Genomic characterization of developmental and adaptation genes: insights into vertebrate evolution

Siby Philip

PhD Thesis

CIIMAR,
Centro Interdisciplinar de Investigação Marinha e Ambiental,
Rua dos Bragas, n.289,
4050-123, Porto,
Portugal.

March, 2013.

Acknowledgements

First of all, I sincerely thank my supervisor, Agostinho Antunes, for his continuous support and guidance. Thank you, Agostinho, for your encouragement and for all the excellent comments on the results I presented and for the stimulating discussions which keep me motivated. My co-supervisor Prof. Vitor Vasconcelos, was always an encouraging presence, sometimes coming late to lab and seeing him in front of microscope would make me feel bad but at the same time make me understand the value of hard work and dedication.

João Paulo helped me a lot during the course of the PhD, he was there right from the beginning and introduced me to many things that I would face in Porto, in the lab and in science. Thank you JP for your support and thanks for the intellectual company during these four years. I wish to also thank my co-authors Filipe Castro, Rute Fonseca, Miguel Santos, Warren E. Johnson, Stephen J. O'Brien and Emanuel Maldonado for great suggestions and additions to the papers we jointly published, their contributions where so vital that sometimes I felt that *yes this is what I wanted to write*, and which made my life a lot easier, thanks all of you. Allen, Rajeev, Lijo, Mithun, Anvar, Neelesh, Josin, Fibin and all co-authors for the last two chapters should also be thanked for all their help, contribution and discussions.

Colleagues from LEGE, Zé Carlos, Pedro, Susana, Micaela, João Morais, Guillermin, Antonio, Marisa, Viviana and all others have helped me a lot, was kind to me and would make everything comfortable for work, thanks to all of you for your kind help. Two guys who need a special mention and better credit are Imran and Anoop, colleagues in LEGE and fellow Indians in Porto, we had common "enemies" in food, language and climate, we survived together here, their contribution has made this thesis possible and words would not suffice to thank them. I am sure that they would be happy to see this thesis, and I wish them all the best for their thesis as well.

Others who need special thanks are the members of Conservation Research Group, Cochin, Kerala. Rajeev, Anvar, Josin, Ambily, Krishnakumar, Fibin, Ramprashanth, Dr. Benno and all of them have been a constant help and inspiration. Time spent in the field sampling and writing papers was fruitful and have immensely helped me in my scientific career up until now. Help from them especially for the last two chapters of this thesis and the future works should be sincerely acknowledged.

Dr. Ralf Britz, from the Natural History Museum, London, a great scientist and inspiration

needs a special thanks while I write this. The valuable lessons learnt, about fishes, with him should help me in the future and it is a special pride for me that I could publish papers with him.

Lot of friends, back in India, would be proud and happy to know that I wrote a thesis at last. Jobin Jose, Allen, Renny, Priyesh, Mahesh, Arun, Rojan, Diljo, Binoj, Ranjini, Deepna, Sherin, Ashin, it is a long list, all of them have helped me a lot during my past and they need a special thanks. Special thanks also to Dr. Joselet, Shaji sir, Rajeesh sir, Dr. G. Prasad, my teachers and people who have been of immense support before and during this thesis.

Apart from all these special thanks should go to all those scientists who answered my questions about various methods and softwares that I used during this thesis work. A big thanks and support should go towards the open software and open-source movement, without which this thesis and work in this thesis would not have been possible, almost everything used to make this work and thesis a reality has been contributed by the open-source community and scientists who made their work freely available. I wish that one day I could significantly contribute to the community with my work.

My parents have been a constant source of inspiration and motivation. Whenever I rang back home, I was be asked about publications and progress, their knowledge, support and hard work has been instrumental for making this a reality. They instilled in me a habit of hard work and taught me to dream big. Since my schooldays, I was advised to study well and get into a professional college/course, and warned that, or else I would end up doing a PhD. I chose the harder path and am just fulfilling the first steps, their blessings and prayers were always with me. My brothers have been of support in making this a reality in their own way, sometimes collecting specimen, papers, letting me know research news from local press etc., their support have been immense and a big thanks to them. My family back in India, uncles, aunts, cousins and in-laws all have been a constant inspiration and a big thanks to them.

My wife, a constant source of inspiration to "write fast" and finish the PhD soon, needs a special mention and had a special role in maintaining my sanity while writing the thesis and doing the last couple of works. She understood my problems well, disturbed me when I was stuck in front of the computer, cheered me up and would let me know that I was alive. Words would not suffice to thank her.

Funding from the Portuguese foundation for science and technology (**FCT**), who funded my stay and study in Portugal, with a PhD grant **SFRH/BD/47938/2008** is sincerely acknowledged.

This thesis is dedicated to my parents.

Contents

List of Tables

List of Figures

Abstract

In this thesis molecular evidences related with the adaptive evolution of vertebrates are presented. In the first part, we studied the retinoid X receptors (RXR), rhodopsin 1 (RH1) and superoxide dismutases (SOD) and the results highlighted adaptive evolution as an important mechanism during the evolution of these essential development and adaptation related vertebrate genes. In the second part of the thesis, evidence is provided on how adaptive evolution may influence teleost fish radiations, namely with studies of the evolutionary diversification of 34 lateral line development genes, and two detailed case studies on the radiations of the endangered red lined torpedo barbs and the malabar snakehead fish, both endemic to the Western Ghats of India.

Results

In the second chapter we studied the RXR's which are transcription factors with important roles in development, reproduction, homeostasis, and cell differentiation. Different types of vertebrate RXRs (α, β and γ) have arisen from multiple duplication events. Here, we investigated various aspects of vertebrate RXR evolution. Codon based tests of positive selection identified that RXR was under significant positive selection immediately after the whole genome duplications in vertebrates. Amino acid based rate shift analysis also revealed significant rate shifts immediately after the whole genome duplications and functional divergence between all the pairs of RXRs. However, the extant RXR genes are highly conserved, particularly the helix involved in dimerization and the DNA-binding domain, but positively selected sites can nevertheless be found in domains involved in the RXR regulation.

In the third chapter, we used complementary codon based positive selection analysis in conjunction with amino acid physicochemical property based analysis to identify positive selection in the teleost RH1. By using new and powerful methods we could identify 30% of the sites involved in teleost rhodopsin's spectral tuning to be under positive selection. We also found that all the sites involved in the spectral tuning of the protein to be evolving non-neutrally and that those were the sites that could tolerate more number of substitutions. The presence of positive selection in 20% of the protein length, raises the question if the use of the rhodopsin marker would be suitable for phylogenetic inference at the taxonomic level of teleostei? However,

we find that the base composition and the codon bias of rhodopsin sequences from different superorders overlap, which makes them still a reliable marker for phylogenetic studies.

In the fourth chapter, we studied the SOD genes from a comparative genomic perspective. We analyzed the three SOD paralogs separately in 46 avian and non-avian reptilian genomes in comparison to 30 mammalian genomes. Our codon based positive selection analysis could identify positive selection in two out of the three SOD genes in birds. The avian SOD's were evolving at a higher evolutionary rate when compared to either mammals or reptiles, as evidenced by our base-by-base conservation-acceleration analysis and dN/dS based compartmentalization analysis.

In the fifth chapter, we conducted evolutionary genomic analyzes of 34 genes associated with lateral line system development in fishes to elucidate the significance of contrasting evolutionary rates and changes in the protein coding sequences. We find that duplicated copies of these genes are preferentially retained in the teleost genomes, and that episodic events of positive selection have occurred in 22 of the 30 post-duplication branches. In general, teleost genes evolved at a faster rate relative to their tetrapod counterparts and the mutation rates of 26 of the 34 genes differed among teleosts and tetrapods. We conclude that following whole genome duplication, evolutionary rates and episodic events of positive selection on the lateral line system development genes might have been one of the factors favoring the subsequent adaptive radiation of teleosts into diverse habitats.

In the final two chapters we studied teleost fish speciations in two endemic and endangered/threatened taxa from the Western Ghats biodiversity hotspot in India. The red lined torpedo barbs (RLTBs) (Cyprinidae: Puntius) and the malabar snakehead fish *Channa diplogramma*.

The RLTBs endemic to the Western Ghats, are popular and highly priced freshwater aquarium fishes. Two decades of indiscriminate exploitation for the pet trade, restricted range, and continuing decline in quality of habitats has resulted in their ‘Endangered’ listing. Here, we determined the species boundaries of allopatric RLTB populations, and demonstrated the effect of geographic barriers on its diversification. Multivariate morphometric analysis using 24 size-adjusted characters could delineate all allopatric populations as distinct. Similarly, the species tree highlighted a phylogeny with 12 distinct RLTB lineages corresponding to each of the different riverine populations. Bayesian species delimitation and generalized mixed yule coalescence methods identified eight evolutionary distinct lineages. Divergence time analysis points to recent independent vicariance events around 5 million years ago, after the lineages were split into two ancestral stocks in the Paleocene, on north and south of a major geographical gap in the hill ranges of Western Ghats.

In the last chapter we studied the snakehead fishes *C. diplogramma* and *C. micropeltes*, often confused as a single species. Our morphometric and meristic analysis provided conclusive evidence to separate them as two distinct species. Number of caudal fin rays, lateral line

scales, scales below lateral line, total vertebrae, pre-anal length and body depth were the most prominent characters that can be used to differentiate both the species. Finally, the genetic distance between both species for the partial mitochondrial 16S rRNA and COI sequences is also well above the inter-specific genetic distances between nine other channid species compared in this study. The current distribution of *C. diplogramma* and *C. micropeltes* is best explained by vicariance. The significant variation in the key taxonomic characters and the results of the molecular marker analysis points towards an allopatric speciation event or vicariant divergence from a common ancestor, which molecular data suggests to have occurred as early as 21.76 million years ago. The resurrection of *C. diplogramma* from the synonymy of *C. micropeltes* has hence been confirmed 146 years after its initial description and 134 years after it was synonymised, establishing it is an endemic species of peninsular India and prioritizing its conservation value.

Conclusions

In short, different methods to study adaptive evolution both at the gene (codon) level and protein level were employed revealing the prevalence of positive selection in the genes studied. In the case of duplicated genes, episodic events of positive selection and evolutionary rate alteration on the ancestral paralogs and functional divergence between the paralogous proteins have been revealed. We propose that these episodic signatures of adaptive evolution and functional divergence immediately after duplication might be one of the main mechanisms by which the paralogs escape from adaptive conflict in the organism. The fish lateral line, which facilitates a sense of “touch at a distance” is genomically characterized here. We check for patterns of adaptive evolution in the genes that have contributed to the lateral line development, which in turn might have an important role in facilitating the species radiations. We also highlight two cases of teleost radiations, which might have been due to ecological divergence that occurred during vicariance. Our use of complementary morphological and molecular methods along with state-of-the-art phylogenetic analysis could reveal hidden diversity, which should aid in devising better conservation plans for the species.

Overall this thesis provide evidence of pervasive episodic adaptive evolution and evolutionary rate variations in vertebrate genes and gene families like RXRs, SODs, RH1 and different genes related to the lateral line system development. Adaptive evolution of RH1, SODs or the lateral line genes might have enabled the species to exploit the "ecological opportunities" presented to them during evolution. We have also revealed the existence of hitherto undescribed diversity in teleost fishes, which arose as a result of vicariance related diversification by exploiting novel opportunities in new habitats.

Resumo

Nesta tese são apresentadas evidências moleculares relacionadas com a evolução adaptativa dos vertebrados. Na primeira parte, foram estudados os receptores X do ácido retinóico (RXR), rodopsina (RH1) e superóxido dismutase (SOD), e os resultados sugerem a evolução adaptativa como um mecanismo importante na evolução de genes envolvidos no desenvolvimento e adaptação dos vertebrados. Na segunda parte da tese, são fornecidas evidências sobre como a evolução adaptativa pode influenciar as radiações de peixes teleósteos, nomeadamente com o estudo da diversificação evolutiva de 34 genes envolvidos no desenvolvimento da linha lateral, e dois estudos detalhados sobre as radiações do barbo torpedo de linhas vermelhas e o peixe cabeça de cobra, ambas espécies endémicas do Ghats Ocidental da Índia e em perigo de extinção.

Resultados

No segundo capítulo, foram estudados os RXRs, que são factores de transcrição com funções importantes na reprodução, desenvolvimento, homeostase e diferenciação celular. Os diferentes tipos de RXRs nos vertebrados (α, β e γ) surgiram de eventos múltiplos de duplicação. Aqui foram investigados os diferentes aspectos da evolução do RXR em vertebrados. Testes de selecção positiva baseados na análise de codões sugeriram que o RXR esteve sob selecção positiva imediatamente após as duplicações completas do genoma dos vertebrados. A análise da variação de aminoácidos também revelou significativas alterações imediatamente após as duplicações completas do genoma e divergência funcional entre todos os pares de RXRs. No entanto, os genes de RXR existentes são altamente conservados, especialmente a hélice envolvida na dimerização e o domínio de ligação ao ADN, mas os locais positivamente seleccionadas podem, contudo, ser encontrados em domínios envolvidos na regulação do RXR.

No terceiro capítulo, foram utilizadas análises complementares de codões em conjunto com análises da variação das propriedades físico-química de amino ácidos para identificar selecção positiva no gene RH1 em peixes teleósteos. A utilização de metodologias recentes e robustas permitiram identificar que, 30% dos locais envolvidos no ajuste espectral da rodopsina em teleósteos estão sob a influência de selecção positiva. Descobrimos também que todos os sites envolvidos no ajuste espectral da proteína estão a evoluir não-neutralmente e que esses

foram os locais que melhor poderiam tolerar um maior número de substituições. A presença de selecção positiva em 20% do comprimento da proteína, levanta a questão se seria adequado a utilização da rodopsina como marcador genético para efectuar inferências filogenéticas ao nível dos teleostei? No entanto, descobrimos que a composição de bases e o enviesamento de codões das sequências de rodopsina sobrepõem-se, o que as torna fiáveis para estudos filogenéticos.

No quarto capítulo, os genes SOD foram estudados numa perspectiva de genómica comparativa. Analisaram-se os três genes SOD parálogos separadamente em 46 genomas de aves e répteis, em comparação com 30 genomas de mamíferos. A análise de codões permitiu identificar selecção positiva em dois dos três genes SOD em aves. A taxa evolutiva da SOD nas aves foi mais rápida quando comparada tanto com os mamíferos como com os répteis, como evidenciado pela análise base-por-base de conservação-aceleração e análise de compartimentalização dN/dS.

No quinto capítulo, foram realizadas análises genómicas evolutivas em 34 genes associados ao desenvolvimento do sistema da linha lateral em peixes para elucidar a importância das taxas evolutivas e das alterações das sequências que codificam proteínas. Verificou-se que as cópias desses genes duplicados foram preferencialmente retidos nos genomas dos teleósteos, e que os eventos episódicos de selecção positiva ocorreram em 22 dos 30 ramos pós-duplicação. Em geral, os genes de teleósteos evoluíram a uma taxa mais rápida relativamente aos seus homólogos em tetrápodes e as taxas de mutação de 26 dos 34 genes diferiu entre teleósteos e tetrápodes. Conclui-se que depois da duplicação completa do genoma, as taxas evolutivas e os eventos episódicos de selecção positiva nos genes de desenvolvimento do sistema da linha lateral podem ter estado entre os factores que favorecem a posterior radiação adaptativa de teleósteos em diversos habitats.

Nos dois últimos capítulos, foi estudada a especiação em espécies endémicas de peixes teleósteos ameaçadas de extinção no Ghats Ocidental na Índia: o barbo torpedo de linhas vermelhas (RLTBs) (Cyprinidae: Puntius) e o peixe cabeça de cobra *Channa diplogramma*.

Os RLTBs endémicos do Ghats Oriental são peixes de aquário de água doce muito populares. Duas décadas de exploração indiscriminada para o comércio de animais de estimação e contínuo declínio na qualidade dos habitats resultou na listagem desta espécie como "em vias de extinção". Neste trabalho, foram determinados os limites específicos das populações RLTB alopátricas, e demonstrado o efeito das barreiras geográficas na sua diversificação. A análise morfométrica multivariada utilizando 24 caracteres de tamanho ajustados permitiram delinear todas as populações alopátricas como distintas. Da mesma forma, a árvore de espécies destacou uma filogenia com 12 linhagens RLTB distintas correspondentes a cada uma das diferentes populações ribeirinhas. A delimitação Bayesiana de espécies e métodos mistos de coalescência generalizadas identificaram oito linhagens evolutivas distintas. A análise de divergência temporal sugere que os recentes eventos independentes de vicariância ocorreram

à cerca de 5 milhões de anos atrás, depois das linhagens se terem divido em dois grupos ancestrais no Paleoceno, a norte e a sul de uma grande lacuna geográfica definida pelos limites das montanhas do Ghats Ocidental.

No último capítulo, foram estudados os peixes cabeça de cobra *C. diplogramma* e *C. micropeltes*, muitas vezes confundidos como uma única espécie. A análise morfométrica e merística forneceu provas conclusivas para a separação das duas espécies como distintas. O número de raios da barbatana caudal, escamas da linha lateral, escamas abaixo da linha lateral, vértebras totais, comprimento pré-anal e profundidade corporal foram os caracteres mais proeminentes utilizados para diferenciar as duas espécies. Finalmente, a distância genética entre as duas espécies para as sequências mitocondriais parciais do 16S rRNA e COI são bem acima das distâncias genéticas inter-específicas entre outras nove espécies de channideos comparados neste estudo. A actual distribuição de *C. diplogramma* e *C. micropeltes* é melhor explicada por vicariância. A variação significativa nos caracteres taxonómicos chave e os resultados das análise de marcadores moleculares indicam um evento de especiação alopátrica ou divergência vicariante de um ancestral comum, o que os dados moleculares sugerem ter ocorrido à cerca de 21,76 milhões de anos atrás. A ressurreição de *C. diplogramma* a partir da sinonímia de *C. micropeltes* foi aqui confirmada, portanto, 146 anos após sua descrição inicial e 134 anos depois de ter sido sinonimizada, afirmando-a como uma espécie endémica da península da Índia e valorizando o seu valor de conservação.

Conclusões

Em suma, diferentes métodos foram utilizados para estudar a evolução adaptativa simultaneamente ao nível do gene (codão) e da proteína, revelando a prevalência de selecção positiva nos genes estudados. No caso dos genes duplicados, foram identificados eventos episódicos de selecção positiva e de modificação da taxa de evolução dos parálogos ancestrais e divergência funcional entre as proteínas parálogas. Propôs-se que estas assinaturas de evolução adaptativa episódica e divergência funcional imediatamente após a duplicação poderão ser um dos principais mecanismos pelos quais os parálogos escaparam ao conflito adaptativo no organismo. A linha lateral dos peixes, o que proporciona uma sensação de "toque à distância" foi aqui caracterizada genomicamente. Foram verificados os padrões de evolução adaptativa nos genes relacionados com o desenvolvimento da linha lateral, que por sua vez podem ter tido um papel importante nas radiações de espécies. Destacamos também dois casos de radiações em teleósteos, em que divergência ecológica poderá ter ocorrido durante vicariância. O uso de métodos complementares morfológicos e moleculares, juntamente com análises filogenéticas sofisticadas poderão revelar diversidade biológica adicional, o que pode ajudar na elaboração de melhores planos de conservação específica.

Em geral, esta tese fornece evidências da evolução episódica adaptativa e as variações evolutivas em genes de vertebrados e famílias de genes como RXRs, SODs, RH1 e diferentes

genes relacionados com o desenvolvimento do sistema da linha lateral. Evolução adaptativa do RH1, SODs ou dos genes da linha lateral poderá ter habilitado as espécies para explorar novas oportunidades "ecológicas" que lhes foram apresentados durante a evolução. Foi também revelada a existência de diversidade não descrita até então em peixes teleósteos, que surgiram como resultado da diversificação por vicariância, permitindo explorar novas oportunidades em novos habitats.

1 Introduction

1.0.1. Background

The central theme in evolutionary biology is to answer the question of how species and the species diversity arise. Charles Darwin in his landmark treatise *"On the Origin of Species"* [1], invoked the theory of evolution by the means of *natural selection*. The successful establishment of an organism to a specific habitat or environment is attributed to phenotypes favoring its survival, selected among from various similar traits, during the course of generations. In evolutionary sense, these traits are called as adaptations or an adaptation is a trait favorable in the given environmental condition (e.g., lateral line system in fishes which are innovations responsible for its survival and radiation in aquatic environment). In short, traits selected for a specific environment, causes the “origin of species” or speciation which in turn, is responsible for the diversity of life.

Studies on the formation of the adaptive complexity (traits and its formation) or the mechanism of speciation/diversification enable us to explain the evolutionary mechanism. Early studies of evolutionary biology took advantage of comparative anatomy, morphology and physiology [2]. However, it should be noted that the evolutionary mechanism acts at different levels of an organism, and studies with traits are overtly complex, nevertheless huge progress in understanding the complex process of evolution, and relationship of organisms or its systematics, have been due to the early comparative studies of traits.

Until the beginning of last century it was not clear as to what composed the basic units of inheritance. Pioneering genetic studies, during the late 1800's and early 1900's, observed that

variations from parents were transmitted to offsprings as discrete traits through the gametes. The term gene was put forward by Johannsen in 1909 [3], to denote the “unit-factors” or “elements” in gametes. The collection of all genes in a gamete was called as the genotype [4]. The definitions of genes have varied a lot ever since and the current universally accepted definition is that genes are *"a union of genomic sequences encoding a coherent set of potentially overlapping functional products"* [5] and the whole haploid DNA compliment in a gamete is known as the genome [6].

Since genes are the principle information transmitted from parents to offspring (but see [7]), it is considered as the basic unit of selection (or replicator – in the gene centric view of evolution) [2, 8]. The advent of the Polymerase Chain Reaction (PCR) and the flood of information in the form of DNA sequences, helped make rapid strides in the field of evolutionary biology, which enabled studies comparing the genes as a proxy for the phenotypes (or traits) they produce. The sequence information yields unparalleled precision, ease, generality and reproducibility to the study of evolutionary biology. However as more and more genetic information started to flow, in the form of DNA sequence information or genomes [9, 10] it is also being clear that natural selection affects not only the genes but also the non-coding elements in the genome [11, 12] and that these information are also inherited.

The work in this thesis takes shape by testing the hypothesis that genes and genomes (the base of any phenotypic characteristic) could be compared at a within and between species basis to identify factors governing the evolution of the phenotypes and morphological innovation, and test the importance of natural selection. I highlight the evolutionary mechanism and its influence on gene family and gene evolution - in chapters 2, 3 and 4 - in Part 1 of this thesis. In the second part using genes as a proxy for phenotypic innovation I check for adaptive innovation (of phenotypes) in teleost fishes in chapter 5 and unravel hidden diversity within teleost fishes using phylogenetic methods applied to genetic (gene sequence) data - in chapters 6 and 7.

Gene and genome evolution:

There are different ways in which a gene can change. The basic building blocks of the deoxyribo nucleic acid (DNA), and in turn genes, are four nucleotides the adenine, guanine, cytosine, and thymine. Sequences of these nucleotides contain and "codes" the information. Changes to the pattern of nucleotides (mutation) and duplication of the sequences, are responsible for new types of genes in organisms. Frequency variations in the two alleles of a gene in the population could also be fixed. Taken together the frequency variation and novel genotypes (any defective genotype or deleterious mutation should be wiped off) could explain one basic facet of evolution. Since genes are responsible for each traits or phenotypes, adaptive evolution (evolution by natural selection) of the genes should explain most of the variation in phenotypes [8] allowing us to comprehend the evolutionary pattern and understand how species arose.

Although the use of DNA sequences to trace the evolutionary history or relationships of organisms is a very common practice today, the view that genes (or genome) evolves due to natural selection is controversial. The neutral theory of molecular evolution [13, 14] posits that *"random fixation of mutations"* with apparently *"no fitness effects"* are responsible for the variation between the genes and genomes of organisms, while the deleterious mutations are wiped off by purifying (negative) selection. This could be easily understood, since if advantageous mutations were responsible for gene or genome evolution, then all the functionally important genes should be evolving at a higher rate, due to fixation of advantageous mutations, but this is not the case.

Neutral theory says that the nucleotide substitution rate (fixation) is equal to the rate of neutral mutations or total mutation rate times the proportion of neutral mutations. Neutral theory also could explain the molecular-clock hypothesis where it was thought that if the mutation rate was constant among organisms (per clock or per generation time), it could be employed to calculate the molecular divergence times as a proxy for the organism's divergence times [15]. However, later studies show that these can be violated [16]. It is also noteworthy that neutral theory pertains only to the genes and genomes, and not to the morphological characteristics. In practice neutral theory forms the null hypothesis to check if any gene or genomic segments evolve according to adaptive evolutionary patterns.

Now-a-days it is also possible to compare the whole genomes of organisms in addition to the genes, thus providing the evolutionary biologist with another tool in their armory. Present day biological synthesis appreciate that genes, their patterns and expression, are responsible for phenotypes, although it is not necessary that a single gene "codes" for a single phenotype. Thus a genome wide or genome scale study would be essential to fully understand the evolutionary force at the basic level. It is an onerous task to relate the adaptive (Darwinian) evolutionary (and neutral or negative selection) mechanisms at the molecular level to the phenotypes.

There are four principle kinds of changes that occur in the DNA (genes), generally called as mutations, the insertions, deletions, inversions and substitutions. Of these the first three, insertions, deletions and inversions, could be true of one or more (a block of) nucleotides, while substitutions refers to one nucleotide changing to another. Substitutions in amino acid coding genes, could be of two kinds, some substitutions that does not change the amino-acid which the codon codes for (synonymous, therefore could be equal to the neutral evolutionary rate) or those that changes the amino acid the codon codes for (non-synonymous, should reflect the positive Darwinian process) due to the degeneracy of the genetic code. If a mutant allele/gene stabilizes and produce a successful adaptation, it will be transmitted to the future generations, thus if we create a phylogeny using a gene, of the organism and its relatives, we could map the emergence of the trait or the mutation (substitution) and its effects, all these help us to comprehend the evolutionary history of the gene as well as that of the organisms.

Utility of Phylogeny in evolutionary explorations:

A phylogenetic tree forms the basic hypothesis about the evolution of the organisms studied, or the history of the gene family studied. Creating a phylogeny is possible by finding out the amount of differences among a group of sequences, and modeling their evolution using mathematical methods. The simplest phylogenetic tree could be one that shows the "distance" between two sequences. For example if there are *n* nucleotides (or amino acids) in a given sequence alignment (of two sequences sharing a common ancestor) and if n_d is the number of differences between any two sequences (arising from a common ancestor), then $P = n_d/n$, denotes the uncorrected *"p-distance"* between two sequences.

However, while accounting for silent or backward substitutions and parallel substitutions (same kind of substitutions on different lineages), uncorrected p-distances may not hold true (reflect the evolutionary process) so we need to invoke mathematical models of nucleotide (or amino acid) evolution. There are various mathematical models of nucleotide and amino acid evolution [8, 2, 17, 18], however it is out of scope to provide a detailed account of each of them here. The distance information or the substitution information (generated using explicit mathematical models) could be used to create a phylogeny using simple distance methods or parsimony methods [8] or the more complex maximum likelihood [19] or bayesian methods [20].

Phylogenetic trees could also form the basic hypothesis regarding the relationship among species (species trees), could be used to calculate the divergence times and also to find the relationship among homologous genes in gene family (gene trees). Information from a phylogenetic tree can be used to explain the evolutionary process at different levels, genes, species and at higher taxonomic levels. We could sometimes correlate a single mutation to a trait (see [21, 22] etc), or compare the substitution frequencies (*dN/dS*) to check if the gene evolution can be explained by positive Darwinian selection (see chapters 2-5). If we identify that the gene is positively selected we could map the substitutions that were responsible for (and thus correlate them to) the phenotype.

Gene duplications:

Moving from the topic of gene level (or protein level) mutations and substitutions, at genomic level there are processes like gene duplications and genome duplications. By the time the first definition for gene was proposed rapid strides were made in the field of genetics. Perhaps the first report of duplication (chromosomal), and its importance for morphological variation was observed in maize as early as 1911 [23]. In the early part of the last century itself importance of duplication (of chromosomal parts) and morphological variation was observed in Jimsonweed's [24], and in Drosophila [25].

The fate of duplicate genes, and the selection patterns on each copy were discussed as early

as 1938 [26] and later by different authors [27, 28, 29, 30, 31, 32]. It was Ohno [27] who first explicitly suggested that gene duplication was the major cause of morphological innovation in organisms. The most important consequence of gene/genome duplications are multiple copies of genes, these multiple copies form gene families. As early as 1970's [33, 34, 35] gene family evolution studies had started. Genomic comparisons could enable us to find the evolutionary pattern of the genes families and copy number variations of the genes among species, and enable us to understand why more copies of a gene were essential for the adaptive success of an organism.

1.0.2. Brief introduction to the Methods used for adaptive evolutionary explorations

A simple work flow (Figure 1.1) for an evolutionary exploration would involve three steps: i) identify a phenotype of interest (be it adaptive or developmental); does it contain enough variation? ii) What are the genes involved? iii) can positive Darwinian selection explain the gene's evolution?

If the gene(s) evolve in a positively selected manner it should shed light that the corresponding phenotype is also undergoing natural selection. In fact with slight modifications to the above work flow a gene family's evolution can also be studied. If there are paralogs of a gene in an organism's genome (e.g., globin family, retinoid x receptor family, etc.) then we could check if positive selection, or more importantly functional diversification, has been a reason for the retention of copies.

Another interesting avenue of using sequence information from multiple individuals of organisms is to check their evolutionary divergence times (time of split since the most recent common ancestor). Reciprocal monophyly could also answer questions like whether the sequenced individuals consists of a single or multiple species (or distinct lineages). Such studies could solve the evolutionary history of a group of organisms and even bring to light how environment and geography can be responsible for the evolution of organisms. If we find multiple species (in the above step) using our above mentioned work flow we could check diverging phenotypes, the genes responsible for those phenotypes and check if those genes (and in turn phenotypes) evolve by positive selection.

Generating sequence data:

The basic prerequisite for any phylogenetic or evolutionary study is a set of homologous sequences. Homologous sequences can be of two types based on their appearance in the species. Those genes that were split into two due to speciation (two sister species carrying the same gene) are called as orthologs. If there are multiple copies of a gene in same species, due to gene or genome duplication, they are called paralogs. It is essential that we select

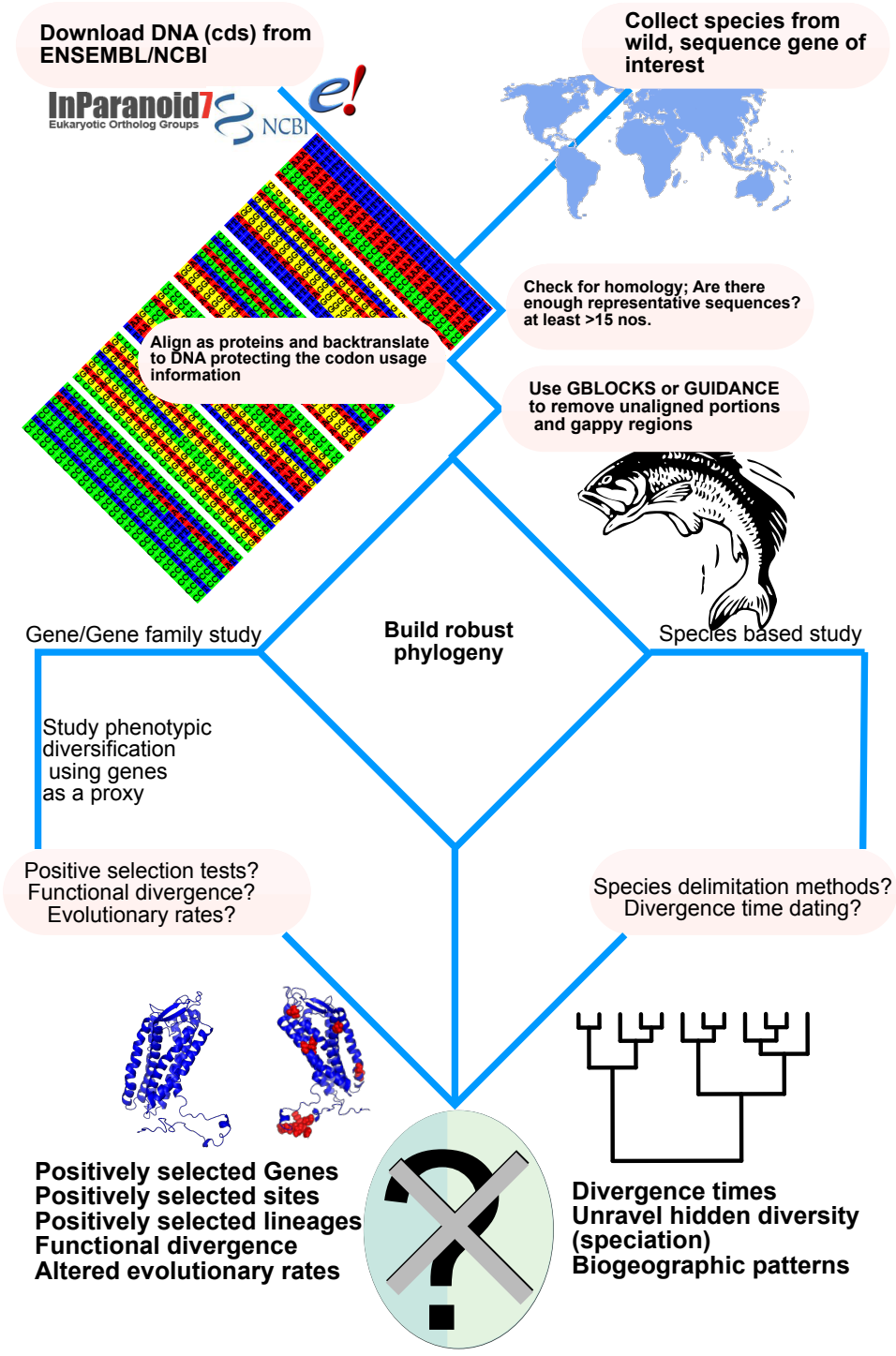


Figure 1.1.: Simple workflow for evolutionary explorations

orthologous sequences to generate a phylogenetic tree when studying evolution since it is the only way that we could reflect correctly the organism's speciation and thus evolutionary path. Paralogous genes from different species could be used to create phylogenies as well but when the intention is to study gene family (not a single gene's) evolution.

Thus the first step is to distinguish the orthologs of a gene, there are different sources like ensembl (www.ensembl.org), inparanoid [36] and genbank [37], where we can find the ortholog/paralog information. If orthologs (and paralogs) are not reported for a given species that we intend to study then we should use sequence similarity based alignment methods like BLAST [38], and reciprocal blast hit methods or profile based methods like HMMER (www.hmmer.janelia.org/) to find genes from genomes of different species which we intend to study.

Once a representative sequence dataset is compiled next step is to create a sequence alignment. There are different alignment programs, a user can use simple alignment methods like clustal [39], muscle [40], or more advanced methods like mafft [41] or prank [42], or even use a protein 3D structure to thread the translated DNA sequences to produce an alignment, programs like t-coffee [43] also has facility to create a consensus alignment produced with different alignment methods (m-coffee or d-coffee mode). Methods are available to filter out the un-aligned segments from the alignment like Gblocks [44] or GUIDANCE [45]. These alignment filtering methods have been shown to be particularly effective and useful for downstream applications like positive selection analysis which minimizes the false positives from the downstream applications [46, 47].

Codon models and testing for positive selection:

As stated earlier, to test for adaptive evolution, neutral theory and its assumptions form a valid null hypothesis. There are various methods to test for positive selection, Tajima's D statistic [48] and McDonald-Kreitman test [49] are the notable ones used at the population level. In this thesis I have primarily studied the lineages and sites (not at population level) undergoing positive selection from a phylogenetic perspective, thus each chapter will contain more elaborate discussions about them. Given an alignment of orthologous sequences, ratio of non-synonymous substitutions per non-synonymous sites (dN) and synonymous substitutions per synonymous sites (dS) could be calculated and their ratio (dN/dS) the omega (ω), signals if the lineages or sites evolve under positive ($\omega > 1$) or neutral ($\omega = 1$) or negative ($\omega < 1$) selection.

Models of substutions for codons [50, 51, 52], considers that codons are the basic units of evolution, instead of nucleotides or amono-acids. The codon-models can be used to calculate the dN/dS accurately (see [52] for a detailed overview on codon models). The most simple forms of dN/dS estimation could be by a counting method. However complex methods using maximum likelihood modeling are used widely since they are found to produce less false positives and provide robust inferences.

First set of models are the site-wise methods where each site's ω is tested, SLR [53] or

SLAC [54] are computer programs that do this kind of tests these are more similar to the dN/dS counting methods. Another category is the random-sites models popularized by PAML [55][56], and the REL methods in HYPHY [54], where the sites in a protein (alignment) are partitioned into using explicit statistical tests. Of note are also the fixed effects likelihood (FEL) models popularized by HYPHY.

By applying the codon models to the branches of the tree one could test the branches under positive selection, similarly applying it to the sequence alignment they could test site-specific positive selection [8]. We could test if lineages evolve at a *dN/dS* ratio greater than 1, by doing two tests one with constraining all lineages to evolve at neutrality (one-ratio) and another allowing all lineages to evolve with their own ω values (free-ratio). The use of free-ratio models are discouraged since they are parameter rich, instead the user could test their branch of interest by invoking two-ratio models, if the user considers only one (or a single set) of branch other than the neutrally evolving branches, similarly the user could test any number of branches (*n-ratio*) of their liking and use a Likelihood Ratio Test (LRT) to check if the one-ratio model or the two-ratio (or *n-ratio*) model fits the phylogenetic tree better. Site-specific tests of positive selection falls into various classes according to the partitioning of the sites into different rate categories.

If the user wants to test if some specific lineage was positively selected and also needs to know the sites that were selected, there are branch-site models [57], which removes the problems of branch models where all ω values on sites (alignment length) are averaged and site-models where ω values on all branches are averaged. Recent publications on the variations of conventional branch-site tests like the "branch site REL" method in HYPHY [58] and mixed effect models of evolution (MEME) [59], are promising and are applied on the teleost rhodopsin sequences (chapter 3) studied here.

Relative rate ratio tests for positive selection [60], capturing the codon models information as well as the amino-acid property information, which is more similar to a McDonald-Kreitman's Test, also exists although not as popular as maximum likelihood methods. There are also methods to test for positive selection at the amino-acid level (using protein alignment opposed to the codon alignments mentioned earlier) [61, 62].

Functional divergence and evolutionary rates:

Gene duplication is thought to be an important evolutionary mechanism by which morphological evolution happens. Once a gene is duplicated there are two copies in the organism which are capable to do a same function, thus the redundant copy should be lost in due course since the organism does not need it. However, if some evolutionary mechanism like positive selection or increased evolutionary rates (and relaxed purifying selection) could modify one copy to perform a new function (neo-functionalization) or if both the copies share functions (syn and/or sub-functionalization) both copies could be maintained.

There are different methods to find functional divergence between proteins, of note are the covarion/heterotachy (type I) models and constant-but-different (type II) models of functional divergence [63]. In type I functional divergence we could check for increased rates in each sites of proteins, if a site in one protein (clade/ortholog) evolves faster (has many different amino acids in the alignment column) and the same site in the other protein (clade/ortholog) has constant sites, it is an indication that one ortholog (or clade/paralog) is evolving faster and function is shifting.

The type II models test for columns of amino acids with different properties preserved in different paralogs, but in each paralog (clade) that amino acid (column) is constant. Computer programs like DIVERGE [64, 65] and RASERv2 [66], can carry out the type I analysis, while the former is also capable of doing the type II analysis. RASERv2 carries the advantage that it can work on alignments without *a priori* specification of clusters. It is thought that constant-but-different models are good to test evolution of paralog and covarion models to capture functional divergence among orthologous clusters [63]. It should be noted that expression divergence [67, 68], selection [30], dosage compensation [28, 29, 27], etc could be also responsible for preservation of duplicates in addition to functional divergence.

Divergence time dating:

As early as 1960's it was observed that the evolutionary rate of protein sequences between lineages were constant or evolution was constant over time [15], better known as the molecular-clock hypothesis. Later with the advent of the neutral theory this was thought to be definite. It was possible to estimate the time of divergence between two lineages separated by a common ancestor.

However, nowadays it is known that molecular clock holds true just for closely related species, and as the time of separation increases the sequences starts to evolve in different rates. Global clock models using fossil dates could be useful for finding the divergence times between closely related species. Local clock models with fossil calibrations, such as non-parametric rate smoothing method [69] or similar algorithms [70] could also be used. It is known that fossils give only minimum ages, also fossil ages could have an amount of uncertainty, which could be incorporated into bayesian markov chain monte-carlo (MCMC) based analysis of divergence times, which are now considered state-of-the-art.

Species delimitation:

DNA based species delimitation methods are varied [71], however the General mixed yule coalescence [72] model (GMYC) and the bayesian species delimitation [73] methods (BPP) based on the coalescent theory are of much importance and could be the standard techniques of the future systematic studies.

GMYC uses the knowledge that there are changes in the branching rates at the species boundaries. Tests (using GMYC model) employs a maximum-likelihood approach to test for the predicted change in branching rates and uses the GMYC model to estimate the species boundary by identifying the transition from coalescent to speciation branching patterns on an ultrametric tree. The GMYC exploits the predicted difference in branching rate under the two modes of lineage evolution, where the branching patterns within each genetic cluster reflects a neutral coalescent process and the branching patterns between two genetic clusters reflects timing of speciation events and by assessing the point of highest likelihood of the transition [72] it differentiates the evolutionary distinct lineages.

Monaghan and co workers [74] developed a modified GMYC model that allows for a variable transition from coalescent to speciation among lineages by identifying multiple thresholds reflecting the variable lineage divergence. The likelihood values of the GMYC models are compared to a null model which assumes a single branching process for the tree using a Likelihood Ratio Test (LRT). The null model is that the sample derives from a single interacting population: the pattern of variation should conform to the genealogy of a single population, for example a neutral coalescent.

Bayesian species delimitation is implemented using the Bayesian Phylogenetics and Phylogeography software (bpp v. 2.1a; [73]). This method requires a multi-species multi-locus dataset and also requires that we assign the candidate groups prior to our analysis and provide a phylogeny showing the relationships between the groups. The bayesian species delimitation (BPP) method accommodates the species phylogeny as well as lineage sorting due to ancestral polymorphism.

The parameters in the model include the species divergence times τ, measured by the expected number of mutations per site, and population size parameters $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per site per generation so that θ is the average proportion of different sites between two sequences sampled at random from the population. The prior distributions on the ancestral population size (θ) and root age (τ) can affect the posterior probabilities for models, with large values for θ and small values for τ favoring conservative models containing fewer species [75]. Thus it is a common practice to use different priors for population size and root age while doing this analysis [75, 73].

Following the above given methodology, we could study the speciation process, adaptations and the importance of evolutionary forces on the genes that shape them, also we could study the gene and gene family evolutionary history. Evolutionary studies of genes, in addition to satisfying our evolutionary quests, could point at mutations/modifications or processes in ancestral or sister species that could reveal important clues in understanding the diseases in humans and domestic animals. Basic studies of evolution of genes and gene families are important also to understand the adaptive radiation process of the organisms. Biodiversity and speciation studies, using genetic methods, could shed light on the history of the planet as well

as help us to conserve the precious biodiversity and aid in different fields related to human welfare [76].

In short evolutionary explorations of genes and phylogeny based studies could enable us to contribute to various fields like evolutionary medicine [77, 78], biodiversity and conservation [79] and in different aspects of human welfare [76]. This thesis deals with the most basic unit of evolution the genes and their evolution. Results from the publications arising from this thesis could enable further research, characterizing the positively selected sites benefiting medicine and physiology research, or studying new species identified adding to our knowledge of biodiversity.

1.0.3. Structure of the thesis

As mentioned earlier, in this thesis I tested for adaptive evolution of genes involved in vertebrate development and adaptation, as well as evaluated dispersal in vertebrates, specifically teleosts, from a biodiversity hotspot. The main idea was that *by comparing the genes/genomes of closely related species* we could *identify the basis of morphological/phenotypic innovations and adaptive radiations*. It should be noted that morphological and phenotypic innovations lead to (adaptive) radiation of species into various ecological niches, which indeed is the basis of the biodiversity. Thus, these two topics, adaptive evolution and species radiations, could be dealt side by side, when we attempt sampling of species and genes for a study like the one done in this thesis. The thesis is divided into two major parts, the first relates to adaptive evolution of developmental and adaptation genes in vertebrates and the second to species radiation in vertebrates (and influence of adaptive evolution), specifically in teleosts.

I investigated several gene families involved in vertebrate development and adaptation, like Retinoid X Receptors (RXR), Superoxide dismutases (SOD), Rhodopsin-1 and 39 genes involved in the teleost lateral line system. When an adaptive radiation of a group occur, there are innumerable morphological, physiological and behavioral innovations, I try to relate the RXR gene family evolution (additional duplicates in teleosts/ostariophysi) to the adaptive radiation in teleosts, the teleost rhodopsin-1 evolution to the adaptive radiation of teleosts to varied habitats, evolution of avian SODs to the mode of locomotion and physiology (and subsequent adaptive radiation) of the birds and lateral line system gene evolution to the mode of locomotion, physiology, behavior and adaptive radiation of the fishes. In the last (two) chapters I look at two cases of teleost radiations, which could be targets to study the genes explored in the earlier four chapters, for detailed evolutionary explorations, these species are closely related (sibling species/cryptic species) and evolutionary explorations could be insightful in such closely related species since we already have results from distantly related groups.

Part I: Adaptive Evolution of Vertebrate Developmental and Adaptation Genes

In the first section of the thesis, I look at the evolutionary forces acting at the level of genomes or genes of organisms. I look at genes important in adaptation and development of vertebrate species. I check for the evolutionary forces involved in the genome level (duplications and functional divergence) important for the maintenance of these genes in the genomes of the organisms. I also check for adaptive evolution of these genes and correlate the evolutionary pattern to the adaptive innovations of the organisms. This part includes three chapters checking adaptive evolution of developmental and adaptation genes of vertebrates.

Chapter two, deals with the Retinoid X Receptor in vertebrates. They are transcription factors with important roles in development, reproduction, homeostasis, and cell differentiation (see chapter two). There are three copies of this gene in tetrapods, while the teleosts posses four to six copies of this gene. This gene family is interesting from the gene evolution point of view since it is important in key processes like patterning, detoxification and homeostasis. This gene family is also interesting from the genome evolution point of view since it presents us a case where we have asymmetric distribution of genes in different groups of vertebrates allowing us to check for the evolutionary forces at the genomic level.

Chapter three, deals with the teleost Rhodopsin-1 gene. The teleost Rhodopsin 1 (*RH1*), is peculiar by possessing a single exon [80, 81], while all the other opsins (sister genes evolved from a same ancestral gene) posses five to six exons. As we know rhodopsins are important for the scotopic vision of organisms [82], thus studying these genes from an adaptive evolutionary perspective allows us to relate the adaptive benefits conferred to the organism by this gene to its mode of life [83] (in low light environments like caves, deep sea environments etc). This gene is also used as a phylogenetic marker now-a-days, owing to its size and due to the lack of introns. In this chapter I look at the adaptive evolution of the teleost rhodopsins as well as check for the phylogenetic utility of the gene. Checking for the adaptive evolution allow us to relate the evolutionary forces to the adaptive radiation of the teleosts in varied habitats.

Chapter four, deals with the avian Super Oxide Dismutase (SOD) genes. Superoxide dismutases are the first line of defense against the Reactive Oxygen Species (ROS) [84, 85]. In vertebrates there are three copies of SOD genes thought to be products of genome duplication in the vertebrate ancestor. In this chapter, I look at the adaptive evolutionary patterns of the avian SODs and compare it with its other tetrapod relatives (reptiles and mammals). As we know exercise and physical activity accelerates metabolism, thus increasing the amount of ROS, we could question *if SOD is related with the adaptations of the organisms regarding physical activities, did the adaptive radiation of birds have a direct link to the evolutionary tinkering of their detoxification genes.* Here, preliminary evidence regarding their evolution and importance in avian adaptation and radiation are provided.

Part II: Adaptive Radiation in Vertebrates: insights from the Teleosts

This section deals with the teleost radiations, in the first of the three chapters I look at the genes involved in the development of a key sense organ and correlate the results to the adaptive radiation of teleosts, the last two chapters deal with freshwater teleosts sampled in the Western Ghats biodiversity hotspot in southern India. We leverage molecular genetics and related analytical tools to aid their systematics and subsequent conservation and management. Our study of these two groups (in the last two chapters) is important in two major fronts, it facilitated the proper systematic cataloging of the species and it also allowed proper conservation management plans to be designed.

Chapter five, deals with a sensory organ in the teleosts, the mechanosensory lateral line system. While lateral line is a developmental innovation only found in aquatic vertebrates (fishes and amphibians), it is also important in the adaptation of the organisms to the water borne mode of life and also directly linked to the adaptive radiations of these organisms. In this chapter, I looked at 39 genes found to be involved in the teleost lateral line system development. I checked for the evolutionary forces at the nucleotide level, codon level and the protein level. I also mention about the evolutionary forces at the genomic level, and look at the duplication of the genes studied during the teleost specific genome duplication [10, 86, 9]. This developmental and morphological innovation (lateral line) has direct implications in the adaptive radiation of teleosts, since lateral line is important for different processes related to their behavior and biology (see chapter five for a detailed treatment of the topic).

Chapter six, deals with the cryptic speciation in one of the worlds most popular aquarium fishes the red lined torpedo barbs. As discussed earlier the genetic data could be used to check for models of speciation, and calculate the divergence times, which could be of help in systematics. Here by using recently developed species delimitation methods, cryptic diversity in this group of teleost fishes are studied.

Chapter seven, deals with another teleost cryptic species complex from the family channidae. In this chapter the focus is to differentiate the two species confused as a single species. Most importantly we re-described one of the species leading to proper biodiversity cataloging and yielding benefits in conservation and management.

In addition to the results from these studies (in chapter six and seven), future explorations with these species could be interesting. The species groups studied here are of peculiar nature, *Channa diplogramma* studied in chapter seven, possess a very different "branched" lateral line pattern, how could this relate to their habit and habitats, results from chapter five could be of immense use to understand this teleost's adaptations as well as more on its lateral line. The red lined torpedo barbs (chapter six) are cyprinids with a benthopelagic habitat living in rocky pools in fast flowing rapids, does this lifestyle require a fine-tuning of the rhodopsin-1 gene? in this case we could compare the study with the results of chapter four.

1.1. Brief conclusions and future directions

In this thesis divided into two parts and seven chapters, I study vertebrate evolution, from a molecular genetic perspective. The first part deals with adaptive evolution of genes and evolutionary forces at genome level (comparative genomics), impacting development and adaptation. The second part relates to the utility of molecular genetic techniques in biodiversity and conservation management (molecular systematics and conservation genetics in addition to comparative genomics), and characterizes potential species that could be used for checking the results of the initial chapters.

Our results suggest that positive selection has been a major reason for teleosts to retain more copies of genes than tetrapods, following the teleost specific genome duplication (chapter two and five). However adaptive evolution is not the only evolutionary force at gene level (for preservation of paralogs), it works in combination with functional divergence and expression shuffling (chapter two and five). The results of the explorations on the teleost rhodopsin-1 gene unravel adaptive evolution of the gene in species with different low light habitats, as well as confirm the phylogenetic utility of the gene as a phylogenetic marker at the level of teleosts (chapter four). The study of avian SODs again reveal the importance of adaptive evolution, at the same time reveal that avian SOD genes have an altered evolutionary rate compared to mammals and reptiles (chapter five). The last two chapters unravel hidden diversity among teleost fishes and provide evidence of additional species which warrant proper species descriptions according to the code of the International Code of Zoological Nomenclature (ICZN).

Future studies on other detoxification genes (catalyses and glutathione synthases) in birds, could be interesting to assess similar evolutionary patterns to SODs and reveal a role of detoxification genes in their higher longevity and fine-tuning of their locomotion. Sequencing and evolutionarily characterizing the red lined barbs' rhodopsin-1 genes could be revealing and could check the evolutionary patterns of the gene in cryptic species and could hint to the exploitation of "ecological opportunities" by these species. In addition, sequencing RXRs from the species we taxonomically characterized would be important, since one (*Channa diplogramma*) belongs to perciformes, where only four RXRs are known (see chapter two) and the other (red lined torpedo barbs) are cyprinids close relatives of *Danio rerio* where we found six copies of RXRs.

The results obtained during my PhD program studying the *adaptive evolution of genes involved in adaptation and developmental innovation in vertebrates* is presented here. Each chapter has its own introduction, methods, results and discussion sections. In studies that I have collaborated with groups outside University of Porto, my role and contribution is explicitly stated.

I.

Adaptive Evolution of Vertebrate Developmental and Adaptation Genes

2

Adaptive Evolution of the Retinoid X Receptor in Vertebrates

Papers arising from this chapter


Genomics 99 (2012) 81–89

Contents lists available at SciVerse ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Adaptive evolution of the Retinoid X receptor in vertebrates

Siby Philip [a,b], L. Filipe C. Castro [a], Rute R. da Fonseca [a], Maria A. Reis-Henriques [a], Vítor Vasconcelos [a,b], Miguel M. Santos [a,b], Agostinho Antunes [a,b,*]

[a] CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Rua dos Bragas, 177; 4050-123 Porto, Portugal
[b] Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Rua do Campo Alegre, 4169-007 Porto, Portugal

Figure 2.1.: Paper arising from the chapter, **Genomics** 99 (2012) 81-89.

2.1. Abstract

Retinoid X receptors (RXR) are transcription factors with important roles in development, reproduction, homeostasis, and cell differentiation. Different types of vertebrate RXRs (α, β and

γ) have arisen from multiple duplication events. The adaptive evolution mechanism that has preserved duplicate RXR paralogs, as well as their role in development and adaptation, is thus far unknown. In this work, we have investigated different aspects of vertebrate RXR evolution. Codon based tests of positive selection identified that RXR was under significant positive selection immediately after the whole genome duplications in vertebrates. Amino acid based rate shift analysis also revealed significant rate shifts immediately after the whole genome duplications and functional divergence between all the pairs of RXRs. However, the extant RXR genes are highly conserved, particularly the helix involved in dimerization and the DNA-binding domain, but positively selected sites can nevertheless be found in domains for RXR regulation.

Keywords:

Adaptive evolution, Retinoid X Receptor, Positive selection

2.2. Introduction

Retinoid X receptors (RXR; NR2B), are transcription factors that mediate an array of extracellular signals in a ligand dependent manner, to regulate the target gene by binding to response elements within the promoter region of those genes. RXR regulates many biological functions in vertebrates such as development, reproduction, homeostasis and cell differentiation [87, 88, 89]. Their disruption has been associated with a vast array of developmental and reproductive abnormalities (e.g. reduced testicular development and fertility, masculinization of female gastropods – imposex) in wildlife and humans [90].

