayn, A. and Mirkin, S. M. (1993) Suicidal nucleotide sequence for DNA polymerization. *EMBO J.,* 12, p4975.

Samollow, P. B., Gouin, N., Miethke, P., Mahaney, S. M., Kenney, M., VandeBerg, J. L., *et al.* (2007) A microsatellite-based, physically anchored linkage map for the gray, short-tailed Opossum (Monodelphis domestica). *Chromosome Res.,* 15, p269.

Samollow, P. B., Kammerer, C. M., Mahaney, S. M., Schneider, J. L., Westenberger, S. J., VandeBerg, J. L., *et al.* (2004) First-generation linkage map of the gray, short-tailed opossum, Monodelphis domestica, reveals genome-wide reduction in female recombination rates. *Genetics,* 166, p307.

Sanger, F. and Coulson, A. R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.,* 94, p441.

Sanger, F., Nicklen, S. and Coulson, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.,* 74, p5463.

Sarre, S. D., Georges, A. and Quinn, A. (2004) The ends of a continuum: genetic and temperature-dependent sex determination in reptiles. *Bioessays,* 26, p639.

Sasaki, T., Shimizu, A., Ishikawa, S. K., Imai, S., Asakawa, S., Murayama, Y., *et al.* (2007) The DNA sequence of medaka chromosome LG22. *Genomics,* 89, p124.

Sasson, O., Kaplan, N. and Linial, M. (2006) Functional annotation prediction: All for one and one for all. *Protein Sci.,* 15, p1557.

Schibler, L., Roig, A., Mahe, M. F., Laurent, P., Hayes, H., Rodolphe, F., *et al.* (2006) High-resolution comparative mapping among man, cattle and mouse suggests a role for repeat sequences in mammalian genome evolution. *Bmc Genomics,* 7.

Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., *et al.* (2000) PipMaker - A Web server for aligning two genomic DNA sequences. *Genome Res.,* 10, p577.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.,* 26, p1135.

Shepherd, G. M. (1972) Synaptic organization of the mammalian olfactory bulb. *Physiol. Rev.,* 52, p864.

Shetty, S., Griffin, D. K. and Graves, J. A. M. (1999) Comparative painting reveals strong homology over 80 million years of bird evolution. *Chromosome Res.,* 7, p289.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B. L., Smith, M. J., *et al.* (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature,* 346, p240.

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature,* 423, p825.

Smith, C. A., Roeszler, K. N., Hudson, Q. J. and Sinclair, A. H. (2007) Avian sex determination: what, when and where? *Cytogenetic and Genome Research,* 117, p165.

Smith, C. A., Roeszler, K. N., Ohnesorg, T., Cummins, D. M., Farlie, P. G., Doran, T. J., *et al.* (2009) The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature.*

Spencer, J. A., Sinclair, A. H., Watson, J. M. and Marshall Graves, J. A. (1991a) Genes on the short arm of the human X chromosome are not shared with the marsupial X. *Genomics,* 11, p339.

Spencer, J. A., Watson, J. M. and Graves, J. A. M. (1991b) The X chromosome of marsupials shares a highly conserved region with eutherians. *Genomics,* 9, p598.

Srivastava, P. K., Desai, D. K., Nandi, S. and Lynn, A. M. (2007) HMM-ModE - improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. *BMC Bioinformatics,* 8, p104.

St George-Hyslop, P., Haines, J., Rogaev, E., Mortilla, M., Vaula, G., Pericak-Vance, M., *et al.* (1992) Genetic evidence for a novel familial Alzheimer's disease locus on chromosome 14. *Nat. Genet.,* 2, p330.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.,* 12, p1611.

Steiger, S. S., Fidler, A. E. and Kempenaers, B. (2009a) Evidence for increased olfactory receptor gene repertoire size in two nocturnal bird species with well-developed olfactory ability. *BMC Evol Biol,* 9, p117.