RXRs are members of the nuclear receptor super-family. The canonical structure of a nuclear receptor follows a common pattern including the N-terminal 'A/B domain', a DNA binding domain and a ligand-binding domain. RXRs bind to their targets often called response elements, which may be single elements or repeats, arranged in a direct, inverted or everted manner, of a consensus sequence 'AGGTCA'. These repeat elements require the formation of dimers, and RXR is an important heterodimerization partner for many other nuclear receptors [91]. RXR can also form homodimers, suggesting an independent signaling pathway, but its exact biological role remains elusive [91]. The key role of RXR in the heterodimerization with other nuclear receptors, and thus its interference in multiple signaling pathways, makes it an interesting therapeutic target for treatment of diseases like cancer and metabolic syndrome [92]. The other major roles of RXRs are in the embryonic development, differentiation, organogenesis and cell proliferation. Due to its ubiquitous presence within metazoans and its activation by low molecular weight ligands, RXR is a prime target of environmental pollutants, both in vertebrates and invertebrates, which may be a cause of cancer and endocrine disruption [93, 94].

RXR has been found to bind 9-cis retinoic acid (9-cis RA) with high affinity [95], suggesting a role in the retinoic acid signaling pathway in addition to its heterodimerization partner status.

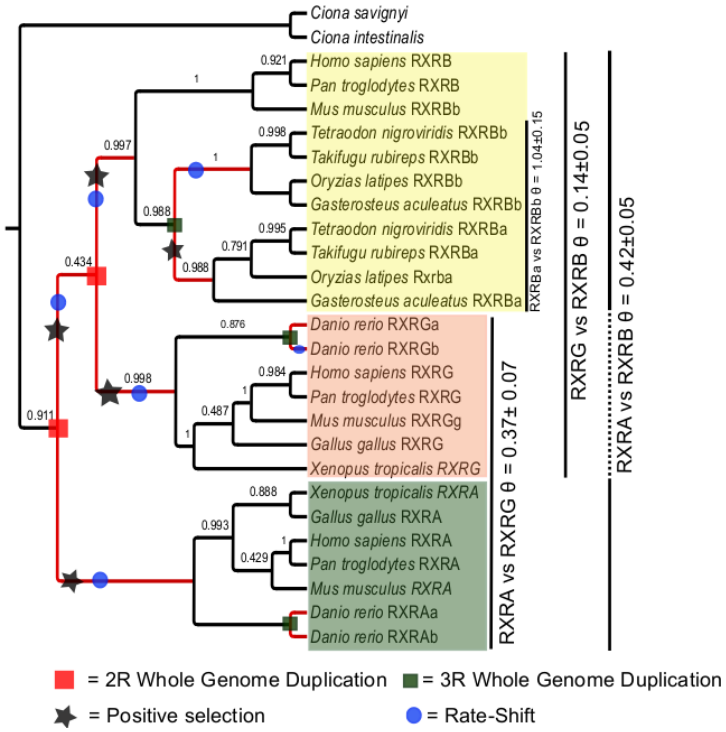


Figure 2.2.: Schematic RXR gene family tree showing the orthologous relationships of the Teleost RXR genes: the post-duplication branches tested are shown in red, positive selected branches are marked with a star symbol and rate-shifting branches are marked with a blue circle, the values for functional divergence (θ) between different clades of RXRs are also shown.

Although 9-cis RA has not been clearly detected in mammalian cells [96], it has recently been identified in teleost fish [97] and invertebrate tissues [98]. Since 9-cis RA can also act as a ligand to the retinoic acid receptors (RARs) its role as the natural ligand of the RXR has been questioned. Also phytanic acid and docosahexaenoic acid have been proposed as ligands of the RXR [99, 100].

One or more RXR genes are found in most metazoan taxa from placozoans to vertebrates [101]. In vertebrates, there are typically three copies of the gene, RXRA (NR2B1 / RXRα), RXRB (NR2B2 / RXRβ) and RXRG (NR2B3 / RXRγ), which arose due to the two rounds of genome duplications (2R) in the vertebrate ancestor [102]. In fishes, owing to a specific whole genome duplication (3R) there is an additional RXRB gene [102], but only in zebrafish (among the sequenced teleost genomes), additional copies of the RXRA and the RXRG genes are found [103](Figure 2.2).

In vertebrates, RXRB is the ubiquitously expressed subtype [95], RXRA is mainly expressed in liver, kidney, epidermis, intestine and dominates the RXR expression in the skin, while RXRG shows a restricted expression in muscles, pituitary gland and certain regions of the brain (see www.nursa.org/10.1621/datasets.02001 for more details about expression of different RXR subtypes).

The main goal of this study was to assess the adaptive evolution of the RXR genes in vertebrates using a comparative genomics framework. We compared the rates of synonymous (silent; dS) and non-synonymous (amino-acid replacement; dN) substitutions and conducted functional divergence analyzes of the vertebrate RXR genes. Moreover, we studied the synteny of teleost RXR genes to retrace its evolutionary history in the vertebrates and we further evaluated the changes of the RXR expression patterns across vertebrates. We found that all the branches immediately following the first round (1R) and 2R of genome duplication were under positive selection and all the pairs of RXRs produced from these two rounds of duplication were functionally divergent. Similarly, the paralogs that resulted from the 3R (third round/fish specific) genome duplication were also functionally divergent. Positive selected sites were identified mainly in the N-terminal region and the Ligand-Binding domain (LBD), which harbors regions responsible for the expression of the protein. To clarify the effect of genome duplication on the RXR expression we evaluated expression databases for the patterns of zebrafish RXRs (duplicates) comparatively to the mouse RXRs (singeltons), which suggested expression shuffling among the paralogs of zebrafish. Finally, we provide evidence that the asymmetric distribution of RXR genes in teleosts (in comparison to zebrafish) was due to secondary gene loss events.

2.3. Results

Analysis of synteny: gene loss in medaka, fugu, stickleback and tetraodon

The synteny analysis between the fish (medaka and zebrafish) chromosomes containing the RXR genes and their human counterparts (Appendix 1 Figure 11.1), confirmed the orthologous relationship between the genes from both groups. The existence of other co-orthologous genes in the fish chromosomes (four chromosomes in the case of medaka and six in the case of zebrafish) containing the RXR gene when comparing them to the three human chromosomes (each with one RXR) suggests that the fish chromosomes are products of an ancient duplication event, supporting that the additional RXR genes in the teleost genomes are actually products of the ancestral teleost specific genome duplication [102, 103].

The teleost chromosome evolutionary model [9] suggests that the ancestral teleost had 24 chromosomes post teleost specific genome duplication. The medaka and the fugu genomes have preserved that same condition till date (for 270 million years; Appendix 1 Figure 11.2) while the zebrafish genome suffered lineage specific chromosomal rearrangements after the ancestral teleost fish genome duplication event. The chromosome 20 of the zebrafish and the chromosome 24 of the medaka genome are thought to be the products of a same ancestral chromosome and no ancestral rearrangement events have been documented leading to the chromosome 20 in zebrafish [9]. However, while the chromosome 20 of zebrafish harbors one RXR gene (rxrgb), the chromosome 24 of the medaka genome does not (Figure 2.3a). Our

analysis of conserved synteny for the chromosomes, between the region 33.5 Mb to 36.7 Mb of the chromosome 20 of the zebrafish (rxrgb is found on 33.90 – 33.94 Mb) and the chromosome 24 of the medaka genome revealed several neighboring orthologous (co-orthologous) genes but the RXRG was missing from the medaka chromosome 24 (Figure 2.3b and Appendix 1 Figure 11.3), signaling a gene loss event.

This trend of conserved synteny of co-orthologs (genes in the neighboring regions) is also evident in the comparison between the zebrafish chromosome 20 and the stickleback group XVIII (Appendix 1 Figure 11.3). The searches in the synteny database [104] for syntenic clusters of both *rxrga* and *rxrgb* of zebrafish with medaka and stickleback identified only, the chromosome 4 of medaka, which contains RXRG genes (clusters: 351475 and 349229) and the group VII of stickleback, which contains RXRG gene (clusters: 372199 and 374691), consistently proving absence of an additional gene in these species. Thus, we conclude that some teleosts, such as the medaka, the three-spine stickleback, the fugu and the tetraodon (all with only 4 RXR genes) have lost an additional copy of the RXRG produced during the whole genome duplication.

The chromosomes 5 and 21 of the zebrafish are products of genome rearrangements in the zebrafish ancestor and these are the chromosomes harboring the *rxrab* and *rxraa*, respectively. The reciprocal blast hit based synteny analysis between the zebrafish and the medaka and stickleback genomes consistently identified the chromosome 12 of medaka and group XIV of stickleback using both *rxrab* and rxraa as query. The chromosomes of these two species share several neighboring co-ortholog genes, for example: the co-orthologous genes of *rxraa, gsna, vav2* and *wdr5*, and co-orthologous genes of *rxrab, gsnb* and *anxa* where identified on these chromosomes (see Appendix 1 Figure 11.4), but no other chromosomes containing an additional RXRA gene could be identified. The absence of a second copy of RXRA gene in medaka and stickleback (on any other chromosome), suggests that an additional copy of RXRA has been secondarily lost in these species.

The circular plots of the chromosomes containing RXR genes for zebrafish and medaka also show that there are several genes orthologous on these chromosomes (Appendix 1 Figure 11.1), enabling us to conclude that the teleost specific genome duplication, not lineage specific gene gains, has been the reason for additional RXR genes in teleosts, while there has been lineage specific gene losses in some of the acanthopterygiian teleosts (medaka, fugu, stickleback and tetraodon) analyzed in this study, there was retention in the only ostariophysan teleost (zebrafish) analyzed.

Positive selection on post-duplication branches

The topology testing revealed that the correct trichotomy of RXR was ((RXRG, RXRB), RXRA), which was selected ahead of the other two competing topologies (Appendix 1 SI Table 1). This topology was used for the branch-site and branch models implemented in PAML

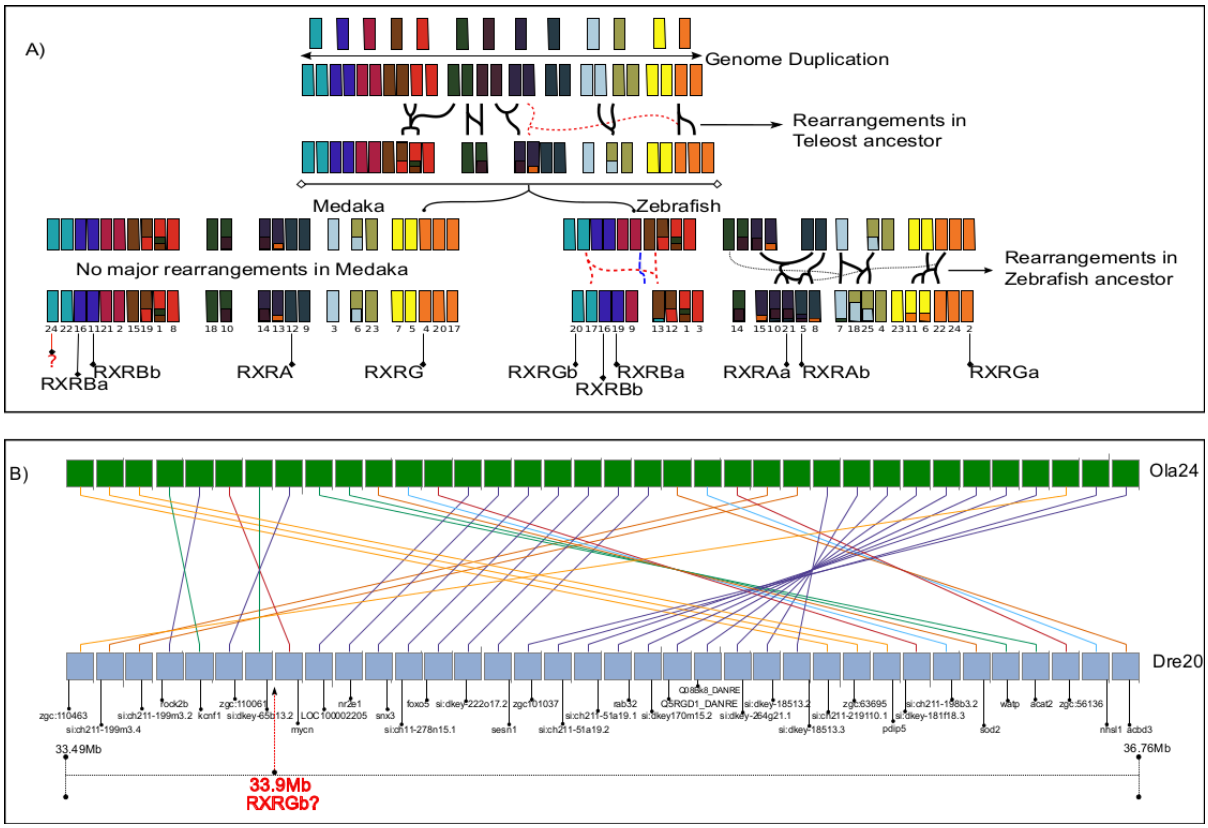


Figure 2.3.: a) The hypothesis of chromosome evolution in teleosts [20] (note that the chromosomal blocks are not up to scale), the ancestor to vertebrates possessed 13 chromosomes which duplicated during the teleost specific genome duplication event producing 24 chromosomes in the teleost ancestor, following the genome duplication there was 8 major genome rearrangements in the teleost ancestor, the medaka genome is almost unchanged after that event 350 Million years ago, whereas the zebrafish ancestor had some major and minor genome rearrangements which gave rise to the present day zebrafish genome; The present chromosomes in the species are placed directly below the ancestral chromosomes in the figure. The major rearrangements are marked by solid lines and minor rearrangements are marked by dotted lines, Chromosome 20 of zebrafish (which possess the *rxrgb*) and Chromosome 24 of medaka are products of the same parental chromosome, but the medaka lack and additional copy of *rxrgb*) The synteny (gene-trace image – a scale free representation of the orthologous clusters) showing the chromosome 20 of zebrafish (33.5 mb to 36.7 mb) to the chromosome 24 of medaka, *rxrgb* (33.9mb) has no ortholog in the medaka genome, whereas there are several neighboring co-orthologs, highlighting a gene loss event.

[56]. For the branch models, the likelihood ratio test (between the alternate and null model likelihoods from PAML shows that the two ratio model fits the data better (Table 2.1). The average ω value (ω_0) identified by the one ratio model was 0.049. The comparison between the unconstrained two ratio model and the constrained two ratio model supported the null model favoring the post-duplication branches not to be under positive selection.

Thus the branch models suggest that the post-duplication branches are under relaxed selection constraints. However, the unconstrained two ratio model identified significantly increased ω values (ω_{PD} = 1.11) in the post-duplication branches. For the RXRB gene duplication in the teleosts, initially the two ratio model was found to fit the data better against the one ratio model (LRT = 10.88). The analysis for positive selection against the constrained Two Ratio (ω_{PD} = 1), also favored the unconstrained two ratio model (LRT = 98.4), however the ω_{PD} value for the unconstrained two ratio model was 0.109 which supports a scenario of relaxed selection constraints in the post-duplication branches (Table 2.1), similar to the post-duplication branches resulting from 2R.

The branch site analysis also revealed that each of the ancestral branches of the RXR genes resulting from the second round of whole genome duplication (2R-WGD), RXRA, RXRG and RXRB were strongly positive selected using the LRTs (Table 2.2), and an ω value of 999 (infinity in the case of each of the post duplication branch tested, Values of 999 for dN/dS indicate dS = 0, so dN/dS is undefined) which signals that there are no synonymous substitutions at the few codons (10%) that appear to have come from site classes 2a and 2b (positively selected site class in the foreground branches). However, only the ancestral branch leading to the *rxrba* gene that resulted from the third round of whole genome duplication in fishes (3R) was positive selected (Table 2.3).

In addition, the branch-site test segregates the amino acid positions/codons into four different categories. Two describe sites for which selective pressure does not change over time, either under purifying selection (site class 0, $\omega_0 < 1$) or under neutral evolution (site class 1, $\omega_1 = 1$). The two other categories (site classes 2a and 2b) are sites potentially evolving under positive selection only in the foreground branches ($\omega_2 > 1$), and evolving in the background branches under purifying selection (site class 2a, background branches $\omega_0 < 1$) or neutral evolution (site class 2b, background branches $\omega_1 = 1$).

On average 10% of the sites were under the site class 2a and 1% of sites were under site class 2b (Table 2.2 and Table 2.3 and Appendix 1 SI Table 2), which signals that majority of the sites were under the influence of strong purifying selection during the evolution of RXR, while immediately after duplication a handful of sites were positive selected (on the foreground branches), signaling episodic events of positive selection during the evolution of RXR. Most of the positive selected sites (P > 0.95) were located either in the N-terminal region of the protein or the ligand binding region of the protein (Figure 2.4 and Appendix 1 SI Table 2), except one site in the RXRB ancestral branch (204-L relative to human RXRB), which was located just one

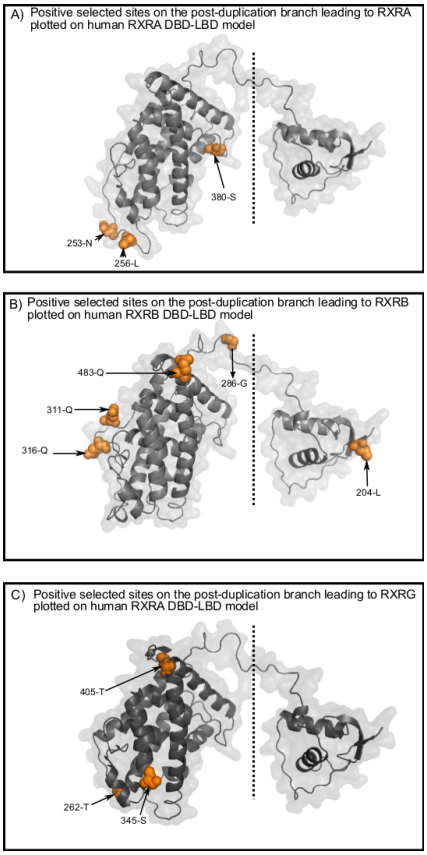


Figure 2.4.: Positive selected sites in the DNA-binding domain and the Ligand-Binding domain are plotted on the RXR DBD-LBD modeled in SWISS-PROT; A) positive selected sites on the ancestral branch leading to RXRA plotted on the human RXRA DBD-LBD model; B) positive selected sites on the ancestral branch leading to RXRB plotted on the human RXRB DBD-LBD model; C) positive selected sites on the ancestral branch leading to RXRG plotted on the human RXRG DBD-LBD model.

site before the beginning of the DNA binding domain. In the post-duplication branch leading to *rxrba* only one site was found to be under positive selection, which was in the N-terminal region (Appendix 1 SI table 2).

Significant rate shifts after RXR duplication

Codon models can suffer from the saturation of substitutions especially in deep branches, like the branches immediately after whole genome duplication, to eliminate any such problems and to give confidence to our codon based analysis we employed amino acid based rate shift analysis, since amino acid based analysis is recommended for divergent sequences [105]. To identify functional divergence between the RXRs we used DIVERGE [64]. Significant evidence for functional divergence (altered evolutionary rate) was observed between all the pairs of RXRA, RXRB and RXRG (Figure 2.2 and Appendix 1 SI Table 3), and between *rxrba* and *rxrbb* the paralogs that resulted from the teleost specific whole genome duplication.

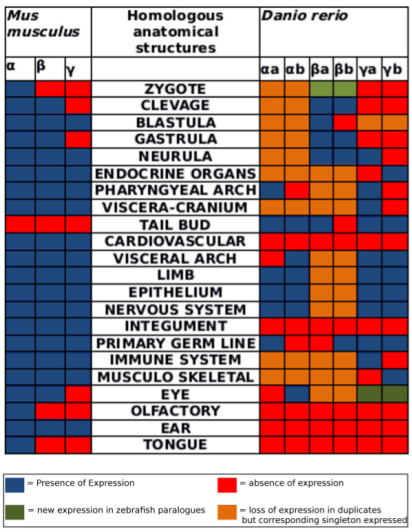


Figure 2.5.: Comparison of the RXR gene expression patterns during the embryonic development stages of *Mus musculus* and *Danio rerio* for multiple homologous anatomical structures (BGEE, retrieved on 10/12/2009), showing subfunctionalization and neofunctionalization of RXR genes in zebrafish following duplication events; blue = presence of expression; red = absence of expression; green = new expression in zebrafish paralogs; orange = loss of expression in duplicates but corresponding singleton is expressed

Using RASER2 [66], we detected significant rate shifts on all the ancestral branches (post-duplication), which lead to the major RXR lineages. In the case of zebrafish specific duplicates, *rxrgb* was the only gene that showed a significant rate shift (Table 2.4). The ancestral branch leading to the *rxrbb* teleost gene was also found to be under significant rate shift; however, *rxrbb* branch had only weak signals of positive selection using the codon models (Table 2.4).

In addition to the lineage specific models RASER2 implements the empirical bayes test to identify the sites evolving under a covarion-like model or heterotachy, which is similar to the site class 2a in the branch-site test of PAML. The details of the sites under rate shift are presented in Appendix 1 SI Table 2.

Changes in the expression pattern of RXR paralogs

We evaluated the expression dataset of BGEE [106] to assess the shifts in the RXR expression patterns during the embryonic stages of *Danio rerio* (paralogous duplicates) and *Mus musculus* (orthologous singletons). Our assessment of the RXR expression patterns during the embryonic development stage in the various homologous anatomical structures (Figure 2.5) was in agreement with an earlier study [107].

However, they had not specifically tested the duplicates of RXRG in zebrafish and our study further revealed that: (i) each of the zebrafish duplicates (paralogs) have at least one distinct expression pattern when compared to the sister paralog and RXRB and RXRG have distinct expression patterns (neofunctionalization) when compared to their singleton mouse orthologue,

and (ii) the zebrafish genes (each of the duplicates) had their expression in a lesser number of anatomic structures (subfunctionalization) when compared to the corresponding mouse singleton orthologue. We suggest that this 'functional shuffling' among the duplicate genes in zebrafish after the whole genome duplication, due to the positive selection on the ancestor, contributed to the conservation of additional gene copies during the course of evolution.

Different domains, different constraints: asymmetric selective pressures in the vertebrate RXR protein

The 3D structure of the ligand-binding domain of amphioxus' RXR (PDB ID: 3EYB) was used to map ProPhylER's [108] results, revealing that helix 10 of the ligand-binding domain is the most conserved substructure (Figure 2.6). Helix 10 interacts with RXRs' dimerization partners, along with the helices 7 and 9. The conservation of helix 10 suggests that the dimerization partners of RXRs are well preserved throughout evolution.

In vertebrates, the ancestral genome duplications have produced three RXR genes that are different in overall protein length and specific domain size (Appendix 1 Figure 11.5). The median P-values obtained from the analysis for each column of the alignment, from multivariate analysis of protein polymorphism (MAPP) [109], was plotted (Appendix 1 Figure 11.5). The DNA-binding domain and the ligand-binding domains were found to be more constrained than the N-terminal region.

2.4. Discussion

RXR gene is found in most metazoan taxa from placozoans to vertebrates [101]. In vertebrates, there are three copies, RXRA, RXRB and RXRG, which is due to the two rounds of genome duplications (1R and 2R-WGD) in the vertebrate ancestor. In the teleost ancestor there has been another whole genome duplication (3R-WGD) and, accordingly, we identified a higher number of RXR genes in fishes. In all the fish species studied, we could identify two copies of the RXRB gene, but the zebrafish showed additional copies of the RXRA and the RXRG. Thus, to insightfully evaluate if the additional RXR gene copies in teleosts have resulted from a scenario of whole genome duplications, we performed detailed synteny analyzes to ascertain if the teleost RXRs were indeed products of whole genome duplication.

Since we did not find evidences of additional RXR gene copies other than the four genes reported here, in medaka, stickleback, tetraodon and fugu genomes, the most parsimonious explanation retrieved from the synteny analysis is that these fishes have underwent trough events of gene loss. Thus, the simplest explanation for the retention of duplicates in vertebrates would be the dosage balance model [110] pertaining to the preservation of duplicates following whole genome duplications.

In addition to the duplicate retention and gene loss patterns, our major finding was the presence of positive selection in the vertebrate RXR right after 2R-WGD and 3R-WGD. This wave of positive selection impacted all of the duplication branches resulting from 2R-WGD and one of the ancestral branches resulting from 3R-WGD; in addition 10% of the sites were under positive selection (site classes 2a and 2b). Interestingly, most of the significantly positive selected sites were located in the LBD and the N-terminal region of the protein which are known to be the variable regions in the nuclear receptors [111], and most of sites in RXR was evolving under strong purifying selection which is in concordance with earlier findings [112].

While all the branches following 2R-WGD were evolving under positive selection, only one of the paralog was evolving under positive selection after the 3R-WGD, similarly significant rate-shift was observed in one paralog of RXRG in zebrafish and only weak signals of rate shift was observed between the paralogs of RXRA. The presence of positive selection/rate-shift on post-duplication branches could signal to a mechanism of escaping from adaptive conflict [28] in the case of the 2R gene copies, and neofunctionalization models explains positive selection only in one post-duplication branch [110] in teleosts, to be the mechanism for preservation of the duplicates following gene duplication.

It has been also suggested that retention of duplicates after WGD is favored by dosage balancing selection, expression (or regulation) divergence and subfunctionalization, and that duplicates (from WGDs) least commonly diverge in their biochemical function [110]. Our results corroborate this hypothesis since the presence of positive selection is only found in the N-terminal region and the LBD of the RXR protein, which are thought to harbor the activation function 1 and 2 domains (AF-1 and AF-2) [113], containing phosphorylation sites for proline-dependent kinases. The presence of positive selection only in the N-terminal region and ligand binding domain makes it possible that the positive selection would have affected the activation function 1 and 2, which is responsible for the spatial and temporal variation of the RXR expression. Thus, significant shifts in DNA-binding function have not occurred due to the positive selection.

When looking at the expression data in homologous anatomic structures of embryonic stages of zebrafish and mouse, expression shuffling is observed between the zebrafish paralogs. An earlier study [107], has found RXRB to be neofunctionalized and RXRA to be subfunctionalized in teleosts, but they have not analyzed the duplicates of RXRG in zebrafish. We found evidence for subfunctionalization and neofunctionalization events in the zebrafish paralogs and we further detected a novel expression (neofunctionalization) pattern in RXRG (Figure 2.5). These results demonstrate that the effect of positive selection or rate shifts on the RXR genes should have enabled the altered expression pattern, which in turn may explain the retention of the duplicates following the increase in the dosage of the protein. This suggests that perhaps the evolution of RXR transcription factors is linked to an increased complexity of its regulation in different cell types. This could be a strategy to fine tune the triggering of the same pathway by

different agents and/or with different intensities in different cell types.

The different regions of the RXR proteins are under different selective pressures. Helix 10, located at the dimerization interface, is the most constrained helix in RXRs, which is consistent with RXR binding many different partners. A mutation in this interface could interfere with various cellular processes [112]. It is known that retinoic acid has a central role in basic biological processes such as cell's fate, survival and growth or the apoptosis, depending on the heterodimeric partners that is activated, i.e., PPAR beta or delta and RAR (alpha, beta or gamma) respectively, in which RXR is the “indispensible dimerization partner” of the nuclear receptors involved [114]. The conservation of the dimerization interface is also consistent with the recent findings that many of the RXR heterodimeric partners, such as RAR, TR, VDR, LXR are not chordate novelties, but were already present in the ancestor of all bilateria [101]. While the ability of RXR to dimerize with other nuclear receptors in basal bilateria remains to be elucidated, the findings of the present study favor this hypothesis.

We conclude that the ability of RXR to bind a conserved set of partners is one of its most important functions, along with its roles in RA signaling [91] and as a lipid sensor [92]. A recent study in insects [115] identified positive selection on the ancestral branch leading to lepidoptera and diptera, specifically in the helix-9 of RXR, which is another helix involved in the dimerization interface (along with helix 7 and 10) with the ecdysone receptor. In our study, we detected positive selection on the post-duplication branch leading to RXRG in one site (405T respective to the human RXRG) of the helix-9 (Appendix 1 SI Table 2).

While dN/dS methods may sometimes retrieve false positives, the power of branch-site models is well recognized [116], showing that our results should not be misinterpreted. Although a large dataset like ours encompassing representatives of the major vertebrate lineages poses challenges in alignment, we overcame such difficulties by removing the non-aligned regions using Gblocks [44], with the resulting dataset showing ample phylogenetic signal (Appendix 1 Figure 11.6) and no saturation bias (Appendix 1 Figure 11.7 and Appendix 1 SI Table 4). The use of dN/dS methods allowed us to detect positive selected sites, most of them in the hitherto known variable regions of the nuclear receptors, consistent with earlier findings [111, 112].

In this study our main goal was to find the major evolutionary factor during the preservation phase of the duplicated RXR paralogs. We provide multiple evidences for the post-duplication evolutionary mechanism of RXR genes. Using the amino acid based methods the retrieved results support altered evolutionary rates in the post-duplication branches (paralogs), while the dN/dS methods find instances of positive selection, both likely contributing to the observed expression pattern changes and functional divergence of the two post-duplication gene copies. Finally, our results from the amino-acid based rate-shift/conservation detection methods and the codon models are consistent, which are further supported by the expression patterns of the genes, as well as by other studies [111, 112, 115]

Further investigation to characterize the RXR from the different orders of Ostariophysi (here

represented by the zebrafish) would be interesting to assess the level of retention of the six RXR genes across this super-order, which could provide valuable insights into the evolution and adaptation of this group of fishes. Since RXR is known to exert its action in the development of the organisms and 28% of the known freshwater fish species belong to Ostariophysi, the remarkable diversity, adaptability and morphological variations among these fish species, especially cypriniformes and siluriformes, makes it an interesting target group to study the evolution of the RXR gene. Interestingly, at least two fish species displaying different number of RXR genes, i.e., medaka and zebrafish, are known to respond differently to the high affinity RXR agonist tributyltin (TBT) [117, 118]. Whereas TBT exposure during the sex differentiation period leads to an almost 100% male zebrafish population, no effects on the sex ratio of medaka were observed upon TBT exposure. Although the experimental demonstration of a link between RXR and the reported differences is still lacking, the data indicates that some caution should be taken in cross-teleost extrapolations.

During the last decade, the need of detailed chemical hazard assessment of a large group of compounds, together with ethical concerns of animal welfare, has prompted the use of zebrafish and other teleosts in large scale chemical risk assessment and drug discovery. However, the fact that teleosts possess more RXR genes than humans should be treated as a major point while generalizing the findings since the expression of the genes and the response of the teleosts could be misleading if generalized to a mammalian context. In contrast to vertebrates, most invertebrates display a single RXR gene. Hence, invertebrate-specific impacts of environmental pollutants acting through RXR and their heterodimeric partners cannot be excluded. Indeed, the TBT-induced imposex observed in female prosobranch gastropods and the synergistic impact of TBT and 20-hydroxyecdysone acting through RXR/EcR in daphnids seems to support this hypothesis [119].

2.4.1. Conclusion

Our results indicated highest constraint in the dimerization helix, allowing us to conclude that the dimerization partners are maintained throughout the evolution of this nuclear receptor. The DNA-binding domain is highly conserved, however, the N-terminal and ligand binding domain which harbor the phosphorylation sites responsible for activation function shows lower constraints and harbours positive selected sites. Thus the evolution of RXR could be linked to an increase in complexity of the organism, where different types of cells that require the same basic biochemical process are triggered by different agents and/or with different intensities. The presence of positive selection/accelerated rates in the paralogs, coupled with the evidences of altered expression of paralogous genes explains the preservation of the additional RXR copies following genome duplication.

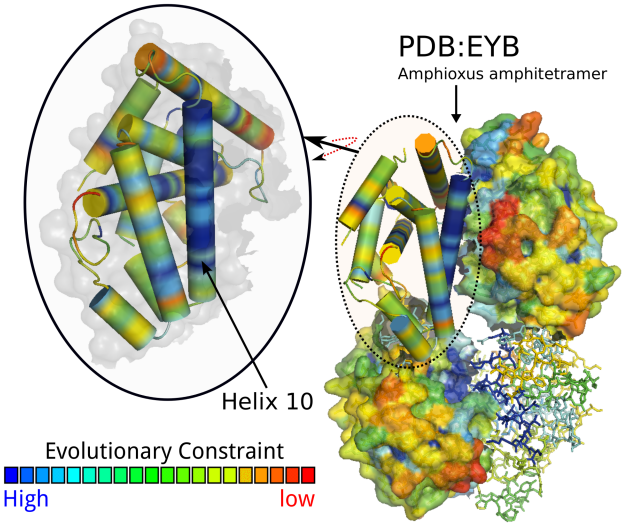


Figure 2.6.: Evolutionary conservation of the RXR Ligand Binding domain: The 3D structure of the amphioxus RXR-LBD 3D structure (PDB ID - 3EYB) was used to map the evolutionary conservation of the metazoan RXRs in PROPHYLER, the helix-10 of the ligand-binding domain is the most conserved cooler colors (blue - green) specify evolutionary conservation and warmer colors (yellow-red) specify lesser constraints.

2.5. Materials and Methods

2.5.1. Sequence alignment and phylogenetic analyses

Vertebrate RXR sequences were downloaded from ENSEMBL database [120]. Sequence alignment was done using MUSCLE [40] and the alignments were viewed and edited in SEAVIEW [121]. Given our large dataset, with representatives of all major vertebrate lineages, we increased the alignment quality by removing the non-aligned regions using Gblocks [44]. A total of 84 sequences of vertebrates were selected for the final analyses. Before proceeding to the evolutionary analyses the dataset was checked for phylogenetic signal with likelihood mapping in TREE-PUZZLE [122] and for saturation bias in DAMBE [123].

The phylogeny was estimated using maximum likelihood (ML) method as implemented in PHYML v.3.0 [124]. The best fit model for the phylogenetic analyzes (GTR+G+I) was chosen with MrAIC [125]. The correct topology of the RXR genes was ascertained using the various topology testing methods implemented in TREE-PUZZLE [122]. For the branch site tests of RXRB teleost paralogs and rate shift analysis of RXRA and RXRG zebrafish paralogs, independent ML phylogenies for each genes were produced.

2.5.2. Synteny analysis: gene gain and gene loss

The synteny analyses, circular plots and the related (orthologs') analyses were done using the Synteny database [104]. The tools provided in this database, such as the Reciprocal

Best Hit BLAST algorithm and the sliding window analysis, allow the detection of gene loss or gene gain in chromosomes by providing evidences of tightly linked genes on the chromosomal segments. The gene orientation and order, clusters connecting between the query and the outgroup species are used to denote orthologous syntenic conservation.

2.5.3. Detection of positive selection

Two different methods were used to evaluate adaptive evolution in the nucleotide sequences. First, a likelihood ratio test between the branch models implemented in PAML 4 [56] was employed. Initially, we compared the log likelihood values of, a two ratio model (where all the post-duplication branches have a different evolutionary rate relative to other branches; $\omega_{PD} \neq \omega_0$) against a one ratio model (where all branches are supposed to evolve at a same rate; $\omega_{PD} = \omega_0$) to find out which model fitted the data better.

Then the two ratio model (unconstrained two ratio model) if found to fit the data better was tested against another null (constrained two-ratio) model where the ω value in the foreground branch was constrained to 1 ($\omega_{PD} = 1$) to check for the prevalence of positive selection. Post-duplication branches leading to RXRBa and RXRBb were tested for positive selection, with a smaller alignment for RXRB gene and the corresponding gene tree, similarly to the above mentioned branch models.

Second, to find out if positive selection acted on a specific post-duplication branch and to identify the sites that were positive selected, we used the modified branch site model implemented in PAML 4. The modified branch site model (branch-site test 2) [57] has been found to be a conservative test of positive selection, which allows the omega value (dN/dS ratio) to vary among the branches and the sites, supporting the hypothesis that the substitutions may vary according to the time and space, rather than averaging the values along all the sites as in the branch models.

In the branch-site test 2, the alternate model assigns two ω values ($0 < \omega_0 > 1$ and $\omega_1 = 1$) for the background branches (all other branches except the foreground branches) and the foreground branches (the branch of interest) is assigned an additional ω value ($\omega_2 > 1$), the alternate model is compared to the null model where the ω value in the foreground branch is constrained to 1, a likelihood ratio test (LRT) is used to check if the foreground branches are evolving under the influence of positive selection. The sites under the influence of positive selection are identified by a Bayes Empirical Bayes [126] (BEB) test.

We tested the post-duplication branches (one branch at a time) to detect episodic events of positive selection, care was taken to ensure that at-least four branches were present on either side of the labeled foreground branch to minimize false positives (due to this requirement the RXRA and RXRG paralogs of zebrafish were not tested using the branch site models). Post-duplication branches leading to *rxrba* and *rxrbb* were tested for positive selection with a smaller alignment for RXRB gene and the corresponding gene tree.

2.5.4. Detection of rate shifts among sites using protein-based methods

We used two methods to check for rate shifts on the RXR genes. First, a likelihood ratio test based method implemented in DIVERGE [64] was used to check for functional divergence in the duplicated/paralogous proteins. DIVERGE calculates a coefficient of functional divergence (θ) between two clades. A value of θ > 0 indicates an altered evolutionary rate at some sites between those clusters. It is of interest (if θ > 0) to identify which sites were evolving at an altered rate; however, we did not look at the site-specific rate-shifts at this stage since we were interested in the post-duplication branches and the sites involved in rate-shifts at those branches.

Secondly, rate shifts of evolution in branches immediately after genome duplication (ancestral branches of RXRA, RXRB, RXRBa, RXRBb, RXRG, RXRG-RXRB, and zebrafish paralogs of RXRA and RXRG) were checked using the RASER2 (RAte Shift EstimatoR version 2) [66]. RASER2 uses the stochastic mapping of mutations [127] to calculate the probability that a rate-shift occurred at a specific branch. We used RASER2 to compare the alternate lineage-specific model to the null model, which does not enable rate shifts. The program was run separately for each of the post-duplication branches. Likelihood-ratio tests (LRT) were performed to determine whether the lineage-specific model fit the data significantly better than the null model.

2.5.5. Expression dataset analysis

To assess the action of whole genome duplication in the expression patterns of the RXR genes we compiled a dataset of RXR expression, for mouse (three RXR genes produced during the first two rounds of whole genome duplications in the vertebrate ancestor or 2R-WGD) and zebrafish (where the RXR compliment is six following a third round of whole genome duplication, in the teleost ancestor, or 3R-WGD), during the embryonic stages from the BGEE [106] database. The mouse genes were considered as singeltons which retain the ancestral functions in comparison to the zebrafish genes, which were considered as duplicates, for which the whole genome duplication (3R-WGD) has increased the dosage of the RXR protein. This dataset was further checked for any altered expression patterns manually.

2.5.6. Evolutionary conservation analysis

ProPhylER (Protein Phylogeny and Evolutionary Rates) [108] was used to evaluate the evolutionary conservation of the RXR protein (the cluster 1731 represents the RXR genes on the server). ProPhylER is a curated database that allows the identification of the evolutionary conservation of protein sequences. This program relies on the assumption that the closely related homologs have not changed in function over time. The evolutionary conservation of the ligand-binding domain was mapped on the three-dimensional (3D) structure of the amphioxus RXR [PDB:3EYB].

We then used MAPP (multivariate analysis of protein polymorphism) [109], to evaluate the evolutionary variation in single columns of the alignment, predicting the impact of all possible variants on the structure and function of the RXR protein. Since the three RXR genes are different in overall protein length and specific domain size, each protein alignment was analyzed separately. MAPP generates impact scores based on six physicochemical properties for all possible variants from the observed evolutionary variation and provides the correspondent P-values - the lower the P-value, the higher the chance that the substitution will be deleterious for the structure or the function of the protein.

2.5.7. Protein tertiary structure modeling

Relevant amino acid sites were mapped on the tertiary structures (homology model of the RXR DBD-LBD region) using SWISS-MODEL (http://swissmodel.expasy.org/). The visualization and editing of the 3D structures was performed in PyMOL [128].

2.6. Tables

Table 2.1.: Likelihood parameter estimates under lineage specific models.

Model	*lnL*	ω_0	ω_{PD} [a]	LRT ($2\Delta lnL$)
One Ratio(1Ra)	-32986.908	0.049	NA	NA
Two Ratio constrained	-32951.918	0.046	1	NA
Two Ratio	-32951.912	0.047	1.11	Vs One ratio 69.99($p<0.01$) [b] Vs Two ratio constrained 0.013 (NS) [c]
PD branch (RXRBa and RXRBb)				
Model	*lnL*	ω_0	ω_{PD}	LRT ($2\Delta lnL$)
One Ratio	-10522.007	0.057	NA	NA
Two Ratio constrained	-10565.773	0.049	1	NA
Two Ratio	-10516.565	0.053	0.109	Vs One ratio 10.88 ($p < 0.01$) [b] Vs Two ratio constrained 98.4 ($p < 0.01$) [c]

[a] Values of 999 for dN/dS indicate dS = 0, so dN/dS is undefined
[b] Vs One ratio
[c] Vs Two ratio constrained

Table 2.2.: Likelihood parameter estimates under branch site models on Post-Duplication branches (RXRA, RXRB and RXRG).

Model	*lnL*	Proportion of sites	ω_{PD}[a]	LRT (2Δ*lnL*)
BS null model - whole vertebrate tree	-32495.157	P0=0.78391 p1=0.05835 p2a=0.14681 p2b=0.01093	1	NA
BS test for positive selection on ancestral branch of RXRA	-32501.17	P0=0.83949 p1=0.06194 p2a=0.09180 p2b=0.00677	999	12.027 (P < 0.01)
Branch-site test for positive selection on ancestral branch of RXRB	-32485.585	P0=0.85621, p1=0.06275, p2a=0.07550, p2b=0.00553	999	19.14 (P < 0.01)
Branch-site test for positive selection on ancestral branch of RXRG	-32482.212	P0=0.77362, p1=0.05798, p2a=0.15666, p2b=0.01174	999	25.89 (P < 0.01)
Branch-site test for positive selection on ancestral branch of RXRB and RXRG	-32501.17	P0=0.83949, p1=0.06194, p2a=0.09180, p2b=0.00677	999	12.027 (P < 0.01)

[a] Values of 999 for dN/dS indicate dS = 0, so dN/dS is undefined

Table 2.3.: Likelihood parameter estimates under branch site models on Post-Duplication branches (RXRBa and RXRBb in teleosts).

Model	*lnL*	Proportion of sites	ω_{PD}[a]	LRT (2Δ*lnL*)
Branch-site null model for the RXRB tree	-10470.676	P0=0.84682 p1=0.03646 p2a=0.11190 p2b=0.00482	1	NA
Branch-site test for positive selection on ancestral branch of RXRBa	-10472.312	p0=0.89881 p1=0.03430 p2a=0.06443 p2b=0.00246	33.97754	3.27 (P < 0.05)
Branch-site test for positive selection on ancestral branch of RXRBb	-10469.822	P0=0.88937 p1=0.03803 p2a=0.06963	2.41428	1,708 (NS)

[a] Values of 999 for dN/dS indicate dS = 0, so dN/dS is undefined

Table 2.4.: Likelihood parameter estimates under lineage specific models for rate shifts among sites between different RXRs.

Model	*lnL*	LRT (2Δ*lnL*)
Null model	-10355.6	
Rate shift on Ancestral branch of RXRA	-10305.8	99.6
Rate shift on Ancestral branch of RXRB	-10285.5	140.2
Rate shift on Ancestral branch of RXRG	-10285.5	140.2
Rate shift on Ancestral branch of RXRB and RXRG	-10285.5	140.2
Analysis of post duplication branches leading to teleost paralogs		
Model	*lnL*	LRT (2Δ*lnL*)
Null model RXRB	-4218.13	
Rate shift on Ancestral branch of RXRBa	-4217.14	1.98 [a]
Rate shift on Ancestral branch of RXRBb	-4213.46	9.34
Null model for RXRA	-3564.79	
Rate shift on Ancestral branch of RXRAa	-3565.87	2.16 [a]
Rate shift on Ancestral branch of RXRAb	-3563.48	2.62 [a]
Null model for RXRG	-4493.22	
Rate shift on Ancestral branch of RXRGa	-4493.22	0 [a]
Rate shift on Ancestral branch of RXRGb	-4485.59	15.26

[a] Not significant

3

Adaptive evolution of the Rhodopsin 1 (*RH 1*) in teleost fishes: spectral tuning for scotopic vision

3.1. Abstract

Teleost fishes can be found in diverse photic environments and molecular changes in rhodopsin (RH1), a pigment predominantly found in rod photoreceptor cells, could influence their scotopic vision. To evaluate such hypothesis, we have used complementary codon based and amino acid physicochemical property/structure based analyzes to identify signatures of positive selection in the teleost RH1. Among the positive selected sites identified (20% of the protein length), four were involved in spectral tuning and other five were related with structural and inter-molecular interactions of RH1. We also found that all the spectral tuning sites in the protein were evolving non-neutrally and were the sites that could tolerate changes. The presence of fast evolving (positive selected) sites in 20% of the protein length raises the question whether rhodopsin could still be a suitable phylogenetic marker in teleostei. However, we find that rhodopsin sequences from different teleosts superorders have an overlap of the base composition and the codon usage patterns providing reliability to their use as phylogenetic marker.

3.2. Introduction

Vertebrate visual pigments can be broadly classified into six evolutionarily distinct groups, which includes rhodopsin, long, medium, short and non-retinal (pineal gland specific) opsins (RH1, RH2, LWS/MWS, SWS1, SWS2, and P) [80, 81]. Among these, rhodopsin (RH1) is a distinct group, found predominantly in the rod photoreceptor cells and involved in the scotopic vision [82]. The teleost RH1 differs characteristically by possessing just one exon 1 kilo base pairs (bp) long [129, 81], while all other opsin (visual pigment) genes possess either five (RH1, RH2, SWS1, SWS2 and P) or six (LWS/MWS) exons.

Visual pigment (opsin), which is an integral heptahelical transmembrane protein (or G-protein coupled receptor), absorbs a photon to start the process of vision [130]. The spectral tuning (wavelength of maximum absorption) of each pigment is based on the interaction of the amino acids in the visual pigment to a chromophore, which it is linked to [131, 132]. The spectral tuning could be achieved at the physiological level by exchanging the chromophores (A1 to A2) or at the DNA level by adaptive mutations [130, 80].

The behaviour and interaction of the animal to its photic environment could be directly linked to the arrangement and density of the rod and cone cells (photoreceptor cells) in their eyes and in turn the amount of visual pigment (opsins – rhodopsins and cone-opsins) and its maximal absorption.

Since the photic environments define the visual pigment in the animals [133, 83, 134], they could be excellent targets to study adaptive molecular evolution [83]. Studies involving the vision genes (or visual pigments) are thus particularly interesting since they could point to the molecular adaptive mechanisms or provide insight into the adaptive radiation of animals.

Since teleosts are found in diverse photic environments they have been the target of various studies linking their adaptation with visual pigment amino-acid changes and sensitivity [135, 136, 82, 137, 131, 83, 134]. Also owing to their peculiar characteristic of possessing just one exon numerous studies have used teleost RH1 gene as a major phylogenetic marker [138, 139, 140, 141].

In this study, we analyzed the coding regions of the teleost RH1 genes in a comparative genomics framework. We focus on the variation in the RH1 gene within and between major teleost superorders. Specifically, we assessed differential selective pressures of the teleost rhodopsin protein to infer adaptive evolution, related with possible changes in habits and habitats of the studied organism. Since RH1 has been used frequently as a common molecular marker in teleost phylogenetic studies, we also assessed if positive selection on the RH1 could influence phylogenetic reconstructions. Our analyzes detected 20 sites to be under the influence of positive selection using codon models and 55 sites when using amino-acid physicochemical property based models. Out of the positively selected sites, four have been related with the spectral tuning of fishes. We also found that all the sites involved in spectral tuning are evolving non-neutrally or under relaxed selection pressures. Eight of the positively selected sites in

teleosts have their corresponding (homologous) sites in humans implicated in visual disorders and related phenotypes. Base compositional bias and codon usage patterns among different teleost superorders supported the suitability of RH1 for phylogenetic inference in teleosts.

3.3. Materials and Methods

3.3.1. Data-mining alignment and tree building

All sequences of teleost RH1 in the GenBank were downloaded to prepare two datasets for downstream analyzes. We used a cut off of 80% sequence coverage (length) for one dataset and the second dataset comprised complete coding sequences of the RH1. The first dataset consisted of 765 sequences of >800 bp length (>80% of the protein), comprising seven of the 12 super-orders of teleostei, and those sequences were used to calculate the basic sequence characteristics. The second dataset consisted of 61 sequences comprising the complete coding sequence of teleost RH1 and were used for the detailed adaptive evolutionary analyzes. Similar sequences (zero genetic distance) of closely related (cryptic) species were removed to avoid bias of the adaptive evolutionary analyzes.

Complete sequences of rhodopsin were searched in the GenBank database using different BLAST [142] approaches like *tblastn* or *pblast*. We also annotated novel rhodopsin genes (previously un-reported) using the protocol detailed below: the whole genome databases at GenBank for fishes was searched with *pblast* or other methods from the *BLAST* protocol (*Danio rerio* sequence was used as query) to find out the contig (or scaffold) that contained a putative rhodopsin gene. This contig (or scaffold) was used as a query for the "genescan" [143, 144] (genes.mit.edu) gene finding program, to detect the genes coded by that contig. The results of *genescan* were manually checked for the presence of rhodopsin; and the probable rhodopsin sequence was used as a query in BLAST (to find if the annotated sequences reciprocally hit on rhodopsins) and CD-search [145] in the NCBI conserved domain database [146], to ensure they had the conserved domains of rhodopsin sequences. The eels *Anguilla japonica* and *Conger myiaster* and the scabbardfish *Lepidopus fitchii* have two paralogs of RH1, but paralogs have not been identified in other species [147].

The datasets were translated into proteins and aligned using MUSCLE [40] backtranslated and manually checked for errors in SEAVIEW [121]. The dataset for basic sequence characteristic analysis were aligned super-order wise (for Ostariophysi, Elopomorpha, Scopelomorpha, Stenopterygii, Paracanthoperygii, Protacanthopterygii and Acanthopterygii) and analyzed separately in MEGA5 [148]. Base compositions for each codon positions (for all sequences in the super-order specific alignment), percent codon usage and the relative synonymous codon usage were calculated.

The best fitting nucleotide substitution model for the complete coding sequence dataset was found using MrAIC [125], and the best maximum likelihood tree search was carried out, using

this model, in GARLI [149], where 10 searches of two replicates each were conducted to find the best likelihood tree. Support values at the nodes of the best likelihood tree were annotated using *sumtrees* program [150] following 100 bootstrap replicate runs in GARLI [149]. For tree building and positive selection analyzes, four chondrichthyes species' sequences were used as out-groups.

3.3.2. Evolutionary genetic analyzes

Extensive selection analyzes were done for the complete sequences dataset using the HYPHY software package [54] and the *datamonkey* web-server [151](www.datamonkey.org). To identify any positively selected sites in the rhodopsin sequence, we used five methods: REL, FEL, SLAC, FUBAR and MEME [152, 153, 154, 155]. To identify the positively selected branches (episodic positive selection) on the tree we used the recently developed Branch-Site-REL approach [156], which is found to be more robust than the classical branch-site models of positive selection and does not need *a priori* specification of branches of interest. However this method does not identify the sites (positive selected), on the positive selected branches. The most recent method of this family of positive selection detection methods – MEME – finds sites subjected to episodic events of positive selection, thus our usage of Branch-SiteREL in conjunction with MEME circumvents the problems of not using the Branch-Site test proposed earlier [57] on (every branch) a branch-by-branch basis.

SLAC is an improved derivative of the Suzuki-Gojobori counting approach [54, 151] that accounts for changes in the phylogeny to estimate selection on a site-by-site basis. SLAC also calculates the number of non-synonymous and synonymous substitutions that have occurred at each site using ML reconstructions of ancestral sequences [54, 151]. The FEL model estimates the ratio of non-synonymous to synonymous without assuming an *a priori* distribution of rates across sites in a site-by-site analysis. FUBAR is a method to model and detect selection much faster than existing methods and to leverage Bayesian MCMC to robustly account for parameter estimation errors [151]. MEME is capable of identifying instances of both episodic and pervasive positive selection at the level of an individual site [154]. REL is an extension of familiar codon-based selection analyzes in PAML [157, 158], however, REL allows synonymous rate variation [153].

To find out the exact sites that were positively selected on each of the branches identified to be under positive selection by Branch-siteREL method, we used the CODEML program from the PAML package [56]. The branches identified earlier were labeled as the foreground branches independently (one at a time) in the branch-site test 2. Here two ω values ($0 < \omega_0 > 1$ and $\omega_1 = 1$) are assigned for the background branches and the foreground branches are assigned a ω value ($\omega_2 > 1$). This model is compared with the null model where the ω value in the foreground branch is constrained to 1, and a LRT is used to check if the foreground branches are evolving under the influence of positive selection. The branch-site test segregates the

amino acid positions/codons into four different categories. Two categories describe sites for which selective pressure does not change over time, either under purifying selection (site class 0, $\omega_0 < 1$) or under neutral evolution (site class 1, $\omega_1 = 1$). The two other categories (site classes 2a and 2b) are sites potentially evolving under positive selection only in the foreground branches ($\omega_2 > 1$), and evolving in the background branches (all the remaining branches) under purifying selection (site class 2a, background branches $\omega_0 < 1$) or neutral evolution (site class 2b, background branches $\omega_1 = 1$). The sites under the influence of positive selection were identified by a Bayes Empirical Bayes (BEB) test [126] applying a threshold of >0.95 posterior probability.

Since dN/dS methods can be biased in detecting positive selection across conserved sequences [159], we used a complementary approach to detect positive selection at the level of amino acids with TreeSAAP [61]. TREESAAP categorizes the physico-chemical changes owing to amino acid replacements into eight magnitude categories ranging from 1 to 8, with 1 being the most conservative and 8 the most radical, and then determines whether the observed magnitude of amino acid changes deviates significantly from neutral expectations. Significant positive z-scores indicate that higher magnitude non-synonymous substitutions are more frequent than expected under neutrality, implying change owing to positive selection [160]. A sliding window of 15 codons was used, and 31 amino acid properties were included for the selection analysis [159]. We only considered amino acid replacements with magnitude categories 7 and 8 and significant ($p < 0.01$), positive z-scores as being under positive selection. We also imposed an empirical cut-off of at least three physicochemical properties to be selected to categorize the amino acid as positive selected. The positively selected sites were mapped onto the 2D structure of bovine rhodopsin [161].

Multivariate analysis of Protein Polymorphism

We used the multivariate analysis of protein polymorphism (MAPP) [109] at the protein level to check for the evolutionary conservation and acceleration of amino-acid sites. The MAPP is based on two premises, (i.) physico-chemical variation of amino-acids (at a site) result in the formation of a deleterious/impaired protein instead of a normal protein; (ii.) the evolutionary variation at an amino-acid site among orthologs points towards all the variations that could be tolerated at that position.

The analysis in MAPP takes into account five key physicochemical properties:

- Hydropathy
- Polarity
- Charge
- Volume

- Free energy in alpha-helix conformation
- Free energy in beta-strand conformation

The analysis estimates the physicochemical constraints for each of these properties at each position. Finally impact scores (MAPP scores) are generated de-correlating the scores of each property and by a principal component transformation. If a site has a low MAPP score it hints at a higher rate of physicochemical variation (tolerate more changes; see Figure 3.1a), hence positive selected sites and adaptive sites should fall in this category. If the MAPP score for a particular site is high it points to a scenario where the site does not tolerate much variation in physicochemical properties and hence these sites could be considered to be under the influence of purifying selection.

Identifying non-neutrally evolving nucleotide positions

We employed the PHAST package [162, 163] to identify non-neutrally evolving sites in the nucleotide alignment of fish RH1. First, *phyloFit* was used to fit the tree model to the multiple alignment of DNA sequences by maximum likelihood using the specified tree topology and the REV [164] substitution model with four rate categories (option "--subst-mod REV –nrates 4"). This model was used to generate "Phylogenetic p-values" using *phyloP* [162], which computes lineage specific or global p-values for conservation or acceleration from the alignment, a phylogenetic tree and a model of neutral evolution. *PhyloP* identifies the departures from neutrality in either direction (conservation or acceleration), for each nucleotide column in the alignment using methods of similar statistical power [162]. We used the likelihood ratio test analysis (LRT), since it is based on the full likelihood function and is expected to make better use of the substitution pattern, is more robust if there were periods of extreme selection and since this is the preferred method for testing each site of the alignment (see [162]). We generated two-sided p-values such that a small p indicates an unexpected departure from neutrality using the NNEUT option.

Homology modeling and annotation of positively selected sites

Tertiary structures of RH1 protein were (homology) modelled in SWISS-MODEL [165]. The protein sequences of all the branches (tips) found to be positively selected by the Branch-siteREL method were used to generate tertiary structures. The 3D structures were visualized and positively selected sites were annotated using PYMOL [128]. Conservation scores generated from MAPP was mapped onto the 3D structure of *Gadus morhua* (homology modeled) protein.

3.4. Results

Our dataset for positive selection methods included 68 complete rhodopsin sequences from cartilaginous and bony fishes. We annotated rhodopsin for four species (*Gadus morhua*, *Xiphophorus maculatus*, *Callorhinchus milli* and *Leucoraja erinicea*) from the genomes hitherto un-annotated but available in NCBI, after confirming the exact match with rhodopsin using conserved domain search. Using different codon based positive selection methods, 20 sites were found to be positive selected in the teleost rhodopsin protein. Out of the 20 sites except two (96 and 261) which were found to be involved in spectral tuning by an earlier study [147], all the other sites are novel identifications and could be functionally important albeit most probably not in spectral tuning. Two sites were found to be positively selected by three different methods (Sites 54 and 165; Figure 3.1b). Site 54 was selected by SLAC, REL and FEL methods, while site 165 was selected using FEL, MEME and FUBAR, site 277 was selected with both REL and MEME. Among the sites, MEME identified 12 as under the influence of positive selection.

Using the improved Branch-Site REL method, we could identify nine branches to be positively selected, out of which seven were terminal branches (Figure 3.2), $p < 0.08$ after correcting for false discovery rates using the Benjamini-Yekutieli method [166].

Using the amino acid based method of positive selection (TreeSAAP) we could identify 55 sites to be under positive selection (Figure 3.3) with at least three properties under the influence of positive selection. Of these 55 sites, (positively selected sites) two sites 122 and 194 were found to be involved in spectral tuning (functional differentiation) of rhodopsins in an earlier study [147]. Of the 55 sites, eight sites (116, 119, 158, 165, 205, 213, 217, 277) were found to be positively selected by an earlier study on rock fishes [134]. Two sites, 165 and 217, found to be positively selected by TreeSAAP (165 also by MEME) were found to be positively selected in sand goby fishes by an earlier study [132].

Since the dN/dS methods could only identify a couple of sites found to be involved in the spectral tuning of the protein (identified by mutagenesis studies) [147], we studied the protein's site-by-site rate variation at the protein and nucleotide level. Our analysis of amino-acid site wise physicochemical property variation (MAPP), also presented similar results to the PHAST (nucleotide based) analysis. The analysis shows that the positions found to be positively selected by TreeSAAP, codon models and the positions found to be involved in spectral tuning can tolerate more variation, thus are less deleterious/constrained (Figure 3.3 and Figure 3.4).

The PHAST analysis gave p-values (two-sided; 0 and above) showing the chance of each site departing from the assumption of neutrality. We imposed an empirical cut-off value of 1 below which we considered that the site showed "strong" tendency to deviate from neutrality. The results could identify eight out of the 12 sites (Figure 3.4) found to be involved in the spectral tuning to be evolving non-neutrally under this criterion imposed by us. For site 96 the p-values were less than 2.77 (the mean of all the p-values) at all codon positions and for sites 102 and 317 the third codon position was found to have a p-value lesser than 2.77. Thus, these three

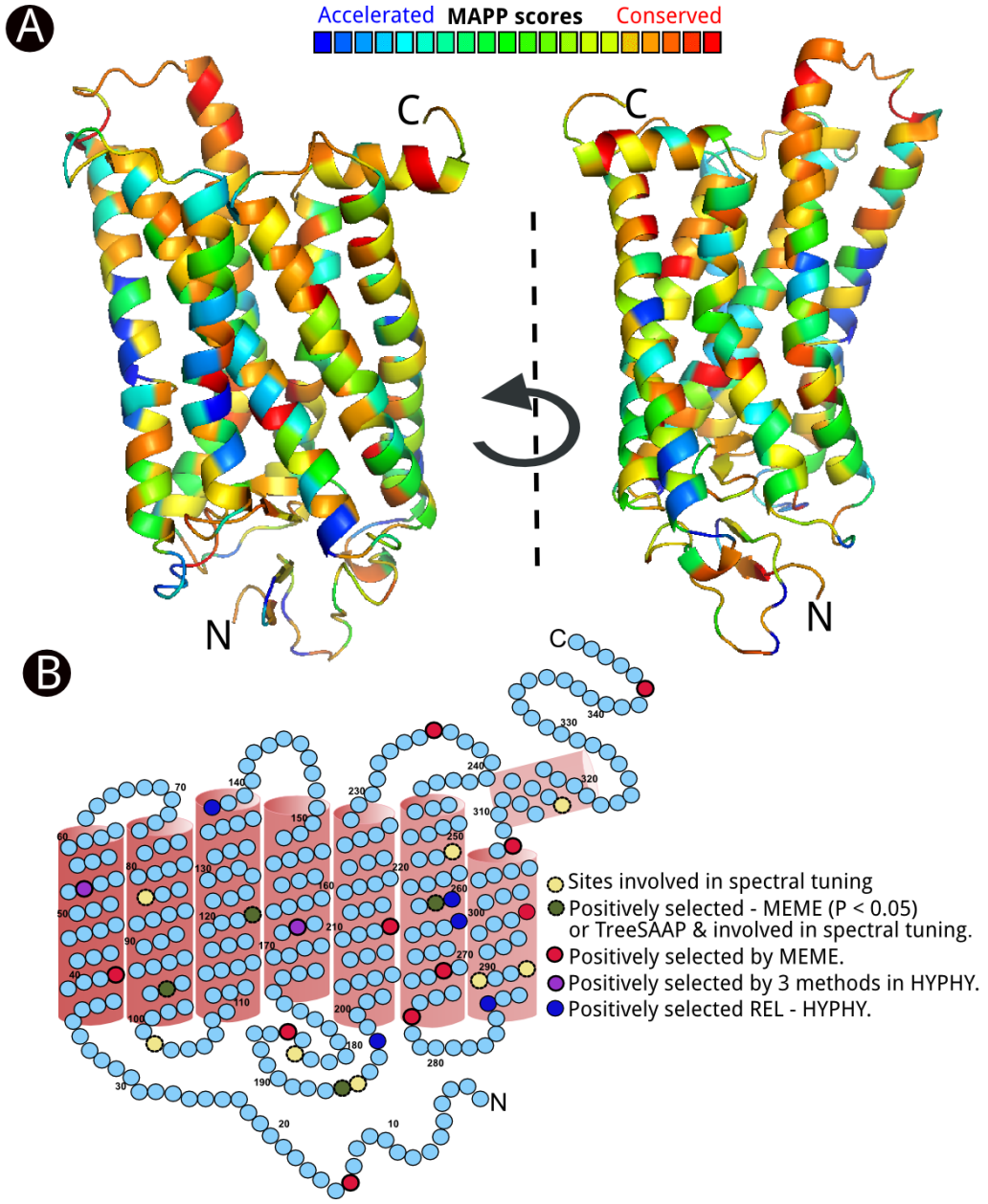


Figure 3.1.: A) Evolutionary conservation (MAPP scores) of the teleost rhodopsins are plotted onto the homology model of *Gadus morhua* rhodopsin protein; B) positively selected sites using the codon models are plotted on the 2D structure of the bovine rhodopsin [161].

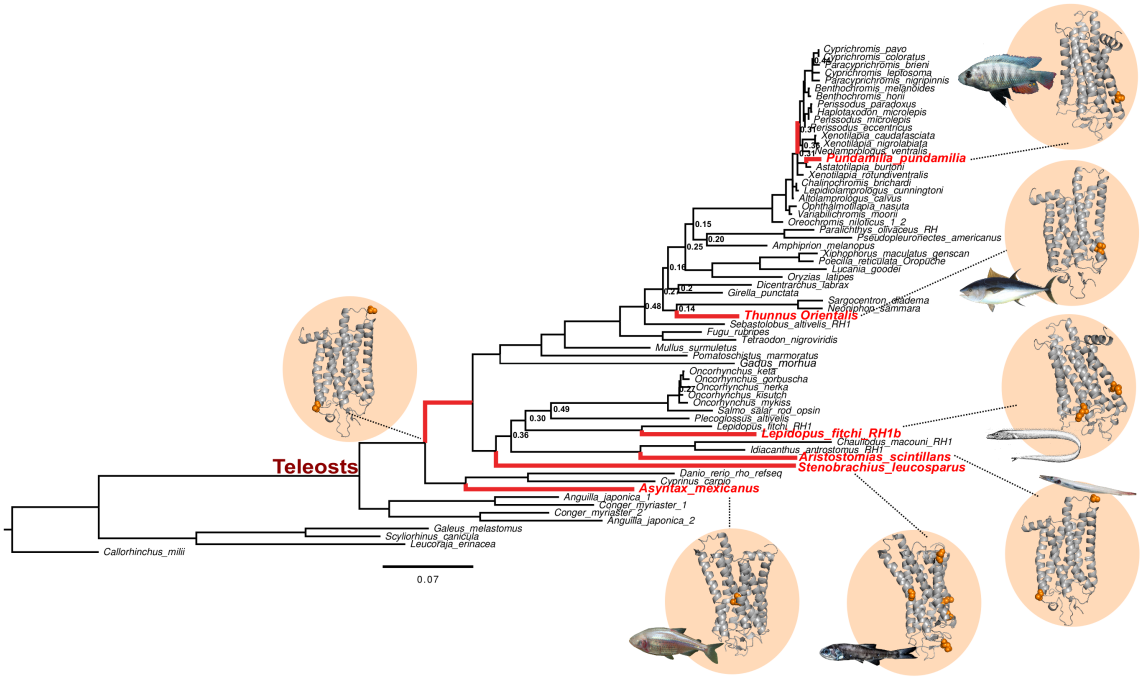


Figure 3.2.: Results of the positive selection analysis of the teleost RH1 gene is shown, the bold (red) branches on the phylogenetic tree are the ones undergoing episodic positive selection (Branch-siteREL). The homology model of each species' protein is annotated with the sites found to be positively selected on that branch by the CODEML branch site method. The *Gadus morhua* protein model was used to annotate the sites on internal branch found to be positively selected by CODEML. Note that two branches, one for *Gadus morhua* and another for the ancestral chiclidae when labeled as foreground branches did not yield any sites as positively selected in CODEML and thus are not represented by 3D structures here.

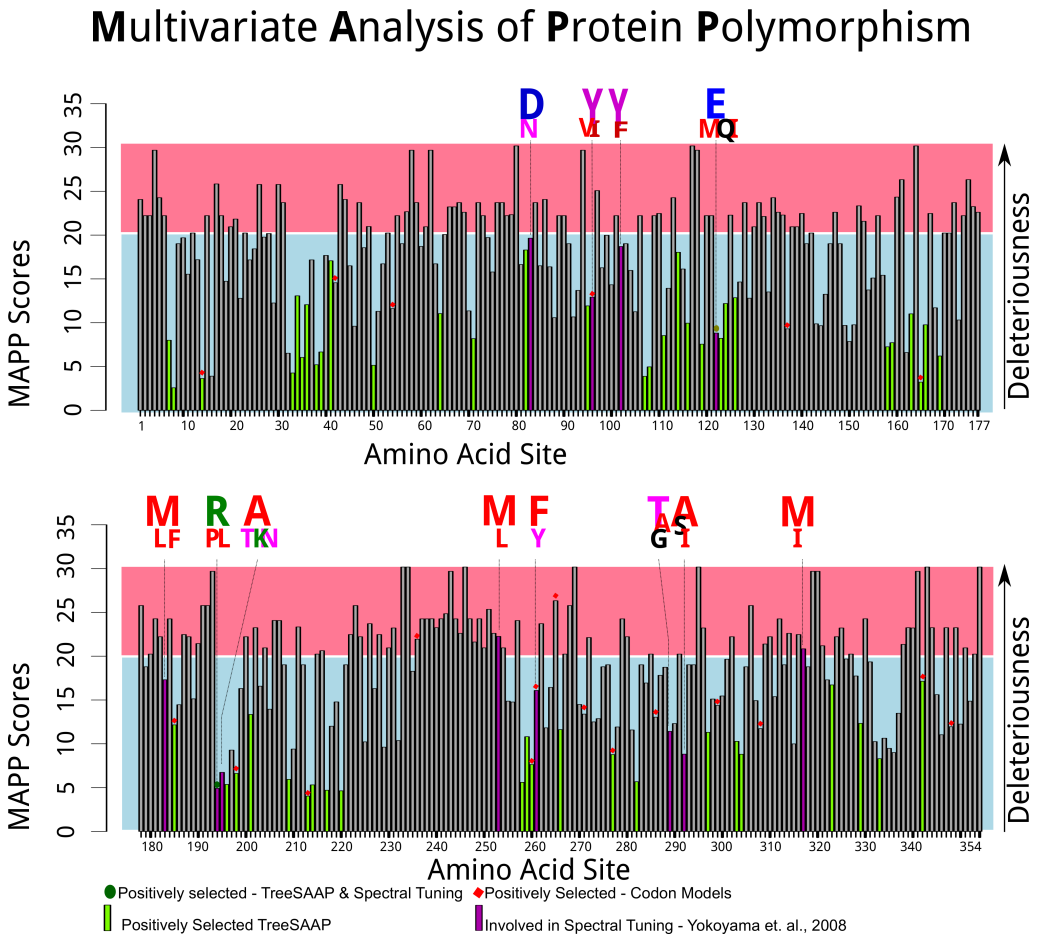


Figure 3.3.: Results of multivariate analysis of protein polymorphism, the sites involved in spectral tuning [147] are plotted as purple bars and have the different amino-acids found in that position shown above, the positively selected sites identified by TreeSAAP are shown as green bars and the positively selected sites according to the codon models have a red spot annotated above the respective bars. Note that as the MAPP value increases fewer substitutions are tolerated at that site.

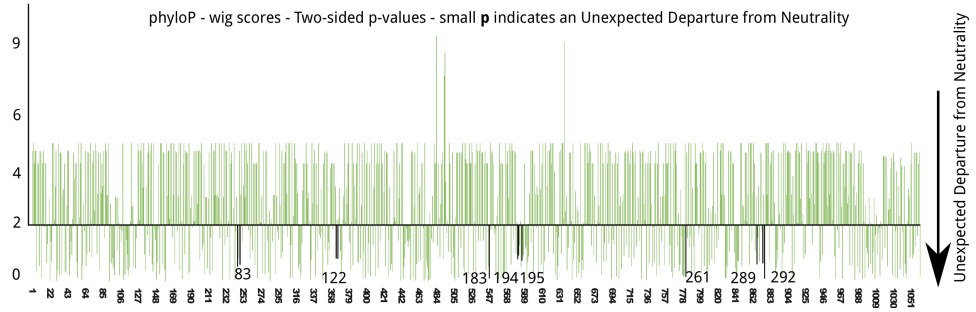


Figure 3.4.: Results of the test for non-neutral evolution conducted in PHAST. Barplot of the *phyloP* p-values for non-neutrality (NNEUT) are shown, the p-values for the sites identified earlier to be important for spectral tuning are in bold with the site number given near the bars. Smaller p-value denote a significant deviation from the assumption of neutral evolution.

sites, 96, 102 and 317, can also be considered to be deviating from the neutral evolutionary rate. In short, with the exception of one site (253) all the other sites involved in the spectral tuning [147], were found to evolve non-neutrally.

To conclude we identified several sites under positive selection using different complementary codon-based and amino-acid based methods (Figure 3.1, 3.2 and 3.3). A previous major criticism against codon models stemmed from the fact that it could not identify any of the functional (spectral tuning) sites to be under positive selection [147]. Here, using more powerful and recently developed methods we have identified some of those functional sites to be under positive selection using dN/dS metrics (Figures 3.1, 3.2 and 3.3). We also identified that the sites involved in spectral tuning evolve non-neutrally. This could point to a scenario where the sites involved in spectral tuning are actually under relaxed selective constraints since we found them to be evolving non-neutrally but most of them are not positively selected. Using protein based and nucleotide based evolutionary rate analysis (PHAST and MAPP) we identified that those sites were actually evolving under non-neutrality.

Suitability of RH1 as a phylogenetic marker

A total of 765 sequences with a length of more than 800 base pairs were used for the basic sequence composition analysis. The majority of the sequences were from the ostariophysi and acanthopterygii groups. A discriminant analysis of the principal components [167] of the base composition data (grouped according to the super-orders) showed the overlap of the clusters and thus pointed to very less base compositional difference among superorders (Figure 3.5a).The relative synonymous codon usage pattern also shows overlapping values for all superorders (Figure 3.5b).

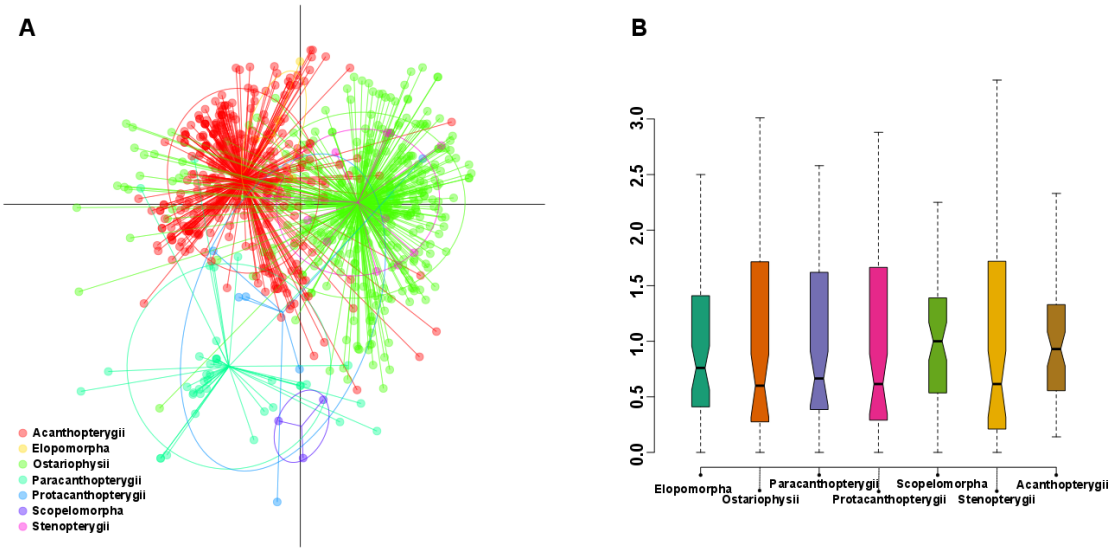


Figure 3.5.: Base composition and Codon Bias of the RH1 in teleost super orders:(a) the differential analysis of the principal components (DAPC) plot of the superorder wise base composition data; (b) violin-plots of relative synonymous codon usage of different teleost superorders. The values for different superorders are overlapping for both the metrics making them suitable as a phylogenetic marker at the level of teleostei.

3.5. Discussion

Rhodopsin gene in teleosts has only one exon [129, 81] of 1000 bp long making it a preferred phylogenetic marker. However, previous studies have not taken into account its basic sequence characteristics and evolutionary rates. In this study, we looked at the basic sequence characteristics like base composition and codon bias, which are important metrics while performing a phylogenetic reconstruction. In addition to the basic sequence characteristics, we also looked at the selection at the species level and protein level. Selection on rhodopsin could be directly linked to the habitat preference of the organism (like dim light vision, deep sea life, cave dwelling, etc.), and thus important in characterizing the evolutionary history of teleosts as a group.

The results showed interesting trends where rhodopsin gene could be positively selected due to the habitat and habit of the species, which included blind/cave dwelling fish (*Asyntax mexicanus*), deep sea fishes with diurnal migration (*Stenobrachius leucosparus* and *Aristomias scintillans*), migratory fishes (*Thynnus orientalis* and *Gadus morhua*), and a rocky shore dwelling fish (*Pundamilla pundamilla*) (Figure 3.2). One of the paralogs of the rhodopsin in *Lepidopus fitchii* was also positively selected pointing to a relaxation of the selection pressure on one copy following gene duplication aiding the preservation of both copies.

Our usage of the mixed effects model of evolution (MEME) identified 12 sites (Figure 3.1) subjected to episodic events of positive selection during the teleost evolution. MEME has been previously used to test for positive selection [154], and has been proved to be a powerful

approach. MEME identified 12 sites under positive selection while REL and FEL identified seven and three sites, respectively.

Out of the seven sites identified by REL only two were identified by MEME and FEL. Out of the 20 sites identified (by different codon models) two sites, 54 and 165, were identified to be positively selected by more than three methods (54 by SLAC, REL, FEL and MEME at $p< 0.1$ level; 165 by FEL, MEME and FUBAR), site 54 has been identified as positive selected by FEL and MEME in an earlier study of vertebrate rhodopsins [154]. Site 96 identified to be positive selected by MEME, was shown to be positively selected by the same method in an earlier study of vertebrates [154] and was found to be involved in the spectral tuning of rhodopsins [147]. Site 261 (found to be positively selected by MEME) is also involved in spectral tuning of rhodopsin [147] (Figure 3.1), the other positively selected sites do not play a role in spectral tuning but may be involved in some other activities of adaptive significance.

Fifty five sites were identified to be positively selected at more than three physicochemical properties by TreeSAAP (Figure 3.3). Sites 122 and 194 found to be involved in the functional differentiation [147] were detected as positively selected by TreeSAAP.

Many of the positive selected sites identified in our analyzes have been implicated in structural and functional modifications of rhodopsin and in rhodopsin related phenotypes, such as autosomal dominant retinitis pigmentosa (ADRP) [168, 169]. Sites 137 and 299 found by codon models are implicated in retina malfunctioning and blindness, V137M cause ADRP in humans while A299S is a polymorphism related to blindness [168]. It should be noted that at site 137 either Valine or Methionine can be found in fishes but this mutation is a reason for ADRP in humans. Sites G114D, Q184P, S186P(T), V209M, F220C(L), and S297A are implicated in human ADRP [168, 169], which were found to be under positive selection in fishes.

Most of these mutations (114,184,186) form abnormal proteins that are retained in the endoplasmic reticulum and do not easily reconstitute with *11-cis-retinal*. Mutation at site 220 affect dimerization capability of the protein and a site 137 mutated rhodopsin showed increased activity for *transducin* [169]. In addition, sites 144, 122 and 186, which were detected as under positive selection in our analyzes, are implicated in *retinal* interaction of the pigment and site 50, which was also positively selected, is involved in the exit of *trans-retinal* from the pigment [169]. The positively selected sites, 64, 71, 236, 329, 333, 342 etc., fall on the cytoplasmic side while sites 7, 8, 34, 185, 186, 194, 198 and 282, fall on the extracellular portion of the rhodopsin protein, which could have important functional consequences.

Thus, many of the sites identified to be positively selected in the fish rhodopsin may have functional or structural implications. Indeed, four positively selected sites are involved in the spectral tuning of the protein and another five sites are involved in the structural integrity and interaction with other molecules. In addition, eight of the positively selected sites in fish RH1 are implicated in ADRP in humans.

Overall we identified many sites to be positively selected by different methods. Five sites

(14, 198, 213, 260 and 277) were found to be positively selected by at least one codon model and the amino-acid based method. In addition, two sites under positive selection (96 and 261) using codon models and two sites under positive selection (122 and 194) using amino-acid models were found to be involved in spectral tuning of teleost rhodopsin in earlier functional characterizations [147].

Although the majority of the positively selected sites are not involved in spectral tuning, it should not be mistaken that the (positively selected) sites have no significance in the function of rhodopsin. As discussed above, most of the sites found to be positively selected could be involved in functions other than spectral tuning of the pigment, as eight of those sites have been related with ADRP in higher vertebrates and five other sites in inter-molecular interactions. Only 3% have been referred previously to be involved in spectral tuning of the protein (pigment) [147], but our study identified another 5% and 15% of sites as positively selected by codon models and by variations of amino-acid physicochemical properties, respectively. Those sites found to be positively selected and without any known functional or structural implications could be the targets of further biochemical experiments.

Our analyzes have confirmed the results of earlier studies (e.g, [59]), and showed that the sites involved in spectral tuning are non-neutrally evolving (Figure 3.4), although most of them were not found to be positively selected (either by codon-models or amino-acid based methods). Our analyzes also highlight the importance of using complementary methods while looking at the evolution of proteins, since some could be less powerful to detect certain molecular signals, which may be obvious to others. The positive selection analyzes identified adaptive evolution in species with dim light vision and uncovered many new sites under episodic positive selection, which could not be previously detected due to lack of powerful methods.

In conclusion, we have characterized the evolution of the teleost rhodopsin, showing that almost 20% of the sites evolve under positive selection. However, the teleost sequences of different super-orders showed similar basic sequence characteristics (Figure 3.5), still providing accuracy for phylogenetic analysis. All sequences available for teleosts (adhering to the quality check we imposed) have been analyzed.

Further studies related to the functional characterization of positively selected sites identified here could provide insights to understand the adaptive mechanisms of rhodopsin in fish, specifically the advantages other than spectral tuning. Sequencing of species specialized for dim light vision, like cave dwelling catfishes and eels, and deep sea fishes, could be of major interest to uncover parallel adaptive mutations in organisms under different habitat pressures, which would highlight the important sites other than the spectral tuning site that are important in species adaptation.

4

Did natural selection of detoxification genes in birds influence in fine-tuning the mode of locomotion and delay the onset of aging? Adaptive evolution of the avian Superoxide Dismutases

4.1. Abstract

Superoxide Dismutase (SOD) is an important detoxifying enzyme forming the first line of defense against reactive oxygen species (ROS) produced in the cell. Owing to the antioxidant properties, SODs are often considered as enzymes important in delaying the onset of aging and mitigating the ill effects of free radicals. Avian species with a same basal metabolic rate as its mammalian counterparts are known to have a higher lifespan, even though flight, a strenuous exercise, increases the amount of ROS produced in birds. It is also know that birds have an efficient respiratory chain, which minimize the production of ROS compared to its mammalian counterparts. Thus, in normalcy birds produce lesser ROS but during flight they produce higher amounts of ROS, which represents the typical pressure to the avian detoxifying enzymes such as SODs. Here, we assessed if positive selection or accelerated evolutionary rates have been

important during the evolution of avian SODs, namely in imparting the birds with a higher lifespan when compared to their mammalian counterparts and also in fine-tuning their mode of locomotion enabling their adaptive radiation into varied habitats. We find that the avian SODs have a significantly higher evolutionary rate when compared to the reptilian or mammalian SOD orthologs. Mann-Whitney U test for conservation-acceleration scores of SOD orthologs identified birds as having significantly lower scores (suggesting accelerated evolution) at $p<0.01$ level when compared to their tetrapod orthologs. We also detected several positively selected sites in the avian SOD genes, which could be targets of future biochemical characterization experiments to understand the aging related physiology and the exercise related stress.

4.2. Introduction

During cellular respiration, in the presence of *Cytochrome C*, oxygen gets reduced to form water in the last step of the oxidative phosphorylation. Most of the oxygen molecules are stable, and Cytochrome c can retain all the oxygen until all their electrons are transferred. However, some of the oxygen molecules (2-3%) are converted into reactive oxygen species (ROS) [84] or superoxide anions by other components of the respiratory chain in the mitochondria [170, 171]. Reactive oxygen species are at the same time beneficial to the organisms by being present in macrophages and phagocytes [172, 173, 174, 175], but also detrimental to the cell by inducing aging, apoptosis, alcohol induced cell damage and cancer [176, 177, 171].

Superoxide Dismutase (SOD) [178] an enzyme found in the cell cytoplasm and mitochondria is the first line of defense against ROS [84, 85]. Vertebrates have three SOD gene copies, SOD1, SOD2 and SOD3, which while important in neutralizing the threat of superoxide anions are also implicated in several other processes. Mutations on SOD1 gene are the cause of Lou Gherig's disease [179, 180] and increased amounts of SOD along with other enzymes can delay the onset of aging [176, 181, 182]. The most relevant consequences of ROS include cell damage, lipid peroxidation and DNA modifications, which might be the cause of aging [183], an idea purported by the free radical theory of aging [183].

It has been observed that birds of same weight (or basal metabolic rate) as mammals have a higher maximum lifespan [183, 181]. Birds have evolved a capacity to produce less ROS and minimize free radical leak in the mitochondria [183, 184], thought to be a major factor contributing to their higher life span. Birds use flight as their major mode of locomotion, which is a strenuous exercise and can increase the metabolic rate by hundred times [185, 85]. The increase in metabolic rate causes proliferation of ROS [85]. Thus, SODs in birds should have a crucial role in neutralizing the effects of increased free radical production due to exercise and imparting a higher lifespan in flying birds [181]. The evolution of SODs in birds should have been presented with different kind of challenges (compared to other vertebrates) of low free radical production in normalcy and more free radicals during flight.

We hypothesize that the avian SODs during their evolution have been subjected to some tinkering, providing them an adaptive benefit. Evolution enabled bird SODs to mitigate the ill effects of increased amounts of ROS produced during flight, which was not simply achieved by increasing SODs' expression [181], as birds produce less ROS during normalcy likely responsible for their higher lifespan. Here, we assessed the evolutionary rates and signatures of positive selection in the avian SOD genes, relatively to their mammalian and reptilian orthologs, to test the hypothesis that bird SODs underwent adaptive evolution to accommodate a new mode of lifestyle.

The results from this study, though preliminary, revealed that SODs have an accelerated evolutionary rate in birds when compared to mammals and reptiles. Avian SODs have also been subjected to positive selection during the course of their evolution. These results are insightful to different fields of research and could provide new exploration avenues in aging and physiology.

4.3. Methods

4.3.1. Dataset compilation and gene finding:

Peptide and coding sequences (CDS) annotations were retrieved from 44 avian and two non-avian reptilian genomes provided by BGI (see Appendix 2 Table 12.1). Orthologs of all vertebrate, SOD1, SOD2 and SOD3, genes were downloaded from *Ensembl* [120]. HMMER (hmmer.org), was used to build a HMM profile (from vertebrate SODs) for each SOD protein, which was used to search against the peptide databases of each bird species. The corresponding CDS of the best-hit peptide (annotation or accession) was extracted from the CDS database. Each CDS was used as a query for BLAST [142] against the NCBI Genbank, to confirm its integrity and correctness of annotation.

4.3.2. Sequence Alignment and Phylogeny Construction

The CDSs were translated into proteins and aligned using MUSCLE [40], backtranslated into nucleotides and checked manually in SEAVIEW [121]. Mis-aligned regions were removed from the resulting alignment using GBLOCKS [44] with the “relaxed” parameter [186, 47]. The best-fit nucleotide substitution model for each alignment was found using MrAIC [125]. A maximum likelihood phylogenetic tree was constructed in GARLI v.2.0 [149] using the best-fit substitution model. Analysis was repeated for five times with each analysis replicated four times, and the best scoring tree was selected for downstream analysis. One hundred bootstrap replicates were performed and the bootstrap values were plotted on the nodes of the best scoring likelihood tree using *sumtrees* program [150].

4.3.3. Nucleotide Level—Evolutionary Rate Analysis:

Differences in evolutionary rates in the bird clade were identified using the PHAST (PHylogenetic Analysis with Space/Time models) computer package [163]. Two different phylogenetic trees were analyzed (separately), the first tree comprised birds and mammals, while the second phylogeny consisted of reptiles and birds. First, *phyloFit* was used to fit the tree model to the multiple alignment of DNA sequences by maximum likelihood using the specified tree topology and the REV substitution model [164] with four rate categories (option ``–subst-mod REV –nrates 4").

This model was used to generate “Phylogenetic P values” using *phyloP* [162], which computes lineage specific or global P-values for conservation or acceleration from the alignment, a phylogenetic tree and a model of neutral evolution. *PhyloP* identifies the departures from neutrality in either direction (conservation or acceleration), for each nucleotide column in the alignment using methods of similar statistical power [162]. We used the likelihood ratio test (LRT) analysis, because it is based on the full likelihood function, is expected to make better use of the substitution pattern, is more robust if there were periods of extreme selection and given this is the preferred method for testing each site of the alignment (see [162]).

We ran the program twice for each tree, first for the bird sub-tree compared with the corresponding super-tree (mammalian clade/reptilian clade) and secondly with the whole tree to determine conservation–acceleration (CONACC) scores, where positive values indicate conservation and negative values indicate acceleration of evolution at the given site. To visualize the distribution of scores in the bird lineage and the whole tree, cumulative distribution frequencies (CDF) of the CONACC scores were plotted as relative frequencies of the fractions of scores in R [187].

4.3.4. Codon Level: Compartmentalization Analysis

To assess evolutionary rate variation in birds, we used the dN/dS based compartmentalization analysis in the HYPHY software [54] with the *SelectionLRT.bf* batch file, using the best-fitting GTR model [164, 188] (identified earlier using MrAIC) crossed with the MG94 codon model [51]. We used two different sets of phylogenies, one comprising only birds and mammals, and the other comprising only reptiles and birds, since the program allows only three compartments.

LRTs were used to compare five evolutionary models where dN/dS estimates were either independent or assumed to be equal [189] among the 1) bird clade, 2) reptile/mammalian clade, and 3) ancestral branch leading to the birds and mammals/reptiles (as the separating branch) and among five evolutionary models that used a: 1) global dN/dS estimate; 2) constrained dN/dS estimate for birds and mammals/reptiles with independent estimate for the separating branch; 3) constrained dN/dS estimate for mammals/reptiles and the separating branch with

independent estimate for birds; 4) constrained dN/dS estimate for birds and the separating branch with independent estimate for mammals/reptiles; and 5) independent dN/dS estimates for birds, mammals/reptiles, and the separating branch. *SelectionLRT.bf* employs Akaike Information Criterion (AIC) statistics to determine which model best explains the data.

4.3.5. Codon level: Positive selection analysis

PAML [55] was used to check for positive selection in the avian SOD genes. Models, M0 vs M3 were used to check for the variable ω ratios at each site. Models M1a vs M2a and M7 vs M8, were used to check if the genes were evolving under positive selection. We used two LRT's based on site-specific models comparing the nested models: M1a-M2a and M7-M8. The first LRT was performed comparing M1a (nearly neutral: p0, p1, $\omega_0 < 1$, $\omega_1 = 1$, NS sites = 1) against M2a (positive selection: p0, p1, p2, $\omega_0 < 1$, $\omega_1 = 1$, $\omega_2 < 1$, NSsites = 2); the second LRT was comparing M7 (beta: p, q, NS sites = 7) with M8 (beta and ω: p0, p1, p, q, $\omega_s > 1$, NS sites = 8).

An Empirical Bayes (EB) approach to calculate the posterior probability (PP) that a given site comes from the class with ω > 1 was used to identify the positively selected sites. Sites with a PP above the defined cut-off value (e.g. p > 95%) [126] were considered to be under positive selection. A Bayes Empirical Bayes (BEB) [126] method was used to accommodate the uncertainties in the maximum likelihood estimates of parameters in the ω distribution.

HYPHY [54] was also used to check for positive selection in the SOD proteins. Five models employed in the HYPHY web-server were used for the present study. SLAC, REL, FEL, FUBAR and MEME [151, 54, 153, 152, 154, 155].

SLAC is an improved derivative of the Suzuki-Gojobori counting approach [54, 151] that accounts for changes in the phylogeny to estimate selection on a site-by-site basis. SLAC also calculates the number of non-synonymous and synonymous substitutions that have occurred at each site using ML reconstructions of ancestral sequences [54, 151]. The FEL model estimates the ratio of non-synonymous to synonymous without assuming an *a priori* distribution of rates across sites in a site-by-site analysis. FUBAR is a method to model and detect selection much faster than existing methods and to leverage Bayesian MCMC to robustly account for parameter estimation errors [151]. MEME is capable of identifying instances of both episodic and pervasive positive selection at the level of an individual site [154]. REL is an extension of familiar codon-based selection analyzes in PAML [157, 158], however, REL allows synonymous rate variation [153].

The positive selected sites identified were annotated on the tertiary structure of *Gallus gallus* SOD protein modeled in SWISS-MODEL [165] server. Visualization and annotation of the sites were carried out using the PyMOL software [128].

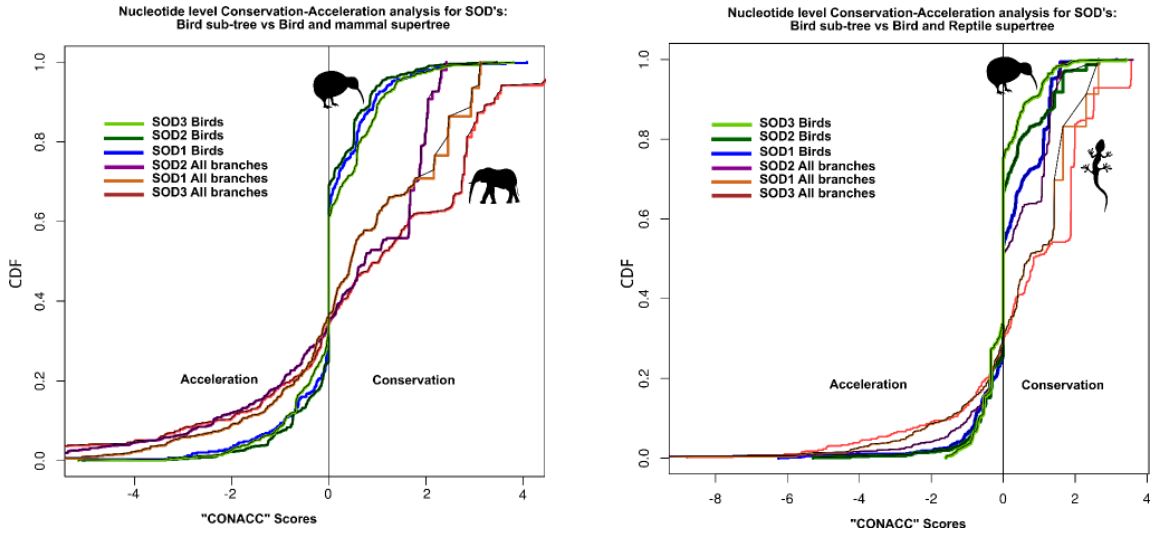


Figure 4.1.: Cumulative distribution frequencies of the "*wig scores*", calculated using *phyloP*.

4.4. Results

4.4.1. Results of Nucleotide Level analysis

Evolutionary Rate Analysis:

The results of conservation acceleration analysis are presented in Figure 4.1. The bird SOD genes (sub-trees) were found to be evolving at an accelerated rate when compared to conservation-acceleration values for the whole tree (comprising mammals and birds). A similar trend was observed in the case of the bird-reptile phylogeny. Two-sided *Mann-Whiteny U* test done to check if the wig-scores of bird clade were similar to the wig-scores for the whole tree rejected the null hypothesis of equality at $p < 0.01$ for all three ortholog sets. In addition, a Mann-Whiteny test to check if the CONACC scores of the bird clade was significantly lesser (alternative hypothesis) than the whole tree scores rejected the null hypothesis that values are equal $p < 0.01$ for all three SOD genes. Thus, the CONACC scores for birds are non-equal (in fact lesser) to the whole tree even when the graphical curves (Figure 4.1) look similar.

4.4.2. Results of Codon Level analyses

Compartmentalization Analysis:

SelectionLRT analysis implemented in HYPHY found that one gene, in both the cases of bird-mammal - SOD1 - and bird-reptile - SOD3, were evolving with a similar ω values throughout the phylogeny. The codons in SOD2 gene had a distinct evolutionary rate in both cases. All the compartments (2 clades and a separating branch) of SOD3 gene in the bird-mammal phylogeny were found to evolve distinctly. Detailed results of the compartmentalization analysis

Table 4.1.: Results of Compartmentalization analysis comparing the codon evolutionary rate of Bird clade with mammals and reptiles. AIC values are presented for each models[c] (phases) Bold values indicate the selected model.

		Bird–Mammal[a]			
Gene	PHASE 1	PHASE 2	PHASE 3	PHASE 4	PHASE 5
SOD1BM	**17816.29**	17817.929	17817.85	17818.03	17819.58
SOD2BM	18953.51	18954.17	**18948.30**	18950.16	18949.67
SOD3BM	18059.91	18056.76	18037.37	18031.13	**18029.26**
		Bird–Reptile[b]			
Gene	PHASE 1	PHASE 2	PHASE 3	PHASE 4	PHASE 5
SOD1RB	11224.31	**11223.27**	11225.02	11223.70	11224.49
SOD2RB	10801.30	10803.14	10801.05	**10801.02**	10802.99
SOD3RB	**12195.32**	12197.31	12196.36	12196.59	12198.36

[a] In the bird-mammal phylogeny mammalia clade was selected as the clade of interest;
[b] In the bird-reptile phylogeny aves clade was selected as the clade of interest; note that this selection is for the purpose of analysis and does not alter the results.
[c] Phase1: $\omega_{Birds} = \omega_{Separatingbranch} = \omega_{Mammals/Reptiles}$
Phase 2: $\omega_{Birds} = \omega_{Mammals/Reptiles} \neq \omega_{Separatingbranch}$
Phase 3: $\omega_{Mammals} = \omega_{Separatingbranch} \neq \omega_{Birds}$
Phase 4: $\omega_{Birds} \neq \omega_{Separatingbranch} = \omega_{Reptiles}$
Phase 5: $\omega_{Birds} \neq \omega_{Separatingbranch} \neq \omega_{Mammals/Reptiles}$

are presented in Table 4.1.

Positive selection analyses:

According to the likelihood ratio tests conducted in PAML [55], SOD2 was not positively selected, nevertheless the (HYPHY and PAML) analysis identified sites under positive selection and the selectionLRT analysis found that SOD2 had a significantly different evolutionary rate compared to the other clades. The M0-M3 comparison points to a site-to-site variable ω (dN/dS) ratio and was highly significant for all genes (see Table 4.2). The M7-M8 comparison was highly significant for SOD1 and SOD3 ($p < 0.01$), and significant at $p<0.5$ level for SOD2. The likelihood parameter estimates are presented in Table 4.2. HYPHY analysis identified 15 sites as positive selected in SOD1, six sites in SOD2, and 22 sites in SOD3 (Table 4.2 and Figure 4.2). Three sites each in SOD1 and SOD3 were found to be positively selected by more than three methods implemented in HYPHY, while SOD2 showed the weakest signal of site-wise positive selection.

4.5. Discussion:

The nucleotide level conservation acceleration and codon level evolutionary rate analyses reveal that birds have an altered evolutionary rate compared to mammals and reptiles. While our *phyloP* analysis specifically shows that birds have a slightly accelerated rate, the dN/dS based compartmentalization analysis shows that they evolve at a different rate. Also according to the PAML codon based positive selection analysis we find that SOD1 and SOD3 are strongly

Table 4.2.: Likelihood parameter estimates of PAML analysis for positive selection[a] and positively selected sites as per PAML and HYPHY analysis.

Gene	M0-M3 (2Δ*lnL*)	M1-M2 (2Δ*lnL*)	M7-M8 (2Δ*lnL*)	Positively Selected sites – P>0.95; In parenthesis HYPHY selected sites
SOD1	756.96	95.892766	93.485774	23Q, 24Q, 105S (25, 31, 35, 37, 38, 40[b], 43, 59, 69, 89, 98[b], 103, 111[b], 112, 128)
SOD2	40.95	**0.000004**	**3.93**	221S (15, 16, 32, 82, 181, 183)
SOD3	408.09	10.60	32.81	28Y, 164S (2, 7, 13, 28[b], 31, 37, 39[b], 41, 46, 47, 91, 100, 103, 105, 121, 124, 155, 157[b], 164, 165, 166, 168)

[a] Site numbers and amino acids according to the *Gallus gallus* protein; non significant LRT values (2Δ*lnl*) are in bold. In parenthesis are the sites found to be positively selected by HYPHY methods,
[b] More than 3 methods in HYPHY identified the site.

positively selected. While the SOD2 LRTs did not provide evidence of positive selection, we recovered one site (212) with a p>0.95, and two sites with p>0.60, thought to be positively selected.

While there are reports that birds produce less free radicals compared to other vertebrates, thus having a longer lifespan, it should also be acknowledged that birds resort to flight, which is a strenuous exercise and produce comparably higher amount of free radicals (although lesser in ratio to the inhaled oxygen - compared to other vertebrates due to lesser ROS leak in avian mitochondria) [181] or reactive oxygen anions. While, it is known that birds have an efficient respiratory system adapted to flight and the ratio of free radical production per oxygen molecule inhaled is lesser compared to non-flying mammals, the role of antioxidants in their longevity has been demonstrated [182].

Thus, the physiological pressure for the efficiency of SODs, catalyses, peroxidases and other detoxifying enzymes in birds is different from that faced by other vertebrates. During the evolution of the ancestral birds and the fine-tuning of the flight adaptation, evolutionary modifications of the detoxifying enzymes must have occurred. We provided evidence that there have been evolutionary rate acceleration and positive selection in SOD during the evolution of birds, which might have helped them to fine tune their enzymatic activity while adapting to a new mode of locomotion and subsequent adaptive radiation.

Our results should be considered as preliminary and should be followed by more detailed evolutionary analyzes. As an extension of this study, it will be interesting to compare the detoxifying enzymes of flightless birds and flying birds (migratory and non-migratory flying birds) against each other in an evolutionary framework. It will also be important to assess if flight adaptation have led to the evolutionary tinkering of other detoxifying enzymes. Interestingly, positive selection in the genes involved in DNA damage checkpoint-DNA repair pathway has been proposed as a significant innovation during the bat flight adaptation [190]. Thus, it would be intriguing to

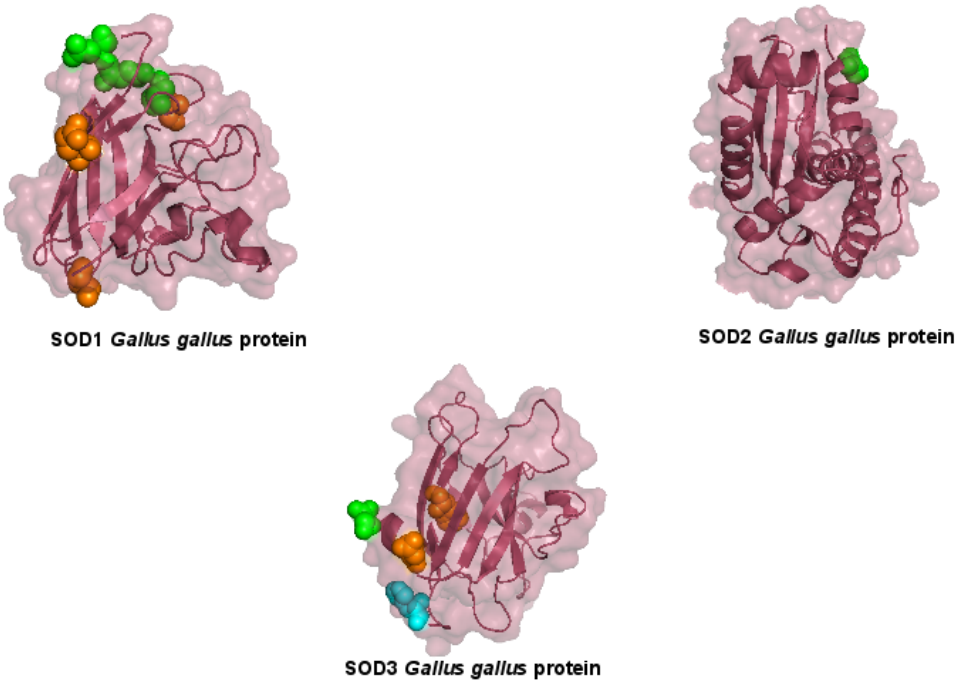


Figure 4.2.: Positively selected sites on the avian SOD1, SOD2 and SOD3 are plotted on the homology models of the *Gallus gallus* SOD protein tertiary structures. Green sites represent those selected by PAML, orange sites the ones selected by HYPHY (>3 methods) and cyan sites those selected by both PAML and HYPHY.

check whether the DNA damage-repair pathway related genes are positively selected in birds to find out instances of parallel adaptation in birds and bats during the flight adaptation.

II.

Adaptive Radiation in Vertebrates: insights from the Teleosts

5

Fish lateral line innovation: insights into the evolutionary genomic dynamics of a unique mechanosensory organ

Papers arising from this chapter

Philip et al. · doi:10.1093/molbev/mss194 MBE

Fish Lateral Line Innovation: Insights into the Evolutionary Genomic Dynamics of a Unique Mechanosensory Organ

Siby Philip,[1,2] João Paulo Machado,[1,3] Emanuel Maldonado,[1] Vítor Vasconcelos,[1,2] Stephen J. O'Brien,[4,5] Warren E. Johnson,[4] and Agostinho Antunes*[,1,2,4]

[1]CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal
[2]Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal
[3]Instituto de Ciências Biomédicas Abel Salazar (ICBAS), Universidade do Porto, Porto, Portugal
[4]Laboratory of Genomic Diversity, National Cancer Institute, Frederick, Maryland
[5]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia
*Corresponding author: E-mail: aantunes@ciimar.up.pt.
Associate editor: John H. McDonald

Figure 5.1.: Papers arising from this chapter: ***Molecular Biology and Evolution***,10.1093/molbev/mss194.

5.1. Abstract

The mechanosensory lateral line, found only in fishes and amphibians, is an important sense organ associated with aquatic life. Lateral line patterns differ among teleost, the most diverse vertebrate taxa, hypothetically in response to selective pressures from different aquatic habitats. In this paper we conduct evolutionary genomic analyses of 34 genes associated with lateral line system development in teleosts to elucidate the significance of contrasting evolutionary rates and changes in the protein coding sequences. We find that duplicated copies of these genes are preferentially retained in the teleost genomes, and that episodic events of positive selection have occurred in 22 of the 30 post-duplication branches. In general, teleost genes evolved at a faster rate relative to their tetrapod counterparts and the mutation rates of 26 of the 34 genes differed among teleosts and tetrapods. We conclude that following whole genome duplication, evolutionary rates and episodic events of positive selection on the lateral line system development genes might have been one of the factors favouring the subsequent adaptive radiation of teleosts into diverse habitats. These results provide the foundation for further detailed explorations into lateral line system genes and the evolution of diverse phenotypes and adaptations.

5.2. Introduction

The mechanosensory lateral line sense organ is unique to aquatic vertebrates (fishes and the amphibians) [191]. The lateral line consists of surface neuromasts located on the skin's surface that detect slow moving water and canal neuromasts that are embedded in the lateral line canals and sense rapidly moving water [192]. The lateral line system provides a sense of "touch at a distance" for the fishes [191, 193] and is important in processes such as rheotaxis [194], schooling [195], courtship and sexual behaviour [196, 197], feeding and prey detection [198, 199, 200] and navigation [201], even in the absence of vision as exemplified in blind cave fishes [202].

In fishes (comprising Chondrichthyes, Actinopterygii and Sarcopterygii) the lateral line varies in configuration [191] (Figure 5.2), number [203, 193], and size [204] among [205, 206] and within species [193, 207] and changes depending upon habitat. Because the lateral line is ubiquitous in fishes, even in deep-sea fishes and cave dwelling fishes without eyes, and because the lateral line system is absent in the terrestrial vertebrates, it is probably a primary sense organ that is closely linked with adaptation to aquatic life.

The teleost lateral line consists of an Anterior Lateral Line (ALL) and a Posterior Lateral Line (PLL) (Figure 5.2). The ALL is located in the head region and is innervated by sensory neurons clustered with pre-otic ganglion, while the PLL is innervated by sensory neurons clustered with post-otic ganglion [208]. During the embryonic PLL development of primitive Actinopterygians

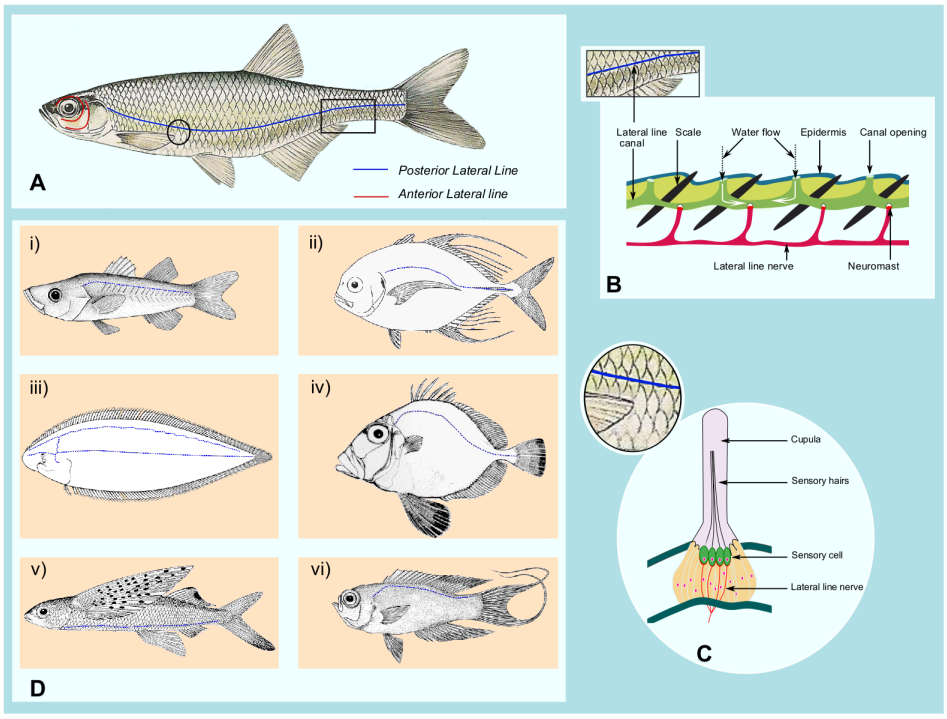


Figure 5.2.: Schematic representation of the anterior and posterior lateral line: A) Anterior (red) and posterior (blue) lateral lines are highlighted on *Alburnus alburnus*, B) schematic representation of the lateral line canal system; C) structure of a single neuromast. D) representative teleosts with different lateral line patterns are shown i) *Microinchthys sanzoni*, ii) *Carangoides hedlandensis*, iii) *Cynoglossus cynoglossus*, iv) *Cyttopsis rosea*, v) *Cheilopogon spilopterus* and vi) *Grammatonotus crosnieri*; (note that there are many other patterns of lateral lines and there are some species with lateral line completely absent or truncated).