Steiger, S. S., Fidler, A. E., Mueller, J. C. and Kempenaers, B. (2009b) Evidence for adaptive evolution of olfactory receptor genes in 9 bird species. *J. Hered.*

Steiger, S. S., Fidler, A. E., Valcu, M. and Kempenaers, B. (2008) Avian olfactory receptor gene repertoires: evidence for a well-developed sense of smell in birds? *Proceedings of the Royal Society B: Biological Sciences,* 275, p2309.

Stevenson, B., Iseli, C., Panji, S., Zahn-Zabal, M., Hide, W., Old, L., *et al.* (2007) Rapid evolution of cancer/testis genes on the X chromosome. *BMC Genomics,* 8, p129.

Sturtevant, A. H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *J. Exp. Zool.,* 14, p43.

Swingley, W. D., Blankenship, R. E. and Raymond, J. (2008) Integrating markov clustering and molecular phylogenetics to reconstruct the cyanobacterial species tree from conserved protein families. *Mol. Biol. Evol.,* 25, p643.

Tatusov, R., Fedorova, N., Jackson, J., Jacobs, A., Kiryutin, B., Koonin, E., *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics,* 4, p41.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.,* 25, p4876.

Vallender, E. J. and Lahn, B. T. (2004) How mammalian sex chromosomes acquired their peculiar gene content. *Bioessays,* 26, p159.

van Dongen, S. (2000), Graph clustering by flow simulation, PhD, Centre for Mathematics and Computer Science, Universiteit Utrecht, Amsterdam

Van Dongen, S. (2008) Graph clustering via a discrete uncoupling process. *Siam Journal on Matrix Analysis and Applications,* 30, p121.

Vanoorschot, R. A. H., Porter, P. A., Kammerer, C. M. and Vandeberg, J. L. (1992) Severly reduced recombination in females of Sounth-American marsupial *Monodelphis domestica. Cytogenet. Cell Genet.,* 60, p64.

Venta, P. J., Brouillette, J. A., YuzbasiyanGurkan, V. and Brewer, G. J. (1996) Gene-specific universal mammalian sequence-tagged sites: Application to the canine genome. *Biochem. Genet.,* 34, p321.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., *et al.* (2001) The sequence of the human genome. *Science,* 291, p1304.

Versteeg, R., van Schaik, B. D. C., van Batenburg, M. F., Roos, M., Monajemi, R., Caron, H., *et al.* (2003) The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes. *Genome Res.,* 13, p1998.

Veyrunes, F., Waters, P. D., Miethke, P., Rens, W., McMillan, D., Alsop, A. E., *et al.* (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res.,* 18, p965.

Vilella, A. J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.,* 19, p327.

Volff, J. N., Nanda, I., Schmid, M. and Schartl, M. (2007) Governing sex determination in fish: regulatory putsches and ephemeral dictators. *Sexual Development,* 1, p85.

Warner, D. A., Lovern, M. B. and Shine, R. (2007) Maternal nutrition affects reproductive output and sex allocation in a lizard with environmental sex determination. *Proceedings of the Royal Society B: Biological Sciences,* 274, p883.

Warren, W. C., Hillier, L. W., Graves, J. A. M., Birney, E., Ponting, C. P., Gr√ºtzner, F., *et al.* (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature,* 453, p175.

Warrington, J. A. and Bengtsson, U. (1994) High-resolution physical mapping of human 5q31-q33 using three methods: radiation hybrid mapping, interphase fluorescence in situ hybridization, and pulsed-field gel electrophoresis. *Genomics,* 24, p395.

Waters, P. D., Delbridge, M. L., Deakin, J. E., El-Mogharbel, N., Kirby, P. J., Carvalho-Silva, D. R., *et al.* (2005) Autosomal location of genes from the conserved mammalian X in the platypus (Ornithorhynchus anatinus): implications for mammalian sex chromosome evolution. *Chromosome Res.,* 13, p401.

Wienberg, J. and Stanyon, R. (1997) Comparative painting of mammalian chromosomes. *Curr. Opin. Genet. Dev.,* 7, p784.