(fish taxa existing prior to the emergence of teleosts) and the Sarcopterygians (lobe finned fishes and amphibians), the neuromast primordium (which arises from the placode in the post-otic region) deposits a continuous stream of neuromasts along the lateral line while migrating towards the tail [209]. In contrast, during PLL development in teleosts (advanced Actinopterigyians), the neuromast primordium from the placode deposit only a few proto-neuromasts/primary neuromasts on the borders of somites at specific intervals as it moves towards the caudal region [209, 204]. Once the primary neuromasts are deposited during the post-embryonic development stage, secondary neuromasts appear between the primary neuromasts and form stitches [210]. It is striking that these two very different patterns of PLL embryonic development evolved after the emergence of the teleosts and that they are still highly conserved among teleosts, even when adult PLL patterns differ (see [209] for details about phylogenetic separation of the PLL development pattern).

The path of deposition of the neuromast by the primordium is guided by the expression of the chemokine *cxcr4b* and its ligand *cxcl12a* and not by external cues [211]. Although numerous studies have described the genes expressed during lateral line development (eg: [212, 213,

214]) (see also the Gene Ontology database [215, 216] process GO:0048925 for a list of genes involved), the evolutionary dynamics of these genes have not been extensively analysed.

Understanding patterns of protein evolution is important for determining the molecular basis of adaptation [217] as changes at the protein level have a strong influence on phenotypic divergence [218]. Yet, there have been relatively few evolutionary studies of genes involved in teleost phenotypes [137, 219, 131], or that compare the phenotypic [220, 221] or genomic evolution [222, 223] between teleosts and tetrapods. Molecular evolutionary analyses of tetrapod and teleost gene evolution would provide insights on the remarkable adaptive radiation and species diversity of teleost fishes.

We speculate that the wide range of lateral line systems observed among and within teleosts was an important adaptive novelty and that it is one of the main reasons these taxa now comprise one of the most speciose and diversified group of vertebrates. Here we describe the evolutionary changes in protein coding sequences of the genes involved in the teleost lateral line system. Our results indicate that these genes were generally retained as paralogs following the fish-specific whole genome duplication and that subsequently episodic periods of positive selection led to differences in lateral line development patterns among primitive actinopterygians and teleosts. Generally teleost genes have evolved at faster rates relative to the tetrapods. Our study provides the basis for future evolutionary explorations of the patterns of diversification of lateral line system diversification in fishes.

5.3. Materials and Methods

5.3.1. Dataset compilation and preparation

The genes involved in lateral line system development were compiled from the Gene ontology (GO) database [215, 216] (Appendix 3 Table S1) using *"lateral line system development"* as the biological process search term and filtered for zebrafish (GO:0048925). The corresponding human ortholog for each zebrafish gene (Appendix 3 Table S2) was used to download coding sequences of the one-to-one orthologs (one-to-many or possible-orthologs in the case of duplicate genes in teleosts) from Ensembl version 62 [120] using the PyCOGENT [224] Ensembl database query interface. Gene orthology and instances of genome/gene duplication were examined using micro-synteny analysis in the Genomicus server [225], and duplicated genes were classified as either strict fish specific genome duplication (FSGD) products or lineage specific duplication products. A rose-window plot of the strict FSGD duplicates (all the teleosts possessed two paralogs of the gene) in zebrafish was prepared using circos version 0.55 [226]. For duplicated genes here we assume that prior to FSGD the ancestral gene had a function in lateral line development because of the absence of a duplicated copy in *Xenopus tropicalis*. Thirty-nine genes were identified that contributed to the development of the lateral line system, of which 34 had sufficient taxonomic representation for further analyzes. Sequences with

substantial gaps were discarded at this stage.

5.3.2. Sequence alignment and phylogeny construction

The coding sequences for each gene were translated into proteins and aligned using MUSCLE [40], backtranslated into nucleotides and checked manually in SEAVIEW [121]. Misaligned regions were removed from the resulting alignment using GBLOCKS [44] with the 'relaxed' parameter [186, 47]. The best-fit nucleotide substitution model for each alignment was found using MrAIC [125]. A maximum likelihood phylogenetic tree was constructed in PHYML v.3.0 [227, 124] using the NNI tree searches. Clade support values (SH-aLRT branch support [19, 228]) were produced using the non-parametric version of aLRT [229]. This nucleotide substitution model and phylogeny were used for all subsequent analyzes. All alignments and phylogenetic trees used for this study are available for download from dx.doi.org/10.6084/m9.figshare.92411 (figshare).

5.3.3. Nucleotide level – evolutionary rate analysis

Differences in evolutionary rates in the teleosts were identified using the PHAST (PHylogenetic Analysis with Space/Time models) computer package [163]. First, phyloFit was used to fit the tree model to the multiple alignment of DNA sequences by maximum likelihood using the specified tree topology and the HKY substitution model with four rate categories (option "--subst-mod HKY –nrates 4") [230, 231]. This model was used to generate "Phylogenetic p-values" using *phyloP* [162], which computes lineage specific or global p-values for conservation or acceleration from the alignment, a phylogenetic tree and a model of neutral evolution. *PhyloP* identifies the departures from neutrality in either direction (conservation or acceleration), for each nucleotide column in the alignment using methods of similar statistical power [162]. We used the likelihood ratio test analysis (LRT), since it is based on the full likelihood function and is expected to make better use of the substitution pattern, is more robust if there were periods of extreme selection and since this is the preferred method for testing each site of the alignment (see [162]). We ran the program twice, first for the teleost subtree compared with the corresponding super-tree (tetrapod clade) and then with the whole tree to determine conservation-acceleration ("CONACC") scores, where positive values indicate conservation and negative values indicate acceleration of evolution at the given site. To visualize the distribution of scores in the teleost lineage and the whole tree, cumulative distribution frequencies (CDF) of the CONACC scores were plotted as relative frequencies of the fractions of scores.

5.3.4. Codon level - compartmentalization analysis

To assess evolutionary rate variation in the teleost lineage we used the dN/dS based compartmentalization analysis in the HYPHY software [54] with the SelectionLRT.bf batch file using

the best-fitting GTR model [164, 188] (identified earlier using MrAIC) crossed with the MG94 codon model [51]. Likelihood ratio tests (LRTs) were used to compare five evolutionary models where dN/dS estimates were either independent or assumed to be equal [189] among the 1) teleost clade, 2) tetrapod clade and 3) ancestral branch leading to the teleosts (as the separating branch) and among five evolutionary models that used a: 1) global dN/dS estimate; 2) constrained dN/dS estimate for teleosts and tetrapods with independent estimate for the separating branch; 3) constrained dN/dS estimate for tetrapods and the separating branch with independent estimate for teleosts; 4) constrained dN/dS estimate for teleosts and the separating branch with independent estimate for tetrapods; and 5) independent dN/dS estimates for teleosts, tetrapods, and the separating branch. SelectionLRT.bf employs Akaike Information Criterion (AIC) statistics to determine which model best explains the data.

5.3.5. Codon level and amino acid level analysis - post-duplication branches

To identify the post duplication evolutionary dynamics of the strict FSGD duplicated genes, we used the Branch-Site models [158, 232, 57] as implemented in PAML 4 [56]. Using the codon alignment and the corresponding maximum likelihood tree, constructed in PHYML v.3.0 [19] after determining the best-fit nucleotide substitution model in MrAIC [125], the two branches immediately after the duplication event (see Figure 5.3a and 5.3b) were labelled as the foreground branches independently (one at a time) in the branch-site test 2, the recommended test for identifying positive selection. Here two ω values ($0 < \omega_0 > 1$ and $\omega_1 = 1$) are assigned for the background branches and the foreground branches are assigned a ω value ($\omega_2 > 1$). This model is compared with the null model where the ω value in the foreground branch is constrained to 1, and a LRT is used to check if the foreground branches are evolving under the influence of positive selection. The branch-site test segregates the amino acid positions/codons into four different categories. Two categories describe sites for which selective pressure does not change over time, either under purifying selection (site class 0, $\omega_0 < 1$) or under neutral evolution (site class 1, $\omega_1 = 1$). The two other categories (site classes 2a and 2b) are sites potentially evolving under positive selection only in the foreground branches ($\omega_2 > 1$), and evolving in the background branches (all the remaining branches) under purifying selection (site class 2a, background branches $\omega_0 < 1$) or neutral evolution (site class 2b, background branches $\omega_1 = 1$). The sites under the influence of positive selection were identified by a Bayes Empirical Bayes (BEB) test [126] applying a threshold of >0.95 posterior probability.

Codon models can be biased by saturation of synonymous substitutions. To eliminate such problems and to provide confidence in our codon-based analysis, we employed amino acid based rate shift analysis, since this approach is recommended for divergent sequences [105, 233]. The same codon alignments of the strict FSGD duplicated genes used in the codon-based analysis were translated into amino acids. Rate shifts of evolution in branches immediately after genome duplication were checked using the RASER2 (Rate Shift EstimatoR version 2)

[66]. RASER2 uses the stochastic mapping of mutations [127] to calculate the probability that a rate-shift occurred at a specific branch, and implements the empirical bayes test to identify sites evolving under a covarion-like model or heterotachy, which is similar to the site class 2a in the branch-site test of PAML. We used RASER2 to compare the alternate lineage-specific rate-shift model with the null model, which does not enable rate shifts. The program was run separately for each of the post-duplication branches. Likelihood-ratio tests (LRT) were performed to determine whether the lineage-specific model fits the data significantly better than the null model. The program was run twice for each branch of interest to check for known convergence problems [233]. Tertiary structures of CXCR4 (human), *cxcr4a* and *cxcr4b* (zebrafish) were built using I-TASSER server [234, 235], and positive selected sites on cxcr4b were visualized and annotated using PYMOL [128].

5.4. Results

5.4.1. Synteny analysis – duplication of genes involved in lateral line system development

Of the 34 genes involved in lateral line system development (Appendix 3 Table S1), 12 were strict FSGD duplicates where all the teleosts possessed two paralogs each (Figure 5.3c). Micro-synteny analysis identified seven genes (*cldnb, kal1, atoh1, cxcr7, cdh4, dkk1* and *pcsk5*) involved in the lateral line system development that had an asymmetric distribution of paralogs among the teleosts (Appendix 3 Figures 13.1 & 13.2). Cases of both lineage-specific gene retention and lineage-specific gene loss were identified in the zebrafish (Appendix 3 Figure 13.2). The *cxcr7a* and *cxcr7b* were found only in the zebrafish, and in close proximity to *cxcr4a* and *cxcr4b* on chromosome 9 and chromosome 6, respectively, and CXCR4 is duplicated in all the teleosts analyzed for this study. Based on synteny analysis there is lineage-specific gene retention of *cxcr7* in zebrafish after the FSGD event. CDH4 has two paralogs in all teleosts except zebrafish, which suggests a lineage specific gene loss in zebrafish post FSGD. In addition, *dkk1* was only duplicated in zebrafish.

5.4.2. Lineage-specific acceleration of evolution in teleosts

To identify cases of accelerated evolution in the teleosts at the nucleotide level we used *phyloP* CONACC scores. P-values for each base position were obtained using the LRT method for the whole vertebrate phylogeny and the teleost subtree (in comparison with the corresponding super-tree) for each gene. Cumulative distribution frequency of the scores clearly show acceleration (negative scores indicate acceleration) of evolution in the teleost sub-tree (Figure 5.4a), which had lower scores relative to those of the whole tree (Figure 5.4b).

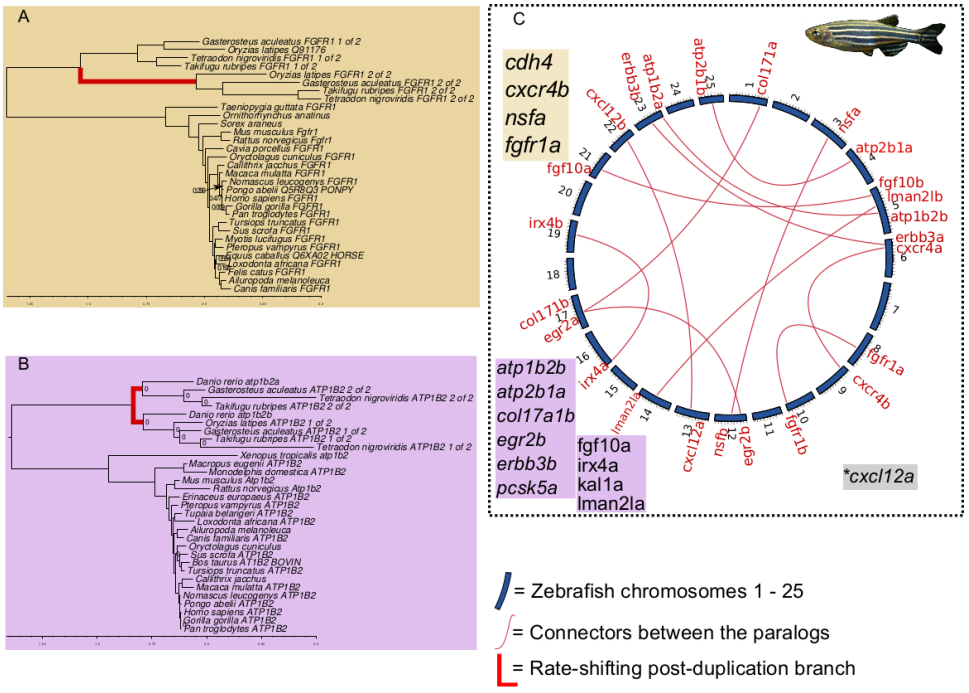


Figure 5.3.: The results of rate-shift analysis on the post duplication branches. a) Schematic representation of a phylogenetic tree showing rate shift in only one post-duplication branch, b) Schematic representation of the phylogenetic tree where both the post-duplication branches faced rated-shift, c) rose window representation of the strict FSGD duplicates in zebrafish plotted, sets of genes in each models of rate-shift are shaded with the corresponding color.

After identifying that teleosts genes involved in lateral line system development were generally evolving at a faster rate, we analyzed the data using dN/dS methods. The compartmentalization analysis implemented in the selectionLRT.bf batch file of the HYPHY program allowed us to separate the tree into two sub-trees (clades/groups). For each we analyzed rate shifts using five models. Models 2-5 are the alternative models and they are compared to a globally homogenous dN/dS in the model 1, which is the null hypothesis. Significantly different dN/dS ratios for three groups (model 5) were found in 29 of 34 genes (Figure 5.5 and Appendix 3 Table S2). However, only eight genes had significant AIC scores for model 5 (or phase-5). The AIC scores favored either a phase-3, phase-4 or phase-5 alternate model in 26 of 34 genes, confirming that in most cases teleosts have an altered evolutionary rate. The alternate hypothesis of different evolutionary rates in the compartments was rejected for *pcsk5a, hcn1, fgf3, cxcr7b* and *tmie*, of which the last three are the genes involved in the PLL neuromast primordium migration (Biological Process GO: 0048920) in zebrafish. The results from the compartmentalization and the *phyloP* analyses suggest that there is an accelerated evolutionary rate in the teleost clade compared with tetrapods in this set of genes involved in lateral line system development.

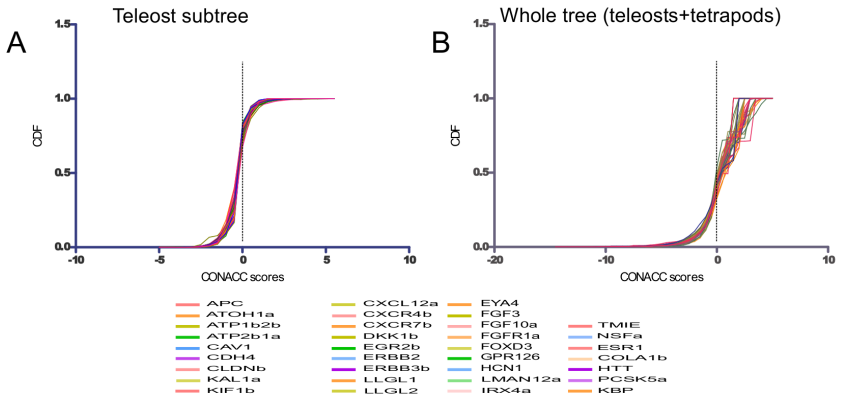


Figure 5.4.: Cumulative Distribution Frequency of the CONACC scores for the a) teleost subtree and b) whole tree; the plots shows a significant skew of the CONACC scores for teleosts towards less than zero (< 0) which indicates acceleration of evolution.

5.4.3. Rate-shift among paralogs

The rate-shift analysis using the amino acid alignment yielded similar results as the codon-based positive selection analysis. Of the 30 branches tested for rate shifts (2 post-duplication branches in each of the 12 strict FSGD duplicates and *kal1, cdh4* and *pcsk5* genes), only six branches were not significant at the P < 0.05 level (Table 5.1, Figure 5.3a and 5.3b). Three genes, *cdh4, fgfr1a*, and *nsfa* (all strict FSGD duplicates), showed no rate-shift on the post-duplication branch leading to the sister paralog of the gene. One gene, *cxcl12* had no rate shifts in either of the post-duplication branches, and *cxcr4* had no rate shift in the post-duplication branch leading to the paralog (*cxcr4b*) implicated in lateral line system development.

5.4.4. Positive selection in the post-duplication branches

To check if the altered rates of evolution on the post-duplication branches were due to positive selection, we employed the branch-site models of the *codeml* program from the PAML package. Fifteen genes were analyzed including *kal1, cdh4* and *pcsk5*, as well as the 12 strict FSGD duplicates and 22 of 30 branches showed evidence of positive selection (Table 5.2 and 5.3). Two genes, *cxcl12a* and *lman2la*, showed no positive selection in both post-duplication branches. The *cxcl12a* gene (also known as *sdf1a*/stromal derived factor 1a) codes for the chemokine responsible for primordium migration along the “lateral line” on the horizontal myoseptum of the trunk of fishes. However, the ancestral branch to the receptor for this ligand, *cxcr4b* shows 13 residues under positive selection (Figure 5.6). Four genes showed positive selection in only one post-duplication branch (Table 5.2 and 5.3), and all of them (*atp1b2a, egr2b, cxcr4b* and *irx4a*) had evidence of positive selection in the post-duplication branch leading to the lateral-line related paralog.

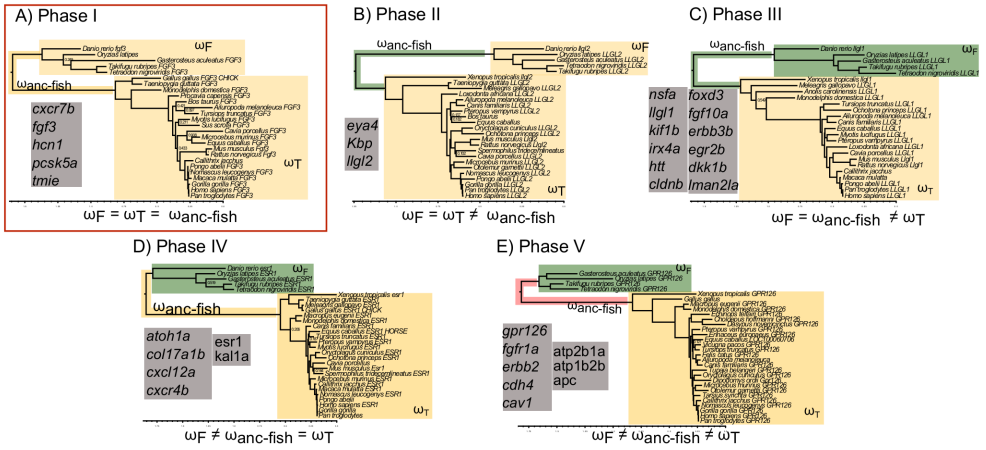


Figure 5.5.: The results of the compartmentalization analysis. All five phases (A - E) are shown and the genes selected in each phase are listed beside the tree, the three compartments used for the analysis are labeled on the tree and differences in evolutionary rates are shaded with different colors. Note that selectionLRT.bf considers the separating branch as a single compartment, so the two branches emanating from the root form a single branch in this analysis.

5.5. Discussion

Our results highlight three major evolutionary patterns observed in genes involved in the lateral line system development, including higher duplicate retention, accelerated rates of evolution of the teleost genes, rate-shifts and positive selection in the post-duplication paralogs.

5.5.1. Duplicate retention of genes involved in lateral line system development

We analyzed the 34 genes involved in lateral line system development. Approximately 35% (12/34) were strict FSGD duplicates, for which all the teleosts had two paralogs that were retained after the whole genome duplication event. Two genes (5.8%), *cdh4* and *cxcr7*, were also duplicated as part of the FSGD, but they had an asymmetric distribution of paralogs in fishes (some species did not retain the duplicated paralog). Four genes (11.7%) that were duplicated during FSGD also underwent a lineage-specific duplication event in the individual species. Thus, for this biological process/development process the duplicate retention is >50%, while the normal retention rate of duplicate copies post whole genome duplication in teleost genomes is less than 24% [236, 237, 238].

Fishes [239, 240, 10, 9] experienced a whole genome duplication (FSGD) around 350 million years ago, just before the emergence of the teleosts [241] and very likely related with their subsequent radiation into >27,000 very diverse and disparate species [242, 86, 243, 244, 238]. Many of the resulting duplicate developmental genes were retained in the newly evolved species [245]. Our results confirm that genome duplication played a major role in shaping the current repertoire of the genes involved in the lateral-line system development, which in turn

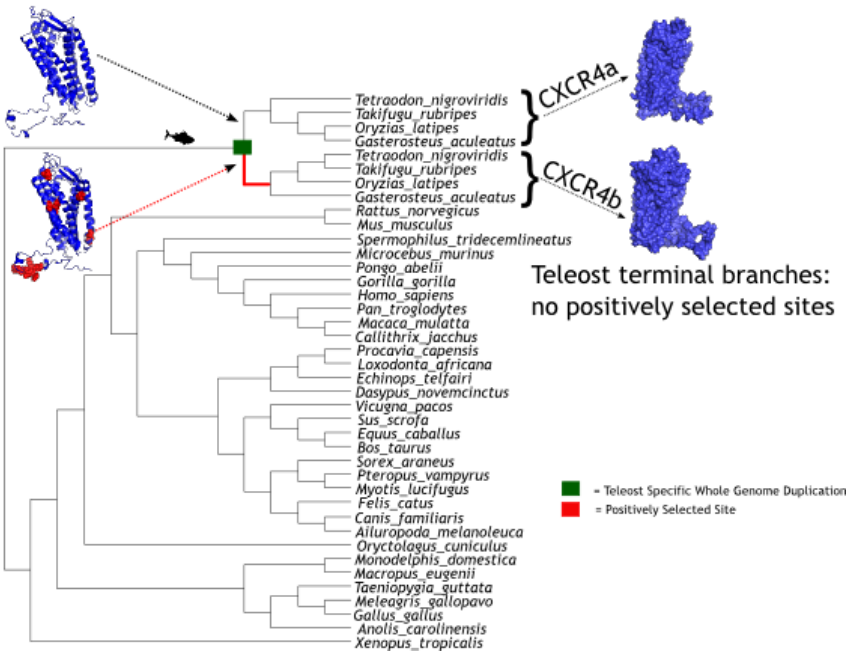


Figure 5.6.: Schematic representation of episodic positive selection in CXCR4 which was duplicated in teleosts during FSGD. Only the branch leading to *cxcr4b* in teleosts had signal of positive selection based on the branch-site model in PAML and teleost terminal branches showed no positive selection in the branch-site model comparison (positively selected sites on the ancestral branch leading to the teleost *cxcr4b* paralog are 30V, 44G, 109T, 145L, 151G, 295Y, 321S, 325R, 326S, 327S, 328H, 329K and 332T).

has contributed to the diversification of the lateral line morphology/phenotype in teleosts and differentiation of PLL development compared with the sarcopterygians [209]. Since the FSGD occurred just before the origin of the teleosts [10, 86, 241], and given our finding of higher duplicate retention in the genes involved in lateral line system development, we propose that genome duplication facilitated the evolution and diversification of the lateral line, and helped promote the radiation of teleosts into diversified environments.

5.5.2. Accelerated rate of evolution in the teleost genes

The genes implicated in the lateral line system development in teleosts evolved at a faster rate relative to their tetrapod orthologs (26 out of the 34 teleost genes using a dN/dS approach), either if having a single or a duplicate copy (Figure 5.4, 5.5 and Appendix 3 Table S2). This is consistent with other studies that have shown that teleost paralogs of duplicated genes evolve asymmetrically [222, 223] and that most of the teleost genes (irrespective of whether they are singletons or duplicated) evolve faster than their tetrapod counterparts [222, 10, 245, 223]. Thus, the altered rate of evolution among the lateral line system development genes is in concordance with earlier studies. Most of the genes have extra functions in addition to lateral line system development, since none of the genes are teleost innovations and half did not retain a

duplicate (paralog) post FSGD. The genes with constant rates in both teleosts and tetrapods, apart from their role in the lateral-line system development, were those involved in cellular activities such as cytokinesis and ion-transport (*hcn1* and *kbp*), DNA repair (*eya4*), cell-fate specification and development (*fgf3*), cytoskeletal organization and exocytosis (*llgl2*), signaling (*cxcr7b*) and proteolysis (*pcsk5a*). This pattern suggests that these genes are under higher constraints than the other genes involved in the developmental process.

Rate-shift and positive selection in post-duplication branches

At least one of the post-duplication branches in the strict FSGD duplicates had a significant rate-shift due to positive selection in all genes except *cxcl12* (Table 5.1, 5.2 and 5.3, Figure 5.3). This episodic event of positive selection is revealing since the teleost clade as a whole did not show evidence of positive selection with the branch site models for all genes (Appendix 3 Table S3). Only three genes (*htt, apc* and *gpr126*) showed positive selection when all the extant branches were tested as foreground branches in branch-site test of positive selection (Appendix 3 Table S3).

Some major caveats for interpreting the results of codon models concern the uncertainty implicit in phylogeny and parameter estimates [246]. Branch-site models could also be sensitive to the absence or low number of synonymous substitutions on foreground branches, which can reduce the accurate estimation of dN/dS values and inflate omega values (see Table 5.2 and 5.3). As this can be a problem when using deep internal branches as foreground branches, we used complementary analyzes at nucleotide, codon and amino-acid levels to obtain general concordance among the results, further validating the evolutionary trends observed in this study.

Three category 1 models have been postulated describing the evolution of duplicated genes [30], including the neofunctionalization model [27], the duplication degeneration complementation (DDC) hypothesis [29] and the specialization model or escape from adaptive conflict (EAC) model [28]. The general feature of these models is that a fate-determination phase occurs rapidly after duplication [30], followed by a preservation phase that precludes the pseudogenisation of one of the copies [247, 30].

Ohno's model [27] suggests neofunctionalization of one of the copies, and that the molecular evolution in the duplicated copy is accelerated, and the other two models [28, 29] propose subfunctionalization in both copies. The DDC model [29] assumes that the ancestral function of the gene will be shared between the two post-duplication daughter genes, and that degenerating neutral mutations that accumulate in the paralogs result in subfunctionalization with neither copy being able to carry out the original functions and thus promoting preservation of both copies.

The EAC model [28] predicts that if the parent gene were performing two functions that could not be independently improved, then after duplication each gene copy could be driven by pos-

itive selection to become more specialized. In fact, none of the three models mentioned here is a perfect fit for the results of this study. However, our results point to a most likely scenario where the paralog is released from the functional constraints of the ancestral gene (due to accelerated evolutionary rate and rate shift) and can become more specialized in lateral line development thereby facilitating evolution of the trait.

From our results we conclude that higher duplicate retention, followed by rate shifts and positive selection, may have contributed to the evolution and remarkable diversity of the teleost lateral line system. We also confirm that the teleost lateral line system development genes generally evolve at a faster rate compared to their tetrapod counterparts in concordance with previous studies. Our results based on comparative genomic analyzes provide, the basis for future confirmation studies on the evolution, development and function of lateral-line system development genes.

5.6. Tables

Table 5.1.: The Likelihood ratio statistics of the amino-acid based rate-shift analysis on the post-duplication branches of the strict FSGD duplicated genes using Rate Sift Estimation in RaSER [66].

Genes[a]	*lnL* null model	lnL alternate model	LRT (2Δ*lnL*)
atp1b2b	-4807.7	-4784.75	45.9
atp1b2a		-4786.44	42.52
atp2b1a	-11947.1	-11949.9	5.6
atp2b1b		-11934.4	25.4
cdh4	-12902.6	-12910	14.8
cdh4 (1 of 2)		-12902.6	0 [b]
col17a1b	-36134	-35992.9	282.2
col17a1a		-36556.5	845
cxcl12a	-1776.82	-1778.51	3.38 [b]
cxcl12b		-1777.37	1.1 [b]
cxcr4b	-6055.45	-6055.45	0 [b]
cxcr4a		-6040.66	29.58
egr2b	-7733.71	-7722.94	21.54
egr2a		-7719.87	27.68
erbb3b	-24022.3	-23941	162.6
erbb3a		-23891.2	262.2
fgf10a	-3234.41	-3232.47	3.88
fgf10b		-3225.88	17.06
fgfr1a	-8963	-8960.81	4.38
fgfr1b		-8963	0 [b]
irx4a	-9353.42	-9374.39	41.94
irx4b		-9303.89	99.06
kal1a	-12649.9	-12609.9	80
kal1b		-12646.7	6.4
lman2la	-6662	-6639.21	45.58
lman2lb		-6647.82	28.36
nsfa	-7955.48	-7951.86	7.24
nsfb		-7955.8	0.64 [b]
pcsk5a	-34988.5	-34898	181
pcsk5b		-34898.9	179.2

[a] paralog listed first is the one involved in lateral line system development.
[b] Not significant.

Table 5.2.: likelihood ratio statistics for codon-models testing for positive selection on the post-duplication branches of the strict FSGD duplicated genes (part-1).

Genes[c]	lnL alternate model	lnL null model	LRT	sig[a]	p|ω	Site class 0	Site class 1	Site class 2a	Site class 2b
atp1b2b	-10314.4	-10315.8	-2.7	NS	p	0.9	0.0	0.0	0.0
					ω			13.0	13.0
atp1b2a	-10311.2	-10313.4	-4.4	S	p	0.9	0.0	0.1	0.0
					ω			999.0	999.0
atp2b1a	-34309.8	-34322.7	-25.7	S	p	0.9	0.1	0.0	0.0
					ω			171.3	171.3
atp2b1a	-34260.3	-34274.2	-27.9	S	p	0.9	0.0	0.1	0.0
					ω			6.2	6.2
cdh4	-31269.3	-31280.4	-22.3	S	p	0.9	0.0	0.0	0.0
					ω			999.0	999.0
cdh4(1 of 2)	-31267.0	-31276.4	-18.8	S	p	0.9	0.0	0.0	0.0
					ω			23.8	23.8
col17a1b	-69370.1	-69408.7	-77.2	S	p	0.6	0.3	0.1	0.0
					ω			888.8	888.8
col17a1a	-69380.7	-69416.4	-71.4	S	p	0.6	0.3	0.1	0.0
					ω			999.0	999.0
cxcl12a	-3790.2	-3790.2	0.0	NS	p	0.8	0.2	0.0	0.0
					ω			1.0	1.0
cxcl12b	-3790.2	-3790.2	0.0	NS	p	0.8	0.2	0.0	0.0
					ω			1.0	1.0
cxcr4b	-14707.0	-14722.2	-30.3	S	p	0.8	0.1	0.1	0.0
					ω			999.0	999.0
cxcr4a	-14729.5	-14729.5	0.0	NS	p	0.6	0.1	0.3	0.0
					ω			1.0	1.0
egr2b	-17304.5	-17305.5	-1.9	NS	p	0.8	0.1	0.1	0.0
					ω			999.0	999.0
egr2a	-17299.1	-17301.3	-4.3	S	p	0.8	0.1	0.1	0.0
					ω			43.9	43.9
erbb3b	-47764.3	-47798.3	-68.0	S	p	0.7	0.2	0.1	0.0
					ω			999.0	999.0
erbb3a	-47779.5	-47793.3	-27.6	S	p	0.8	0.2	0.0	0.0
					ω			198.8	198.8

[a] Significance; S= significant; NS = Not Significant, Values of 999 for dN/dS indicate dS = 0, so dN/dS is undefined.
[b] paralog listed first is the one involved in lateral line system development

Table 5.3.: likelihood ratio statistics for codon-models testing for positive selection on the post-duplication branches of the strict FSGD duplicated genes (part-2).

Genes[c]	lnL alternate model	lnL null model	LRT	sig[a]	p|ω	Site class 0	Site class 1	Site class 2a	Site class 2b
fgf10a	-6782.4	-6785.9	-7.1	S	p	0.8	0.1	0.1	0.0
					ω			132.4	132.4
fgf10b	-6786.1	-6788.6	-5.0	S	p	0.9	0.1	0.0	0.0
					ω			999.0	999.0
fgfr1a	-22543.4	-22556.7	-26.5	S	p	0.8	0.0	0.1	0.0
					ω			999.0	999.0
fgfr1b	-22545.7	-22550.6	-9.8	S	p	0.9	0.0	0.1	0.0
					ω			17.5	17.5
irx4a	-18250.5	-18251.6	-2.2	NS	p	0.8	0.2	0.0	0.0
					ω			69.0	69.0
irx4b	-18244.3	-18247.6	-6.7	S	p	0.7	0.1	0.1	0.0
					ω			8.6	8.6
kal1a	-25846.2	-25852.3	-12.0	S	p	0.9	0.1	0.0	0.0
					ω			17.8	17.8
kal1b	-25815.3	-25825.1	-19.6	S	p	0.6	0.1	0.3	0.0
					ω			5.0	5.0
lman2la	-14426.2	-14426.5	-0.7	NS	p	0.9	0.1	0.0	0.0
					ω			166.2	166.2
lman2lb	-14424.9	-14425.6	-1.4	NS	p	0.8	0.1	0.1	0.0
					ω			999.0	999.0
nsfa	-23022.6	-23027.7	-10.3	S	p	0.9	0.0	0.1	0.0
					ω			999.0	999.0
nsfb	-23022.2	-23030.6	-16.8	S	p	0.9	0.0	0.0	0.0
					ω			999.0	999.0
pcsk5a	-64402.8	-64418.5	-31.5	S	p	0.7	0.3	0.0	0.0
					ω			999.0	999.0
pcsk5b	-64390.6	-64422.0	-62.8	S	p	0.7	0.3	0.0	0.0
					ω			439.0	439.0

[a] Significance; S= significant; NS = Not Significant, Values of 999 for dN/dS indicate dS = 0, so dN/dS is undefined.
[b] paralog listed first is the one involved in lateral line system development

6

Morphological and genetic evidence for multiple evolutionary distinct lineages in the endangered Red lined torpedo barbs endemic to the Western Ghats of India

6.1. Abstract

Red lined torpedo barbs (RLTBs) (Cyprinidae: Puntius) endemic to the Western Ghats Hotspot of India, are popular and highly priced freshwater aquarium fishes. Two decades of indiscriminate exploitation for the pet trade, restricted range, and continuing decline in quality of habitats has resulted in their 'Endangered' listing. Here, we tested whether the isolated RLTB populations demonstrated considerable variation qualifying to be considered as distinct conservation targets. Multivariate morphometric analysis using 24 size-adjusted characters delineated all allopatric populations. Similarly, the species-tree highlighted a phylogeny with 12 distinct RLTB lineages corresponding to each of the different riverine populations. However, coalescence-based methods using mitochondrial DNA markers identified only eight evolutionary distinct lineages. Divergence time analysis points to recent separation of the populations, owing to the

geographical isolation, around 5 million years ago, after the lineages were split into two ancestral stocks in the Paleocene, on north and south of a major geographical gap in the Western Ghats. Our results revealing the existence of eight evolutionary distinct RLTB lineages calls for the re-determination of conservation targets for these cryptic and endangered taxa.

6.2. Introduction

Of the 5 ± 3 million species on earth, only 1.5 million have names [248]. Accelerating the description of unknown biodiversity continues to be a major challenge as extinction rates increase [249] and modern taxonomy is far from reaching a scientific consensus on species concept and delimitation [250, 251]. Traditional methods for detecting and describing species are still slow and as a result, distinctive units, such as evolutionarily significant units (ESUs) or designatable units (DUs), which are appropriate targets for conservation may remain undetected for long periods of time [252]. This is a critical impediment particularly for regions harboring exceptionally high biodiversity, that face a high risk of anthropogenic impacts [253] and also among speciose yet poorly known taxa, such as reptiles [75, 254, 255] and freshwater fish [256, 257, 258].

The order cypriniformes is a monophyletic group of primary freshwater fishes containing over 3500 species, with a wide distribution in North America, Europe, Africa and Asia [259, 244]. These fishes are an essential protein source for many societies, are highly valued in recreational fisheries and constitute a major component of the tropical fish trade [260]. Being a taxonomically diverse group exhibiting a remarkable and fascinating array of morphologies, cypriniform fishes present many challenges to systematists and evolutionary biologists [244]. Such challenges are particularly severe in biogeographic 'Hotspots' such as the Western Ghats (WG) of India, where endemic lineages have evolved in several taxa, including fishes, due to extended geographical isolation [261, 262, 263].

Several small (<220km) and isolated (not inter-connected) west flowing rivers between 8° and 12° latitudes in the WG harbor a unique assemblage of endemic freshwater fishes, sometimes as high as 129 species within a sub-basin [264]. This remarkable diversity is nevertheless known to be a gross under-representation [265], as around 10-20% of fish species in any basin of reasonable size in this region are likely to be undescribed [266]. Connections and divisions between rivers affect opportunities for dispersal, which while allowing the gene flow between some populations may promote the isolation of others [267, 268].

The endemic Red Lined Torpedo Barbs (RLTBs) are represented by two extremely popular aquarium species, *Puntius denisonii* and *P. chalakkudiensis*, significant numbers of which are being collected from the wild [269, 270]. RLTBs occur as fragmented populations (figure 6.1; Table 6.1) in 14 small rivers in the WG [271, 272]. However, due to their restricted distribution, unregulated exploitation, decrement in habitat quality and population decline, both species are currently listed as Endangered in the IUCN Red List [271, 272]. In spite of their public appeal,

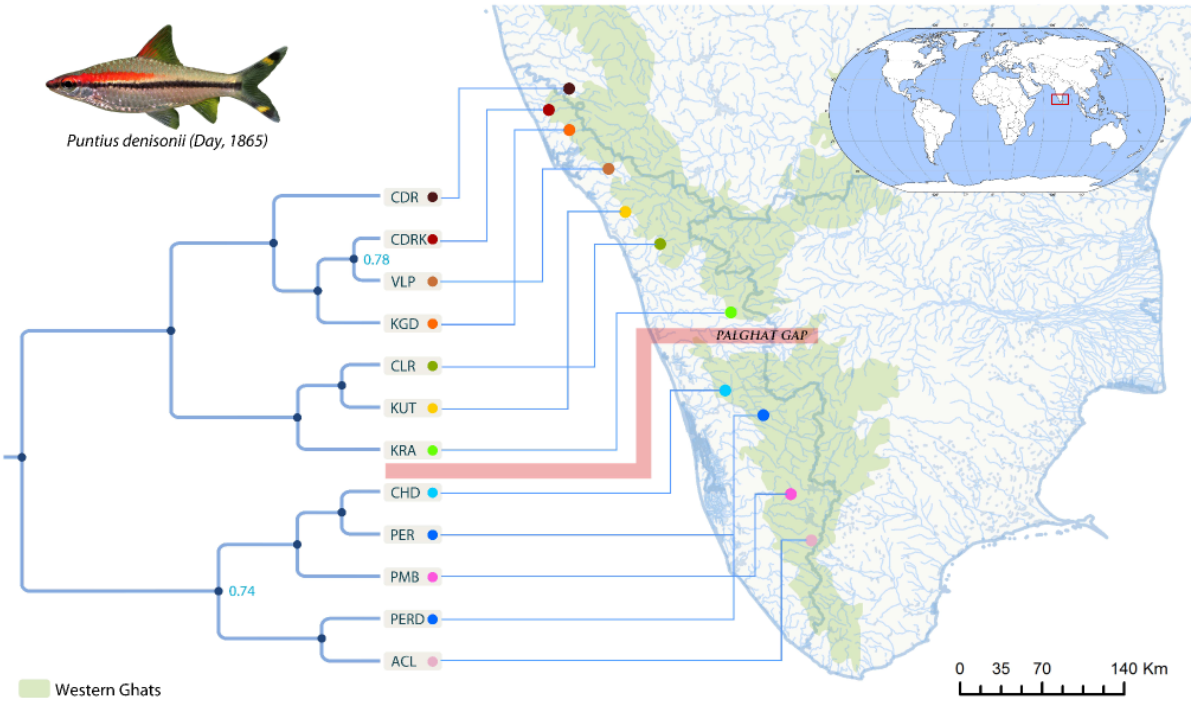


Figure 6.1.: Map showing distribution range of RLTBs and rivers from where samples were collected, the species tree built in *BEAST is shown on the left side. Posterior probability values below 1 are shown at the nodes. Photograph of a specimen considered as *Puntius denisonii* is shown; notice the absence of a black spot on the dorsal fin which is the current diagnostic character for distinguishing it from its congener *Puntius chalakkudiensis* found at location CHD in the map, the tip label codes are explained in Table 6.1.

popularity and conservation importance, the RLTBs have, however, received little scientific attention. Uncertainty still exist on whether the RLTBs comprise one [273, 274], two [275, 276] or more species [258].

Here, we tested whether the RLTB populations demonstrated considerable variation in the absence of gene flow due to geographic isolation. We uncover eight evolutionary distinct lineages that advance our understanding of cyprinid evolution in the WG of India, but at the same time raising numerous conservation and management challenges for one of the world's most popular freshwater aquarium species.

6.3. Results

6.3.1. Morphological analyses

Univariate analysis of normality suggested that 24 out of 28 characters were normally distributed. After removing these four variables the resultant matrix of 24 characters did not deviate significantly from multivariate normality (Doornik and Hansen [277] omnibus, Ep = 56.68,

P = 0.1829). All size-adjusted characters were significantly different for the 12 studied populations (Appendix 4 Table S1). MANOVA/CVA [278] extracted 11 factors out of which the first two axes explain 61.63% of the total variation. The null hypothesis that the mean vectors of the 12 groups are equal was rejected (Pillai's trace = 6.361, F308,726 = 3.232, P < 0.0001) and Fisher's distances between the groups suggested that all 12 populations formed significantly different clusters (figure 6.2 and Appendix 4 figures 14.1 and 14.2).

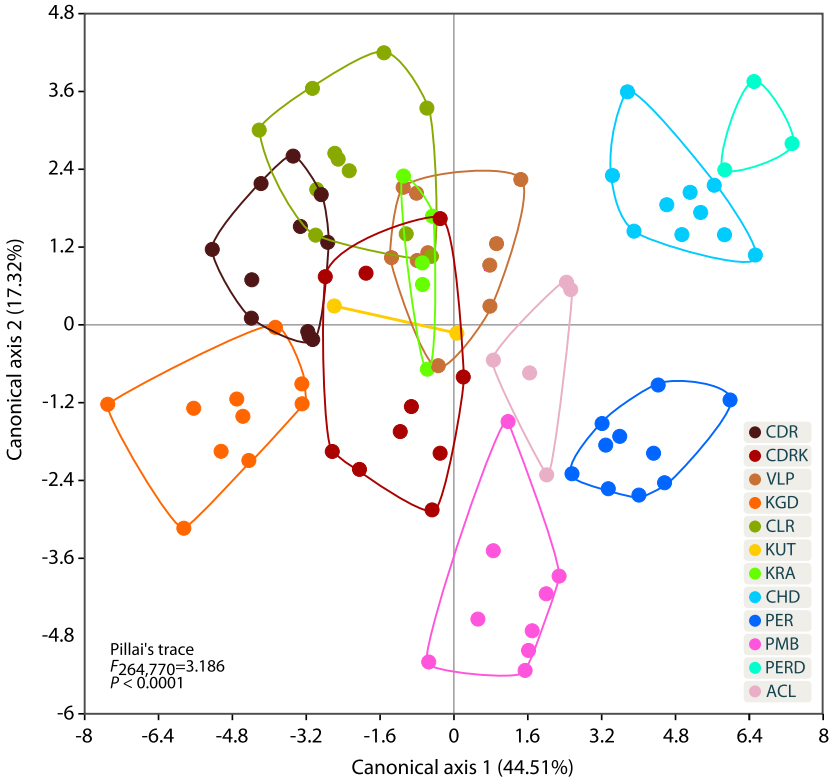
Based on the distribution of the populations along the first canonical axis, the 12 populations formed two feeble clades (figure 6.2a), one comprising the populations north (CDR, CDRK, VLP, KGD, CLR, KUT and KRA) and the other south (CHD, PER, PERD, PMB and ACL) of the Palghat (or Palakkad) gap, a major geographical discontinuity in the WG at 11°N [see 4]. Among the multiple variables separating the northern populations from the southern ones (Appendix 4 Table S2), the two most prominent characters were comparatively higher head length and lower caudal peduncle depth in the northern populations. This distinction of two separate clades of northern and southern populations was also supported by non-metric multidimensional scaling (NMDS) of the centroids where the southern populations were distributed along the negative axis while northern populations were distributed along the positive axis of the first NMDS axis (Appendix 4 figure 14.1). Species discrimination in different RLTB populations could only be resolved with complex linear discriminant functions (Appendix 4 Table S3), but not by univariate comparison between populations (Appendix 4 figure 14.3). Removing the four non-normally distributed characters from the parametric analysis did not substantially influence the statistical analysis, and NPMANOVA performed on all 28 size adjusted variables suggested that our results were qualitatively similar with significant difference among 12 populations (number of permutations = 100000, F = 7.999, P = 0.00001).

6.3.2. Genetic analyses

The initial evaluation of our data suggested ample phylogenetic signal as evidenced by having more than 90% of the quartets resolved in the likelihood mapping procedure [279] for both the alignments (see materials and methods and Appendix 4 figures 14.4, 14.5 and 14.6). The species-tree from *BEAST [280] identified each of the 12 a priori designated groups as distinct clusters with high posterior probability (>70) (figure 6.1). Only one terminal split (CDRK-VLP) (figure 6.1) had a posterior probability less than one. However, an initial maximum likelihood phylogenetic tree constructed using the concatenated alignment produced only eight distinct clades (Appendix 4 figure 14.4). Thus, the first method supported a scenario where all 12 populations were distinct, while the second indicated only eight distinct clades.

Five evolutionary distinct lineages (figure 6.3a) were distinguished by a fixed distance threshold of ≥3% [281]. GMYC method [282] suggested that the single threshold model was a better fit to the data than the null model (LRT = 9.03e-10 for cytb, and 4.78e-7 for concatenated tree). Similarly the multiple threshold model fit the data better than the null model (LRT = 3e-3 for cytb

(a)



(b)

	CDR	CDRK	VLP	KGD	CLR	KUT	KRA	CHD	PER	PERD	PMB	ACL
CDR		5.079	5.142	2.060	4.726	3.891	2.816	12.142	12.589	10.226	9.778	4.793
CDRK	< 0.0001		3.640	4.255	3.954	1.798	4.958	11.200	10.080	9.401	9.536	2.516
VLP	< 0.0001	< 0.0001		6.217	2.702	2.354	3.971	10.508	9.006	8.889	12.703	2.240
KGD	0.011	< 0.0001	< 0.0001		6.145	3.625	4.331	15.719	13.603	12.699	9.507	5.300
CLR	< 0.0001	< 0.0001	0.00077	< 0.0001		2.131	6.241	14.535	15.789	11.653	17.154	4.823
KUT	< 0.0001	0.031	0.00322	< 0.0001	0.008		3.839	5.926	5.018	6.237	6.057	2.359
KRA	0.000	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001		6.691	6.734	5.497	6.299	2.552
CHD	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001		5.499	3.797	7.679	4.164
PER	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001		4.893	6.246	2.973
PERD	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001		8.100	5.292
PMB	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001	< 0.0001		4.627
ACL	< 0.0001	0.002	0.005	< 0.0001	< 0.0001	0.003	0.001	< 0.0001	0.000	< 0.0001	< 0.0001	

Figure 6.2.: MANOVA/CVA of 24 size adjusted biometric characters of 12 RLTB populations. (a) Clusters of all 12 populations on the first two canonical axis and (b) pair wise matrix of Fisher's distances between the centroids of the clusters (upper diagonal) and P values for Fisher's distances (lower diagonal). Percent discrimination by each canonical axis is shown in parenthesis.

and 9.84e-7 for concatenated tree). The multiple threshold model distinguished six lineages based on the cytb tree and nine lineages based on the concatenated ultrametric tree (figure 6.3b; see also Appendix 4 Table S4).

When assuming 12 populations (tips), based on a guide tree produced using *BEAST [280], bayesian species delimitation (bpp) [73] supported eight distinct lineages with posterior probabilities of >0.98 on 7 out of the 11 nodes on the guide tree (figure 6.1). Different prior distributions on the ancestral population size (θ) and root age (τ) did not affect these results (figure

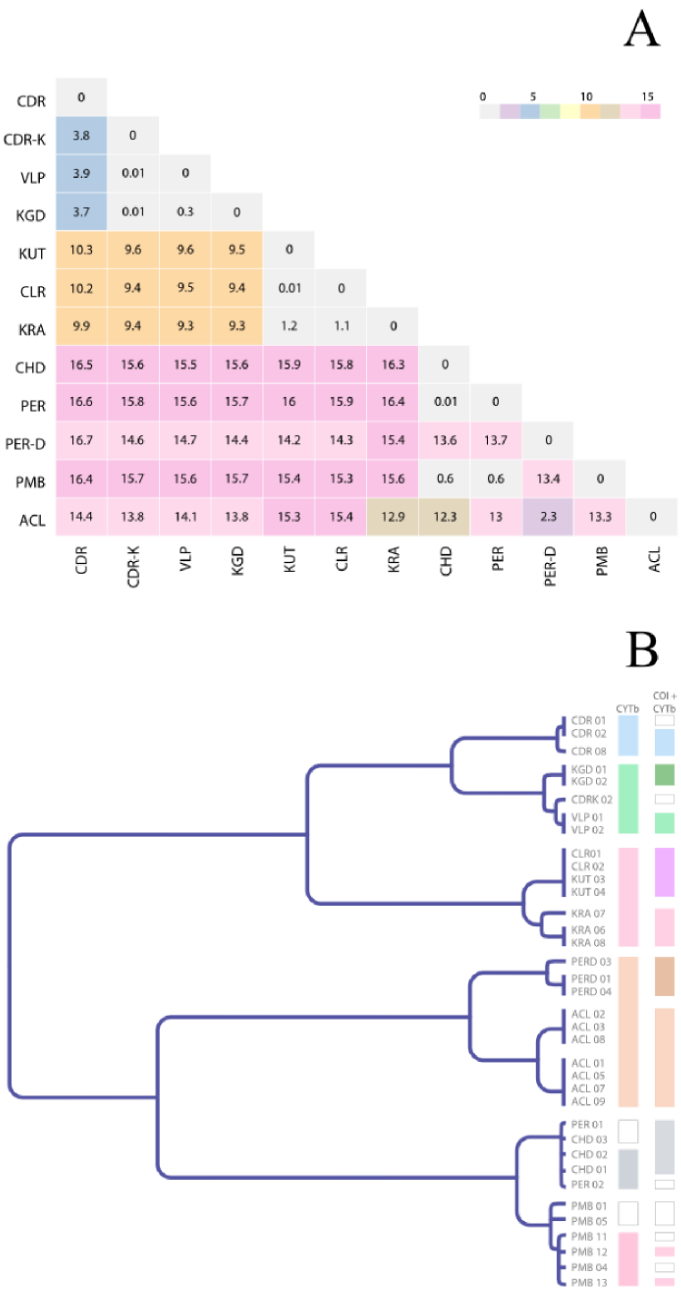


Figure 6.3.: Results of a) Fixed distance threshold methods and b) GMYC methods.

6.4). Thus, the multiple-lineage model explained the data better than the single lineage model as evidenced by the higher posterior probabilities for a multiple species-tree and high (>0.98) speciation probabilities on the nodes of the guide tree.

In short, the Bayesian coalescent analysis and the ML tree identified eight lineages with high probability. The multivariate analysis (based on morphology) and the GMYC methods (based

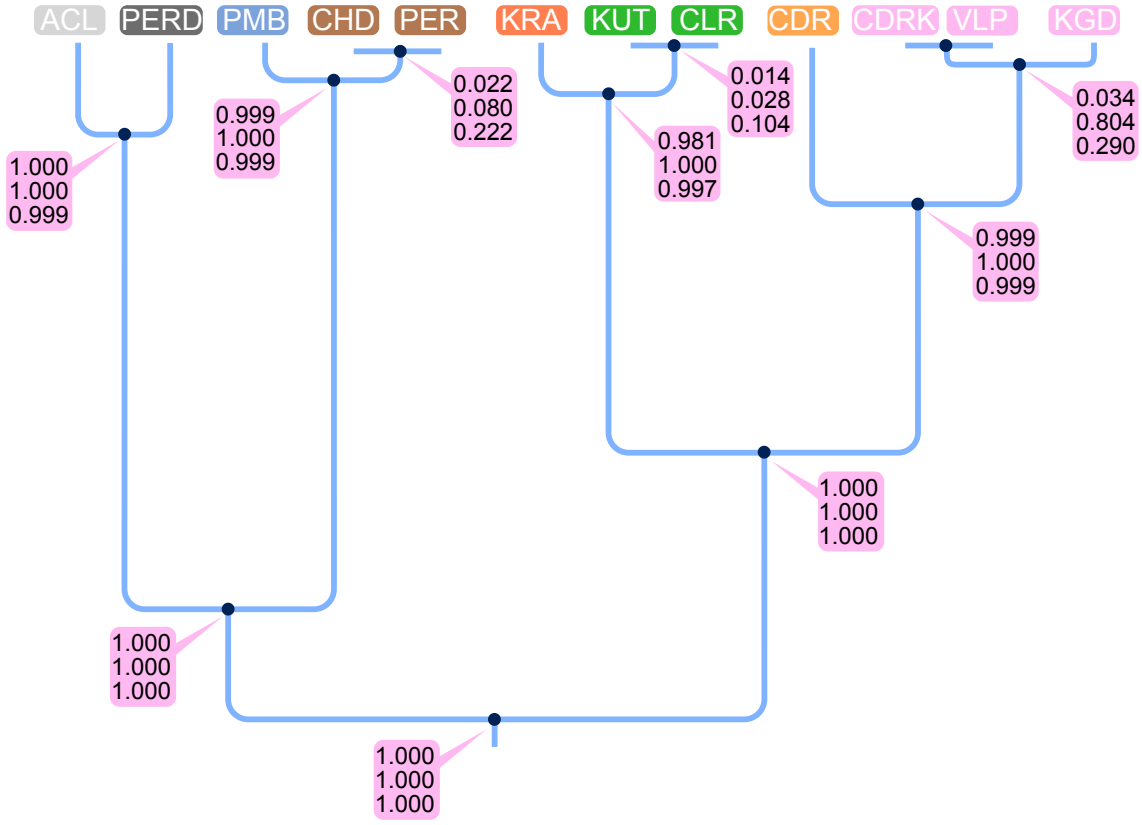


Figure 6.4.: Result of the bayesian species delimitation method applied on the RLTB species tree, values at nodes are the speciation probabilities using three different prior settings (see materials and methods for details); each distinct species (tip) is denoted with a distinct colour.

on concatenated dataset) support those eight clades, and also identify others in addition. Thus, by integrating both morphological and molecular results we propose that RLTBs consists of at least eight evolutionarily distinct lineages, i.e., the number of distinct populations identified with high probabilities and corroborated by both morphological and molecular methods (see Table 6.1).

6.3.3. Divergence time analysis

A phylogenetic tree with *Chanos chanos* (Anotophysi) and *Apteronotus albiforns* (Gymnotiformes) as out-groups was calibrated using fossil ages, with two calibration points as soft bounds: (i) at the base of cyprinidae; we set a minimum age of 49 Million years (Ma) and maximum age of 59 Ma based on the oldest known cyprinid fossil [283, 284], and (ii) at the base of Ostariophysi, a constraint of 146 Ma was set based on the oldest available ostariophysian fossil [285, 286]. In addition to the fossil calibrations, the root of the tree, at the base of Gymnotiformes, was constrained with a loose upper bound, to a maximum age limit of 239 Ma

Table 6.1.: Micro-level distribution of the eight evolutionary distinct lineages including the two recognized species of RLTBs in the Western Ghats

Lineage	Distribution	Remarks
CDR	tributaries of Chandragiri River in Karnataka part of WG and Northern most distribution range of RLTBs	a putative species
CDRK,KGD,VLP	tributaries of Chandragiri River in Kerala part of WG, Karyangode and Valapattanam Rivers	a putative species
KUT, CLR	Kuttyadi and Chaliyar Rivers	a putative species
KRA	Bharatapuzha River	a putative species
CHD, PER	Chalakudy and Periyar Rivers	Type locality of the currently recognized species *Puntius chalakkudiensis* [287] is located in the Chalakudy River.
PERD[a]	Periyar River	Occurs in sympatry with *Puntius chalakkudiensis*; a putative species
PMB[a]	Pampa River	a putative species
ACL[a]	Achankovil River	Southern most distribution range; a putative species

[a] The precise type locality of *P. denisonii* is still unclear. Three river systems, Periyar, Pampa and Achankovil drain the larger landscape in and around from where Francis Day described *P. denisonii* [288].

based on the results of a recent study [285] as the basal time of emergence of Ostariophysi. The ancestor of the RLTBs was estimated to have given rise to two lineages around 59 Ma on north and south of the Palghat gap. Further splits around 28-40 Ma in the Eocene due to vicariance of the lineages from two ancestral stocks eventually gave rise to eight evolutionary distinct lineages at around 5 Ma in its present distribution pattern (figures 6.1 and 6.5; Table 6.1; Appendix 4 Table S5 and Appendix 4 figure 14.5).

6.4. Discussion

Using morphological and various DNA based delimitation methods we provided significant new knowledge on the taxonomic status of RLTBs. Morphometric and initial ML analysis suggested that all 12 populations are distinct. However, the Bayesian coalescent method (and the ML tree) supported only eight lineages with high posterior probabilities (also corroborated by multivariate methods and the GMYC method based on concatenated data), which could signal to a scenario where some populations, even though geographically separated into different river systems, have not genetically diverged significantly. Thus, parsimoniously we have considered RLTBs to be composed of eight evolutionarily distinct lineages (Table 6.1). Our study also validates preliminary claims on cryptic diversity within the RLTBs (e.g. [258]).

Morphometric analysis delineated all allopatric populations of RLTBs as distinct. However, it should be noted that the morphological variation observed during the detailed examination

is not obvious for a layman. Despite the fact that the populations formed different clusters, a univariate analysis of the different size adjusted parameters (Appendix 4 figure 14.3) could not extract distinct character(s) to separate any one population from the rest. However, multivariate discriminant functions (Appendix 4 Table S3) could identify an individual belonging to each population with utmost certainty except in one case where an individual of CDRK population was assigned to CLR in the confusion matrix (Appendix 4 Table S6). This illustrates the complexity in discriminating cryptic populations using morphological analysis indicating that morphological segregation among/between populations can be understood only by a combination of characters.

The Palghat gap has been suggested as a biogeographic barrier [262], which has separated species and/or genetic lineages of several taxa including plants [289], amphibians [290, 291], birds [292] and elephant [293]. Our findings support previous studies and indicate that this biogeographic barrier might have played an important role in the distribution of freshwater fishes. Interestingly, the morphological analysis (figure 6.2a, Appendix 4 figure 14.1 and Appendix 4 figure 14.2) also suggests that the RLTB populations south of the gap have diverged from each other more than those north of the gap.

Further, the divergence time analysis provided evidence that all RLTB populations were separated less than 5 Ma (figure 6.5). We argue that this recent separation event (in evolutionary sense) could explain why different populations exhibited strikingly similar visual morphology. A second argument is that most of the evolutionary significant lineages identified in this study (except one pair of PER vs. PERD) were products of vicariance events around 5 Ma that precluded gene flow among these populations. Allopatric speciation is often observed in populations inhabiting geographically isolated areas with similar ecological characteristics and those events are mostly non-adaptive (as opposed to adaptive radiations), where accelerated evolution of traits and phenotypic divergence are typically absent [294, 295]. Our analysis provides the first evidence for population segregation among the different isolated populations of RLTBs, which should be further validated with a wider sampling and extended molecular markers (e.g., both mitochondrial and nuclear DNA loci).

While our morphological and species-tree methods differentiated all the 12 allopatric RLTB populations, we used the coalescence-based methods to add confidence and to determine the exact number of evolutionarily distinct lineages. The Bayesian method used here is based on the assumption of a biological species concept, where gene flow stops at a speciation event [73]. This method is based on a population genetic perspective and accommodates uncertainty in the phylogeny, as well as lineage sorting due to ancestral polymorphism (see supplementary material). When the user provides a guide tree, which is fully resolved, the program evaluates subtrees by collapsing or splitting nodes (without branch swapping). Under this method, we expect strong support for populations/species isolated for an extended period of time, and weak support for populations/species that have experienced extensive gene flow [75].

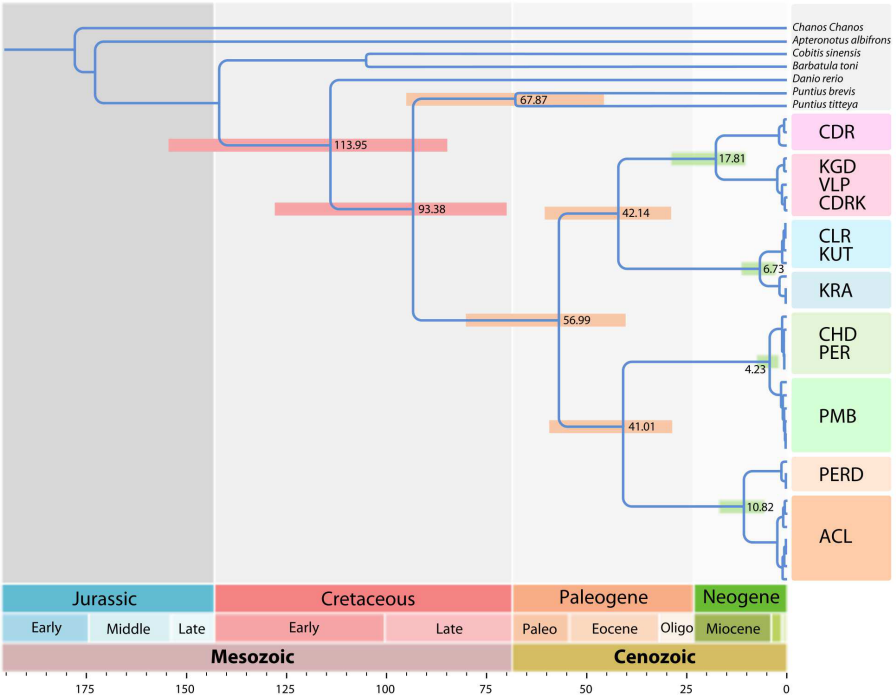


Figure 6.5.: Timetree showing the divergence times of the major RLTB lineages, node bars denote the 95% credibility interval; values at nodes indicate the mean age in million years.

Analysis of our species-tree with three different combinations of priors for population size and divergence time led to concordant results (figure 6.4). The GMYC model supported the eight distinct lineages identified by Bayesian method. GMYC operates on the idea that the branching rates differ at species boundaries. Branch lengths at the species level are determined by the macro-evolutionary process of speciation and extinction, while those at population level by the micro-evolutionary process like coalescence [282, 296]. However, GMYC method could not classify some of the branches as distinct and classified KGD (figure 6.3b) as an independent lineage, which could be due to the fact that this method is sensitive to sample size [297].

While the morphological data analysis differentiated each population as distinct, DNA based methods could identify only eight distinct evolutionary lineages with high confidence (bpp and ML tree; figures 6.3, 6.4 and Appendix 4 figure 14.4). Thus, we propose that the different isolated populations of RLTBs consist of a minimum of eight differentiated lineages, the minimum number of lineages agreed by all methods, which should receive separate conservation attention and be considered as eight distinct management units.

Studies with small sample sizes like the present one are inevitable, when dealing with endangered species with populations distributed even inside protected areas. Future use of multilocus nuclear markers with an increased sample size and the application of coalescence-based methods [71] should yield confidence to the present results. Moreover, detailed taxonomic studies should validate the species status of the evolutionary distinct lineages recognized in

this study.

Although the coalescence-based techniques used here have been useful for species delimitation including description of new species (see [71] and the references therein), there have been concerns on the use of mtDNA for such purposes. We have overcome such problems by testing the phylogenetic signal [279] of our mtDNA dataset (Appendix 4 figure 14.6) that revealed an ample phylogenetic signal. Furthermore, we have been cautious in not overemphasizing our results and considered that while the discrete populations identified here could indeed be distinct species, they should at present be only considered as 'Evolutionary Significant Units' [298].

Our findings of the unrecognized diversity in the RLTBs in the form of evolutionary distinct lineages have considerable impacts for conservation at both local and global scales. Millions of RLTBs are collected (from wild) and exported from the WG since the 1990s. Conservation plans, such as ranching, stock enhancement, translocations and reintroductions, require the ability to distinguish populations, and their evolutionary and ecological boundaries [299]. Our study provides the required information for planning and executing such strategies. The conservation/management units identified in this study can also form the basis for future Red List assessments for *P. denisonii* and *P. chalakkudiensis* [271, 272].

6.4.1. Conclusion

Using RLTBs as a case study, we unravel unrecognized diversity among poorly known yet threatened tropical endemic freshwater fish species. coalescence-based methods led us to discover eight evolutionarily distinct lineages among the isolated RLTB populations. While the advantages and limitations of coalescent-based methods have been discussed recently [71], this method can be extremely useful to supplement biodiversity and taxonomic investigations, and facilitate conservation planning in tropical regions facing the taxonomic impediment. Collecting multilocus datasets could, nevertheless, be prohibitively expensive and, turn away researchers in resource poor (developing and under-developed) nations from using such methods [71]. However, our study demonstrates that even with low sample sizes and few loci, this technique can be adopted by researchers with minimum resources, provided they are used in conjunction with morphological data and with a wide range of samples. Overall, this study advances our understanding of diversity and distribution of freshwater fishes, which comprise one of the world's most threatened vertebrate groups.

6.5. Materials and Methods

A dataset of two mitochondrial gene sequences (cox1 and cytb) and 28 morphometric characters of RLTBs collected from ten rivers throughout its distribution range (figure 6.1, Table 6.1)

was generated. The molecular dataset consisted of an average of 2.9 individuals per population and the morphological data consisted of an average of 7.9 individuals per population.

6.5.1. Ethics statement

Specimens were procured from aquarium collectors and/or directly collected from the wild. Permits for collection of fish inside Protected Areas (PA's) were provided by Kerala State Forest and Wildlife Department (No.WL12-8550/2009) (applicable to four of the sampled sites: ACL, PER/PERD, PMB, CHD; see figure 6.1). Two of the sites from where fishes were collected (KRA and KUT; figure 6.1) fell outside PA's and therefore no permits were required. From the remaining sites, fishes were procured from local aquarium collectors. Fishes were captured by backpack electro-shocker (in sites from where we collected directly) and eco-friendly seine and bag nets (in case of material collected by aquarium fish collectors). For downstream molecular biology protocols, a small piece of tissue from the lower lobe of the caudal fin (fin clip) was excised, and subsequently (whenever possible) the fishes were released back into the same habitat. For morphometric analyses, fishes were transferred to ice slurry post anaesthetization (in 200mg/L Tricaine Methane Sulphonate (MS222)) and transported to the laboratory. We chose ice slurry because the fish had to be transported to long distances (in some cases around 300kms) without compromising on the morphological characters (shape, color) that are essential for taxonomic investigations. Details of samples used for the study is provided in Appendix 4 Table S7. Institutional ethics committee of St. Albert's College, Kochi, Kerala, India (SAC-IAEC 2005-01) approved the design and implementation of the study.

6.5.2. Morphological measurements and analysis

Measurements were made point to point with dial calipers to the nearest 0.1 mm. Counts and measurements were made as far as possible on the left side of specimens following standard methods followed for cyprinid taxonomy [300]. To nullify the effect of size, size adjusted measurements were obtained by expressing subunits of body as percent of standard length (SL) and subunits of head as percent of head length (HL).

Size adjusted morphometric measurements were used for morphometric analysis of the data. Univariate normality for each variable was checked using Shapiro-Wilk test. Variables that were not normally distributed were removed from further parametric analysis; however, all characters were used in non-parametric analysis. ANOVA was performed to understand whether standardized morphometric characters differed among the populations. Since multiple tests were performed on the same data we applied sequential Bonferroni correction to the α wherever applicable. Multivariate normality of the final data was checked using Doornik and Hansen omnibus [277]. MANOVA (Multivariate Analysis of Variance)/CVA (Canonical Variates Analysis) was performed to check whether the populations form significantly distinct clusters morpho-

metrically [278]. MANOVA/CVA explicitly attempts to model the difference between the groups of data by extracting factors that maximize inter group variation and minimize intra group variations. MANOVA/CVA was chosen as a more appropriate technique than Principle Component Analysis (PCA), which gives equal weight to all the variables and as a result cannot reveal the differences among closely related clusters in less number of dimensions. This is true especially when the groups do not have highly diverged morphological structures. However, since MANOVA/CVA considers prior groups, we tested for intra-group homogeneity by two methods so as to account for the bias created by the grouping method itself. (1) The null hypothesis, which states that the mean vectors of the 12 populations are equal, was tested using Pillai's trace [301]. (2) We calculated the Mahalanobis distances among the individuals and computed Fisher's distances between 12 populations (as the distance between the centroids of the two clusters, divided by the sum of their standard deviations) to check if the clusters formed by 12 populations are significantly different. Distances between the centroids of the 12 populations were visualized by performing Non-metric Multidimensional Scaling [302]. To account for any loss of information from the characters, which were not normally distributed, we performed non-Parametric MANOVA (NPMANOVA) [303] [59] on all size adjusted characters to test the null hypothesis that the populations are the same. Statistical analysis was performed in Microsoft EXCEL ®, Systat 12 ® and the freeware PAST [304].

6.5.3. Genetic analyses

To yield confidence to the results from the morphometry based analysis, we attempted various DNA methods, which are described below.

Total genomic DNA from the specimens was isolated using a modified salting out protocol [305]. Partial sequences of two mitochondrial genes, cytochrome b (cytb) and cytochrome oxidase 1 subunit (cox1), were amplified using universal primers published earlier [306, 284]. The amplifications were performed in 25µl reactions containing 1X assay buffer (100 mM Tris, 500 mM KCl, 0.1% gelatin, pH 9.0) with 1.5 mM MgCl2, 10 p moles/µL of primer mix, 10 mM dNTPs), 1.5 U Taq DNA polymerase and 20 ng of template DNA. To evaluate the reliability of the DNA amplification, a negative control was set up by omitting the template DNA from the reaction mixture. The reaction mixture was initially denatured at 95°C for 5 minutes followed by 29 cycles [denaturation at 94°C for 45 seconds, annealing at 50°C (for CYTb) or 54°C (for COI) for 30 seconds and 72°C for 45 seconds]. Reaction was then subjected to a final extension at 72°C for 5 minutes. The PCR products were then cleaned up and subsequently sent for sequencing.

The DNA sequences were edited using BIOEDIT [307] and aligned using MUSCLE [40]. Relationships among the mtDNA haplotypes were assessed using the maximum-likelihood (ML) method implemented in TREEFINDER [308]. Phylogenetic signal of the datasets were analyzed using the likelihood-mapping procedure [279]. Before carrying out the Maximum likeli-

hood analysis the best-fit nucleotide substitution models were determined using TREEFINDER [308]. Sequences generated for this study are deposited in Genbank (Appendix 4 Table S8). Trace files for each of the sequences are available from dx.doi.org/10.6084/m9.figshare.95635.

DNA barcoding methods use the mitochondrial cox1 sequence based fixed distance thresholds to delineate distinct lineages [281, 306]. We calculated the maximum likelihood distances for the concatenated dataset of cox1 and ctyb sequences. The dataset was divided into two partitions and distance calculated based on the best-fit nucleotide substitution models HKY+G for cytb partition and TVM+G for cox1 partition, with five rate categories. Maximum likelihood distance calculation was done in TREEFINDER [308].

General mixed yule coalescence [282] model is based on the knowledge that there are changes in the branching rates at the species boundaries. The GMYC exploits the predicted difference in branching rate under the two modes of lineage evolution, where the branching patterns within each genetic cluster reflects a neutral coalescent process and the branching patterns between two genetic clusters reflects timing of speciation events, and by assessing the point of highest likelihood of the transition [282, 309] it differentiates the evolutionary distinct lineages. Monaghan and co workers [296] developed a modified GMYC model that allows for a variable transition from coalescent to speciation among lineages by identifying multiple thresholds reflecting the variable lineage divergence. The likelihood values of the GMYC models are compared to a null model, which assumes a single branching process for the tree, using a Likelihood Ratio Test (LRT).

GMYC clustering was performed using the package splits (SPecies' LImits by Threshold Statistics, http://r-forge.r-project.org/projects/splits/) implemented in R [187]. A Maximum Likelihood tree using the concatenated dataset and the cytb tree (separately) was used to generate the ultrametric tree. The ultrametric tree was constructed using a penalised likelihood method [310] employed in the R package ape [311], polytomies were resolved and outgroup tips were dropped before employing the GMYC model.

Bayesian Phylogenetics and Phylogeography software (bpp v. 2.1a; [73]) was used to identify distinct evolutionary lineages. This method requires a multi-species multi-gene dataset and also requires that the user assign the candidate groups prior to the analysis, and a phylogeny showing the relationships between the groups. We assumed that each sampling location was a distinct population, since each of the sampling locations are isolated drainages and no gene flow is possible among the populations, except in two cases of PER-PERD and CRD-CDRK. PERD was a morphological variant compared to the commonly occurring specimens PER in river Periyar, while the second group CDR and CDRK occurred in two distant tributaries of River Chandragiri (figure 6.1, Appendix 4 Table S7). Thus we had samples from 10 isolated rivers, which we assigned as 12 distinct clusters for the bpp analysis (figure 6.4).

Our first strategy was to construct a species-tree using the multilocus data for the different populations, for this we employed the program *BEAST [280]. *BEAST estimates the species-

tree directly from the sequence data, and incorporates uncertainty associated with gene trees, nucleotide substitution model parameters and the coalescent process [280]. The MCMC analysis was run twice and a total of 50 million generations (sampling trees every 1000 generations), first 25% trees were discarded as burnin and the convergence was examined TRACER v. 1.4.1 (http://beast.bio.ed.ac.uk/Tracer). The species tree was summarised using the tree-annotator program from the BEAST package [312] and the tree was visualised and edited using figtree (http://tree.bio.ed.ac.uk/software/figtree/).

The Bayesian (bpp) method accommodates the species phylogeny as well as lineage sorting due to ancestral polymorphism. The parameters in the model include the species divergence times τ, measured by the expected number of mutations per site, and population size parameters θ = 4Nμ, where N is the effective population size and μ is the mutation rate per site per generation so that θ is the average proportion of different sites between two sequences sampled at random from the population.

The prior distributions on the ancestral population size (θ) and root age (τ) can affect the posterior probabilities for models, with large values for θ and small values for τ favouring conservative models containing fewer species [73]. We evaluated the influence of these priors by considering three different combinations of prior, similar to an earlier study [75].

The first combination of priors was to set a relatively large ancestral population size θ G (1, 10) and deep divergence time and τ G (1, 10) both with a mean of 0.1 and variance of 0.01. The second combination was to set a small ancestral population θ G (2, 2000) and shallow divergence time τ G (2, 2000), both with a prior mean 0.001 and variance of 5e-07. The third prior combination set a large ancestral population θ G (1, 10), with shallow divergence time τ G (2, 2000). The rjMCMC algorithm-0 was run with a fine tune parameter of 15 and 20 and was run twice to confirm consistency between runs. The species tree and the sequence alignment used for the analysis are available for download from (http://dx.doi.org/10.6084/m9.figshare.95635). The program outputs the speciation probabilities at each nodes of the maximum posterior probability tree (MAP) tree. A posterior (speciation) probability of >0.95 was considered as a strong evidence of speciation at the node, we also ensured that all the three different priors used produced consistent results.

6.5.4. Divergence time estimation

We estimated the divergence times at each node of the phylogenetic tree using MCMCtree [313] with a log-normal rate prior and birth-death time prior. Independent rates for each branch was considered, and maximum likelihood estimation of branch lengths was done using HKY85 model [231].

The node at the base of cyprinidae was based on the oldest cyprinid fossil [283, 284] and was set to 49-59 million years ago. The root node (figure 6.5) was constrained to an upper bound of 239 million years ago and a lower bound of 146 million years ago [285, 286]. The MCMC

algorithm was run for 5 X 20000 iterations, and first 2000 samples were discarded as burnin. The outgroups for cypriniformes used for the phylogeny construction and divergence time estimation were *Chanos chanos* (Anotophysi) and *Apteronotus albiforns* (Gymnotiformes). The gamma prior for the overall rate parameter μ was set to G (2,7), with a mean of 0.29 and variance of 0.04. The rates for individual loci were calculated using baseml program implemented in PAML package (v4.4a; [56]), with global clock assumption and fossil calibrations as specified above.

6.6. Papers arising from this chapter

Lijo John*, Siby Philip*, Neelesh Dahanukar, Anvar Ali P. H., Josin Tharian, Rajeev Raghavan@ and Agostinho Antunes@. 2012. **Cryptic diversity in the endangered Red Lined Torpedo Barbs: morphological and molecular species delimitation methods reveal eight putative species**. In review.

* Equally contributed to the work, joint first authors. @ Joint corresponding authors.

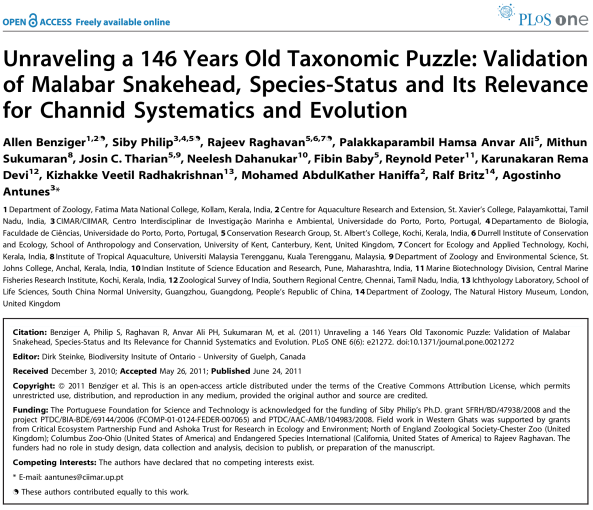
Contribution

I participated in designing the work and participated in part of the sampling. Carried out all the (in silico) phylogenetic, divergence time and DNA based species delimitation analysis, and jointly drafted the paper with AA, ND and RR.

7

Unraveling a 146 Years Old Taxonomic Puzzle: Validation of Malabar Snakehead, Species-Status and its Relevance for Channid Systematics and Evolution

Papers arising from this chapter



OPEN ACCESS Freely available online

PLoS one

Unraveling a 146 Years Old Taxonomic Puzzle: Validation of Malabar Snakehead, Species-Status and Its Relevance for Channid Systematics and Evolution

Allen Benziger[1,2☯], Siby Philip[3,4,5☯], Rajeev Raghavan[5,6,7☯], Palakkaparambil Hamsa Anvar Ali[5], Mithun Sukumaran[8], Josin C. Tharian[5,9], Neelesh Dahanukar[10], Fibin Baby[5], Reynold Peter[11], Karunakaran Rema Devi[12], Kizhakke Veetil Radhakrishnan[13], Mohamed AbdulKather Haniffa[2], Ralf Britz[14], Agostinho Antunes[3]*

1 Department of Zoology, Fatima Mata National College, Kollam, Kerala, India, **2** Centre for Aquaculture Research and Extension, St. Xavier's College, Palayamkottai, Tamil Nadu, India, **3** CIMAR/CIIMAR, Centro Interdisciplinar de Investigação Marinha e Ambiental, Universidade do Porto, Porto, Portugal, **4** Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, **5** Conservation Research Group, St. Albert's College, Kochi, Kerala, India, **6** Durrell Institute of Conservation and Ecology, School of Anthropology and Conservation, University of Kent, Canterbury, Kent, United Kingdom, **7** Concert for Ecology and Applied Technology, Kochi, Kerala, India, **8** Institute of Tropical Aquaculture, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia, **9** Department of Zoology and Environmental Science, St. Johns College, Anchal, Kerala, India, **10** Indian Institute of Science Education and Research, Pune, Maharashtra, India, **11** Marine Biotechnology Division, Central Marine Fisheries Research Institute, Kochi, Kerala, India, **12** Zoological Survey of India, Southern Regional Centre, Chennai, Tamil Nadu, India, **13** Ichthyology Laboratory, School of Life Sciences, South China Normal University, Guangzhou, Guangdong, People's Republic of China, **14** Department of Zoology, The Natural History Museum, London, United Kingdom

Citation: Benziger A, Philip S, Raghavan R, Anvar Ali PH, Sukumaran M, et al. (2011) Unraveling a 146 Years Old Taxonomic Puzzle: Validation of Malabar Snakehead, Species-Status and Its Relevance for Channid Systematics and Evolution. PLoS ONE 6(6): e21272. doi:10.1371/journal.pone.0021272

Editor: Dirk Steinke, Biodiversity Institute of Ontario - University of Guelph, Canada

Received December 3, 2010; **Accepted** May 26, 2011; **Published** June 24, 2011

Funding: The Portuguese Foundation for Science and Technology is acknowledged for the funding of Siby Philip's Ph.D. grant SFRH/BD/47938/2008 and the project PTDC/BIA-BDE/69144/2006 (FCOMP-01-0124-FEDER-007063) and PTDC/AAC-AMB/104983/2008. Field work in Western Ghats was supported by grants from Critical Ecosystem Partnership Fund and Ashoka Trust for Research in Ecology and Environment; North of England Zoological Society-Chester Zoo (United Kingdom); Columbus Zoo-Ohio (United States of America) and Endangered Species International (California, United States of America) to Rajeev Raghavan. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: aantunes@cimar.up.pt

☯ These authors contributed equally to this work.

Figure 7.1.: Papers arising from this chapter:***PloS one***, 6(6), e21272.

7.1. Abstract

7.1.1. Background:

The Malabar snakehead Channa diplogramma is one of the most enigmatic and least understood species within the family Channidae, which comprise one of the most important groups of freshwater food fish in tropical Asia. Since its description from peninsular India in 1865, it has remained a taxonomic puzzle with many researchers questioning its validity, based on its striking similarity with the South East Asian *C. micropeltes*. In this study, we assessed the identity of the Malabar snakehead, *C.diplogramma*, using morphological and molecular genetic analyses, and also evaluated its phylogenetic relationships and evolutionary biogeography.

7.1.2. Methodology/Principal findings:

The morphometric and meristic analysis provided conclusive evidence to separate *C.diplogramma* and *C. micropeltes* as two distinct species. Number of caudal fin rays, lateral line scales, scales below lateral line; total vertebrae, pre-anal length and body depth were the most prominent characters that can be used to differentiate both the species. *Channa diplogramma* also shows several ontogenic color phases during its life history, which is shared with *C. micropeltes*. Finally, the genetic distance between both species for the partial mitochondrial 16S rRNA and COI sequences is also well above the intra-specific genetic distances of any other channid species compared in this study.

7.1.3. Conclusions/Significance:

The current distribution of *C.diplogramma* and *C. micropeltes* is best explained by vicariance. The significant variation in the key taxonomic characters and the results of the molecular marker analysis points towards an allopatric speciation event or vicariant divergence from a common ancestor, which molecular data suggests to have occurred as early as 21.76 million years ago. The resurrection of *C. diplogramma* from the synonymy of *C. micropeltes* has hence been confirmed 146 years after its initial description and 134 years after it was synonymised, establishing it is an endemic species of peninsular India and prioritizing its conservation value.

7.2. Introduction

Freshwater fishes comprise one of the most diverse groups of vertebrates with an estimated 13,000 species worldwide, and many more waiting to be described in the tropics, especially in countries where exploratory surveys are still incomplete such as China and India [314]. In the Southern Indian state of Kerala, where this study was based, 10-20% of the fishes in any basin of reasonable size are thought to be undescribed [266]. This slow rate of progress in fish species assessments and identification is largely due to the lack of funding and trained taxonomists in these regions, all of which contribute to the 'taxonomic impediment' [315].

Snakeheads of the genus Channa comprise one of the most important groups of freshwater food fish in tropical Asia [316], with a wide natural distribution extending across the continent from Iran in the West, to China in the East, and parts of Siberia in the Far East [317]. They are one of the most common staple food fish in Thailand, Cambodia, Vietnam and other South East Asian countries where they are extensively cultured [316, 318]. Apart from their importance as a food fish, snakeheads are also consumed as a therapeutic for wound healing as well as reducing post-operative pain and discomfort [319], and collected for the international aquarium pet trade [320].

The taxonomy of the genus Channa remains incompletely known, as a comprehensive revision of the family has not been performed, and more new species continue to be described. Therefore, an uncertainty still exists regarding the total number of species within this genus. Of the 87 nominal species and 4 subspecies that have been described, many are now considered synonyms of recognized species, and there are about 20 names that cannot be associated with any valid taxa [321]. It has also been suggested that as many as five species viz, *C. gachua, C. marulius, C. micropeltes, C. punctata*, and *C. striata* may in fact represent "species complexes" [321, 322, 323]. A recent phylogenetic study has also indicated the likelihood of the existence of more undescribed species of channids in South East Asia [323].

The Malabar snakehead, *Channa diplogramma* is one of the most enigmatic and least known of all channids. Sir Francis Day [288] described *Ophiocephalus diplogramma* in 1865 based on one juvenile specimen (42 mm in length) collected near the mouth of the Cochin River in the port

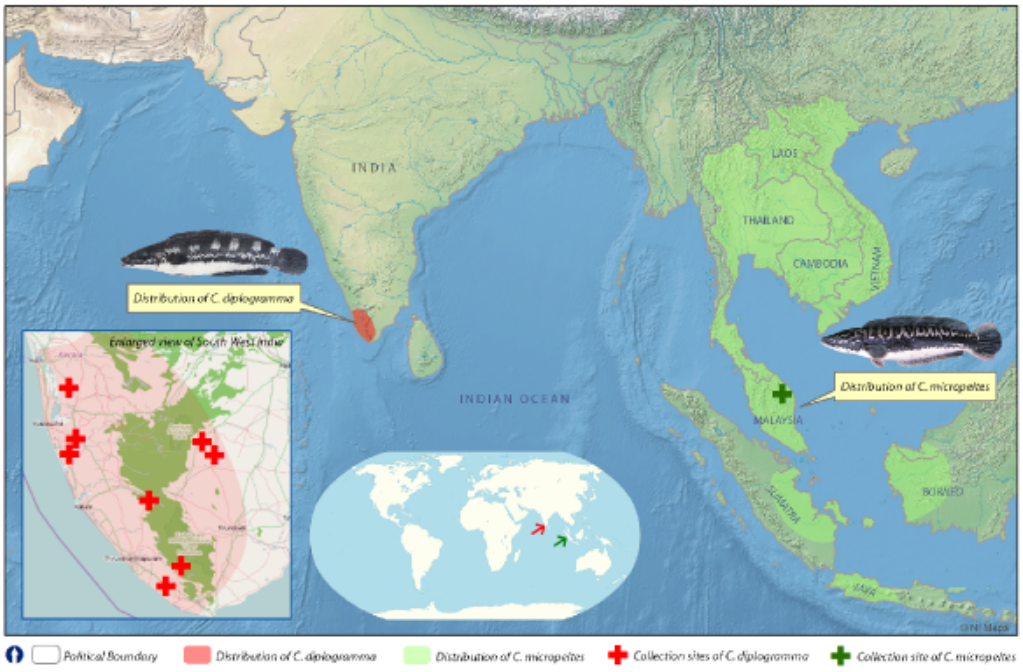


Figure 7.2.: Map showing the distribution range of *Channa diplogramma* (in pink shades) and *C. micropeltes* (in green shades) including sampling sites for the present study.

city of Cochin (Southwestern India), and called it Malabar snakehead (Holotype at the Natural History Museum, London; BMNH 1865.7.17.24). The color pattern of this juvenile matched with that of juveniles of another species of snakehead, *O. micropeltes* originally described by Cuvier and Valenciennes [324] from Java, Indonesia. This possibly led Francis Day to synonymise *C. diplogramma* with *C. micropeltes* in 1878 [325] (Table 7.1). The close similarity, rarity of adult specimens in museum collections, and the fact that no taxonomist has studied this snakehead since its description, resulted in the acceptance of the synonymy by subsequent taxonomists [321, 326, 327, 328]. However, recent researchers [323, 329] suggested that *C. diplogramma* is distinct from *C. micropeltes* and should be considered as a valid species.

In peninsular India, from where *C. diplogramma* was described (Figure 7.2), this species has long been identified and documented as *C. micropeltes* [321, 328, 330, 331, 332, 333, 334]. But there have also been opinions that the species recorded as *C. micropeltes* from India is actually a distinct species [329], and that it is *C. diplogramma* [335]. There are also others who have suggested that both *C. micropeltes* and *C. diplogramma* occur in India [336], while another school of thought was that *C. micropeltes* was introduced, prior to mid 1800's, to South India from South East Asia since Cochin was a major port with trading activity for many centuries [321].

The primary aim of this paper was to resolve the taxonomic ambiguity, and discuss the identity as well as systematic position of the Malabar snakehead,*C. diplogramma*, using morphologi-

cal and molecular genetic (mitochondrial 16S rRNA and COI gene) information, in addition to making an attempt to understand its phylogenetic relationships and evolutionary biogeography. Both morphological and genetic analyses support *C. diplogramma* as a distinct and valid species endemic to peninsular India and reveal its importance for conservation.

7.3. Methods

7.3.1. Biometry

Measurements and counts followed those in standard literature on channid taxonomy [337, 338]. Rays were counted with a binocular microscope and vertebral counts were taken from radiographs. The following abbreviations are used in the text: SL, standard length and TL, total length. Institutional abbreviations: BMNH – Natural History Museum, London, United Kingdom; RMNH - Rijksmuseum van Natuurlijke Histoire RMNH/Naturalis, Leiden, The Netherlands; NHM – Natural History Museum, Vienna, Austria; UMT – Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia; CRG- Conservation Research Group, Department of Aquaculture, St. Albert's College, Kochi, India.

Ten individuals of the Malabar Snakehead were collected from the Rivers Meenachil (9.65°N & 76.59°E) and Pamba (9.36°N & 76.53°E) in Kerala, India and five individuals of *C. micropeltes* collected from Tasik Kenyir Lake (4.96°N & 102.70°E) in Terengganu State, Malaysia. At the first stage, the morphometric and meristic characters of these fresh specimens were matched and confirmed with those of the type specimens of both species (RMNH D2318, BMNH 1865.7.17.24) (see Appendix 5 Table S1 and Figure 7.3 for details and measurements of the type specimen). Since the types of *C. micropeltes* were dry (stuffed) specimens, with missing fin rays and dry/damaged scales, we could not do a complete morphometric assessment. We therefore used only the measurements of fresh specimens to do the statistical analyses. The measurements were compared using a two-tailed unpaired t test. For some of the meristic characters where one species did not show any variation, we performed one-sample t test with the character value of the species showing no variation as the hypothetical mean. Principle Component Analysis (PCA) was performed on the morphometric characters (measured as % TL) and meristic characters using a correlation matrix between the variables to nullify the size and unit effect. The PCA was performed in Statistica 10® and the PCA biplot was plotted using the freeware Biplot 1.1 [339].

Voucher specimens of *C. diplogramma* examined in our study are currently deposited at the museum of CRG, Department of Aquaculture, St. Albert's College, Kochi, India (CRG-CHDIP-20-CRG-CHDIP- 29), while those of C. micropeltes at the Museum of the Institute of Tropical Aquaculture, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia (UMTCM1 to UMTCM5).

7.3.2. DNA extraction, amplification, sequencing and analysis

The total genomic DNA of two individuals each from six of the eight Channa species found in India (*C. aurantimaculata, C. bleheri, C. gachua, C. marulius, C. punctata, C. striata*), six individuals of *C. diplogramma* (River Meenachil, India; 9.65°N and 76.59°E) and one individual of C. micropeltes (Tasik Kenyar, Malaysia; 4.96°N & 102.70°E) were isolated using a modified salting out protocol [305]. Details of the specimens used for the molecular analysis, voucher numbers and museum details are given in Appendix 5 Table S2. Approximately 600 base pair (bp) fragments of the mitochondrial (mtDNA) 16S rRNA and Cytochrome c Oxidase subunit 1 (COI) genes were amplified from each of these eight species of Channa using 1µl of the DNA extract as a template, and using the following primers; L2510 (5'CGC CTG TTT ATC AAA AAC AT 3') and H3080 (5' CCG GTC TGA ACT CAG ATC ACG T 3') for the 16S rRNA gene [340], FishR2-(5' TCA ACC AAC CAC AAA GAC ATT GGC AC 3'), FishR1- (5' TAG ACT TCT GGG TGG CCA AAG AAT CA 3'), FishF2-(5' TCG ACT AAT CAT AAA GAT ATC GGC AC 3'), and FishF1- (5' ACT TCA GGG TGA CCG AAG AAT CAG AA 3') for the COI gene [341]. The amplifications were performed in 25µL reactions containing 1x assay buffer (100mM Tris, 500mM KCl, 0.1% gelatin, pH 9.0) with 1.5mM MgCl2, 10 p moles/µL of primer mix, 10 mMdNTPs), 1.5 U Taq DNA polymerase and 20 ng of template DNA. To evaluate the reliability of the DNA amplification, a negative control was set up by omitting the template DNA from the reaction mixture. The reaction mixture was initially denatured at 95oC for 5 minutes followed by 29 cycles [denaturation at 94°C for 45 seconds, annealing at 50°C (for 16S rRNA) or 54°C (for COI) for 30 seconds and 72°C for 45 seconds]. Reaction was then subjected to a final extension at 72°C for 5 minutes. The PCR products were then cleaned up and subsequently sent for sequencing.

The DNA sequences were edited using BIOEDIT [307] and aligned using MUSCLE [40]. Relationships among the mtDNA haplotypes were assessed using neighbor-joining (NJ) and maximum-likelihood (ML) algorithms in SEAVIEW [121] and PHYML [19], respectively. Before carrying out the Maximum likelihood analysis the best fit nucleotide substitution model was determined using MrAIC [125]. *Notopterus notopterus* was used as an out-group species for all the analyses. A concatenated dataset of both COI and 16S rRNA sequences was prepared to produce a final phylogenetic tree.

7.3.3. Genetic Distance Calculation

Using the best fit nucleotide substitution model the gamma shape parameter was calculated. The estimated value of shape parameter for the discrete Gamma Distribution was 0.2424 for 16S rRNA and 0.2238 for COI. Substitution pattern and rates were estimated under the General Time Reversible model + gamma (GTR+G) with five rate categories. Analyzes were conducted using the Maximum Composite Likelihood method [342] in MEGA5 [148]. The rate

variation among sites was modeled using the previously calculated gamma shape parameter. The differences in the composition bias among sequences were considered in the evolutionary comparisons [343]. All ambiguous positions were removed for each sequence pair.

7.3.4. Phylogenetic tree calibration and divergence time estimation

We used four different tree calibration methods, the Non Parametric Rate Smoothening (NPRS) and its variant NPRS-LOG [69], the Global Rate Minimum Deformation Method (GRMD) and the Local Rate Minimum Deformation Method (LRMD) [308]. The NPRS cost functions have the disadvantage of being asymmetric, but the latter two methods are perfectly symmetric. We implemented 10000 replicates to each method, which produced a two-dimensional array of data replicates, which was then calibrated by rate smoothing. Finally, the mean and confidence limit of rates and divergence times were computed from their observed distribution among the replicate sample.

A calibration file was prepared (expression written in the special purpose Treefinder's language) to implement the calibration constraints in Treefinder [308]. We used two different constraints on the channid phylogenetic tree. The node separating the genus Parachanna from Channa was constrained to 50 million years ago (MYA), which corresponds to the earliest channid fossil records from the early Eocene [344]. The fossils, Kuldana and Chorgali formations of Anchichanna kuldanensis, and another fossil, Eochanna chorlakkiensis, from Chorlakki, both located in the North West Frontier Province of Pakistan, are from deposits believed to be of similar age [345]. The alternative constraint applied of 110 - 84 MYA corresponds to the emergence of the genus Channa [346].

7.4. Results

7.4.1. Taxonomy

Table 7.1.: Taxonomic status of *Channa diplogramma* (Day 1865)

Family:	Channidae
Genus:	Ophiocephalus (Bloch 1793)
Genus:	Channa, Scopoli 1777
	Ophiocephalus diplogramma Day 1865 [288]
	Ophiocephalus diplogramme Day 1865 [347]
	Ophiocephalus micropeltes non Cuvier 1831 [325]
	Channa micropeltes (non Cuvier 1831) [321, 326, 327, 328, 348]
	Channa diplogramma (Day 1865) [323, 329]

Figure 7.3.: Types specimen examined in the study A) *Channa diplogramma* (BMNH 1865.7.17.24) B) *C. micropeltes* (RMNH D2318).

Comparative material

Channa micropeltes - RMNH D2318, 605mm SL, Java (Syntype); RMNH D1131, 210mm SL, Java & D1132 250mm SL, Java (both possible syntypes); four specimens collected from Tasik Kenyar Lake, Terengganu, Malaysia, deposited at the Institute of Tropical Aquaculture, Universiti Malaysia Terengganu, Kuala Terrengganu, Malaysia (UMT CM1 to UMT CM5).

Channa diplogramma - BMNH 1865.7.17.24, 81.6mm SL, Malabar, India (Holotype: Unique); NMW 73835, 352mm SL, Canara, India; NMW 73838, 230mm SL, Mangalore, India; NMW 84220, 380mm SL, Canara, India; Six specimens collected from Meenachil River, Kerala, India and four specimens collected from Pamba River, Kerala, India deposited at the Museum of the Conservation Research Group, St. Albert's College, Kochi, India (CRG-CHDIP 20 to CRG-CHDIP 29).

Diagnosis

Channa diplogramma differs from all other species in the genus by its high number of lateral line scales (103-105 vs. 36-91). It further differs from all other Channa species, except *C. bankanensis, C. lucius, C. micropeltes* and *C. pleurophthalma* by the presence of gular scales, a patch of scales between the anterior tips of the lower jaws, visible in ventral view. *Channa diplogramma* differs from *C. bankanensis, C. lucius*, and *C. pleurophtalma* by having a very different color pattern [338].

From its most closely related species, *C. micropeltes, C. diplogramma* can be distinguished with a combination of characters. As a percentage of standard length, pre anal length of *C. diplogramma* was significantly greater than that of *C. micropeltes* (t = -2.570, df = 13, P = 0.023), while body depth was significantly smaller (t = 2.622, df = 13, P = 0.021) (Table 7.2).

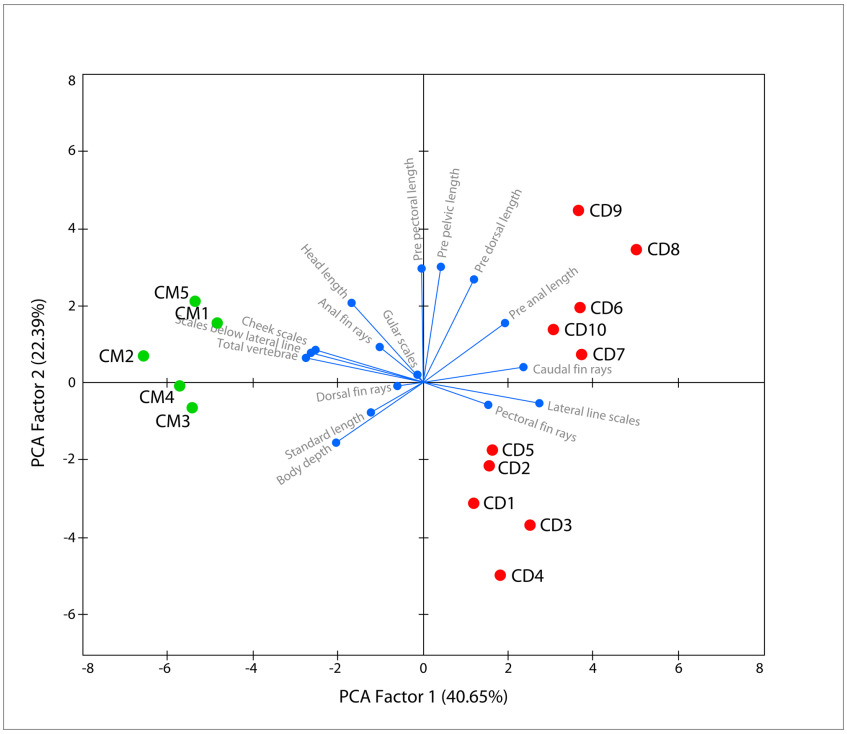


Figure 7.4.: Principle Component Analysis of morphometric and meristic characters of *Channa diplogramma C. micropeltes*.

For the meristic characters, the number of cheek scales (t = 8.529, df = 13, P < 0.0001) and total vertebrae (one-sample t = -20.821, df = 9, P < 0.0001) in *C. diplogramma* was significantly smaller than in *C. micropeltes*, while the number of caudal fin rays (one-sample t = 6.091, df = 9, P < 0.0001) and lateral line scales (one-sample t = 72.962, df = 9, P < 0.0001) was significantly higher (Table 7.3). PCA extracted four factors with eigenvalues higher than 1. Together, these four factors contributed to 86% of the total variation in the data. A clear separation of *C. micropeltes* and *C. diplogramma* was possible along the first PCA axis (Figure 7.4). Variables, namely caudal fin rays, lateral line scales, scales below lateral line, total vertebrae, pre-anal length and body depth, had highest squared cosines on the first PCA factor.

Redescription

Large species, reaching a maximum length of at least 480 mm standard length (SL). Body elongated. Body depth is 14.2-25.6% of SL. Cross section of body is circular in anterior portion, somewhat compressed posteriorly in the caudal area. Body depth is greatest at insertion of dorsal fin. Body width is greatest at insertion of pectoral fin (11.18-21.62% of SL). Head is large, long (25.02-35.06% of SL), dorsally flattened and rounded anteriorly, covered by scales anteriorly up to level of posterior nostrils. Head depth is 52.0-69.3% of head length (HL). Head width is 63.45-86.75% of HL. Inter-orbital region narrow (25.20-40.86% of HL) and slightly

convex. Eye diameter 10.12- 20.83% of HL. Mouth large, upper jaw length 37.9- 51.6% of HL, maxilla extending posteriorly beyond posterior margin of eye. Predorsal scales 21-23. Gular portion covered with 30-31 gular scales. Cephalic sensory pores open via numerous satellite openings in the skin.

Scales on head and body small. Cheek scales 16-20. Lateral line scales small, 103-105. Scale rows above lateral line 10.5, below lateral line 15. Circumpeduncular scales 15-16. Dorsal fin rays 43-44. Anal fin rays 26-28. Pectoral fin rays 17. Pelvic fin with 6 rays. Principal caudal fin rays 15-17. Total vertebrae 53-54. Outer margins of pectoral and caudal fins rounded.

Mouth is big, terminal, with maxilla reaching anteriorly slightly posterior to a vertical through anterior nostril. Many rows of small conical teeth on premaxilla, an additional series of 2–3 times larger conical teeth anteromedially on the premaxilla. Several rows of small teeth at the symphysis, numbers of rows and size of teeth decreasing ventrally along the pre-maxilla towards its posteroventral tip. Vomer and palatine with a series of small teeth marginally, followed medially by several conspicuous, large canines. Dentary with a marginal row of large teeth restricted to the area close to the symphysis, followed medially by several rows of small teeth extending along the dentary and an internal row of conspicuous, large canines. Many variously sized conical teeth on vomer and palatine, those on inner row much larger and canine-like.

Coloration: In life (see subsequent section on ontogenic color phases)

Distribution

Channa diplogramma is endemic to the southern Western Ghats of peninsular India. It is known from the Rivers (including its principal reservoirs) Meenachil, Manimala, Pampa, Achenkovil and Kallada in Kerala state, as well as the Chittar and Tambraparini Rivers (and its reservoirs) in Tamil Nadu state (see Figure 7.2).

Ontogenic color phases of Channa diplogramma

Channa diplogramma shows multiple color phases during its life history (Figure 7.5), which makes local fishers, believe that they are different species. The different specimens are also known by different vernacular names (Pulivaka, Karivaka, Manalvaka, and Charalvaka). We collected eight differently colored specimen (Figure 7.5) of *C. diplogramma* from the rivers Pamba and Meenachil in Kerala, India, which occur sympatrically and utilize the same ecological habitat. *Channa micropeltes* also possess similar ontogenic color phases [349] like *C. diplogramma*. However, local knowledge of the fishers in the Mekong River attributes this color variation of *C. micropeltes* to the differential habitat occupancy of the individuals [349]. Due to logistical difficulties, we were unable to obtain all the morphs of *C. micropeltes* for the present study.

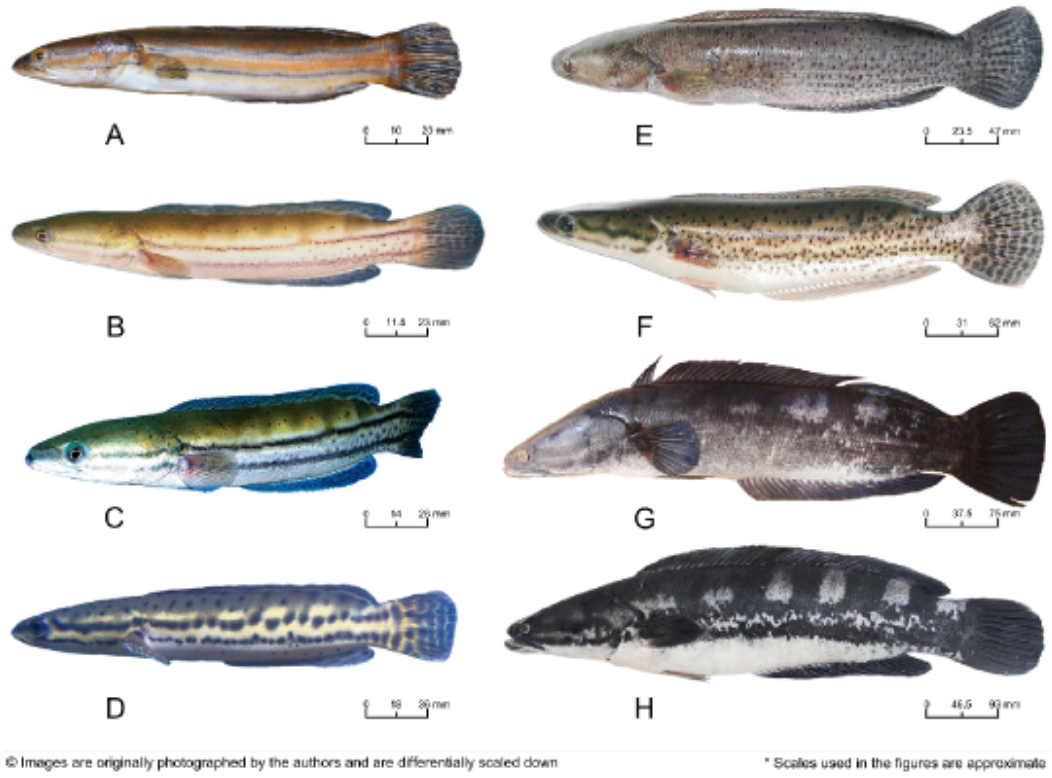


Figure 7.5.: Ontogenetic color phases of *Channa diplogramma* (A: Fingerling; B: Fingerling, C: Juvenile, D: Juvenile, E: Sub-Adult, F: Sub-Adult, G: Adult, H: Adult) (length in millimeters is given as a scale below each specimen). All individuals were collected from the river Pampa in Kerala, India.

We did not observe any individual of the Malabar Snakehead measuring less than 97.1mm TL and so do not have any information on the color pattern or external morphology of early larvae and fry of *C. diplogramma*. In fingerlings and early juveniles, a broad black band passes through the eye straight to the upper half of the caudal fin (Figure 7.5; A-D), and a second black line commences at the angle of the mouth, and proceeds to the lower half of caudal region. An orange colored stripe passes in between these black bands, and the orange color covers most of the dorsal region. During subsequent development (large juveniles), the orange stripe fades and becomes yellow to light brown, and light black; later the black lines fade and black colored spots appear on the body (Figure 7.5; E-F), which changes the color then to off white and grey. From the sub-adult stage, the black colored spots coalesce and four to six white blotches appear on the sides of the body starting from the dorsum downwards up to the lateral line region, later becoming conspicuous in adults (Figure 7.5; G-H). In large adults, the abdomen is pure white, the caudal fin, dorsal surface, cheeks and head in general are black, with a purple tint, while dorsal and anal fins have a grey border.