Wind, A. E.-v. d., Larkin, D. M., Green, C. A., Elliott, J. S., Olmstead, C. A., Chiu, R., *et al.* (2005) A high-resolution whole-genome cattle,Äihuman comparative map reveals details of mammalian chromosome evolution. *Proc. Natl. Acad. Sci. U. S. A.,* 102, p18526.

Womack, J. E. and Moll, Y. D. (1986) Gene map of the cow: conservation of linkage with mouse and man. *J. Hered.,* 77, p2.

Wu, R. and Taylor, E. (1971) Nucleotide sequence analysis of DNA. II. Complete nucleotide sequence of the cohesive ends of bacteriophage lambda DNA. *J. Mol. Biol.,* 57, p491.

Yang, F., Alkalaeva, E. Z., Perelman, P. L., Pardini, A. T., Harrison, W. R., O'Brien, P. C. M., *et al.* (2003) Reciprocal chromosome painting among human, aardvark, and elephant (superorder Afrotheria) reveals the likely eutherian ancestral karyotype. *Proc. Natl. Acad. Sci. U. S. A.,* 100, p1062.

Zechner, U., Wilda, M., Kehrer-Sawatzki, H., Vogel, W., Fundele, R. and Hameister, H. (2001) A high density of X-linked genes for general cognitive ability: a run-away process shaping human evolution? *Trends Genet.,* 17, p697.

Zenger, K. R., McKenzie, L. M. and Cooper, D. W. (2002) The first comprehensive genetic linkage map of a marsupial: The tammar wallaby (*Macropus eugenii*). *Genetics,* 162, p321.

Zhang, X. and Firestein, S. (2002) The olfactory receptor gene superfamily of the mouse. *Nat. Neurosci.,* 5, p124.

Zhang, Z., Schwartz, S., Wagner, L. and Miller, W. (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.,* 7, p203.

8 Appendix

8.1 Local pairwise sequence alignment and searches

Pairwise sequence searches enable to identify homologous sequences from a database of nucleotide or protein sequences. Currently popular pairwise sequence searching programs include BLAST (Altschul *et al.* 1990, Altschul *et al.* 1997), BLAT (Kent 2002) and FASTA (Pearson 1994, Pearson and Lipman 1988). All these programs aim at identifying nearly optimal alignments between a query sequence and database sequences by using heuristics. The workflow of these programs begins with identifying exact matches of specified length in the sequence database. Based on the proximity and number of such exact matches found with a database sequence, an alignment is generated between a query sequence and a database sequence. The pairwise sequence alignments resulting in a score above the specified threshold score are then reported as results. Scoring parameters for match, mismatch, insertion and deletion can be adjusted along with the initial exact match length to increase the specificity and sensitivity of alignments, however, doing so can also increase the processing time.

8.2 Multiple sequence alignments

Multiple sequence alignments for nucleotide and protein sequences are routinely used to understand the evolutionary relationship between groups of sequences. MUSCLE (Edgar 2004a, Edgar 2004b) and CLUSTAL (Ma *et al.* 2007, Thompson *et al.* 1997) programs are widely used programs to locally align more than two sequences. The workflow for generating multiple sequence alignment begins with first obtaining nearly optimal pairwise sequence alignments. A pairwise alignment with the highest score is used as a template for adding more sequences to generate multiple sequence alignment. Subsequently, more sequences are added to the first pair such that each column of the multiple sequence alignment maintains a score above the threshold value. It is imperative that only homologous sequences are subjected to multiple sequence alignments as non-homologous sequences can introduce errors in the final alignment and hence misleading evolutionary relationship of included sequences.

8.3 Hidden Markov model (HMM) searches

Hidden Markov model in bioinformatics describes the Markov process of DNA sequence evolution constrained by selection pressure (Eddy 1998). The multiple sequence alignment of homologous sequences represents possible outcomes as a result of evolution. The transition state and emission probabilities of residues in each column of the multiple

sequence alignments can be used to generate a profile HMM of homologous sequences. These can subsequently be used to search DNA or protein sequence database to detect remote homologs.