The ten individuals of *C. diplogramma* used for morphometric and meristic character assessment (Tables 7.2 and 7.3) included all the range of color morphs previously described (two individuals each of morphs A and H, and one sample each of morphs B, C, D, E, F and G; see Figure 7.5). All these ten individuals have almost identical morphometric and meristic characters. Our analyses of the COI and 16S rRNA gene sequences from different color phases of *C. diplogramma* (morphs A, C, D, E, G and H; see Figure 7.5) also revealed that they are genetically identical (same molecular profile; see Appendix 5 Table S2 for details).

7.4.2. Phylogenetic relationships

The 36 nucleotide sequences of the Indian channids (six sequences each of 16S rRNA and COI for *C. diplogramma* and two 16S rRNA and two COI sequences each for the other six channids used in the study) were submitted to GenBank (Accession Numbers: EU342175 to EU342210; Appendix 5 Table S2). In addition, one sequence each of COI and 16S rRNA from the specimen of C. micropeltes used in the present study has been submitted to NCBI (Accession No: JF900369 and JF900370). The phylogenetic trees constructed using the Maximum Likelihood method yielded well-resolved phylogenies in all the cases. GTR+G+I was found to be the best-fit nucleotide substitution model for both the mtDNA 16S rRNA and COI genes. A phylogenetic tree constructed with the 16S rRNA gene sequences, including a sequence of *C. micropeltes* [350]; DQ532852), *C. marulius* from North East India [351] and Parachanna obscura (AY763726), along with the sequences that we generated, clearly distinguishes *C. diplogramma* from *C. micropeltes* (90% bootstrap support; Appendix 5 Figure 15.1). Similarly, the two species were clearly differentiated in the phylogenetic tree based on the COI sequences (99% bootstrap support; Appendix 5 Figure 15.2). The concatenated dataset produced a similar topology (Figure 7.6) with high bootstrap support values for all clades. The results of our

genetic distance calculations showed that *C. diplogramma* and *C. micropeltes* showed the highest intra-specific genetic distance (2.4-3.0% for 16S rRNA and 21% for COI; Appendix 5 Table S3 and S4), yielding support that *C. diplogramma* is a separate species concordant with the morphometric analysis.

7.4.3. Divergence time estimates

The divergence time for *C. diplogramma* and *C. micropeltes* was calculated as 7.77 MYA using fossil calibration, and 17.68 MYA with the alternate calibration in the LRMD method (assumes local rates for every internal node and it is used when the sequence dataset is assumed to be not clock-like). The mean divergence time values for the node E that correspond to the split between *C. marulius* from North East India and South India (see Figure 7.6) was 6.56 and 15.00 MYA with the two different calibrations, which are very high divergence values for individuals from the same species. The high genetic divergence and divergence time estimates between *C. marulius* from geographically isolated locations points towards the presence of further cryptic species within the genus Channa that should be investigated using comprehensive sampling and detailed taxonomic and genetic analyses. The results of the tree calibrations (Figure 7.6) are presented in Tables 7.4 and 7.5.

7.5. Discussion

After Francis Day's (1865) [288] initial description of *C.diplogramma* he himself synonymised the species with *C. micropeltes* in 1878 [325]. Since then, there have been no collections of *C.diplogramma* for detailed taxonomic investigations, and all subsequent information in the literature [321, 326, 327, 328, 330, 334, 331] was based on Day's (1878) synonymy [347] (Table 7.1). The highly fragmented distribution of *C. micropeltes* and its markedly different adult appearance (with the individuals in peninsular India), based on observation in various public and retail aquariums (Ralf Britz; Rajeev Raghavan Pers. Observation), led us to examine the systematic position of the species in detail.

Color pattern is frequently used as the sole character to distinguish closely related species. This is well justified if it serves as a primary cue in the recognition of con-specifics [352, 353]. However, using coloration as a basis for species identification may turn problematic if color variation is a result of phenotypic plasticity, rather than reproductive isolation [354]. Another concern is that coloration genes [238] may evolve more rapidly [355] than other morphological and genetic characters. Channids are well known for the fact that the colour patterns of their juveniles are very different from that of the adults [356], although the reasons for this difference remain unknown. During the life history of *C.diplogramma*, individuals have multiple color phases. However, it was observed that these individuals, belonging to different life stages, of

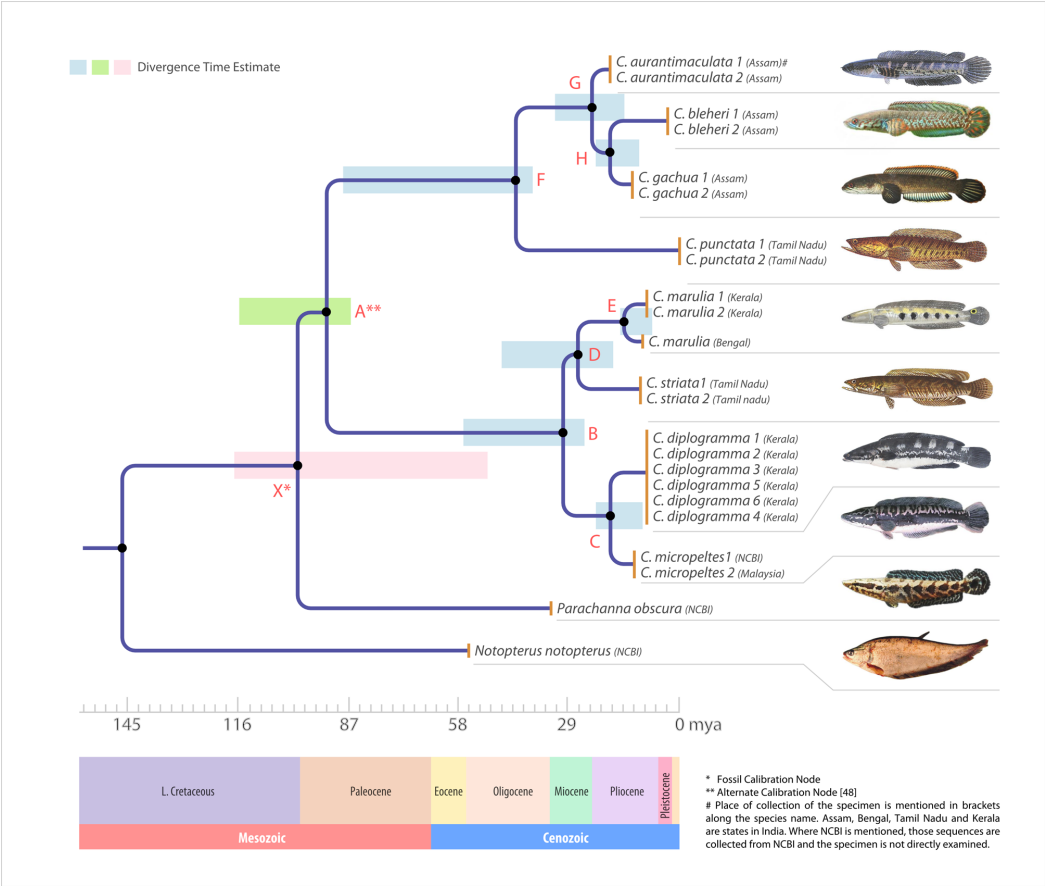


Figure 7.6.: Phylogram showing the relationships of the channids used in this study rooted with *Notopterus notopterus* (AP008925.1). The nodes for which the divergence time is presented in tables 7.4 and 7.5 are labeled as A through H below the branches. The mean time intervals of divergence calculated by the two calibration methods are represented as rectangular bars on the nodes.

C.diplogramma occur sympatrically and utilize the same ecological habitat, unlike the observations from South East Asia, where local knowledge of fishers reveals that the color variation in *C. micropeltes* is linked to the differential habitat occupancy by the individuals [349].

The gular scales [337], a morphological trait that has been hypothesized to be plesiomorphic [349] at the level of the family Channidae has been reported only in four species of channids endemic to South East Asia, *C. bankanensis, C. lucius*, *C. micropeltes*, and *C. pleurophthalma*, apart from the Parachanna of Africa [337, 349]. Our observation of gular scales in *C.diplogramma* makes it the only species of channid from the Indian subcontinent with gular scales, a character shared with its sister species *C. micropeltes* (Appendix 5 Figure 15.3).

The morphometric and meristic analysis of *C. micropeltes* and *C.diplogramma* provided conclusive evidence to separate them as two distinct species. Our analyses indicate that number of caudal fin rays, lateral line scales, scales below lateral line; total vertebrae, pre-anal length and body depth were the most prominent characters that can be used to differentiate both the species.

A high genetic differentiation at the intraspecific level was observed for C. marulius (2.1% for the 16S rRNA gene) that included individuals from Bengal, North East India [351] and from Kerala, South India (present study). All the other species showed lower intraspecific genetic differentiation values. The genetic distance between *C. micropeltes* (sequence [350] and *C. micropeltes* present study) and *C.diplogramma* from South India (present study) was 2.7-3.0% (for 16S rRNA gene sequence comparison), and the genetic distance for the COI gene sequences were 21% between these species - which was well above the average observed for any other intraspecific genetic distances (Appendix 5 Table S3 and S4). This indicates that *C. micropeltes* and *C.diplogramma* cannot be considered conspecific, and results of both morphological and genetic analyses clearly support the existence of two distinct species

Recent studies have estimated the molecular divergence time dates for channids. Some researchers [346] have favored the hypothesis that a vicariant divergence of channids occurred during the Gondwanaland split based on a divergence time calibration using reliable biogeographic scenarios and fossil records. By contrast, others [323] favored the “out of Asia into Africa” hypothesis when calibrating the tree solely based on fossil records. In this study, we calibrated the phylogenetic tree with two alternative constraints, one based on the oldest known fossil of channids and the other based on the available molecular divergence time estimate for the emergence of the genus Channa. Due to the incomplete nature of the fossil record, fossil calibrations can only provide minimum ages and therefore, will tend to underestimate lineage divergence times [357]. To reduce such bias we calibrated the tree a second time with a previously calculated value of 110 - 84 MYA for the mean divergence time of the emergence of the genus Channa [346]. This divergence time value was attained based on the continental breakup of African and South American landmasses (100–120 MYA) and the estimated divergence time between sarcopterygians and actinopterygians (420–500 MYA), which has been

successfully used previously to date old divergence times in actinopterygian fishes [358, 359]. Moreover, the recent identification of channid fossils from Africa in the middle Eocene [360], further supports the use of this additional time constraint, and highlights the incomplete nature of the fossil records.

The fossil records (including the oldest known channid fossil) from Northwest Pakistan had faunal affinities towards both Asia and Africa [345], which could be due to the contact, of the drifting Indian subcontinent, with Africa, during its northward movement allowing the dispersion of African fauna into Asia [361]. Thus, assigning a center for the origin of channids in the Indian subcontinent could be erroneous. We therefore speculate a vicariant divergence of Parachanna and Channa genera during the Gondwana land breakup, with the genus Channa dispersing into Eurasia. It is likely that fishes of the genus Channa could have been widely distributed from South East Asia to the Indian Subcontinent (or vice versa) during the multiple contacts of the two land masses [362, 363, 364] during the drift to the present positions.

Our average divergence time estimates between *C.diplogramma* and C. micropleltes were from 9.52 (with fossil data) to 21.76 MYA (with the alternative calibration). According to the Satpura Hypothesis [365], the westward migration of Malayan fishes deflected southwards in the late Miocene (10-15 MYA) due to the formation of a ridge in the North (the Nepal Ridge) of the Himalayas. Thus, our lower values attained by fossil calibration for the split of *C.diplogramma* and *C. micropeltes* are in concordance with this time frame of migration of fishes from Malaya. However, this may only hold good for torrential freshwater fishes, and the dispersion of channids through this route could be difficult to explain. The mean upper value of 21.76 MYA (early Miocene) makes it highly improbable for this species to have dispersed towards India from South East Asia, or having originated in Northwest India, due to the absence of any geographic connections towards Southern India during this time frame. Another scenario is the dispersal of the most recent common ancestor of these two species from Southern India through North East India to South East Asia, in a reverse direction. However, this scenario can be ruled out due to the above said reasons. Thus, the Satpura hypothesis or the origin of the most recent common ancestor of *C.diplogramma* and *C. micropletes* in Northwestern India cannot conclusively explain the presence of *C.diplogramma* in peninsular India.

Hence, the most plausible scenario for the evolution of channids would be a vicariant divergence after the Gondwanaland split-up, of the genus Parachanna into Africa and the genus Channa into Eurasia. The presence of *C.diplogramma* in South India, also point towards a scenario of the vicariant divergence of the most recent common ancestor of *C. micropeltes* and *C.diplogramma* during the drift of the South East Asian and Indian sub-continental land masses towards its present positions [362, 363, 364].

Our study clearly support the recognition of *C.diplogramma* as an endemic species of peninsular India, subsequently justifying its high conservation value due to its restricted distribution. Like all channids, *C.diplogramma* is a ‘K selected’ species with a slow growth, long time to

reproduce and longer life, which makes them highly vulnerable to overexploitation. Channa diplogramma is a connoisseurs' delight in Central Kerala and locals pay premium prices for sub adult and adult specimens. Local fishers operating in the rivers and reservoirs where this species is known to occur have confirmed its rarity and that populations have declined considerably (> 90%) over the last two decades.

In addition to the indiscriminate exploitation by local fishers, *C.diplogramma* is also severely threatened by the loss of critical riverine habitats due to sand mining and reclamation of riverine areas for the construction boom in Kerala, as well the increasing pollution in existing habitats due to domestic and industrial sewage.

The key to effectively preserving the remaining populations of *C.diplogramma* will therefore need to consider: (i) habitat protection, (ii) fishery management plans (regulation of total allowable catch, restrictions on mesh sizes and closed seasons), and (iii) the development of a captive breeding technology for facilitating large scale ranching and stock enhancement in the rivers and reservoirs where the species occur.

The International Union for Conservation of Nature (IUCN) has recently completed a comprehensive assessment of freshwater biodiversity in the Western Ghats Hotspot. However, the Western Ghats species list does not include *C.diplogramma* as it is still considered to be a synonym of *C. micropeltes* in the Catalog of Fishes [348], the database from which the species list were compiled. The experts at the IUCN Workshop including two of the authors of this paper have however suggested that the "Indian race" of *C. micropeltes* should be considered as distinct and its conservation status categorized as 'Vulnerable'.

7.5.1. Conclusion

The species status of *C.diplogramma* as an endemic species of peninsular India has been confirmed through both morphological and molecular analyses after a period of 146 years since its initial description, and 134 years after it was synonymised. Our results suggest that this species shared a most recent common ancestor with *C. micropeltes*, around 9.52 to 21.76 MYA. An effective conservation effort specifically targeted for this enigmatic and economically important species is highly recommended to avoid endangerment and possible extinction in its restricted range. Also, there is a need for carrying out comprehensive taxonomic and genetic profiling of the Snakeheads in tropical Asia to identify its population structure, and also to evaluate the likelihood of additional species. This is of utmost importance as the Snakeheads are widely exploited as food and ornamental fishes, and their conservation and management is a priority in many Asian countries where their populations are declining.

Contribution

I participated in designing the work and participated in part of the sampling. Carried out all the (in silico) phylogenetic and divergence time analysis, and jointly drafted the paper with AA, ND, RB and RR. Note that the sequences for Indian channids used in this study was taken from NCBI (generated by the first author) while the sequence for *Channa micropeltes* was generated specifically for this paper.

7.6. Tables

Table 7.2.: Morphometric characters of *Channa diplogramma* and *C. micropeltes*

	Channa diplogramma		*Channa micropeltes*	
	Range	Mean (sd)	Range	Mean (sd)
Total length (mm)	107.24 (589.19)	312.45 (184.96)	338.93-654.93	502.30 (128.83)
Standard length (mm)	85.40 (479.15)	251.65 (151.66)	290.87-564.22	415.14 (120.11)
%SL				
Head Length (mm)	25.03 (35.37)	32.12 (2.82)	32.23-39.39	35.28 (2.64)
Pre dorsal length (mm)	31.47 (38.75)	35.04 (2.53)	30.50-37.57	33.25 (2.63)
Pre pectoral length (mm)	30.98 (38.77)	34.73 (3.26)	31.54-38.66	34.03 (3.06)
Pre pelvic length (mm)	31.88 (42.16)	36.93 (3.41)	34.28-41.97	37.01 (2.91)
Pre anal length (mm)	49.86 (60.25)	55.66 (3.42)	46.68-57.08	50.64 (3.88)[a]
Body depth (mm)	14.16 (25.61)	19.48 (3.92)	22.54-26.58	24.35 (1.68)[a]
% TL				
Standard length (mm)	77.14 (81.91)	79.95 (1.48)	72.42-86.15	82.29 (5.68)
Head Length (mm)	20.36 (27.34)	25.66 (1.99)	27.76-30.21	28.93 (0.93)[b]
Pre dorsal length (mm)	25.59 (30.86)	27.98 (1.61)	26.28-28.03	27.25 (0.64)
Pre pectoral length (mm)	25.13 (30.87)	27.73 (2.23)	26.66-30.52	27.89 (1.55)
Pre pelvic length (mm)	25.93 (32.52)	29.49 (2.28)	28.38-31.29	30.34 (1.17)
Pre anal length (mm)	40.55 (47.98)	44.47 (2.23)	40.06-43.64	41.51 (1.31)[a]
Body depth (mm)	11.27 (20.76)	15.60 (3.29)	17.62-22.40	20.05 (2.05)[a]

[a] P <0.05

[b] P <0.01

Table 7.3.: Meristic characters of *Channa diplogramma* and *C. micropeltes*.

	Channa diplogramma		*Channa micropeltes*	
	Range	Mean (sd)	Range	Mean (sd)
Dorsal fin rays	43-44	43.20 (0.42)	43-44	43.40 (0.55)
Pectoral fin rays	17	17.00 (0.00)	16-17	16.60 (0.55)
Pelvic fin rays	6	6.00 (0.00)	6	6.00 (0.00)
Anal fin rays	26-28	27.50 (0.71)	27-29	28.00 (0.71)
Caudal fin rays	15-17	15.30 (0.67)	14	14.00 (0.00)[a, b]
Lateral line scales	103-105	104.20 (0.79)	86	86.00 (0.00)[a, b]
Cheek scales	16-20	17.80 (1.55)	23-25	24.20 (0.84) [a]
Gular scales	30-31	30.60 (0.52)	18-39	30.60 (10.26)
Total vertebrae	53-54	53.60 (0.52)	57	57.00 (0.00) [a, b]

[a] P <0.0001

[b] one sample t-test

Table 7.4.: Results of divergence time estimation in million years for the various nodes of the phylogenetic tree presented in Figure 7.6 the calibration point at node X was the earliest channid fossil age from Eocene (50 MYA; [349]

Node	LRMD	GRMD	NPRS	NPRS-Log	Mean divergence time
X*	50	50	50	50	50
A	40.49	41.72	43.35	40.25	41.425
B	24.09	24.04	28.27	21.56	24.49
C	7.77	8.5	13.9	7.914	9.52
D	19.19	19.74	24.14	17.67	20.185
E	5.317	6.349	9.26	5.301	6.556
F	38.1	37.63	40.27	36.32	38.08
G	10.65	13.64	22.42	11.13	14.46
H	5.633	8.866	16.67	6.781	9.4875

* Calibration Node

Table 7.5.: Results of divergence time estimation in million years for the various nodes of the phylogenetic tree presented in Figure 7.6 the calibration point at node one was the split between Parachanna and Channa calculated by Li et al., (110-84 MYA)[346]

Node	LRMD	GRMD	NPRS	NPRS-Log	Mean divergence time
X	120	116.2	115.2	120.6	118
A*	110-84	110-84	110-84	110-84	110-84
B	56.29	55.86	62.19	52.20	56.64
C	17.68	19.76	30.34	19.25	21.76
D	44.44	45.88	52.96	42.78	46.52
E	12.25	14.76	20.17	12.84	15.00
F	91.01	87.44	89.79	87.63	88.97
G	24.68	31.72	49.66	26.82	32.22
H	12.85	20.61	36.87	16.33	21.67

* Calibration Node

III.

Discussion and Conclusion

8 Discussion

In this thesis, divided into two parts, I present molecular evidences related with the adaptive evolution of vertebrates. In the first part, molecular genetic studies highlight the adaptive evolution of important vertebrate genes. In the second part, evidence is provided on how adaptive evolution may influence teleost radiations, namely with studies of the lateral line genomic diversification and evolution in fishes as well as with detailed case studies of two teleost radiations from the western ghats biodiversity hotspot.

In part one, of the thesis, we used different methods to study adaptive evolution both at the gene (codon) level and protein level. These methods revealed the prevalence of positive selection in the studied genes, which would not have been evident by using simple analytical methods. For example, a previous study of the rhodopsin 1 (RH1) in teleosts [147] had questioned the validity of positive selection methods after finding that positively selected sites are not the ones causing measurable functional differences in the protein. However, we uncover that the use of appropriate state-of-the-art methods could be insightful to identify the functionally relevant sites.

In short, this thesis studies the prevalence of adaptive evolution and other evolutionary mechanisms using a comprehensive array of cutting edge evolutionary methods, which includes studies to uncover positive selection at the gene and protein-level, functional divergence, rate shifts, evolutionary rate variations and also genomic synteny analysis to assess gene evolution.

After exploring adaptive evolution of genes and gene families, it was a natural extension to assess the prevalence of natural selection in species radiations and to relate the evolutionary mechanisms (at gene/protein level) to traits (phenotypes) that can be relevant for species ra-

diations. As mentioned in the introductory chapter, genes are the basic unit of evolution and any changes on them could have an implication at the species/organism level or on their phenotypes. Here we studied adaptive evolution of genes in chapters two - five and correlated adaptive evolution of genes to species radiation (or phenotypic innovations) in chapters three, four and five. Different methods were used to evaluate adaptive evolution on genes important in the evolution of the organisms.

The teleost RH1 had a major importance in the visual specialization to varied habitats. We also studied the avian SOD genes, which might have played an indirect but major influence on the successful radiation of birds. Adaptive evolution on genes important to the evolution of the lateral line, which has been a major innovation in the adaptive radiations of fishes has also been studied. In the final chapters we studied the molecular evidence of teleost radiations, which allowed uncovering cryptic species and resolve taxonomic ambiguities. These radiations are a consequence of a multitude of changes at the gene level and phenotype level as referred in the previous chapters.

This thesis can be divided into four general topics:

- ✓ One relates to the adaptive evolution of genes
- ✓ The second pertains to the gene and genome duplications and the influence of Darwinian selection on and after such events
- ✓ The third deals with evolutionary rate variations and functional divergence of duplicated genes at the level of both paralogs and orthologs
  and
- ✓ The fourth relates to unraveling key evolutionary process by highlighting speciation (or radiation).

Part One is divided into three chapters studying three important gene families in vertebrates from a comparative genomic perspective: the ***RXR***, ***RH1*** and ***SOD***. All three cases are gene families with multiple rounds of duplications leading to more than one copy of genes in vertebrates.

We see that there is an asymmetric distribution of paralogs of RXR and rhodopsin 1 in teleosts, however, it should be noted that in tetrapods there is no major variation in the number of paralogs for all the three genes studied in part one. This can be easily understood by the fact that there was an additional round of genome duplication [239, 240, 10, 9] at the base of teleosts (FSGD), however, what we do not fully understand is that *why there are different number of paralogs in different teleosts*.

In **chapter two**, we looked at the problem of asymmetric distribution of paralogs in teleosts by taking a classic example of the RXR family, which shed light on how the genes are retained

post-duplication. This same topic (asymmetric distribution of paralogs) is revisited in part two (**chapter five**) and our results suggest that positive selection and subsequent functional divergence immediately after duplication events were the main reasons for retention of multiple copies of genes following genome duplication.

Positive selection (adaptive evolution) of genes can also prove adaptive benefits to species, discussed in detail in chapters three and four. In **chapter three** the role of adaptive evolutionary mechanisms on the RH1 genes of teleosts is analyzed. While *RH1* has an asymmetric distribution of paralogs in teleosts it is restricted to only some species (*Anguilla japonica, Conger myaster, Lepidopus fitchii*). However, teleost have radiated into a variety of habitats and some of them into very different low-light environments. Evidence of positive selection acting on species distributed in different dim-light environments like the caves, deep-sea environment, etc., is presented.

In **chapter four** we provided evidence for positive selection in the avian SOD genes, which faces a different physiological pressure relative to other tetrapods. Evolutionary rate variation in a gene or many genes related to a specific function could be a signature of the evolutionary process. We provided evidence that a group of genes, important for a specific function in a clade, faces distinct evolutionary pressures such as evolutionary rate variation and positive selection, when compared to the orthologs from other clades lacking that function. This is exemplified by our studies on genes related to flight as a trait exclusive to birds or genes related to lateral line as a trait exclusive to fishes.

In **chapter five** we showed that the genes involved in the lateral line system development of teleosts have a different evolutionary rate when compared to its one-to-one orthologs in other tetrapods. Episodic events of positive selection and evolutionary rate variation leading to functional divergence of paralogs are shown to be the major trend in all retained paralogs. On the other hand lateral line system related genes are preferentially retained post-duplication when compared to the total amount of duplicate retention in teleosts (see chapter five for details). Thus, higher duplicate retention has been a major reason for phenotypic innovation, which in fact is due to functional divergence because of episodic events of positive selection and beneficial mutations. The final chapters highlighted examples of endangered endemic teleost radiations, which are a result of gene and genome level evolutionary innovations, and our results are crucial in designing proper conservation and management plans for these species.

8.0.1. Role of positive darwinian selection on adaptation related genes: insights into species radiations and adaptive benefits

As discussed throughout we can use different models and methods to find the prevalence of positive selection. The codon models [51, 188] employed to find the non-synonymous substitutions and synonymous substitutions [366] are employed at site (across alignment) [157, 158], branch (among branches) and branch site (across branches and sites on branches) [57, 58]

levels.

In this thesis all these methods were employed, we checked for the presence of positive selection on the ancestral branches of RXR paralogs and different lateral line related genes. Similarly branch-site models were also employed to find the prevalence of positive selection on branches and sites on the ancestral branches of different genes in chapter two, three and five. Site models were employed in chapters three and four to find positively sites which could be involved in functional differentiation of the proteins.

Identifying the positive selected sites can shed light into the functional diversification and its adaptive benefits. As shown in chapter two where sites positively selected are functionally important (but not on conserved regions, which is intuitive) for e.g., positive selection is found on helix 9 of the ancestral branch leading to RXRG which is the helix important for dimerization (function) but there is no positive selection on DNA binding domain, which is highly conserved. In chapter three we saw many sites involved in spectral tuning of the RH1 pigment as positively selected these were fast evolving sites and the ones that have tolerated changes throughout the evolutionary history. Some spectral tuning sites are not positively selected which, although are fast evolving sites, are the ones that have tolerated lesser changes (relatively more conserved so not detected to be positively selected) at that site. In addition many sites implicated in human diseases, like autosomal dominant retinitis pigmentosa, are positively selected in fishes.

It is also notable that the best method to do an adaptive evolution study varies with each and every dataset. Our study with RXR if done on a *site-wise* scale would have provided insights with direct biochemical/physiological consequences (site models). On the other hand when we looked on the selection pattern on the ancestral branches (branch-site model) we could unravel an altogether different perspective where we detail the mechanism of preservation of duplicated copies of genes in addition to the biochemical consequences of mutations.

It does not mean that positive selection identification methods are without caveats. There are different studies detailing the caveats of the methods (for e.g., [232]) and many studies which rectify them (for e.g.,[57]). Our choice to overcome any possible hidden caveat of the positive selection method was to employ complementary analyzes at nucleotide (PHAST), codon (PAML, HYPHY) and protein (TreeSaap, DIVERGE, RASERv2, CONTEST, MAPP, ProPhyler etc.) level [367], so that we produced similar results from different methods adding confidence to our explanations. Indeed, employing positive selection analysis with site-wise models on all the paralogs together as a single alignment (RXRA, RXRB and RXRG) would have likely given inflated ω values. Thus, use of complementary methods and choosing the right models and methods while doing dN/dS analyzes is crucial for the accuracy of comparative evolutionary genetic studies.

8.0.2. Accelerated evolutionary rates and insights into evolutionary innovations

When a gene duplicates during a genome duplication event two copies of the same gene are formed which face a challenge of being relevant to the same organism. On the other hand while a species diversifies into other species the same set of genes in the different species face different challenges from the habit and habitat of the species. There are studies which show that duplicated genes can be beneficial to the diversification of the organisms, essentially meaning that the genes allow the organism to adopt a new mode of life or use a different food substance for which the additional gene copy can specialize. The genes accumulate mutations that have apparently no adaptive benefit until the species is presented with a situation where the gene (mutation) becomes beneficial (for e.g.,[368]).

Thus, evolution (specialization) occurs silently, which may not always be evident for "naked eyes" or even positive selection methods. A more powerful approach is to check for the evolutionary rate variation at nucleotide level - note that dN/dS is a measure of evolutionary rate variation. However, more powerful base-by-base approaches are being introduced, which captures acceleration or deceleration at each nucleotide sites [162] or amino-acid sites [109, 108].

In chapter three, we checked if the positive selection methods could find all the sites involved in spectral tuning of the pigment [147], which was not possible with old less powerful methods and was a major criticism against positive selection methods. By using new and more powerful methods we could identify four of the twelve sites. However, the use of conservation acceleration metrics like MAPP scores, or NNEUT scores (phyloP) could detect that all those spectral tuning sites were evolving non-neutrally but were more conserved than the positively selected sites.

The avian SODs were found to have an accelerated rate of evolution when compared to either reptiles or mammals. Superoxide dismutases are involved in detoxification of ROS and birds have a higher chance of ROS production due to flight as their mode of locomotion. However, birds have devised a method for lower mitochondrial ROS leakage thus having a lower ROS production when compared to other vertebrates. Thus, birds produce low amounts of ROS normally, while during flight they are presented with a larger amount of ROS, which is a different evolutionary pressure when compared to other vertebrates. Thus, our results show that SOD's in birds have an altered evolutionary rate points to a scenario where evolutionary process on the gene has helped in fine tuning the birds' mode of locomotion and subsequent diversification.

We compared the evolutionary rates of the genes in the teleost clade with the tetrapod clade in chapter five, the results show that teleosts have an accelerated evolutionary rate for the lateral line development genes which are crucial for their diversification. As seen earlier evolutionary innovation could occur as episodes as seen in (chapter two) the case of RXR. Positive selection occurs in episodes very frequently, in chapter five we see that positive selection methods find no positive selection on the extant paralogs using branch site models, however finds positive selection on the ancestral branch leading to one of the paralog in most of the cases.

That situation could be discerned because we specifically tested both scenarios. However, this does not mean that there is no evolutionary rate variation on the extant branches when compared to the tetrapods, as shown by the base-by-base conservation acceleration analysis, but positive selection is episodic.

8.0.3. Role of positive darwinian selection and functional divergence on paralogs post genome duplications

When a gene duplicates (e.g., as the result of a whole genome duplication event) immediately there are two copies of the gene to carry out a same function. This presents a serious problem for the organism, where imbalance of the dose of a protein or product occurs. Thus, how duplicated genes retain both their copies (paralogs) in the organism is a major thrust area of evolutionary explorations. We have done detailed investigations into this matter in two different chapters (two and five). We used methods to identify functional divergence, instances of positive selection and evaluate syntenic patterns to gain insights into the duplicate preservation mechanisms.

We employed functional divergence analysis to identify clade specific functional divergence (using DIVERGE [64]) or branch specific (on ancestral paralogous branch) functional divergence (in RASERv2 [66]). Both these methods are powerful and point towards the protein's functional divergence based on amino acid changes and change patterns.

Similarly we found instances of positive selection on the ancestral paralogs. However, when we choose a paralog clade (as a whole/all tips in a paralogous clade) and compare it with its sister paralog, we find that the evidence of positive selection is elusive as seen in the chapter five. This could be a limitation of the method employed for finding positive selection, thus our choice of protein sequence based functional divergence studies was a complimentary to what we found using the codon based positive selection methods. On the other hand this could also point to a case where positive selection occurs immediately after the duplication event and in the extant branches there is relaxed purifying selection, which would also present evidence of functional divergence. We tested this scenario by evaluating positive selection in the branches immediately after duplication and found positive selection in one (most cases) or both of them.

In the case of RXR (chapter two) we present evidence of expression shuffling. This is a major consequence of functional divergence causing amino acid changes and the positively selected mutations. Functional shuffling also intuitively explains how two copies of a gene can co-exist without posing a risk for the organism or invalidating the need of a sister paralog. We present evidences of sub-functionalization and neo-functionalization (see Figure 2.5).

Thus, we present three kinds of evidences that could be useful to understand the mechanism of preservation of duplicated paralogs in an organism. In addition, we also use synteny plots, to check for the gene of interest and the neighboring genes to check for patterns of how the present distribution of genes have occurred. We find cases of gene loss and gene retention in

teleosts when RXR's are studied. Similar patterns were seen when we used synteny analyzes to check for duplicated genes involved in the lateral line system evolution. The normal duplicate retention rate post FSGD is 10-20% we find a higher (50%) duplicate retention for genes related to the lateral line system development (see chapter five). From our results, we could argue that gene (duplicate) retention is favored by positive selection and functional divergence, which is necessitated by the need for the gene to be specialized for the trait.

Three category 1 models have been postulated describing the evolution of duplicated genes [30], including the neofunctionalization model [27], the duplication-degeneration-complementation (DDC) hypothesis [29] and the specialization model or escape from adaptive conflict (EAC) model [28]. The general feature of these models is that a fate-determination phase occurs rapidly after duplication [30], followed by a preservation phase that precludes the pseudogenisation of one of the copies [247, 30].

Ohno's model [27] suggests neofunctionalization of one of the copies, and that the molecular evolution in the duplicated copy is accelerated, and the other two models [28, 29] propose subfunctionalization in both copies. The DDC model [29] assumes that the ancestral function of the gene will be shared between the two post-duplication daughter genes, and that degenerating neutral mutations that accumulate in the paralogs result in subfunctionalization with neither copy being able to carry out the original functions and thus promoting preservation of both copies. The EAC model [28] predicts that if the parent gene were performing two functions that could not be independently improved, then after duplication each gene copy could be driven by positive selection to become more specialized.

Our results from the study of RXR or the different duplicated genes involved in the lateral line system evolution points to a scenario where none of the above stated model is a perfect fit. In fact the most plausible argument that we could make is that when a duplicated copy originated (as a result of duplication) a likely scenario is that one of the paralog is released from the functional constraints of the ancestral gene due to accelerated evolutionary rate and rate shift (functional divergence/positive selection) and can become more specialized in lateral line development or specialize its expression in one organ (in the case of RXR) thereby facilitating evolution of the trait and retention of the paralog.

8.0.4. Species radiations

In the second part of this thesis a lot of emphasis has been placed on species radiations. While genes are the basis of any heritable phenotypic changes, those new phenotypes or traits are the ones directly responsible for the ecological success of the species. Most of the taxonomic literature looks at traits and their differences to classify organisms, however, it is an increasingly difficult task when there are cryptic coloration or traits among different species that have a recent divergence. Thus, species radiation studies could aid us in understanding the adaptive mechanism that formed a key trait in addition to the taxonomic quest behind it. On the

other hand looking at the adaptive evolution of genes involved in the phenotypic diversification of a key trait could provide new insights about species radiations.

In chapter five we look at a main reason for teleost radiations - the lateral line, which facilitates a sense of “touch at a distance” for fishes [191, 193] and is important in processes such as rheotaxis [194], schooling [195], courtship and sexual behavior [196, 197], feeding and prey detection [198, 199, 200] and navigation [201]. We check for patterns of adaptive evolution in the genes that have contributed to the lateral line evolution/development, which in turn facilitated the species radiations.

In the final chapters we provide results for two instance of teleost radiations. Our choice of coalescent based methods ahead of the DNA barcoding based distance metrics has been crucial in identifying cryptics in the case of the endangered red lined torpedo barbs, while the use of divergence time methods have helped in understanding the phylogeography of the cyprinids or the malabar snakehead in both the last chapters. The studies in the last chapters also point to the role of vicariance as a major method by which speciation occurs, which could be evident only with wide range of samples and choice of phylogenetic and divergence time analysis, because vicariance normally result in cryptic species when the evolutionary history (divergence time) is short and the morphological differences become evident only with larger divergence periods. This is also evident by our choice of multivariate methods for analyzing the morphological measurements (in chapter six), which signals that the changes between different populations of red lined torpedo barbs are not obvious by univariate methods and thus for a layman. Similarly the snakehead fishes have been confused due to the similar colouration of the adults, this was solved by detailed morphometric studies and by the use of molecular methods. This again highlights that the final interpretation of a data depends on the choice of the methods and the models. The second part of the thesis provides a holistic view of species radiations, it picks genes responsible for a trait, looks at how it evolve correlate the pattern to species radiations and finally presents examples of species radiations.

Short Synthesis

The result from this thesis show that the evolutionary rate variation, positive selection and functional divergence have been crucial for vertebrate evolution by facilitating either duplicate retention or phenotypic diversification. We capture the signal of evolutionary rate variation using positive selection methods (on ancestral branches), as well as functional divergence (between paralogous clades/genes) based methods or the PHAST CONACC (or NNEUT) and MAPP analysis methods.

The results from the evolutionary genetic explorations done in this thesis can be extrapolated into different branches of biology. The patterns of positive selection on the lateral line system genes could have important revelations on the different inner-ear phenotypes of higher vertebrates especially humans. The study of SODs in birds should open up studies at physi-

ological level (exercise tolerance, detoxification and aging research) and yield support to the avian species radiations, ecological success and biodiversity. Similarly, the RXR and the RH1 evolution studies contribute to the understanding of species' ecological success and genome evolution. The study of species radiations aids in the cataloguing of biodiversity and conservation. The results could have impacts on the fields of evolutionary medicine [77, 78], biodiversity and conservation [79] and in different aspects related to human welfare [76].

9
Conclusion

In this thesis, the main chapters are based on a simple fact that *the genes are the basis of phenotypic variations in organisms*. After detailing the basics of evolutionary explorations and a brief methodology introduction (in chapter one), six chapters are presented with detailed explorations into specific questions regarding the evolution of vertebrates.

We used different methods to check for positive selection at the codon and protein level, as well as dN/dS based and nucleotide sites-wise, evolutionary rate analysis to characterize genes important in vertebrate development and adaptation (chapter two to five). We used methodologies for checking functional divergence and made extensive use of synteny plots (chapter two and five) to check for the evolutionary history of the genes that we chose. In addition we used recently developed species delimitation methods (chapter six) and divergence time analysis (chapters six and seven) to study cryptic species.

Overall we studied (selected) genes important for the development and adaptation of vertebrate species, which may be important to understand distribution or their biodiversity. We also leverage the power of molecular data and cutting edge analytical methods to uncover cryptic species. In short, we obtained important insights to the evidence of *molecular evolution* and its relation to *"how species emerge and evolve"* and further catalogued evolution in action in the formation of cryptic species.

The first part of the thesis deals with adaptive evolution of genes involved in vertebrate development and adaptation. Retinoid X receptors, important for development and detoxification, rhodopsin-1, essential in dim light adaptation in teleosts and Superoxide dismutases critical in detoxification, were studied from a comparative genomic perspective.

The study of RXR revealed that with an increase in organism's complexity, the different types of cells, while requiring the same basic biochemical process, are nevertheless triggered by different agents and/or with different intensities. Presence of assymetric distribution of paralogs allowed us to check for the evolutionary forces at the genomic level, which favored the preservation of the paralogs. We found positive selection and functional divergence (accelerated evolutionary rates), to be the two most important factors along with expression shuffling.

Rhodopsin-1 gene in teleosts is found to be suitable for use as a phylogenetic marker based on the analysis of the basic sequence characteristics. We also uncovered 20% of the protein length to be under positive selection that could be of importance in the adaptation of teleost. We could also find positively selected species (branches), which could be indicative of habitat specialization. In short we evolutionarily characterised the teleost rhodopsin. All sequences available for teleosts (adhering to the quality check we imposed), were analysed. The base composition and codon bias analysis showed an overlap between different superorders, thus making the gene suitable as a phylogenetic marker for “pan-teleost” studies. The positive selection analysis uncovered interesting signals, by identifying positive selection in species adapted to dim light vision and uncovered many new sites under episodic positive selection, which were previously unknown.

The study of superoxide dismutases, using birds as a focal group, showed the presence of positive selection in the avian SOD protein and evolutionary rate analysis comparing birds to mammals and reptiles found accelerated rates of bird SODs. The locomotion of birds and their higher life span could be related to our results of evolutionary tinkering of the avian detoxification genes. The physiological pressure for SODs, catalyses, peroxidases and other detoxifying enzymes in birds are different from those that they face in other vertebrates. During the evolution of the ancestral stock of birds, and the fine-tuning of the flight adaptations, evolutionary modifications of the detoxifying enzymes must have occurred. We present evidence that there have been evolutionary rate acceleration and positive selection during the evolution of birds, which might have helped them fine tune the enzyme (activity) and in adapting to a new mode of locomotion and their subsequent adaptive radiation.

The second part of the thesis deals with the teleost adaptive radiations. In the first chapter in this section (chapter 5), the fish lateral line development related genes are studied. We find compelling evidence related to increased evolutionary rates in fishes and relate it to the, preservation of additional (gene) copies, and the adaptive radiation of teleost fishes. We also find positive selection and altered evolutionary rates (functional divergence) in the ancestral branches leading to the teleost paralogs, which is thought to be the major mechanism that helped the preservation of additional gene copies following the teleost specific genome duplication. We conclude that higher duplicate retention, followed by rate shifts and positive selection, may have contributed to the evolution and remarkable diversity of the teleost lateral line system. We provide initial results confirming the evolutionary innovation in the lateral line

system development genes which could be tested with future studies in a pan-teleost level and with functional and mutational studies.

In the last two chapters we performed taxonomic (molecular genetic) investigations of two teleost cryptic complexes. The red lined torpedo barbs were thought to comprise just 2 species until recently, our study utilizing state-of-the-art species delimitation analysis confirm the presence of cryptic species within this endangered clade. The fact that these fishes are categorized as *endangered* in the IUCN (International Union for Conservation of Nature) Red-list makes our study timely and serves as a basic study which could save the species from extinction. Our utilization of genetic methods and comparative analysis also reveal the utility of genes as a critical tool in saving the biodiversity. In the last chapter of this section (chapter 7) we also re-described a channid from the south Indian Western Ghats, a biodiversity hotspot. In this chapter also we use genetic methods to confirm the presence of a valid species which was misidentified as another. Our study has contributed in the cataloguing of biodiversity and also helped in identifying a rare species which needs to be assessed by the IUCN red listing agencies, which appear to be considerably endangered due to anthropogenic threats.

9.1. Future directions

- ✓ Sequencing RXR paralogs from a species before the fish specific genome duplication and chondrichthyes species could be an interesting further study.
- ✓ The two groups of fishes that we taxonomically characterized could be used to confirm the results of the RXR evolution where cyprinids possessed six paralogs and perciforms possessed four paralogs.
- ✓ Sequencing rhodopsins in the species that we taxonomically characterized; the barbs are benthopelagic, while the channid that we re-described is a benthic species thus inhabiting in different photic environments.
- ✓ Future studies of other avian detoxification genes, like catalyses and glutathione synthases, could be interesting in conjunction with the results from our preliminary study of the avian SOD genes.
- ✓ Studies of the lateral line development genes in teleost species living in varied habitats could be an easy way to highlight the importance of the sense organ for the ``teleost way of life'' and also reveal another major facet of teleost adaptive radiations.

IV.

Bibliography

10
Bibliography

Bibliography

[1] Darwin, C., 1859 *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: Murray.

[2] Nei, M. & Kumar, S., 2000 *Molecular evolution and phylogenetics*. Oxford University Press, USA.

[3] Johannsen, W., 1909 *Elemente der exakten Erblichkeitslehre*. Gustav Fischer, Jena.

[4] Johannsen, W., 1911 The genotype conception of heredity. *The American Naturalist* **45**, pp. 129--159. ISSN 00030147.

[5] Gerstein, M., Bruce, C., Rozowsky, J., Zheng, D., Du, J., Korbel, J., Emanuelsson, O., Zhang, Z., Weissman, S. & Snyder, M., 2007 What is a gene, post-encode? history and updated definition. *Genome research* **17**, 669--681.

[6] Gregory, T., 2005 Genome size evolution in animals. In *The evolution of the genome* (ed. T. Gregory), volume 1, pp. 4--87. Elsevier, San Diego.

[7] Noble, D., 2008 Genes and causation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **366**, 3001--3015. (doi:10.1098/rsta.2008.0086).

[8] Yang, Z., 2006 *Computational molecular evolution*, volume 284. Oxford University Press Oxford.

[9] Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., Yamada, T., Nagayasu, Y., Doi, K., Kasai, Y. *et al.*, 2007 The medaka draft genome and insights into vertebrate genome evolution. *Nature* **447**, 714--719. 10.1038/nature05846.

[10] Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. *et al.*, 2004 Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946--957. 10.1038/nature03025.

[11] Ponting, C. & Lunter, G., 2006 Signatures of adaptive evolution within human non-coding sequence. *Human molecular genetics* **15**, R170--R175.

[12] Bush, E. & Lahn, B., 2008 A genome-wide screen for noncoding elements important in primate evolution. *BMC Evolutionary Biology* **8**, 17.

[13] Kimura, M. *et al.*, 1968 Evolutionary rate at the molecular level. *Nature* **217**, 624.

[14] King, J. & Jukes, T., 1969 Non-darwinian evolution. *Science* **164**, 788--798.

[15] Kumar, S., 2005 Molecular clocks: four decades of evolution. *Nature Reviews Genetics* **6**, 654--662.

[16] Easteal, S., 1992 Problems and paradigms: A mammalian molecular clock? *BioEssays* **14**, 415--419. ISSN 1521-1878. (doi:10.1002/bies.950140613).

[17] Abascal, F., Zardoya, R. & Posada, D., 2005 ProtTest: selection of best-fit models of protein evolution. *Bioinformatics (Oxford, England)* **21**, 2104--2105. ISSN 1367-4803. (doi:10.1093/bioinformatics). PMID: 15647292.

[18] Posada, D. & Crandall, K. A., 1998 MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**, 817--818.

[19] Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O., 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307--321. ISSN 1076-836X. (doi:10.1093/sysbio). PMID: 20525638.

[20] Ronquist, F. & Huelsenbeck, J. P., 2003 MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572--1574. 10.1093/bioinformatics/btg180.

[21] Zhang, J., 2006 Parallel adaptive origins of digestive rnases in asian and african leaf monkeys. *Nature genetics* **38**, 819--823.

[22] Schoenebeck, J., Hutchinson, S., Byers, A., Beale, H., Carrington, B., Faden, D., Rimbault, M., Decker, B., Kidd, J., Sood, R. *et al.*, 2012 Variation of bmp3 contributes to dog breed skull diversity. *PLoS Genetics* **8**, e1002849.

[23] Kuwada, Y., 1911 Meiosis in the pollen mother cells of *Zea Mays* L. *Bototanical Magagazine Tokyo* **25**, 163.

[24] Blakeslee, A., 1934 New jimson weeds from old chromosomes. *Journal of Heredity* **25**, 81--108.

[25] Bridges, C. B., 1936 The bar gene a duplication. *Science (New York, NY)* **83**, 210.

[26] Serebrovsky, A., 1938 Genes scute and achaete in *Drosophila melanogaster* and a hypothesis of gene divergency. *Comptes Rendus de l'Académie I' URSS* **19**, 77--81.

[27] Ohno, S., 1970 *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag. ISBN 0045750157.

[28] Hughes, A. L., 1994 The Evolution of Functionally Novel Proteins after Gene Duplication. *Proceedings of the Royal Society B: Biological Sciences* **256**, 119--124. ISSN 0962-8452.

[29] Force, A., Lynch, M., Pickett, F., Amores, A., Yan, Y. & Postlethwait, J., 1999 Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531. ISSN 0016-6731.

[30] Innan, H. & Kondrashov, F., 2010 The evolution of gene duplications: classifying and distinguishing between models. *Nature Reviews Genetics* **11**, 97--108.

[31] Conant, G. & Wagner, A., 2003 Asymmetric sequence divergence of duplicate genes. *Genome research* **13**, 2052--2058.

[32] Kondrashov, F. A., 2010 Gene Dosage and Duplication. In *Evolution after Gene Duplication* (eds. K. Dittmar, D. Liberles & F. A. Kondrashov). Wiley.

[33] Zuckerkandl, E., 1975 The appearance of new structures and functions in proteins during evolution. *Journal of molecular evolution* **7**, 1--57.

[34] Ohta, T., 1990 How gene families evolve. *Theoretical population biology* **37**, 213--219.

[35] Dayhoff, M. O., 1974 Computer analysis of protein sequences. In *Federation proceedings*, volume 33, p. 2314.

[36] O'Brien, K., Remm, M. & Sonnhammer, E., 2005 Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research* **33**, D476--D480.

[37] Benson, D., Boguski, M., Lipman, D. & Ostell, J., 1997 Genbank. *Nucleic acids research* **25**, 1--6.

[38] Altschul, S., Madden, T., Schäffer, A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D., 1997 Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research* **25**, 3389--3402.

[39] Larkin, M., Blackshields, G., Brown, N., Chenna, R., McGettigan, P., McWilliam, H., Valentin, F., Wallace, I., Wilm, A., Lopez, R. *et al.*, 2007 Clustal w and clustal x version 2.0. *Bioinformatics* **23**, 2947--2948.

[40] Edgar, R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792--1797.

[41] Katoh, K., Kuma, K., Toh, H. & Miyata, T., 2005 Mafft version 5: improvement in accuracy of multiple sequence alignment. *Nucleic acids research* **33**, 511--518.

[42] Löytynoja, A. & Goldman, N., 2005 An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 10557--10562.

[43] Notredame, C., Higgins, D., Heringa, J. *et al.*, 2000 T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* **302**, 205--218.

[44] Castresana, J., 2000 Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* **17**, 540--552.

[45] Penn, O., Privman, E., Ashkenazy, H., Landan, G., Graur, D. & Pupko, T., 2010 Guidance: a web server for assessing alignment confidence scores. *Nucleic Acids Research* **38**, W23--W28.

[46] Jordan, G. & Goldman, N., 2012 The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Molecular biology and evolution* **29**, 1125--1139.

[47] Privman, E., Penn, O. & Pupko, T., 2011 Improving the Performance of Positive Selection Inference by Filtering Unreliable Alignment Regions. *Molecular Biology and Evolution* ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev).

[48] Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by dna polymorphism. *Genetics* **123**, 585--595.

[49] McDonald, J., Kreitman, M. *et al.*, 1991 Adaptive protein evolution at the adh locus in drosophila. *Nature* **351**, 652--654.

[50] Goldman, N. & Yang, Z., 1994 A codon-based model of nucleotide substitution for protein-coding dna sequences. *Molecular biology and evolution* **11**, 725--736.

[51] Muse, S. V. & Gaut, B. S., 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Molecular Biology and Evolution* **11**, 715--724. ISSN 0737-4038. PMID: 7968485.

[52] Yap, V. B., Lindsay, H., Easteal, S., Huttley, G. *et al.*, 2010 Estimates of the effect of natural selection on protein-coding content. *Molecular biology and evolution* **27**, 726--734.

[53] Massingham, T. & Goldman, N., 2005 Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**, 1753--1762.

[54] Pond, S. L. K., Frost, S. D. W. & Muse, S. V., 2005 HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**, 676 --679. (doi:10.1093/bioinformatics).

[55] Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer applications in the biosciences : CABIOS* **13**, 555--556.

[56] Yang, Z., 2007 PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**, 1586--1591.

[57] Zhang, J., Nielsen, R. & Yang, Z., 2005 Evaluation of an Improved Branch-Site Likelihood Method for Detecting Positive Selection at the Molecular Level. *Mol Biol Evol* **22**, 2472--2479.

[58] Pond, S., Murrell, B., Fourment, M., Frost, S., Delport, W. & Scheffler, K., 2011 A random effects branch-site model for detecting episodic diversifying selection. *Molecular biology and evolution* **28**, 3033--3043.

[59] Murrell, B., Wertheim, J., Moola, S., Weighill, T., Scheffler, K. & Pond, S., 2012 Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* **8**, e1002764.

[60] Creevey, C., McInerney, J. *et al.*, 2003 Crann: detecting adaptive evolution in protein-coding dna sequences. *Bioinformatics* **19**, 1726.

[61] Woolley, S., Johnson, J., Smith, M., Crandall, K. & McClellan, D., 2003 Treesaap: selection on amino acid properties using phylogenetic trees. *Bioinformatics* **19**, 671--672.

[62] Dutheil, J., 2008 Detecting site-specific biochemical constraints through substitution mapping. *Journal of molecular evolution* **67**, 257--265.

[63] Studer, R. & Robinson-Rechavi, M., 2009 Evidence for an episodic model of protein sequence evolution. *Biochemical Society Transactions* **37**, 783.

[64] Gu, X. & Vander Velden, K., 2002 DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. *Bioinformatics* **18**, 500--501.

[65] Gu, X., 1999 Statistical methods for testing functional divergence after gene duplication. *Molecular Biology and Evolution* **16**, 1664--1674.

[66] Penn, O., Stern, A., Rubinstein, N. D., Dutheil, J., Bacharach, E., Galtier, N. & Pupko, T., 2008 Evolutionary Modeling of Rate Shifts Reveals Specificity Determinants in HIV-1 Subtypes. *PLoS Computational Biology* **4**, e1000214. (doi:10.1371/journal.pcbi.1000214).

[67] Li, W., Yang, J. & Gu, X., 2005 Expression divergence between duplicate genes. *TRENDS in Genetics* **21**, 602--607.

[68] Huminiecki, L. & Wolfe, K., 2004 Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome research* **14**, 1870--1879.

[69] Sanderson, M. J., 1997 A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**, 1218. ISSN 0737-4038, 1537-1719.

[70] Yoder, A. & Yang, Z., 2000 Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution* **17**, 1081--1090.

[71] Fujita, M. K., Leaché, A. D., Burbrink, F. T., McGuire, J. A. & Moritz, C., 2012 Coalescent-based species delimitation in an integrative taxonomy. *Trends in Ecology & Evolution* **27**, 480--488. ISSN 0169-5347. (doi:10.1016/j.tree.2012.04.012).

[72] Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D. & Vogler, A. P., 2006 Sequence-Based Species Delimitation for the DNA Taxonomy of Undescribed Insects. *Systematic Biology* **55**, 595--609. ISSN 1063-5157, 1076-836X. (doi:10.1080/10635150600852011).

[73] Yang, Z. & Rannala, B., 2010 Bayesian Species Delimitation Using Multilocus Sequence Data. *Proceedings of the National Academy of Sciences* **107**, 9264--9269. ISSN 0027-8424, 1091-6490. (doi:10.1073/pnas.0913022107).

[74] Monaghan, M. T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D. J. G., Lees, D. C., Ranaivosolo, R., Eggleton, P., Barraclough, T. G. *et al.*, 2009 Accelerated Species Inventory on Madagascar Using Coalescent-Based Models of Species Delineation. *Systematic Biology* **58**, 298--311. ISSN 1063-5157, 1076-836X. (doi:10.1093/sysbio).

[75] Leaché, A. D. & Fujita, M. K., 2010 Bayesian species delimitation in West African forest geckos (Hemidactylus fasciatus). *Proceedings of the Royal Society B: Biological Sciences* **277**, 3071--3077. ISSN 0962-8452, 1471-2954. (doi:10.1098/rspb.2010.0662).

[76] Losos, J. B., Arnold, S. J., Bejerano, G., Brodie, E. D., Hibbett, D., Hoekstra, H. E., Mindell, D. P., Monteiro, A., Moritz, C., Orr, H. A. *et al.*, 2013 Evolutionary biology for the 21st century. *PLoS Biology* **11**, e1001466. (doi:10.1371/journal.pbio.1001466).

[77] Antolin, M. F., Jenkins, K. P., Bergstrom, C. T., Crespi, B. J., De, S., Hancock, A., Hanley, K. A., Meagher, T. R., Moreno-Estrada, A., Nesse, R. M. *et al.*, 2012 Evolution and medicine in undergraduate education: A prescription for all biology students. *Evolution* **66**, 1991–2006. ISSN 1558-5646. (doi:10.1111/j.1558-5646.2011.01552.x).

[78] Nesse, R. M. & Stearns, S. C., 2008 The great opportunity: Evolutionary applications to medicine and public health. *Evolutionary Applications* **1**, 28–48. ISSN 1752-4571. (doi:10.1111/j.1752-4571.2007.00006.x).

[79] Moritz, C., 2002 Strategies to protect biological diversity and the evolutionary processes that sustain it. *Systematic Biology* **51**, 238--254. ISSN 1063-5157, 1076-836X. (doi:10.1080/10635150252899752).

[80] Yokoyama, S., 2000 Molecular evolution of vertebrate visual pigments. *Progress in retinal and eye research* **19**, 385--419. ISSN 1350-9462. PMID: 10785616.

[81] Bellingham, J., Tarttelin, E. E., Foster, R. G. & Wells, D. J., 2003 Structure and evolution of the teleost extraretinal rod-like opsin (errlo) and ocular rod opsin (rho) genes: Is teleost rho a retrogene? *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **297B**, 1–10. ISSN 1552-5015. (doi:10.1002/jez.b.18).

[82] Sugawara, T., Terai, Y., Imai, H., Turner, G. F., Koblmüller, S., Sturmbauer, C., Shichida, Y. & Okada, N., 2005 Parallelism of amino acid changes at the RH1 affecting spectral sensitivity among deep-water cichlids from lakes tanganyika and malawi. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 5448--5453. ISSN 0027-8424, 1091-6490. (doi:10.1073/pnas.0405302102).

[83] Larmuseau, M. H. D., Vanhove, M. P. M., Huyse, T., Volckaert, F. a. M. & Decorte, R., 2011 Signature of selection on the rhodopsin gene in the marine radiation of American sevenspined gobies (Gobiidae, Gobiosomatini). *Journal of Evolutionary Biology* **24**, 1618--1625. ISSN 1420-9101. (doi:10.1111/j.1420-9101.2011.02290.x).

[84] Wickens, A. P., 2001 Ageing and the free radical theory. *Respiration physiology* **128**, 379--391. ISSN 0034-5687. PMID: 11718765.

[85] Larcombe, S. D., Coffey, J. S., Bann, D., Alexander, L. & Arnold, K. E., 2010 Impacts of dietary antioxidants and flight training on post-exercise oxidative damage in adult parrots. *Comparative biochemistry and physiology. Part B, Biochemistry & molecular biology* **155**, 49--53. ISSN 1879-1107. (doi:10.1016/j.cbpb.2009.09.009). PMID: 19800412.

[86] Meyer, A. & Van de Peer, Y., 2005 From 2R to 3R: evidence for a fishspecific genome duplication (FSGD). *BioEssays* **27**, 937--945. ISSN 1521-1878. (doi:10.1002/bies.20293).

[87] Mangelsdorf, D. J. & Evans, R. M., 1995 The RXR heterodimers and orphan receptors. *Cell* **83**, 841--850.

[88] Olefsky, J. M., 2001 Nuclear receptor minireview series. *The Journal of biological chemistry* **276**, 36863--36864.

[89] Bowles, J., Knight, D., Smith, C., Wilhelm, D., Richman, J., Mamiya, S., Yashiro, K., Chawengsaksophak, K., Wilson, M., Rossant, J. *et al.*, 2006 Retinoid signaling determines germ cell fate in mice. *Science Signalling* **312**, 596.

[90] Castro, L. F. C., Lima, D., Machado, A., Melo, C., Hiromori, Y., Nishikawa, J., Nakanishi, T., Reis-Henriques, M. & Santos, M., 2007 Imposex induction is mediated through the Retinoid X Receptor signalling pathway in the neogastropod Nucella lapillus. *Aquatic Toxicology* **85**, 57--66. ISSN 0166-445X. (doi:10.1016/j.aquatox.2007.07.016).

[91] Germain, P., Chambon, P., Eichele, G., Evans, R. M., Lazar, M. A., Leid, M., De Lera, A. R., Lotan, R., Mangelsdorf, D. J. & Gronemeyer, H., 2006 International Union of Pharmacology. LXIII. Retinoid X receptors. *Pharmacological reviews* **58**, 760--772.

[92] Shulman, A. & Mangelsdorf, D., 2005 Retinoid x receptor heterodimers in the metabolic syndrome. *New England Journal of Medicine* **353**, 604--615.

[93] Lima, D., Reis-Henriques, M., Silva, R., Santos, A., Filipe C. Castro, L. & Santos, M., 2011 Tributyltin-induced imposex in marine gastropods involves tissue-specific modulation of the retinoid X receptor. *Aquatic Toxicology* **101**, 221--227. ISSN 0166-445X. (doi:10.1016/j.aquatox.2010.09.022).

[94] Novák, J., Beníšek, M. & Hilscherová, K., 2008 Disruption of retinoid transport, metabolism and signaling by environmental pollutants. *Environment international* **34**, 898--913.

[95] Mangelsdorf, D. J., Borgmeyer, U., Heyman, R. A., Zhou, J. Y., Ong, E. S., Oro, A. E., Kakizuka, A. & Evans, R. M., 1992 Characterization of three RXR genes that mediate the action of 9-cis retinoic acid. *Genes & Development* **6**, 329--344. 10.1101/gad.6.3.329.

[96] Ulven, S. M., Gundersen, T. E., Sakhi, A. K., Glover, J. C. & Blomhoff, R., 2001 Quantitative axial profiles of retinoic acid in the embryonic mouse spinal cord: 9-Cis retinoic acid only detected after all-trans-retinoic acid levels are super-elevated experimentally. *Developmental Dynamics* **222**, 341--353.

[97] Gesto, M., Castro, L., Reis-Henriques, M. & Santos, M., 2011 Tissue-specific distribution patterns of retinoids and didehydroretinoids in rainbow trout *Oncorhynchus mykiss*. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* .

[98] Nowickyj, S. M., Chithalen, J. V., Cameron, D., Tyshenko, M. G., Petkovich, M., Wyatt, G. R., Jones, G. & Walker, V. K., 2008 Locust retinoid X receptors: 9-Cis-retinoic acid in embryos from a primitive insect. *Proceedings of the National Academy of Sciences* **105**, 9540 --9545.

[99] Lampen, A., Meyer, S. & Nau, H., 2001 Phytanic acid and docosahexaenoic acid increase the metabolism of all-trans-retinoic acid and CYP26 gene expression in intestinal cells. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression* **1521**, 97--106.

[100] Urquiza, A. M. d., Liu, S., Sjoberg, M., Zetterstrom, R. H., Griffiths, W., Sjovall, J. & Perlmann, T., 2000 Docosahexaenoic Acid, a Ligand for the Retinoid X Receptor in Mouse Brain. *Science* **290**, 2140--2144.

[101] Bridgham, J. T., Eick, G. N., Larroux, C., Deshpande, K., Harms, M. J., Gauthier, M. E. A., Ortlund, E. A., Degnan, B. M. & Thornton, J. W., 2010 Protein Evolution by Molecular Tinkering: Diversification of the Nuclear Receptor Superfamily from a Ligand-Dependent Ancestor. *PLoS Biol* **8**, e1000497.

[102] Tallafuss, A., Hale, L. A., Yan, Y.-L., Dudley, L., Eisen, J. S. & Postlethwait, J. H., 2006 Characterization of retinoid-X receptor genes rxra, rxrba, rxrbb and rxrg during zebrafish development. *Gene Expression Patterns* **6**, 556--565. ISSN 1567-133X. (doi:10.1016/j.modgep.2005.10.005).

[103] Waxman, J. S. & Yelon, D., 2007 Comparison of the expression patterns of newly identified zebrafish retinoic acid and retinoid X receptors. *Developmental Dynamics* **236**, 587--595. ISSN 1097-0177.

[104] Catchen, J. M., Conery, J. S. & Postlethwait, J. H., 2009 Automated identification of conserved synteny after whole-genome duplication. *Genome research* **19**, 1497--1505.

[105] Bielawski, J. P. & Yang, Z., 2003 Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics* **3**, 201--212. ISSN 1345-711X. PMID: 12836699.

[106] Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V. & Robinson-Rechavi, M., 2008 Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In *DILS: Data Integration in Life Sciences. Lecture Notes in Computer Science*, 5109, pp. 124--131. Springer Berlin.

[107] Bertrand, S., Thisse, B., Tavares, R., Sachs, L., Chaumot, A., Bardet, P.-L., Escrivà, H., Duffraisse, M., Marchand, O., Safi, R. *et al.*, 2007 Unexpected Novel Relational Links Uncovered by Extensive Developmental Profiling of Nuclear Receptor Expression. *PLoS Genetics* **3**, e188.

[108] Binkley, J., Karra, K., Kirby, A., Hosobuchi, M., Stone, E. A. & Sidow, A., 2010 ProPhylER: a curated online resource for protein function and structure based on evolutionary constraint analyses. *Genome research* **20**, 142--154.

[109] Stone, E. A. & Sidow, A., 2005 Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome research* **15**, 978--986.

[110] Conant, G. & Wolfe, K., 2008 Turning a hobby into a job: how duplicated genes find new functions. *Nature Reviews Genetics* **9**, 938--950.

[111] Krasowski, M. D., Ni, A., Hagey, L. R. & Ekins, S., 2011 Evolution of promiscuous nuclear hormone receptors: Lxr, fxr, vdr, pxr, and car. *Molecular and cellular endocrinology* **334**, 39--48.

[112] Zhang, Z., Burch, P. E., Cooney, A. J., Lanz, R. B., Pereira, F. A., Wu, J., Gibbs, R. A., Weinstock, G. & Wheeler, D. A., 2004 Genomic Analysis of the Nuclear Receptor Family: New Insights Into Structure, Regulation, and Evolution From the Rat Genome. *Genome Research* **14**, 580--590.

[113] Bastien, J. & Rochette-Egly, C., 2004 Nuclear retinoid receptors and the transcription of retinoid-target genes. *Gene* **328**, 1--16.

[114] Michalik, L. & Wahli, W., 2007 Guiding ligands to nuclear receptors. *Cell* **129**, 649--651.

[115] Hult, E., Tobe, S. & Chang, B., 2011 Molecular evolution of ultraspiracle protein (usp/rxr) in insects. *PloS one* **6**, e23416.

[116] Yang, Z. & dos Reis, M., 2011 Statistical properties of the branch-site test of positive selection. *Molecular biology and evolution* **28**, 1217--1228.

[117] Kuhl, A. J. & Brouwer, M., 2006 Antiestrogens Inhibit Xenoestrogen-Induced Brain Aromatase Activity but Do Not Prevent Xenoestrogen-Induced Feminization in Japanese Medaka (Oryzias latipes). *Environmental Health Perspectives* **114**, 500--506. ISSN 0091-6765. (doi:10.1289/ehp.8211). PMID: 16581536 PMCID: 1440771.

[118] McAllister, B. G. & Kime, D. E., 2003 Early life exposure to environmental levels of the aromatase inhibitor tributyltin causes masculinisation and irreversible sperm damage in zebrafish (Danio rerio). *Aquatic Toxicology* **65**, 309--316. ISSN 0166-445X. (doi:10.1016/S0166-445X(03)00154-1).

[119] Wang, Y. H., Kwon, G., Li, H. & LeBlanc, G. A., 2011 Tributyltin Synergizes with 20-Hydroxyecdysone to Produce Endocrine Toxicity. *Toxicological Sciences* **123**, 71--79.

[120] Hubbard, T. J. P., Aken, B. L., Beal, K., Ballester, B., Caccamo, M., Chen, Y., Clarke, L., Coates, G., Cunningham, F., Cutts, T. *et al.*, 2007 Ensembl 2007. *Nucleic Acids Research* **35**, D610--D617. ISSN 0305-1048. (doi:10.1093/nar).

[121] Gouy, M., Guindon, S. & Gascuel, O., 2010 SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution* **27**, 221--224.

[122] Schmidt, H., Strimmer, K., Vingron, M. & Von Haeseler, A., 2002 Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**, 502--504.

[123] Xia, X. & Xie, Z., 2001 DAMBE: software package for data analysis in molecular biology and evolution. *Journal of Heredity* **92**, 371--373.

[124] Guindon, S., Delsuc, F., Dufayard, J.-F. & Gascuel, O., 2009 Estimating maximum likelihood phylogenies with PhyML. *Methods in molecular biology (Clifton, N.J.)* **537**, 113--137.

[125] Nylander, J. A. A., 2004. MrAIC.pl. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

[126] Yang, Z., Wong, W. S. W. & Nielsen, R., 2005 Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Molecular Biology and Evolution* **22**, 1107--1118.

[127] Nielsen, R., 2002 Mapping Mutations on Phylogenies. *Systematic Biology* **51**, 729 --739. (doi:10.1080/10635150290102393).

[128] Schrodinger, L. L. C., 2010. The PyMOL Molecular Graphics System, Version 1.3r1.

[129] Fitzgibbon, J., Hope, A., Slobodyanyuk, S. J., Bellingham, J., Bowmaker, J. K. & Hunt, D. M., 1995 The rhodopsin-encoding gene of bony fish lacks introns. *Gene* **164**, 273--277. ISSN 0378-1119. PMID: 7590342.

[130] Yokoyama, S., 1995 Amino acid replacements and wavelength absorption of visual pigments in vertebrates. *Molecular biology and evolution* **12**, 53--61. ISSN 0737-4038. PMID: 7877496.

[131] Larmuseau, M. H. D., Vancampenhout, K., Raeymaekers, J. A. M., Van Houdt, J. K. J. & Volckaert, F. A. M., 2010 Differential modes of selection on the rhodopsin gene in coastal Baltic and North Sea populations of the sand goby, *Pomatoschistus minutus*. *Molecular Ecology* **19**, 2256--2268. ISSN 1365-294X. (doi:10.1111/j.1365-294X.2010.04643.x). PMID: 20444083.

[132] Larmuseau, M. H. D., Huyse, T., Vancampenhout, K., Van Houdt, J. K. J. & Volckaert, F. A. M., 2010 High molecular diversity in the rhodopsin gene in closely related goby fishes: A role for visual pigments in adaptive speciation? *Molecular phylogenetics and evolution* **55**, 689--698. ISSN 1095-9513. (doi:10.1016/j.ympev.2009.10.007). PMID: 19822217.

[133] Bowmaker, J. K., 2008 Evolution of vertebrate visual pigments. *Vision research* **48**, 2022--2041. ISSN 1878-5646. (doi:10.1016/j.visres.2008.03.025). PMID: 18590925.

[134] Sivasundar, A. & Palumbi, S. R., 2010 Parallel amino acid replacements in the rhodopsins of the rockfishes (*Sebastes spp.*) associated with shifts in habitat depth. *Journal of evolutionary biology* **23**, 1159--1169. ISSN 1420-9101. (doi:10.1111/j.1420-9101.2010.01977.x). PMID: 20345807.

[135] Parry, J. L., Carboo, A., Bowmaker, J., Seehausen, O. & Carleton, K., 2004 Spectral sensitivity tuning by differential gene expression in african cichlids. *ARVO Meeting Abstracts* **45**, 3633.

[136] Yokoyama, S. & Takenaka, N., 2004 The molecular basis of adaptive evolution of squirrelfish rhodopsins. *Molecular biology and evolution* **21**, 2071--2078. ISSN 0737-4038. (doi:10.1093/molbev/msh217). PMID: 15269277.

[137] Seehausen, O., Terai, Y., Magalhaes, I. S., Carleton, K. L., Mrosso, H. D. J., Miyagi, R., van der Sluijs, I., Schneider, M. V., Maan, M. E., Tachida, H. *et al.*, 2008 Speciation through sensory drive in cichlid fish. *Nature* **455**, 620--626. ISSN 0028-0836. (doi:10.1038/nature07285).

[138] Mayden, R. L., Tang, K. L., Conway, K. W., Freyhof, J., Chamberlain, S., Haskins, M., Schneider, L., Sudkamp, M., Wood, R. M., Agnew, M. *et al.*, 2007 Phylogenetic relationships of danio within the order cypriniformes: a framework for comparative and evolutionary studies of a model species. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* **308B**, 642–654. ISSN 1552-5015. (doi:10.1002/jez.b.21175).

[139] Chen, W.-J. & Mayden, R. L., 2009 Molecular systematics of the cyprinoidea (teleostei: Cypriniformes), the world's largest clade of freshwater fishes: further evidence from six nuclear genes. *Molecular phylogenetics and evolution* **52**, 544--549. ISSN 1095-9513. (doi:10.1016/j.ympev.2009.01.006). PMID: 19489125.

[140] Fang, F., Norén, M., Liao, T. Y., Källersjö, M. & Kullander, S. O., 2009 Molecular phylogenetic interrelationships of the south asian cyprinid genera danio, devario and microrasbora (teleostei, cyprinidae, danioninae). *Zoologica Scripta* **38**, 237–256. ISSN 1463-6409. (doi:10.1111/j.1463-6409.2008.00373.x).

[141] Chen, W.-J. & Mayden, R. L., 2010 A phylogenomic perspective on the new era of ichthyology. *BioScience* **60**, 421--432. ISSN 0006-3568. (doi:10.1525/bio.2010.60.6.6).

[142] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J., 1990 Basic local alignment search tool. *Journal of molecular biology* **215**, 403--410. ISSN 0022-2836. (doi:10.1016/S0022-2836(05)80360-2). PMID: 2231712.

[143] Burge, C. & Karlin, S., 1997 Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**, 78--94. ISSN 0022-2836. (doi:10.1006/jmbi.1997.0951). PMID: 9149143.

[144] Burge, C. B. & Karlin, S., 1998 Finding the genes in genomic DNA. *Current opinion in structural biology* **8**, 346--354. ISSN 0959-440X. PMID: 9666331.

[145] Marchler-Bauer, A. & Bryant, S. H., 2004 CD-Search: protein domain annotations on the fly. *Nucleic acids research* **32**, W327--331. ISSN 1362-4962. (doi:10.1093/nar/gkh454). PMID: 15215404.

[146] Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R. *et al.*, 2011 CDD: a conserved domain database for the functional annotation of proteins. *Nucleic acids research* **39**, D225--229. ISSN 1362-4962. (doi:10.1093/nar/gkq1189). PMID: 21109532.

[147] Yokoyama, S., Tada, T., Zhang, H. & Britt, L., 2008 Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. *Proceedings of the National Academy of Sciences* **105**, 13480--13485. ISSN 0027-8424, 1091-6490. (doi:10.1073/pnas.0802426105).

[148] Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S., 2011 MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**, 2731--2739. ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev/msr121).

[149] Zwickl, D., 2006 *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion.* Ph.D. thesis, University of Texas at Austin.

[150] Sukumaran, J. & Holder, M. T., 2010 DendroPy: a python library for phylogenetic computing. *Bioinformatics* **26**, 1569--1571. ISSN 1367-4803, 1460-2059. (doi:10.1093/bioinformatics/btq228).

[151] Pond, S. L. K. & Frost, S. D. W., 2005 Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. *Bioinformatics (Oxford, England)* **21**, 2531--2533. ISSN 1367-4803. (doi:10.1093/bioinformatics/bti320). PMID: 15713735.

[152] Pond, S. K. & Muse, S. V., 2005 Site-to-site variation of synonymous substitution rates. *Molecular biology and evolution* **22**, 2375--2385. ISSN 0737-4038. (doi:10.1093/molbev/msi232). PMID: 16107593.

[153] Pond, S. L. K. & Frost, S. D. W., 2005 Not so different after all: A comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution* **22**, 1208--1222. ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev/msi105).

[154] Murrell, B., Wertheim, J. O., Moola, S., Weighill, T., Scheffler, K. & Kosakovsky Pond, S. L., 2012 Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics* **8**, e1002764. (doi:10.1371/journal.pgen.1002764).

[155] Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L. & Scheffler, K., 2013 FUBAR : A fast, unconstrained bayesian AppRoximation for inferring selection. *Molecular biology and evolution* ISSN 1537-1719. (doi:10.1093/molbev/mst030). PMID: 23420840.

[156] Kosakovsky Pond, S. L., Murrell, B., Fourment, M., Frost, S. D. W., Delport, W. & Scheffler, K., 2011 A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution* (doi:10.1093/molbev).

[157] Yang, Z., Nielsen, R., Goldman, N. & Pedersen, A. M., 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**, 431--449.

[158] Yang, Z. & Nielsen, R., 2002 Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **19**, 908. ISSN 0737-4038.

[159] Foote, A. D., Morin, P. A., Durban, J. W., Pitman, R. L., Wade, P., Willerslev, E., Gilbert, M. T. P. & Fonseca, R. R. d., 2010 Positive selection on the killer whale mitogenome. *Biology Letters* ISSN 1744-9561, 1744-957X. (doi:10.1098/rsbl.2010.0638).

[160] McClellan, D. A., Palfreyman, E. J., Smith, M. J., Moss, J. L., Christensen, R. G. & Sailsbery, J. K., 2005 Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *Molecular biology and evolution* **22**, 437--455. ISSN 0737-4038. (doi:10.1093/molbev/msi028). PMID: 15509727.

[161] Palczewski, K., Kumasaka, T., Hori, T., Behnke, C. A., Motoshima, H., Fox, B. A., Trong, I. L., Teller, D. C., Okada, T., Stenkamp, R. E. *et al.*, 2000 Crystal structure of rhodopsin: A g protein-coupled receptor. *Science* **289**, 739--745. ISSN 0036-8075, 1095-9203. (doi:10.1126/science.289.5480.739).

[162] Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A., 2010 Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research* **20**, 110 --121. (doi:10.1101/gr.097857.109).

[163] Hubisz, M. J., Pollard, K. S. & Siepel, A., 2011 PHAST and RPHAST: phylogenetic analysis with space/time models. *Briefings in Bioinformatics* **12**, 41 --51. (doi:10.1093/bib).

[164] Simon, T., 1986 Some probabilistic and statisical problems on the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences (American Mathematical Society)* **17**, 57--86.

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

[166] Benjamini, Y. & Yekutieli, D., 2001 The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* **29**, 1165--1188. ISSN 0090-5364. (doi: 10.1214/aos/1013699998). Mathematical Reviews number (MathSciNet): MR1869245; Zentralblatt MATH identifier: 01829051.

[167] Jombart, T., 2008 adegenet: a r package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403--1405.

[168] Briscoe, A. D., Gaur, C. & Kumar, S., 2004 The spectrum of human rhodopsin disease mutations through the lens of interspecific variation. *Gene* **332**, 107--118. ISSN 0378-1119. (doi:10.1016/j.gene.2004.02.037).

[169] Rakoczy, E. P., Kiel, C., McKeone, R., Stricher, F. & Serrano, L., 2011 Analysis of disease-linked rhodopsin mutations based on structure, function, and protein stability calculations. *Journal of Molecular Biology* **405**, 584--606. ISSN 0022-2836. (doi: 10.1016/j.jmb.2010.11.003).

[170] Boveris, A. & Chance, B., 1973 The mitochondrial generation of hydrogen peroxide. general properties and effect of hyperbaric oxygen. *The Biochemical journal* **134**, 707--716. ISSN 0264-6021. PMID: 4749271.

[171] Turrens, J. F., 1997 Superoxide production by the mitochondrial respiratory chain. *Bioscience reports* **17**, 3--8. ISSN 0144-8463. PMID: 9171915.

[172] Johnston, J. R., Godzik, C. & Cohn, Z., 1978 Increased superoxide anion production by immunologically activated and chemically elicited macrophages. *The Journal of Experimental Medicine* **148**, 115--129. ISSN 0022-1007. PMID: 209122 PMCID: PMC2184904.

[173] Nauseef, W. M., 2004 Assembly of the phagocyte NADPH oxidase. *Histochemistry and cell biology* **122**, 277--291. ISSN 0948-6143. (doi:10.1007/s00418-004-0679-8). PMID: 15293055.

[174] Robinson, J. M., 2008 Reactive oxygen species in phagocytic leukocytes. *Histochemistry and Cell Biology* **130**, 281--297. ISSN 0948-6143. (doi:10.1007/s00418-008-0461-4). PMID: 18597105 PMCID: PMC2491708.

[175] Craig, M. & Slauch, J. M., 2009 Phagocytic superoxide specifically damages an extracytoplasmic target to inhibit or kill salmonella. *PLoS ONE* **4**, e4975. (doi:10.1371/journal.pone.0004975).

[176] Warner, H. R., 1994 Superoxide dismutase, aging, and degenerative disease. *Free Radical Biology and Medicine* **17**, 249--258. ISSN 0891-5849. (doi:10.1016/0891-5849(94)90080-9).

[177] Kroemer, G., Petit, P., Zamzami, N., Vayssière, J. L. & Mignotte, B., 1995 The biochemistry of programmed cell death. *The FASEB Journal* **9**, 1277--1287. ISSN 0892-6638, 1530-6860.

[178] McCord, J. M. & Fridovich, I., 1969 Superoxide dismutase an enzymic function for erythrocuprein (hemocuprein). *Journal of Biological Chemistry* **244**, 6049--6055. ISSN 0021-9258, 1083-351X.

[179] Gee, H., 1999 Inside lou gehrig's disease. *Nature News* ISSN 1744-7933. (doi:10.1038/news990422-4).

[180] Cardoso, R. M. F., Thayer, M. M., DiDonato, M., Lo, T. P., Bruns, C. K., Getzoff, E. D. & Tainer, J. A., 2002 Insights into lou gehrig's disease from the structure and instability of the A4V mutant of human Cu,Zn superoxide dismutase. *Journal of molecular biology* **324**, 247--256. ISSN 0022-2836. PMID: 12441104.

[181] Costantini, D., 2008 Oxidative stress in ecology and evolution: lessons from avian studies. *Ecology Letters* **11**, 1238–1251. ISSN 1461-0248. (doi:10.1111/j.1461-0248.2008.01246.x).

[182] Saino, N., Caprioli, M., Romano, M., Boncoraglio, G., Rubolini, D., Ambrosini, R., Bonisoli-Alquati, A. & Romano, A., 2011 Antioxidant defenses predict long-term survival in a passerine bird. *PLoS ONE* **6**, e19593. (doi:10.1371/journal.pone.0019593).

[183] Perez-Campo, R., López-Torres, M., Cadenas, S., Rojas, C. & Barja, G., 1998 The rate of free radical production as a determinant of the rate of aging: evidence from the comparative approach. *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology* **168**, 149--158. ISSN 0174-1578. PMID: 9591361.

[184] Pamplona, R., 2011 Mitochondrial DNA damage and animal longevity: Insights from comparative studies. *Journal of Aging Research* **2011**, 1--9. ISSN 2090-2212. (doi:10.4061/2011/807108).

[185] Leeuwenburgh, C., Hansen, P. A., Holloszy, J. O. & Heinecke, J. W., 1999 Hydroxyl radical generation during exercise increases mitochondrial protein oxidation and levels

of urinary dityrosine. *Free radical biology & medicine* **27**, 186--192. ISSN 0891-5849. PMID: 10443935.

[186] Talavera, G. & Castresana, J., 2007 Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments. *Systematic Biology* **56**, 564 --577. (doi:10.1080/10635150701472164).

[187] R-Core-Team, 2012. R: A language and environment for statistical computing.

[188] Yang, Z., 1994 Estimating the pattern of nucleotide substitution. *Journal of Molecular Evolution* **39**. ISSN 0022-2844. (doi:10.1007/BF00178256).

[189] Frost, S. D. W., Liu, Y., Pond, S. L. K., Chappey, C., Wrin, T., Petropoulos, C. J., Little, S. J. & Richman, D. D., 2005 Characterization of Human Immunodeficiency Virus Type 1 (HIV-1) Envelope Variation and Neutralizing Antibody Responses during Transmission of HIV-1 Subtype B. *Journal of Virology* **79**, 6523--6527. (doi: 10.1128/JVI.79.10.6523-6527.2005).

[190] Zhang, G., Cowled, C., Shi, Z., Huang, Z., Bishop-Lilly, K. A., Fang, X., Wynne, J. W., Xiong, Z., Baker, M. L., Zhao, W. *et al.*, 2013 Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science* **339**, 456--460.

[191] Dijkgraaf, S., 1963 The functioning and significance of the lateral line organs. *Biological Reviews* **38**, 51--105. ISSN 1469-185X.

[192] Northcutt, R. G., 1997 Animal behaviour: Swimming against the current. *Nature* **389**, 915--916. ISSN 0028-0836. (doi:10.1038/40018).

[193] Wark, A. R. & Peichel, C. L., 2010 Lateral line diversity among ecologically divergent threespine stickleback populations. *Journal of Experimental Biology* **213**, 108 --117. (doi:10.1242/jeb.031625).

[194] Montgomery, J. C., Baker, C. F. & Carton, A. G., 1997 The lateral line can mediate rheotaxis in fish. *Nature* **389**, 960--963. ISSN 0028-0836. (doi:10.1038/40135).

[195] Pitcher, T., Partridge, B. & Wardle, C., 1976 A blind fish can school. *Science* **194**, 963. ISSN 0036-8075.

[196] Bleckmann, H., 1993 Role of the lateral line in fish behaviour. In *The behaviour of teleost fishes* (ed. T. J. Pitcher), pp. 201--246. Chapman and Hall.

[197] Satou, M., Takeuchi, H., Nishii, J., Tanabe, M., Kitamura, S., Okumoto, N. & Iwata, M., 1994 Behavioral and electrophysiological evidences that the lateral line is involved in the inter-sexual vibrational communication of the himé salmon (landlocked red salmon,

Oncorhynchus nerka). *Journal of Comparative Physiology A: Neuroethology, Sensory, Neural, and Behavioral Physiology* **174**, 539--549. ISSN 0340-7594.

[198] Montgomery, J. C. & MacDonald, J. A., 1987 Sensory Tuning of Lateral Line Receptors in Antarctic Fish to the Movements of Planktonic Prey. *Science* **235**, 195 --196. (doi:10.1126/science.235.4785.195).

[199] Bleckmann, H. & Bullock, T., 1989 Central nervous physiology of the lateral line, with special reference to cartilaginous fishes. *The Mechanosensory Lateral Line--Neurobiology and Evolution, Springer-Verlag, New York* pp. 387--408.

[200] Janssen, J., Sideleva, V. & Biga, H., 1999 Use of the lateral line for feeding in two Lake Baikal sculpins. *Journal of Fish Biology* **54**, 404--416. ISSN 1095-8649. (doi:10.1111/j.1095-8649.1999.tb00839.x).

[201] Hassan, E., 1989 Hydrodynamic imaging of the surroundings by the lateral line of the blind cave fish *Anoptichthys jordani*. In *The mechanosensory lateral line: Neurobiology and evolution* (ed. Coombs S, Görner P, Münz H), pp. 217--228.

[202] Montgomery, J., Coombs, S. & Baker, C., 2001 The mechanosensory lateral line system of the hypogean form of Astyanax fasciatus. *Environmental biology of fishes* **62**, 87--96. ISSN 0378-1909.

[203] Vischer, H., 1990 The morphology of the lateral line system in 3 species of Pacific cottoid fishes occupying disparate habitats. *Cellular and Molecular Life Sciences* **46**, 244--250. ISSN 1420-682X.

[204] Wada, H., Hamaguchi, S. & Sakaizumi, M., 2008 Development of diverse lateral line patterns on the teleost caudal fin. *Developmental Dynamics* **237**, 2889--2902. ISSN 1097-0177. (doi:10.1002/dvdy.21710).

[205] Webb, J. F., 1989 Gross Morphology and Evolution of the Mechanoreceptive Lateral-Line System in Teleost Fishes. *Brain, Behavior and Evolution* **33**, 34--53. ISSN 1421-9743. (doi:10.1159/000115896).

[206] Carton, A. & Montgomery, J., 2004 A comparison of lateral line morphology of blue cod and torrentfish: two sandperches of the family Pinguipedidae. *Environmental biology of fishes* **70**, 123--131. ISSN 0378-1909.

[207] Trokovic, N., Herczeg, G., Scott McCairns, R. J., Izza Ab Ghani, N. & Merila, J., 2011 Intraspecific divergence in the lateral line system in the ninespined stickleback (Pungitius pungitius). *Journal of Evolutionary Biology* **24**, 1546--1558. ISSN 1420-9101. (doi:10.1111/j.1420-9101.2011.02286.x).

[208] Northcutt, R., 1989 The phylogenetic distribution and innervation of craniate mechanoreceptive lateral lines. In *The Mechanosensory Lateral Line. Neurobiology and Evolution* (ed. Coombs S, Görner P, Münz H), pp. 17--78.

[209] Pichon, F. & Ghysen, A., 2004 Evolution of posterior lateral line development in fish and amphibians. *Evolution & Development* **6**, 187--193. ISSN 1525-142X. (doi:10.1111/j.1525-142X.2004.04024.x).

[210] Ledent, V., 2002 Postembryonic development of the posterior lateral line in zebrafish. *Development (Cambridge, England)* **129**, 597--604. ISSN 0950-1991. PMID: 11830561.

[211] David, N. B., Sapède, D., Saint-Etienne, L., Thisse, C., Thisse, B., Dambly-Chaudière, C., Rosa, F. M. & Ghysen, A., 2002 Molecular basis of cell migration in the fish lateral line: Role of the chemokine receptor CXCR4 and of its ligand, SDF1. *Proceedings of the National Academy of Sciences* **99**, 16297 --16302. (doi:10.1073/pnas.252339399).

[212] Ghysen, A. & Dambly-Chaudière, C., 2004 Development of the zebrafish lateral line. *Current Opinion in Neurobiology* **14**, 67--73. ISSN 0959-4388. (doi:10.1016/j.conb.2004.01.012). PMID: 15018940.

[213] Ghysen, A. & Dambly-Chaudière, C., 2007 The lateral line microcosmos. *Genes & Development* **21**, 2118 --2130. (doi:10.1101/gad.1568407).

[214] Ma, E. Y. & Raible, D. W., 2009 Signaling pathways regulating zebrafish lateral line development. *Current Biology* **19**, R381--386. ISSN 1879-0445. (doi:10.1016/j.cub.2009.03.057). PMID: 19439264.

[215] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. *et al.*, 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics* **25**, 25--29. ISSN 1061-4036. (doi:10.1038/75556). PMID: 10802651.

[216] Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C. *et al.*, 2004 The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**, D258--261. ISSN 1362-4962. (doi:10.1093/nar). PMID: 14681407.

[217] Hurst, L. D., Feil, E. J. & Rocha, E. P. C., 2006 Protein evolution: Causes of trends in amino-acid gain and loss. *Nature* **442**, E11--E12. ISSN 0028-0836. (doi:10.1038/nature05137).

[218] Benner, S. A., Sassi, S. O. & Gaucher, E. A., 2010 Molecular Paleoscience: Systems Biology from the Past. *Advances in Enzymology: And Related Areas of Molecular Biology* **75**, 1--132. (doi:10.1002/9780471224464.ch1).

[219] Sato, Y., Hashiguchi, Y. & Nishida, M., 2009 Evolution of multiple phosphodiesterase isoforms in stickleback involved in cAMP signal transduction pathway. *BMC Systems Biology* **3**, 23. ISSN 1752-0509. (doi:10.1186/1752-0509-3-23). PMID: 19232106.

[220] Yokoyama, S., 2002 Molecular evolution of color vision in vertebrates. *Gene* **300**, 69--78. ISSN 0378-1119. PMID: 12468088.

[221] Braasch, I., Volff, J. & Schartl, M., 2008 The evolution of teleost pigmentation and the fishspecific genome duplication. *Journal of Fish Biology* **73**, 1891--1918. ISSN 1095-8649. (doi:10.1111/j.1095-8649.2008.02011.x).

[222] Robinson-Rechavi, M. & Laudet, V., 2001 Evolutionary Rates of Duplicate Genes in Fish and Mammals. *Molecular Biology and Evolution* **18**, 681 --683.

[223] Steinke, D., Salzburger, W., Braasch, I. & Meyer, A., 2006 Many genes in fish have species-specific asymmetric rates of molecular evolution. *BMC Genomics* **7**, 20. ISSN 1471-2164. (doi:10.1186/1471-2164-7-20).

[224] Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z. *et al.*, 2007 PyCogent: a toolkit for making sense from sequence. *Genome Biology* **8**, R171. ISSN 1465-6914. (doi: 10.1186/gb-2007-8-8-r171).

[225] Muffato, M., Louis, A., Poisnel, C.-E. & Crollius, H. R., 2010 Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119--1121.

[226] Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J. & Marra, M. A., 2009 Circos: an information aesthetic for comparative genomics. *Genome Research* **19**, 1639--1645. ISSN 1549-5469. (doi:10.1101/gr.092759.109). PMID: 19541911.

[227] Guindon, S. & Gascuel, O., 2003 A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology* **52**, 696--704.

[228] Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C. & Gascuel, O., 2011 Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Systematic Biology* **60**, 685 --699. (doi: 10.1093/sysbio).

[229] Anisimova, M. & Gascuel, O., 2006 Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology* **55**, 539 --552. (doi:10. 1080/10635150600755453).

[230] Hasegawa, M., Yano, T. & Kishino, H., 1984 A new molecular clock of mitochondrial DNA and the evolution of hominoids. *Proceedings of The Japan Academy Series B-physical and Biological Sciences* **60**, 95--98. (doi:10.2183/pjab.60.95).

[231] Hasegawa, M., Kishino, H. & Yano, T., 1985 Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160--174. ISSN 0022-2844. (doi:10.1007/BF02101694).

[232] Zhang, J., 2004 Frequent false detection of positive selection by the likelihood method with branch-site models. *Molecular biology and evolution* **21**, 1332. ISSN 0737-4038.

[233] Huang, S., Tian, H., Chen, Z., Yu, T. & Xu, A., 2010 The evolution of vertebrate tetraspanins: gene loss, retention, and massive positive selection after whole genome duplications. *BMC Evolutionary Biology* **10**, 306. ISSN 1471-2148. (doi:10.1186/1471-2148-10-306).

[234] Zhang, Y., 2008 I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**, 40. ISSN 1471-2105. (doi:10.1186/1471-2105-9-40).

[235] Roy, A., Kucukural, A. & Zhang, Y., 2010 I-TASSER: a unified platform for automated protein structure and function prediction. *Nature Protocols* **5**, 725--738. ISSN 1754-2189. (doi:10.1038/nprot.2010.5).

[236] Postlethwait, J. H., Woods, I. G., Ngo-Hazelett, P., Yan, Y.-L., Kelly, P. D., Chu, F., Huang, H., Hill-Force, A. & Talbot, W. S., 2000 Zebrafish Comparative Genomics and the Origins of Vertebrate Chromosomes. *Genome Research* **10**, 1890--1902. ISSN 1088-9051, 1549-5469. (doi:10.1101/gr.164800).

[237] Woods, I. G., Wilson, C., Friedlander, B., Chang, P., Reyes, D. K., Nix, R., Kelly, P. D., Chu, F., Postlethwait, J. H. & Talbot, W. S., 2005 The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Research* **15**, 1307--1314. ISSN 1088-9051. (doi:10.1101/gr.4134305). PMID: 16109975 PMCID: 1199546.

[238] Braasch, I., Brunet, F., Volff, J. & Schartl, M., 2009 Pigmentation pathway evolution after whole-genome duplication in fish. *Genome biology and evolution* **1**, 479.

[239] Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J.-m., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. *et al.*, 2002 Whole-Genome Shotgun Assembly and Analysis of the Genome of *Fugu rubripes*. *Science* **297**, 1301 --1310. (doi:10.1126/science.1072104).

[240] Christoffels, A., Koh, E. G. L., Chia, J.-m., Brenner, S., Aparicio, S. & Venkatesh, B., 2004 Fugu Genome Analysis Provides Evidence for a Whole-Genome Duplication Early

During the Evolution of Ray-Finned Fishes. *Molecular Biology and Evolution* **21**, 1146 --1151. (doi:10.1093/molbev).

[241] Ravi, V. & Venkatesh, B., 2008 Rapidly evolving fish genomes and teleost diversity. *Current Opinion in Genetics & Development* **18**, 544--550. ISSN 0959-437X.

[242] Vandepoele, K., De Vos, W., Taylor, J. S., Meyer, A. & Van de Peer, Y., 2004 Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 1638 --1643. (doi: 10.1073/pnas.0307968100).

[243] Crow, K., Stadler, P., Lynch, V., Amemiya, C. & Wagner, G., 2006 The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Molecular Biology and Evolution* **23**, 121--136.

[244] Nelson, J. S., 2006 *Fishes of the World*. John Wiley & Sons. ISBN 9780471250319.

[245] Brunet, F. G., Crollius, H. R., Paris, M., Aury, J.-M., Gibert, P., Jaillon, O., Laudet, V. & Robinson-Rechavi, M., 2006 Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. *Molecular Biology and Evolution* **23**, 1808--1816. ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev).

[246] Delport, W., Scheffler, K. & Seoighe, C., 2009 Models of coding sequence evolution. *Briefings in Bioinformatics* **10**, 97--109. ISSN 1467-5463. (doi:10.1093/bib).

[247] Sato, Y., Hashiguchi, Y. & Nishida, M., 2009 Temporal pattern of loss/persistence of duplicate genes involved in signal transduction and metabolic pathways after teleost-specific genome duplication. *BMC Evolutionary Biology* **9**, 127. ISSN 1471-2148. (doi: 10.1186/1471-2148-9-127). PMID: 19500364.

[248] Costello, M. J., May, R. M. & Stork, N. E., 2013 Can we name earth's species before they go extinct? *Science* **339**, 413--416.

[249] Puillandre, N., Modica, M., Zhang, Y., Sirovich, L., Boisselier, M.-C., Cruaud, C., Holford, M. & Samadi, S., 2012 Large-scale species delimitation method for hyperdiverse groups. *Molecular ecology* **21**, 2671--2691.

[250] Padial, J., Miralles, A., De la Riva, I. & Vences, M., 2010 The integrative future of taxonomy. *Frontiers in Zoology* **7**, 16. ISSN 1742-9994. (doi:10.1186/1742-9994-7-16).

[251] May, R. M. & Harvey, P. H., 2009 Species uncertainties. *Science* **323**, 687--687.

[252] Schönhuth, S., Hillis, D. M., Neely, D. A., Lozano-Vilano, L., Perdices, A. & Mayden, R. L., 2011 Phylogeny, diversity, and species delimitation of the north american round-nosed minnows (teleostei: Dionda), as inferred from mitochondrial and nuclear dna sequences. *Molecular Phylogenetics and Evolution* **62**, 427--446.

[253] Funk, W. C., Caminer, M. & Ron, S. R., 2012 High levels of cryptic species diversity uncovered in amazonian frogs. *Proceedings of the Royal Society B: Biological Sciences* **279**, 1806--1814.

[254] Spinks, P. Q., Thomson, R. C., Hughes, B., Moxley, B., Brown, R., Diesmos, A. & Shaffer, H. B., 2012 Cryptic variation and the tragedy of unrecognized taxa: the case of international trade in the spiny turtle heosemys spinosa (testudines: Geoemydidae). *Zoological Journal of the Linnean Society* **164**, 811--824.

[255] Daugherty, C., Cree, A., Hay, J. & Thompson, M., 1990 Neglected taxonomy and continuing extinctions of tuatara (sphenodon). *Nature* **347**, 177--179.

[256] Lundberg, J. G., Kottelat, M., Smith, G. R., Stiassny, M. L. & Gill, A. C., 2000 So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. *Annals of the Missouri Botanical Garden* pp. 26--62.

[257] Piggott, M. P., Chao, N. L. & Beheregaray, L. B., 2011 Three fishes in one: cryptic species in an amazonian floodplain forest specialist. *Biological Journal of the Linnean Society* **102**, 391--403.

[258] Collins, R. A., Armstrong, K. F., Meier, R., Yi, Y., Brown, S. D. J., Cruickshank, R. H., Keeling, S. & Johnston, C., 2012 Barcoding and Border Biosecurity: Identifying Cyprinid Fishes in the Aquarium Trade. *PLoS ONE* **7**, e28381. (doi:10.1371/journal.pone.0028381).

[259] Conway, K., Hirt, M., Yang, L., Mayden, R. & Simons, A., 2010 Cypriniformes: systematics and palaeontology. In *Cypriniformes: systematics and paleontology (ed. J. S. Nelson, H.-P. Schultze & M. V. H. Wilson)*, pp. 295--316. Germany: Verlag Dr. Friedrich Pfeil.

[260] Tao, W., Zou, M., Wang, X., Gan, X., Mayden, R. L. & He, S., 2010 Phylogenomic Analysis Resolves the Formerly Intractable Adaptive Diversification of the Endemic Clade of East Asian Cyprinidae (Cypriniformes). *PLoS ONE* **5**, e13508. (doi:10.1371/journal.pone.0013508).

[261] Van Bocxlaer, I., Biju, S., Willaert, B., Giri, V. B., Shouche, Y. S. & Bossuyt, F., 2012 Mountain-associated clade endemism in an ancient frog family (Nyctibatrachidae) on the Indian subcontinent. *Molecular Phylogenetics and Evolution* **62**, 839--847. ISSN 1055-7903. (doi:10.1016/j.ympev.2011.11.027).

[262] Bossuyt, F., Meegaskumbura, M., Beenaerts, N., Gower, D. J., Pethiyagoda, R., Roelants, K., Mannaert, A., Wilkinson, M., Bahir, M. M., Manamendra-Arachchi, K. *et al.*, 2004 Local Endemism Within the Western Ghats-Sri Lanka Biodiversity Hotspot. *Science* **306**, 479--481. ISSN 0036-8075, 1095-9203. (doi:10.1126/science.1100167).

[263] Benziger, A., Philip, S., Raghavan, R., Anvar Ali, P. H., Sukumaran, M., Tharian, J. C., Dahanukar, N., Baby, F., Peter, R., Devi, K. R. *et al.*, 2011 Unraveling a 146 Years Old Taxonomic Puzzle: Validation of Malabar Snakehead, Species-Status and Its Relevance for Channid Systematics and Evolution. *PLoS ONE* **6**, e21272. (doi:10.1371/journal.pone.0021272).

[264] Smith, K., Raghavan, R., Dahanukar, N., Molur, S., Holland, R., Hughes, A. & Allen, D., 2011 Regional Synthesis for all taxa. In *The status and distribution of freshwater biodiversity in the Western Ghats, India. S. Molur, KG Smith, BA Daniel, WRT Darwall (Eds).*, pp. 87--108. International Union for Conservation of Nature (IUCN) Gland, Switzerland and Zoo Outreach Organization (ZOO) Coimbatore, India.

[265] Dahanukar, N., Raghavan, R., Ali, A., Abraham, R. & Shaji, C., 2011 The status and distribution of freshwater fishes of the Western Ghats. In *The status and distribution of freshwater biodiversity in the Western Ghats, India. S. Molur, KG Smith, BA Daniel, WRT Darwall (Eds).*, pp. 21--48. International Union for Conservation of Nature (IUCN), Gland, Switzerland and Zoo Outreach Organization (ZOO) Coimbatore, India.

[266] Pethiyagoda, R. & Kottelat, M., 1994 Three new species of fishes of the genera Osteochilichthys (Cyprinidae), Travancoria (Balitoridae) and Horabagrus (Bagridae) from the Chalakudy River, Kerala, India. *Journal of South Asian Natural History* **1**, 97--116.

[267] Berendzen, P. B., Simons, A. M., Wood, R. M., Dowling, T. E. & Secor, C. L., 2008 Recovering cryptic diversity and ancient drainage patterns in eastern north america: Historical biogeography of the *Notropis rubellus* species group (teleostei: Cypriniformes). *Molecular Phylogenetics and Evolution* **46**, 721--737.

[268] Jones, M. T., Voss, S. R., Ptacek, M. B., Weisrock, D. W. & Tonkyn, D. W., 2006 River drainages and phylogeography: an evolutionary significant lineage of shovel-nosed salamander *Desmognathus marmoratus* in the southern Appalachians. *Molecular phylogenetics and evolution* **38**, 280--287. ISSN 1055-7903. (doi:10.1016/j.ympev.2005.05.007). PMID: 15996487.

[269] Raghavan, R., Prasad, G., Anvar Ali, P. H. & Sujarittanonta, L., 2007 Boom and bust fishery in a Biodiversity Hotspot-Is the Western Ghats (South India) losing its most celebrated ornamental fish, *Puntius denisonii*, Day? *Current Science* **92**, 1671--1672.

[270] Raghavan, R., Prasad, G., Pereira, B., Anvar Ali, P. & Sujarittanonta, L., 2009 `Damsel in distress'- The tale of Miss Kerala, *Puntius denisonii* (Day), an endemic and endangered cyprinid of the Western Ghats biodiversity hotspot (South India). *Aquatic Conservation: Marine and Freshwater Ecosystems* **19**, 67--74. ISSN 1099-0755. (doi:10.1002/aqc.963).

[271] Anvar Ali, P. H., Dahanukar, N. & Raghavan, R., 2012. *Puntius denisonii*. www.iucnredlist.org.

[272] Raghavan, R. & Anvar Ali, P. H., 2012. *Puntius chalakkudiensis*. www.iucnredlist.org.

[273] Easa, P. & Shaji, C., 2003 *Biodiversity Documentation for Kerala. Part 8: Freshwater Fishes. India*. Kerala, India: Kerala Forest Research Institute, Peechi, Kerala, 1 edition.

[274] Thomas, R., 2004 *Habitat and distribution of hill stream fishes of Southern Kerala*. Ph.D. thesis, Mahathma Gandhi University, Kottayam, Kerala, India.

[275] Jayaram, K., 2010 *The Freshwater Fishes of the Indian Region*. Delhi, India: Narendra Publishing House, 1 edition.

[276] Dahanukar, N., Raut, R. & Bhat, A., 2004 Distribution, endemism and threat status of freshwater fishes in the Western Ghats of India. *Journal of Biogeography* **31**, 123--136. ISSN 1365-2699. (doi:10.1046/j.0305-0270.2003.01016.x).

[277] Doornik, J. A. & Hansen, H., 2008 An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics* **70**, 927--939.

[278] Huberty, C. J. & Olejnik, S., 2006 *Applied MANOVA and Discriminant Analysis*. John Wiley & Sons. ISBN 9780471468158.

[279] Strimmer, K. & Von Haeseler, A., 1997 Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proceedings of the National Academy of Sciences* **94**, 6815--6819.

[280] Heled, J. & Drummond, A. J., 2010 Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution* **27**, 570--580. ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev).

[281] Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R., 2003 Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270**, 313--321. ISSN 0962-8452, 1471-2954. (doi:10.1098/rspb.2002.2218).

[282] Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell, S., Kamoun, S., Sumlin, W. D. & Vogler, A. P., 2006 Sequence-Based Species Delimitation

for the DNA Taxonomy of Undescribed Insects. *Systematic Biology* **55**, 595--609. ISSN 1063-5157, 1076-836X. (doi:10.1080/10635150600852011).

[283] Patterson, C., 1993 Osteichthyes: Teleostei. In *The fossil record* (ed. M. Benton), volume 2, pp. 621--656. London: Chapman and Hall.

[284] Rüber, L., Kottelat, M., Tan, H. H., Ng, P. K. & Britz, R., 2007 Evolution of miniaturization and the phylogenetic position of Paedocypris, comprising the world's smallest vertebrate. *BMC Evolutionary Biology* **7**, 38. ISSN 1471-2148. (doi:10.1186/1471-2148-7-38).

[285] Saitoh, K., Sado, T., Doosey, M. H., BART Jr, H. L., Inoue, J. G., Nishida, M., Mayden, R. L. & Miya, M., 2011 Evidence from mitochondrial genomics supports the lower Mesozoic of South Asia as the time and place of basal divergence of cypriniform fishes (Actinopterygii: Ostariophysi). *Zoological Journal of the Linnean Society* **161**, 633--662. ISSN 1096-3642. (doi:10.1111/j.1096-3642.2010.00651.x).

[286] Arratia, G., 1997 Basal teleosts and teleostean phylogeny. *Paleo ichthyologica* **7**, 5--168.

[287] Menon, A., Devi, K. R. & Thobias, M., 1999 *Puntius chalakkudiensis*, a new colorful species of Puntius (family cyprinidae) fish from Kerala, South India. In *Records of the Zoological Survey of India*, volume 97(4), pp. 61--63. Zoological Survey of India.

[288] Day, S. F., 1865 On the fishes of Cochin, on the Malabar coast of India. *Proceedings of the Zoological Society of London* **33**, 286--318. ISSN 1469-7998. (doi:10.1111/j.1469-7998.1865.tb02337.x).

[289] Subramanyam, K. & Nayar, M., 1974 Vegetation and phytogeography of the Western Ghats. In *Ecology and biogeography in India* (ed. M. Mani), volume 23, pp. 178--196. The Hague Netherlands: Dr W Junk Publishers.

[290] Gower, D. J., Dharne, M., Bhatta, G., Giri, V., Vyas, R., Govindappa, V., Oommen, O. V., George, J., Shouche, Y. & Wilkinson, M., 2007 Remarkable genetic homogeneity in unstriped, long-tailed Ichthyophis along 1500 km of the Western Ghats, India. *Journal of Zoology* **272**, 266--275. ISSN 1469-7998. (doi:10.1111/j.1469-7998.2006.00266.x).

[291] Biju, S. D. & Bossuyt, 2005 Two New Philautus (Anura: Ranidae: Rhacophorinae) from Ponmudi Hill in the Western Ghats of India. *Copeia* **1**, 29--37.

[292] Robin, V. V., Sinha, A. & Ramakrishnan, U., 2010 Ancient Geographical Gaps and Paleo-Climate Shape the Phylogeography of an Endemic Bird in the Sky Islands of Southern India. *PLoS ONE* **5**, e13321. (doi:10.1371/journal.pone.0013321).

[293] Vidya, T. N. C., Fernando, P., Melnick, D. J. & Sukumar, R., 2005 Population differentiation within and among Asian elephant *(Elephas maximus)* populations in southern India. *Heredity* **94**, 71--80. ISSN 0018-067X. (doi:10.1038/sj.hdy.6800568).

[294] Kozak, K. H., Blaine, R. A. & Larson, A., 2005 Gene lineages and eastern North American palaeodrainage basins: phylogeography and speciation in salamanders of the Eurycea bislineata species complex. *Molecular Ecology* **15**, 191--207. ISSN 1365-294X. (doi:10.1111/j.1365-294X.2005.02757.x).

[295] Losos, J. & Mahler, L., 2010 Adaptive Radiation: The Interaction of Ecological Opportunity, Adaptation, and Speciation. In *Evolution Since Darwin: The First 150 Years. (ed. M. A. Bell., D. J. Futuyma, W.F. Eanes & J.S. Levinton)*, pp. 381--420. Sunderland, MA: Sinauer Associates.

[296] Monaghan, M. T., Wild, R., Elliot, M., Fujisawa, T., Balke, M., Inward, D. J. G., Lees, D. C., Ranaivosolo, R., Eggleton, P., Barraclough, T. G. *et al.*, 2009 Accelerated Species Inventory on Madagascar Using Coalescent-Based Models of Species Delineation. *Systematic Biology* **58**, 298--311. ISSN 1063-5157, 1076-836X. (doi:10.1093/sysbio).

[297] Sauer, J. & Hausdorf, B., 2012 A comparison of DNA-based methods for delimiting species in a Cretan land snail radiation reveals shortcomings of exclusively molecular taxonomy. *Cladistics* **28**, 300--316. ISSN 1096-0031. (doi:10.1111/j.1096-0031.2011.00382.x).

[298] Zachos, F. E., Apollonio, M., Bärmann, E. V., Festa-Bianchet, M., Göhlich, U., Habel, J. C., Haring, E., Kruckenhauser, L., Lovari, S., McDevitt, A. D. *et al.*, 2012 Species inflation and taxonomic artefacts—a critical comment on recent trends in mammalian classification. *Mammalian Biology-Zeitschrift für Säugetierkunde* **78**, 1--6.

[299] Austin, J., Jelks, H., Tate, B., Johnson, A. & Jordan, F., 2011 Population genetic structure and conservation genetics of threatened Okaloosa darters *(Etheostoma okaloosae)*. *Conservation Genetics* **12**, 981--989. ISSN 1566-0621. (doi:10.1007/s10592-011-0201-5).

[300] Pethiyagoda, R. & Kottelat, M., 2005 A review of the barbs of the puntius filamentosus group (teleostei: Cyprinidae) of southern india and sri lanka. *Contributions to biodiversity exploration and research in Sri Lanka. The Raffles Bulletin of Zoology, Supplement* pp. 127--144.

[301] Harris, R. J., 2001 *A primer of multivariate statistics*. Lawrence Erlbaum.

[302] Legendre, P. & Legendre, L., 2012 *Numerical ecology*, volume 20. Elsevier.

[303] Anderson, M. J., 2001 A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32--46.

[304] Hammer, O., Harper, D. & Ryan, P., 2001 Past: Paleontological statistics software package for education and data analysis. *Paleontologia Electronica* **4**, 1--9.

[305] Miller, S. A., Dykes, D. D. & Polesky, H. F., 1988 A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research* **16**, 1215.

[306] Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N., 2005 DNA barcoding australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1847--1857. ISSN 0962-8436, 1471-2970. (doi:10.1098/rstb.2005.1716).

[307] Hall, T. A., 1999 BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95-98-NT. *Nucleic Acids Symposium Series* **41**, 95–98.

[308] Jobb, G., von Haeseler, A. & Strimmer, K., 2004 TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evololutionary Biology* **4**, 18. Using Smart Source Parsing Jun 28.

[309] Fontaneto, D., Kaya, M., Herniou, E. A. & Barraclough, T. G., 2009 Extreme levels of hidden diversity in microscopic animals (Rotifera) revealed by DNA taxonomy. *Molecular Phylogenetics and Evolution* **53**, 182--189. ISSN 1095-9513. (doi:10.1016/j.ympev.2009.04.011). PMID: 19398026.

[310] Sanderson, M., 1997 A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular biology and evolution* **14**, 1218--1231.

[311] Paradis, E., Claude, J. & Strimmer, K., 2004 Ape: analyses of phylogenetics and evolution in r language. *Bioinformatics* **20**, 289--290.

[312] Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A., 2012 Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution* ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev).

[313] Yang, Z. & Rannala, B., 2006 Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. *Molecular Biology and Evolution* **23**, 212--226. ISSN 0737-4038, 1537-1719. (doi:10.1093/molbev).

[314] Lévêque, C., Oberdorff, T., Paugy, D., Stiassny, M. L. J. & Tedesco, P. A., 2008 Global diversity of fish (Pisces) in freshwater. In *Freshwater Animal Diversity Assessment* (eds. E. V. Balian, C. Lévêque, H. Segers, K. Martens & H. J. Dumont), volume 198 of *Developments in Hydrobiology*, pp. 545--567. Springer Netherlands. ISBN 978-1-4020-8259-7.

[315] Swartz, E. R., Mwale, M. & Hanner, R., 2008 A role for barcoding in the study of african fish diversity and conservation. *South African Journal of Science* **104**, 293--298. ISSN 0038-2353.

[316] Wee K.L, 1982 Snakeheads-their biology and culture. In *Recent advances in aquaculture* (eds. Muir, J. F. & Roberts, R. J), volume 1, pp. 179--213. London: Croom Hlem.

[317] Berra, T. M., 2001 *Freshwater Fish Distribution*. London: Academic Press.

[318] Sinh, L. X. & Pomeroy, R. S., 2010 Farming of snakehead fish (Channa micropeltes and channa striatus) in the mekong delta of vietnam. In *World Aquaculture*, p. 953. San Diego, California: Network of Aquaculture Centres in Asia-Pacific.

[319] Gam, L.H, Leow, C.Y & Baie, S., 2006 Proteomic analysis of snakehead fish (Channa striata) muscle tissue. *Malaysian Journal of Biochemistry and Molecular Biology* **14**, 25--32.

[320] Raghavan, R., 2010 Ornamental fisheries and trade in kerala. In *Fish Conservation in Kerala* (eds. Leonard Sonnenschein & Allen Benziger), pp. 169--197. St. Louis, USA: World Aquariums and Oceans Federation.

[321] Courtenay, W. R. & Williams, J. D., 2004 *Snakeheads (Pisces, Channidae): A Biological Synopsis and Risk Assessment*. US Geological Survey Circular. Denver, CO.: US Geological Survey.

[322] Rainboth, W. J., 1996 *Fishes of the Cambodian Mekong*. Rome: Food and Agricultural Organization of the United Nations.

[323] Adamson, E. A., Hurwood, D. A. & Mather, P. B., 2010 A reappraisal of the evolution of Asian snakehead fishes (Pisces, Channidae) using molecular data from multiple genes and fossil calibration. *Molecular Phylogenetics and Evolution* **56**, 707--717. ISSN 1055-7903. (doi:10.1016/j.ympev.2010.03.027).

[324] Cuvier, G. & Valenciennes, A., 1831 *Histoire naturelle des poissons*, volume 7. Paris: Chez FG Levrault.

[325] Day, F., 1878 *The fishes of India: being a natural history of the fishes known to inhabit the seas and fresh waters of India, Burma, and Ceylon*. New Delhi: Today & Tomorrow's Book Agency.

[326] Roberts, T. R., 1989 The freshwater fishes of western borneo (Kalimantan barat, indonesia). In *Memoirs of the California Academy of Sciences*, volume 14, pp. 1--210. San Francisco: California Academy of Sciences.

[327] Jayaram, K. C., 1981 *The freshwater fishes of India, Pakistan, Bangladesh, Burma and Sri Lanka-a handbook*. Zoological Survey of India.

[328] Talwar, P. K. & Jhingran, A. G., 1991 *Inland fishes of India and adjacent countries*. New Delhi: Oxford & IBH Publishing Co.

[329] Kottelat, M., 1998 Fishes of the nam theun and xe bangfai basins, laos, with diagnoses of twenty-two new species (Teleostei: cyprinidae, balitoridae, cobitidae, coiidae and odontobutidae). *Ichthyological Research* **9**, 1--128.

[330] Johnsingh, A. J. T., 2006 *Field days: a naturalist's journey through South and Southeast Asia*. Hyderabad: Universities Press.

[331] Kurup, B. M., Radhakrishnan, K. V. & Manojkumar, T. G., 2004 Biodiversity status of fishes inhabiting rivers of kerala (S. india) with special reference to endemism, threats and conservation measures. In *Proceedings of LARS2. 2nd Large Rivers Symposium* (eds. Welcome, R. L. & Petr, T.), pp. 163--182. Phnom Penh, Cambodia: Mekong River Commission and Food and Agricultural Organization.

[332] Molur, S. & Walker, S. (eds.), 2001 *Conservation Assessment and Management Plan*. Coiambatore, India: Zoo Outreach Organisation and Conservation Breeding Specialist Group, IUCN.

[333] Unnithan, V. K., 2000 Decline of endemic fish species in selected reservoirs of western ghats. In *Endemic fish diversity of Western Ghats* (eds. Ponniah A. G. & opalakrishnan, A.), pp. 169--170. NBFGR-NATP Publication.

[334] Ebanasar, J. & Jayaprakas, V., 2005 Length weight relationship of the malabar snakehehead channa micropeltes from pechipparai reservoir, kanyakumari district, tamil nadu. *Journal of the Inland Fisheries Society of India* **37**, 60--63.

[335] NBFGR, 2010. Threatened freshwater fishes of india. http://www.nbfgr.res.in/pdf/ThreatenedFreshwaterFishes.pdf.

[336] Haniffa, M. A., 2010 Indian snakeheads. *Fishing Chimes* **30**, 34--36.

[337] Musikasinthorn, P. & Taki, Y., 2001 Channa siamensis (Günther, 1861), a junior synonym of channa lucius (Cuvier in cuvier and valenciennes, 1831). *Ichthyological Research* **48**, 319--324. ISSN 1341-8998. (doi:10.1007/s10228-001-8153-2).

[338] Vishwanath, W. & Geetakumari, K., 2009 Diagnosis and interrelationships of fishes of the genus channa scopoli (Teleostei: channidae) of northeastern india. *Journal of Threatened Taxa* **1**, 97--105.

[339] Smith E. P. & Lipkovich, I. A., 2002. Biplot 1.1: Excel addin freeware. http://www.stat.vt.edu/facstaff/epsmith.html.

[340] Palumbi, S. R., Martin, A., Romano, S., Mcmillan, W. O., Stice, L. & Grabowski, G., 1991. The simple fool's guide to PCR, version 2.0.

[341] Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D., 2005 DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1847--57.

[342] Tamura, K., Nei, M. & Kumar, S., 2004 Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11030--11035. ISSN 0027-8424, 1091-6490. (doi:10.1073/pnas.0404206101).

[343] Tamura, K. & Kumar, S., 2002 Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Molecular biology and evolution* **19**, 1727--1736.

[344] Roe J. L., 1991 Phylogenetic and ecological significance of channidae (Osteichthyes, teleostei) from the early eocene kuldana formation of kohat, pakistan. *Contributions from the Museum of Paleontology* **28**, 93–100.

[345] Murray, A. M. & Thewissen, J. G. M., 2008 Eocene actinopterygian fishes from pakistan, with the description of a new genus and species of channid (channiformes). *Journal of Vertebrate Paleontology* **28**, 41--52. ISSN 0272-4634.

[346] Xia Li, Musikasinthorn, P. & Kumazawa, Y., 2006 Molecular phylogenetic analyses of snakeheads (Perciformes: channidae) using mitochondrial DNA sequences. *Ichthyological Research* **53**, 148--159. ISSN 1341-8998. (doi:10.1007/s10228-005-0321-3).

[347] Day, F., 1865 *The fishes of Malabar*. London: Bernard Quaritch.

[348] Eschmeyer, W. N. & Fricke, R., 2010. Catalog of fishes. http://research.calacademy.org/ichthyology/catalog/fishcatmain.asp.

[349] Adamson, E. A. S., 2010 *Influence of historical landscapes, drainage evolution and ecological traits on patterns of genetic diversity in South East Asian freshwater snakehead fishes*. Ph.D. thesis, Queensland University of Technology, Brisbane, Australia.

[350] Smith, W. L. & Wheeler, W. C., 2006 Venom evolution widespread in fishes: A phylogenetic road map for the bioprospecting of piscine venoms. *Journal of Heredity* **97**, 206--217. ISSN 0022-1503, 1465-7333. (doi:10.1093/jhered/esj034).

[351] Ruber, L., Britz, R. & Zardoya, R., 2006 Molecular Phylogenetics and Evolutionary Diversification of Labyrinth Fishes (Perciformes: Anabantoidei). *Systematic Biology* **55**, 374--397. (doi:10.1080/10635150500541664).

[352] Randall, J.E., 1998 Zoogeography of shore fishes of the indo-pacific region. *Zoological Studies* **37**, 227--268.

[353] McMillan, W. O., Weigt, L. A. & Palumbi, S. R., 1999 Color pattern evolution, assortative mating, and genetic differentiation in brightly colored butterflyfishes (Chaetodontidae). *Evolution* **53**, 247. ISSN 00143820. (doi:10.2307/2640937).

[354] Grady, J. & Quattro, J., 2001 Using character concordance to define taxonomic and conservation units. *Conservation Biology* **13**, 1004--1007.

[355] Endler, J., Westcott, D., Madden, J. & Robson, T., 2005 Animal visual systems and the evolution of color patterns: sensory processing illuminates signal evolution. *Evolution* **59**, 1795--1818.

[356] Weber, M., 1964 Heteromi, solenichthyes, synentognathi, percesoces, labyrinthici, microcyprini. In *The fishes of the Indo-Australian archipelago* (eds. Weber, M., de Beaufort, L.F. & Weber, M.W.C.), volume 1. Leiden: EJ Brill.

[357] Benton, M. & Ayala, F., 2003 Dating the tree of life. *Science* **300**, 1698--1700.

[358] Kumazawa, Y., Yamaguchi, M. & Nishida, M., 1999 Mitochondrial molecular clocks and the origin of euteleostean biodiversity: familial radiation of perciforms may have predated the Cretaceous/Tertiary boundary. In *The biology of biodiversity* (ed. Kato, M.), p. 35–52. Tokyo: Springer.

[359] Kumazawa, Y. & Nishida, M., 2000 Molecular phylogeny of osteoglossoids: A new model for gondwanian origin and plate tectonic transportation of the asian arowana. *Molecular Biology and Evolution* **17**, 1869--1878. ISSN 0737-4038, 1537-1719.

[360] Murray, A. M., 2006 A new channid (teleostei: Channiformes) from the eocene and oligocene of egypt. *Journal of Paleontology* **80**, 1172--1178. ISSN 0022-3360,.

[361] West, R. M., 1980 Middle eocene large mammal assemblage with tethyan affinities, ganda kas region, pakistan. *Journal of Paleontology* **54**, 508--533. ISSN 0022-3360,.

[362] Steenis, C. G. G. J. V., 1979 Plant-geography of east malesia. *Botanical Journal of the Linnean Society* **79**, 97–178. ISSN 1095-8339. (doi:10.1111/j.1095-8339.1979.tb01511.x).

[363] Klaus, S., Schubart, C., Streit, B. & Pfenninger, M., 2010 When indian crabs were not yet asian - biogeographic evidence for eocene proximity of india and southeast asia. *BMC Evolutionary Biology* **10**, 287. ISSN 1471-2148. (doi:10.1186/1471-2148-10-287).

[364] Briggs, J., 2003 The biogeographic and tectonic history of india. *Journal of Biogeography* **30**, 381--388.

[365] Hora, S. L., 1949 Satpura hypothesis of the distribution of the malayan flora and fauna to peninsular india. *Proceedings of the National Institute of Science (India)* **15**, 309--314.

[366] Yang, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution* **15**, 568--573.

[367] Antunes, A. & Ramos, M. J., 2007 Gathering computational genomics and proteomics to unravel adaptive evolution. *Evolutionary Bioinformatics Online* **3**, 207--209. ISSN 1176-9343. PMID: 19461985 PMCID: PMC2684141.

[368] Blount, Z. D., Barrick, J. E., Davidson, C. J. & Lenski, R. E., 2012 Genomic analysis of a key innovation in an experimental escherichia coli population. *Nature* **489**, 513--518. ISSN 0028-0836. (doi:10.1038/nature11514).

V.

Appendix

11

Appendix 1 (Supplementary materials for chapter 2)

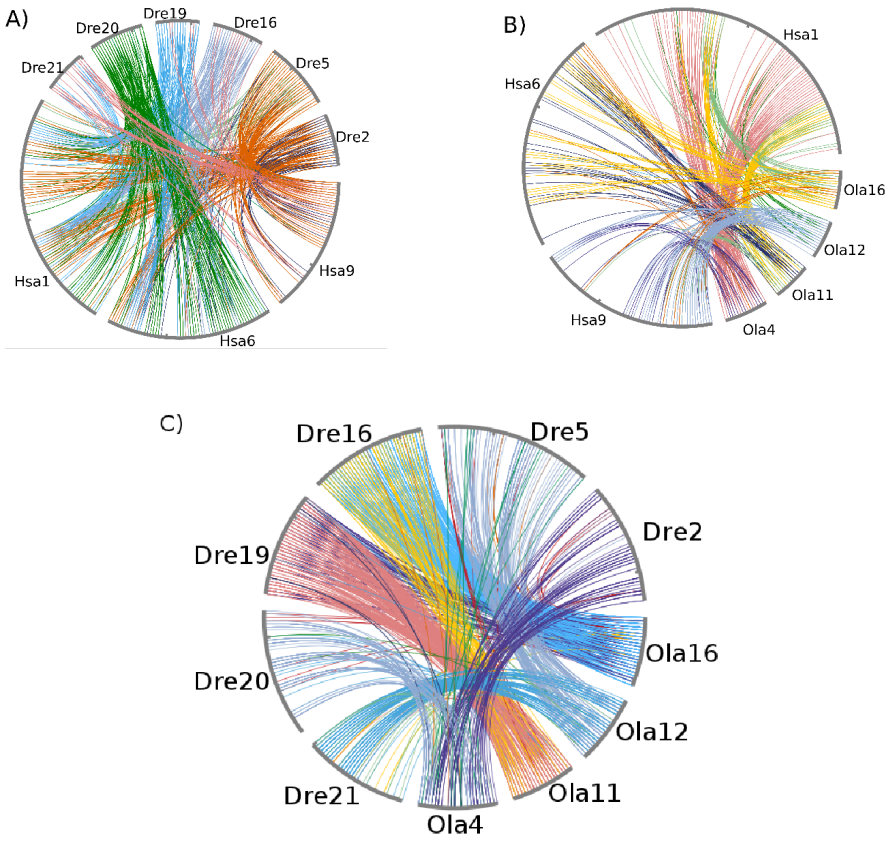


Figure 11.1.: Circular synteny plots between the chromosomes containing RXR genes (A) between Human and Zebrafish; B) between Human and Medaka; C) between Zebrafish and Medaka) showing the presence of numerous co-orthologous genes shared between these chromosomes, supporting the view that all the RXR genes were products of genome duplications

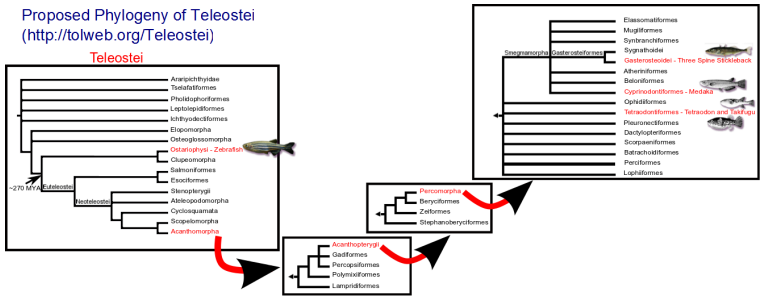


Figure 11.2.: Schematic phylogeny of the teleosts showing the phylogenetic positions of the teleost species used in this study (for which whole genome sequences are available), according to the tree of life (www.tolweb.com/teleostei)

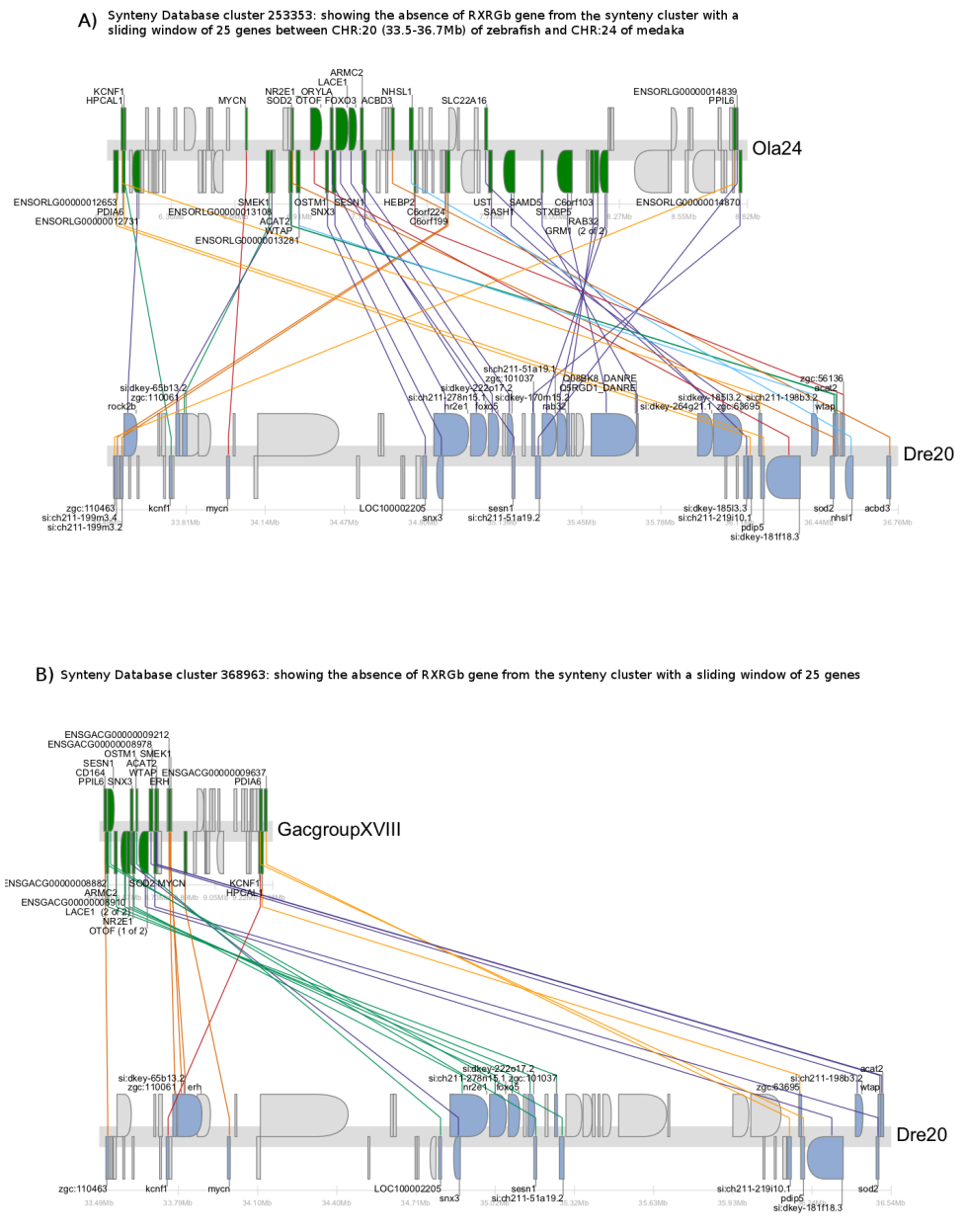


Figure 11.3.: Synteny image showing the gene loss event of RXRG in medaka and threespine stickleback: A) shows the synteny between the chromosome 20 (33.5-36.7 Mb) of zebrafish to chromosome 24 of medaka, 33.9-33.98Mb of chromosome 20 of zebrafish possess rxrgb gene but this is not represented in the synteny cluster, these chromosomes are products of the same ancestral chromosome (46); B) shows the synteny between the chromosome 20 (33.5-36.7 Mb) of zebrafish and scaffold group XVIII of stickleback which also does not possess a copy of RXRG gene

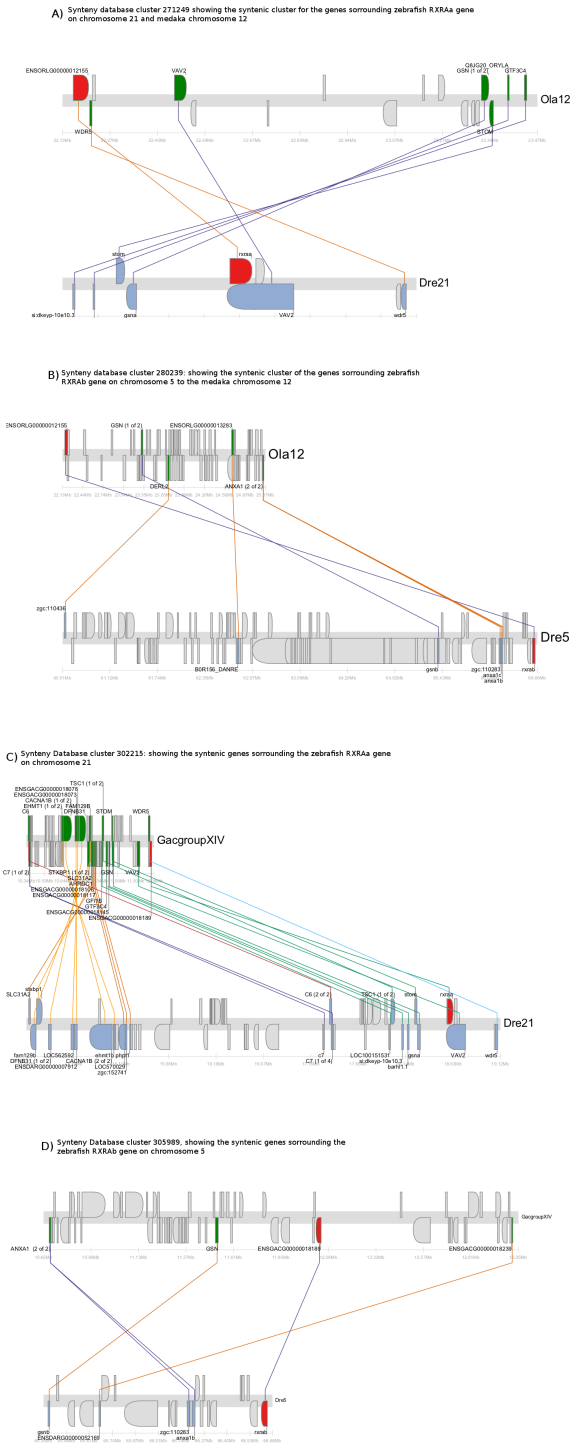


Figure 11.4.: Synteny analysis of the RXRAa and RXRAb genes: A) and B) shows that both rxraa and rxrab of zebrafish forms synteny with the same chromosome of medaka chromosome 12; C) and D) shows that both rxraa and rxrab of zebrafish forms synteny with the same scaffold of threespine stickleback group XIV

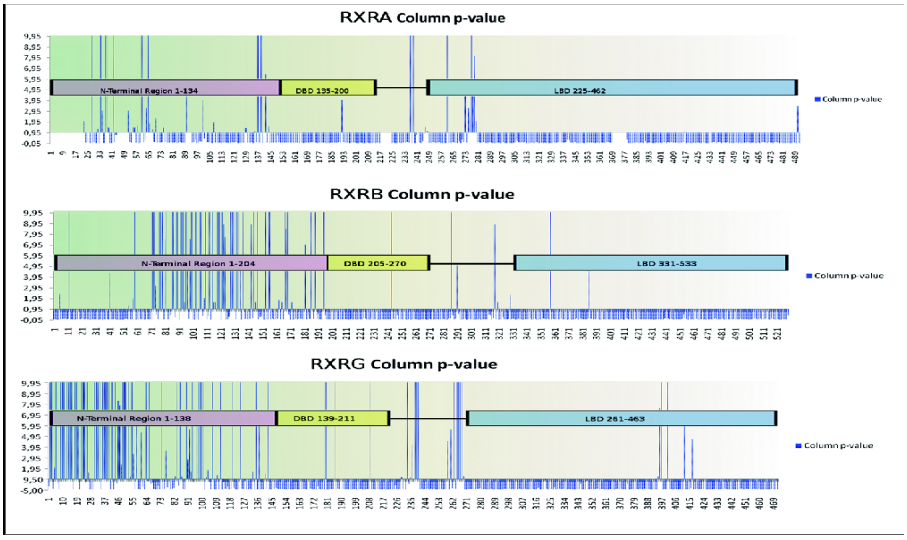


Figure 11.5.: Schematic representation of the major domains of RXRA, RXRB and RXRG superimposed on the Column-P values from MAPP: Schematic representation of the major domains of RXRA, RXRB and RXRG superimposed on the Column-P values from MAPP showing the constraints of amino acid sites, negative values show higher constraints and positive values show less constraints x-axis = amino acid positions and y-axis = column P values; the start and end amino acid site (in Humans) are labeled in the boxes

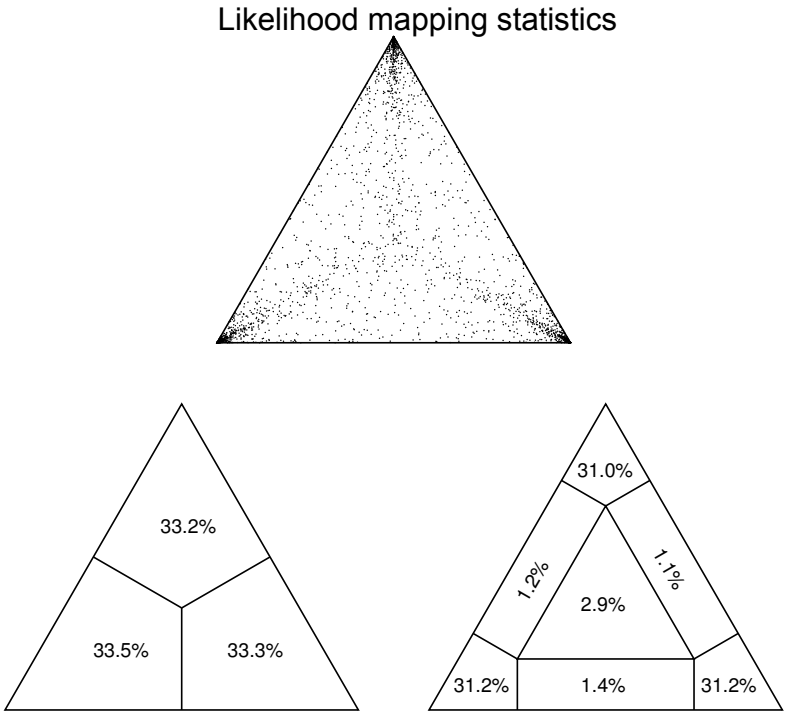


Figure 11.6.: Likelihood ratio statistics for phylogenetic signal analysis of the RXR dataset used in this study: >90 *per cent* of the quartets are resolved, which shows ample phylogenetic signal in the dataset.

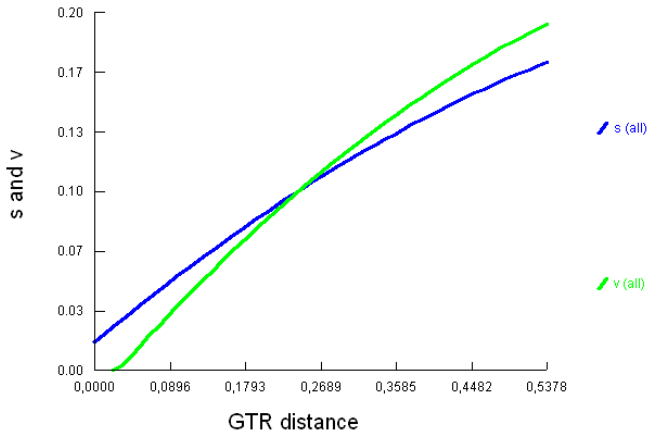


Figure 11.7.: Saturation plot for the RXR dataset, The transitions (s) and transversions (v) are plotted agains the GTR distance, the plot shows that there is no effect of saturation on the dataset

Sl Table1: Topology testing alternate hypothesis of RXR evolution:

Tree	log L	difference	S.E.	p-1sKH [2]	p-SH[3]	c-ELW [4]	2sKH[5]
1	-9932.04	0.36	1.8467	0.4320 +	0.5490 +	0.3374 +	+
2	-9931.68	0.00	<---- best	1.0000 +	1.0000 +	0.4668 +	best
3	-9932.50	0.82	1.3864	0.2770 +	0.4350 +	0.1958 +	+

Alternate topologies used for test were:

1. (((RXRA,RXRB), RXRG),RXR)
2. (((RXRG,RXRB), RXRA),RXR)
3. (((RXRG,RXRA), RXRB),RXR)

The columns show the results and p-values of the following tests:

1sKH - one sided KH test based on pairwise SH tests (Kishino-Hasegawa 1989 [1]; Shimodaira-Hasegawa 2000 [2], Goldman et al., 2001[3]); SH - Shimodaira-Hasegawa test (2000), ELW - Expected Likelihood Weight (Strimmer-Rambaut 2002 [4]) and 2sKH - two sided Kishino-Hasegawa test (1989) [1]. Plus signs denote the confidence sets. Minus signs denote significant exclusion. All tests used 5% significance level. 1sKH, SH, and ELW performed 1000 re-samplings using the RELL method. 1sKH and 2sKH are correct to the 2nd position after the the decimal point of the log-likelihoods.

References:

[1] Kishino, H. and Hasegawa, M. (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J. Mol. Evol., 29, 170–179.

[2] Shimodaira, H. and Hasegawa, M. (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol., 16, 1114–1116.

[3] Goldman, N., Anderson, J. P. and Rodrigo, A. G. (2000) Likelihood-based tests of topologies in phylogenetics. Syst. Biol., 49, 652–670.

[4] Strimmer, K. and Rambaut, A. (2002) Inferring confidence sets of possibly misspecified gene trees. Proc. R. Soc. Lond. B, 269, 137–142.

[5] Abascal F, Zardoya R, Posada D. (2005) ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics:* **21(9)**:2104-2105

SI Table 2: Positive Selected and Rate Shifting sites identified using maximum likelihood methods implemented in PAML (BEB) and RASER2 (empirical bayes) on the post-duplication branches

Positive selected sites on post-duplication branch leading to RXRA identified by NEB (PP >95)*[$] BEB identified no positive selected sites	Positive selected sites on post-duplication branch leading to RXRB (PP >95) *	Positive selected sites on post-duplication branch leading to RXRG(PP >95) *	Positive selected sites on post-duplication branch leading to RXRBa (PP >95)[#]	Rate-shifting sites on post-duplication branch leading to RXRA (PP >90) *	Rate-shifting sites on post-duplication branch leading to RXRB (PP >90) *	Rate-shifting sites on post-duplication branch leading to RXRB and RXRG (PP >90) [c]
48H[%] 71P[%] 75H[%] 116V[%] 253N[&] (L5-6) 256L[&] (L5-6) 380S[&] (L12-13)	99S[%] 115G[%] 144G[%] 204L[%] 287G (Hinge) 313Q (Hinge) 317Q (Hinge) 483Q[&] (L10-11)	70Y[%] 74T[%] 116P[%] 128L[%] 262T[&] (L1-2) 345S[&] (H6) 405T[&] (H9)	15V[%]	44P[%]	88S[%] 155G[%] 157V[%]	57S[%]

*Relative to human sequence; [#]Relative to zebrafish sequence; [$] also the ancestral branch leading to RXRB and RXRG clade; [c] relative to human RXRG; [%] N-terminal region; L = loop between the helices; H = helix; Hinge = hinge region between DBD and LBD; [&] Ligand binding domain.

Sl Table 3: Theta (θ) values for the type I functional divergence among the different clusters of RXR genes; θ values are presented in lower diagonal; values in brackets denote the LRT value.

	RXRA	RXRB	RXRG	RXRBa	RXRBb
RXRA	--	--	--	--	--
RXRB	**0.42±0.05 (11.021)**	--	--	--	--
RXRG	**0.37± 0.07 (20.58)**	**0.14±0.05 (6.08)**	–	--	--
RXRBa	na	NS	na	--	--
RXRBb	na	NS	na	**1.04±0.15 (41.37)**	--

na = not applicable; NS = Not significant

Sl. Table. 4: Test of substitution saturation (Xia et al. 2003; Xia and Lemey 2009)

NumOTU	Iss	Iss.cSym	T	DF	P	Iss.cAsym	T	DF	P
4	0.387	0.829	27.914	1301	0.0000	0.798	25.929	1301	0.0000
8	0.379	0.802	24.407	1301	0.0000	0.700	18.515	1301	0.0000
16	0.391	0.785	22.290	1301	0.0000	0.597	11.630	1301	0.0000
32	0.391	0.766	20.922	1301	0.0000	0.475	4.712	1301	0.0000

Note: two-tailed tests are used. Analysis performed on all sites with gaps treated as unknown nucleotide. Testing whether the observed Iss is significantly lower than Iss.c. IssSym is Iss.c assuming a symmetrical topology. IssAsym is Iss.c assuming an asymmetrical topology.

Interpretation of results:

	Significant Difference	
	Yes	No
Iss < Iss.c	**Little saturation**	Substantial saturation

References:
Xia, X., Z. Xie, M. Salemi, L. Chen, Y. Wang. 2003. An index of substitution saturation and its application. Molecular Phylogenetics and Evolution 26:1-7.
Xia, X. and Lemey, P. 2009. Assessing substitution saturation with DAMBE. Pp. 615-630 in Philippe Lemey, Marco Salemi and Anne-Mieke Vandamme, eds. The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny. 2nd edition Cambridge University Press.

12

Appendix 2 (Supplementary materials for Chapter 4)

Table 12.1.: List of species scanned for SOD genes; Genomes were downloaded from http://phybirds.genomics.org.cn

Abbr	latin	common name
ACACH	*Acanthisitta chloris*	Rifleman
ALLIG[a]	*Alligator mississippiensis*	American Alligator
ANAPL	*Anas platyrhynchos*	Mallard (domestic)
ANTCA	*Antrostomus carolinensis*	Chuck-will's-widow (Nightjar)
APAVI	*Apaloderma vittatum*	Bar-tailed Trogon
APTFO	*Aptenodytes forsteri*	Emperor Penguin
BALRE	*Balearica regulorum*	Grey Crowned Crane
BUCRH	*Buceros rhinoceros*	Rhinoceros Hornbill
CALAN	*Calypte anna*	Anna's Hummingbird
CARCR	*Cariama cristata*	Red-legged Seriema
CATAU	*Cathartes aura*	Turkey Vulture
CHAPE	*Chaetura pelagica*	Chimney Swift
CHAVO	*Charadrius vociferus*	Killdeer
CHEMY[a]	*Chelonia mydas*	Green Turtle
CHLUN	*Chlamydotis undulata*	Houbara Bustard
COLLI	*Columba livia*	Rock Pigeon (domestic)
COLST	*Colius striatus*	Speckled Mousebird
CORBR	*Corvus brachyrhynchos*	American Crow
CUCCA	*Cuculus canorus*	Common Cuckoo
EGRGA	*Egretta garzetta*	Little Egret
EURHE	*Eurypyga helias*	Sunbittern
FALPE	*Falco peregrinus*	Peregrine Falcon
FULGL	*Fulmarus glacialis*	Northern Fulmar
GAVST	*Gavia stellata*	Red-throated Loon
GEOFO	*Geospiza fortis*	Medium Ground-finch
HALAL	*Haliaeetus albicilla*	White-tailed Eagle
HALLE	*Haliaeetus leucocephalus*	Bald Eagle
HUMAN	*Homo sapiens*	Human
LEPDI	*Leptosomus discolor*	Cuckoo Roller
MANVI	*Manacus vitellinus*	Golden-collared Manakin
MELUN	*Melopsittacus undulatus*	Budgerigar
MERNU	*Merops nubicus*	Carmine Bee-eater
MESUN	*Mesitornis unicolor*	Brown Mesite
NESNO	*Nestor notabilis*	Kea
NIPNI	*Nipponia nippon*	Crested Ibis
OPHHO	*Ophisthocomus hoazin*	Hoatzin
PELCR	*Pelecanus crispus*	Dalmatian Pelican
PHACA	*Phalacrocorax carbo*	Cormorant
PHALE	*Phaethon lepturus*	White-tailed Tropicbird
PHORU	*Phoenicopterus ruber*	American Flamingo
PICPU	*Picoides pubescens*	Downy Woodpecker
PODCR	*Podiceps cristatus*	Great Crested Grebe
PTEGU	*Pterocles gutturalis*	Yellow-throated Sandgrouse
PYGAD	*Pygoscelis adeliae*	Adelie Penguin
STRCA	*Struthio camelus*	Ostrich
TAUER	*Tauraco erythrolophus*	Red-crested Turaco
TINMA	*Tinamus major*	Great Tinamou
TYTAL	*Tyto alba*	Barn Owl

[a] Reptiles.

13

Appendix 3 (Supplementary materials for Chapter 5)

Figure 13.1.: Synteny plots of genes with asymmetric distribution of paralogs in teleosts, these genes comprise those which retained duplicates other than from the FSGD, A) Phyloview of the syntenic genes surrounding CLDN4; B) Phyloview of the syntenic genes surrounding PCSK5; C) Phyloview of the syntenic genes surrounding KAL1; D) Phyloview of the syntenic genes

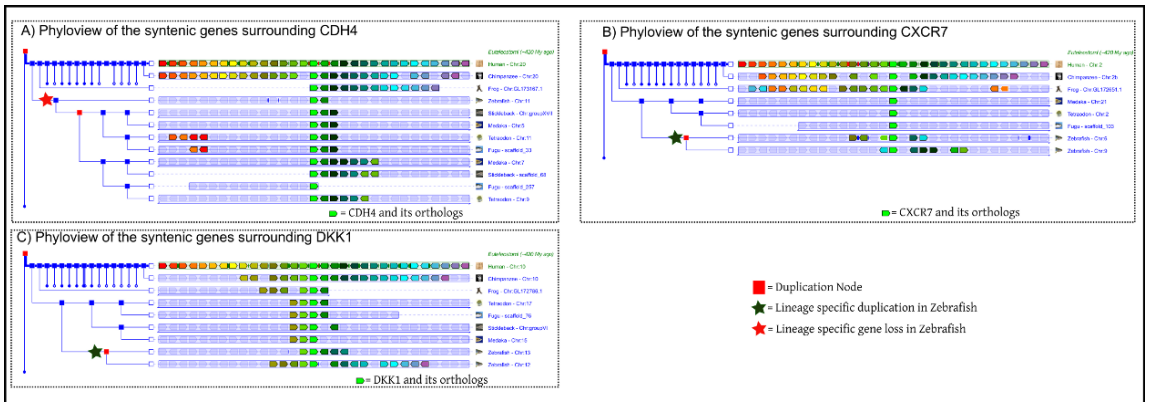


Figure 13.2.: Synteny plots of genes that were duplicated during the FSGD but only some teleosts retained both paralogs, A) Phyloview of the synteny of the genes surrounding CDH4; B) Phyloview of the synteny of genes surrounding CXCR7; C) Phyloview of the synteny of the genes surrounding DKK1.

Table S1: Details of genes used for the study

	Gene (in Homo sapiens)	Ensembl stable ID (Human)	Ensembl ID (Danio rerio) – involved in GO process:0048925a	FSGD-duplicated	Danio rerio Paralogs/Orthologsb	Duplicated only in Danio rerio	Duplicated in other teleosts (except Danio rerio)	Lineage specific duplications in fishes
1	adenomatous polyposis coli – APC	ENSG00000134982	ENSDART00000044432	no	**apc**	NA	NA	NA
2	atonal homolog 1 – ATOH1	ENSG00000172238	ENSDARG00000055294	***yes***	**atoh1a**	NA	NA	NA
					atoh1b	NA		
3	ATPase, Na+/K+ transporting, beta 2 polypeptide – ATP1B2	ENSG00000129244	ENSDARG00000034424	***yes***	**atp1b2b**	NA	NA	NA
					atp1b2a	NA	NA	NA
4	ATPase, Ca++ transporting, plasma membrane 1 – ATP2B1	ENSG00000070961	ENSDARG00000012684	***yes***	**atp2b1a**	NA	NA	NA
					atp2b1b	NA	NA	NA
5	caveolin 1, caveolae protein, 22kDa – CAV1	ENSG00000105974	ENSDARG00000052004	no	**cav1**	NA	NA	NA
6	cadherin 4, type 1, R-cadherin (retinal) – CDH4	ENSG00000179242	ENSDARG00000015002	***yes***	**cdh4**	-not duplicated in zebrafish - gene loss?	"cdh4 1 of 2" & "cdh4 2 of 2"	no
7	Claudin 4 – CLDN4	ENSG00000189143	ENSDARG00000009544	***yes***	**cldnb**	8 paralogs in zebrafish owing to lineage specific duplications	11 paralogs in Tetraodon nigroviridis, 7 paralogs in Oryzias latipes, 9 paralogs for Takifugu rubireps and 13 paralogs in Gasterosteus aculeatus	***yes***

8	collagen, type XVII, alpha 1 – COL17A1	ENSG00000065618	ENSDARG00000079011	***yes***	**col17a1b**	NA		
					col17a1a	NA	NA	NA
9	chemokine (C-X-C motif) ligand 12 – CXCL12	ENSG00000107562	ENSDARG00000037116	***yes***	**cxcl12a**	NA	NA	NA
					cxcl12b	NA	NA	NA
10	chemokine (C-X-C motif) receptor 4 – CXCR4	ENSG00000121966	ENSDARG00000041959	***yes***	**cxcr4b**	NA	NA	NA
					cxcr4a	NA		
11	chemokine (C-X-C motif) receptor 7 – CXCR7	ENSG00000144476	ENSDARG00000058179		**cxcr7b**	***yes***	not duplicated in other teleosts (except zebrafish)	Yes - only in Danio rerio
					cxcr7a			
12	dickkopf homolog 1 – DKK1	ENSG00000107984	ENSDARG00000045219		**dkk1b**	***yes***	not duplicated in other teleosts (except zebrafish)	Yes - only in Danio rerio
					dkk1a			
13	early growth response 2 – EGR2	ENSG00000122877	ENSDARG00000042826	***yes***	**egr2b**	NA	NA	NA
					egr2a			
14	v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog – ERBB2	ENSG00000141736	ENSDARG00000026294	no	**erbb2**	NA	NA	NA
15	v-erb-b2 erythroblastic leukemia viral oncogene homolog 3 – ERBB3	ENSG00000065361	ENSDARG00000036993	***yes***	**erbb3b**	NA	NA	NA
					erbb3a	NA		

16	estrogen receptor 1 – ESR1	ENST00000440973	ENSDARG00000004111	no	**esr1**	NA	NA	NA
17	eyes absent homolog 4 – EYA4	ENSG00000112319	ENSDARG00000012397	no	**eya4**	NA	NA	NA
18	fibroblast growth factor 3 – FGF3	ENSG00000186895	ENSDARG00000068094	no	**fgf3**	NA	NA	NA
19	fibroblast growth factor 10 – FGF10	ENSG00000070193	ENSDARG00000030932	***yes***	**fgf10a**	NA	NA	NA
					fgf10b	NA	NA	NA
20	fibroblast growth factor receptor 1 – FGFR1	ENSG00000077782	ENSDARG00000011027	***yes***	**fgfr1a**	NA	NA	NA
					fgfr1b	NA	NA	NA
21	forkhead box D3 – FOXD3	ENSG00000187140	ENSDARG00000021032	no	**foxd3**	NA	NA	NA
22	G protein-coupled receptor 126 – GPR126	ENSG00000112414	ENSDARG00000054137	no	**gpr126**	NA	NA	NA
23	Hyperpolarization activated cyclic nucleotide-gated potassium channel 1 – HCN1	ENSG00000164588	ENSDARG00000077190	no	**hcn1**	NA	NA	NA
24	Huntingtin – HTT	ENSG00000197386	ENSDARG00000052866	no	**htt**	NA	NA	NA
25	iroquois homeobox 4 – IRX4	ENSG00000113430	ENSDARG00000035648	***yes***	**irx4a**	NA	NA	NA
					irx4b	NA	NA	NA
26	Kallmann syndrome 1 sequence – KAL1	ENSG00000011201	ENSDARG00000012896	yes	**kal1a**	yes	4 paralogs are present in Tetraodon, and 3 paralogs each in Zebrafish and Fugu, only 2 paralogs in Medaka and Stickleback	4 paralogs are present in Tetraodon, and 3 paralogs each in Zebrafish and Fugu, only 2 paralogs in Medaka and Stickleback

27	KIAA1279	ENSG00000198954	ENSDARG00000062053	no	kal1b, kal1 (2 of 2) **Kbp (kif1 binding protein)**	NA	NA	NA
28	kinesin family member 1B – KIF1B	ENSG00000054523	ENSDARG00000037020	no	**kif1b**	NA	NA	NA
29	lethal giant larvae homolog 1 – LLGL1	ENSG00000131899	ENSDARG00000009693	no	**llgl1**	NA		
30	lethal giant larvae homolog 2 – LLGL2	ENSG00000073350	ENSDARG00000023920	no	**llgl2**	NA		
31	lectin, mannose-binding 2-like – LMAN2L	ENSG00000114988	ENSDARG00000018865	**yes**	**lman2la (lman2l - 2 of 2)**	NA	NA	NA
					lman2lb	NA	NA	NA
32	N-ethylmaleimide-sensitive factor – NSF	ENSG00000073969	ENSDARG00000007654	***yes***	**nsfa**	NA	NA	NA
					nsfb	NA	NA	NA
33	proprotein convertase subtilisin/kexin type 5 – PCSK5	ENSG00000099139	ENSDARG00000067537	**yes**	**pcsk5a**	4 paralogs in Zebrafish	3 paralogs in Tetraodon all other fishes there are only 2 paralogs	yes
					pcsk5b, pcsk5 3 of 3, pcsk5 1 of 3			
34	transmembrane inner ear – TMIE	ENSG00000181585	ENSDARG00000069423	no	**tmie**	NA	NA	NA

a The Ensembl ID of the gene involved in the GO Process is shown, ID's of the paralogs if present are not shown;
b When there is a Duplicated copy in all fishes the paralogs are listed, if no duplicates are present the ortholog to the gene listed in column 1 for humans is written the genes in bold are the ones involved in the GO Process of lateral line system development.

TableS2: Results of the compartmentalization analysis (Frost et al., 2005) in HYPHY (Pond et al., 2004), showing the selective pressures on the different compartmentalized clades (fishes and tetrapods); Akike Information Criterion (AIC) ranks are shown for each of the five models, the model which would survive a 4 way multiple test correction is denoted by an asterisk (*) all the models with significant LRT – P value is underlined

Genes	Phase1	Phase2	Phase3	Phase4	Phase5
apc	182017.89769	181990.56304	181857.32598	181793.80887	***181755.27798****
atoh1a	20569.70465	20571.34484	20523.22034	***20522.81318****	20523.81299
atp1b2b	20518.44105	20515.54347	20498.8022	20499.86116	***20497.60387****
atp2b1a	80921.97691	80903.77966	80765.48107	80724.49799	***80722.55449****
cav1	10101.48144	10091.11475	10097.56691	10101.01245	***10090.0612****
cdh4	61944.86001	61938.02667	61899.40012	61908.37409	***61898.08662****
cldnb	13220.37079	13221.69803	***13210.46235****	13211.59887	13212.46343
col17a1b	139746.31916	139748.05866	139540.81662	***139536.92551****	139537.71116
cxcl12a	8327.97096	8329.93375	8320.55069	***8320.27649****	8322.2741
cxcr4b	32909.54127	32909.13047	32895.3864	***32894.05666****	32894.77315
cxcr7b	***11019.85519****	11022.47454	11020.40879	11020.41657	11021.56063
dkk1b	19752.28671	19751.95104	***19743.37615****	19744.5371	19744.50887
egr2b	35140.33607	35141.9239	***35116.37241****	35117.33822	35118.1666
erbb2	87306.86884	87290.07247	87290.16728	87296.4942	***87278.56747****
erbb3b	97126.89439	97128.8004	***96934.4656****	96938.91179	96935.60915
esr1	37869.2784	37870.31772	37841.05336	***37838.93447****	37840.66699
eya4	28166.84936	***28161.12683****	28168.57512	28167.91505	28162.73376
fgf3	***12842.42664****	12844.15047	12843.38833	12843.70645	12845.32737
fgf10a	14576.12569	14576.70509	***14523.34514****	14529.49011	14525.29879
fgfr1a	50861.08502	50830.67831	50662.16063	50709.03808	***50650.53004****
foxd3	15584.65769	15586.17237	***15581.89382****	15582.70811	15583.41135
gpr126	63368.3573	63243.59575	63322.35801	***63229.15335***	***63133.73321****
hcn1	***43730.53136***	43731.38282	43732.51747	43732.1369	43733.00974
htt	189238.95452	189236.75376	***189207.79386****	189212.5155	189209.64783
irx4a	36998.80089	37000.75645	***36986.16791****	36986.22486	36988.14078
kal1a	60317.53079	60319.54535	60294.11329	***60294.09495****	60296.10085
Kbp	38695.91636	***38682.11143****	38697.06938	38695.4681	38682.34938
kif1b	86365.76597	86363.8798	***86342.01013****	86347.59687	86343.65115
llgl1	66767.16659	66767.5862	***66750.54753****	66752.42232	66752.43241
llgl2	66736.76115	***66722.239****	66738.66647	66738.06775	66723.8892
lman2la	29384.51144	29385.34078	***29348.7676****	29350.94285	29350.48439

nsfa	46165.85753	46166.77123	***46031.52127****	46040.82573	46032.82909
pcsk5a	***129682.64098 ****	129683.439	129684.6071	129684.63885	129685.4387
tmie	***10784.90583****	10785.52962	10786.29773	10786.67476	10787.37192

Phase1 = $\omega_F = \omega_T = \omega_{anc\text{-}fish}$
Phase2 = $\omega_F = \omega_T \neq \omega_{anc\text{-}fish}$
Phase3 = $\omega_F = \omega_{anc\text{-}fish} \neq \omega_T$
Phase4 = $\omega_F \neq \omega_{anc\text{-}fish} = \omega_T$
Phase5 = $\omega_F \neq \omega_{anc\text{-}fish} \neq \omega_T$
ω_F = dN/dS ratio of fish clade
ω_T = dN/dS ratio of tetrapod clade
$\omega_{anc\text{-}fish}$ = dN/dS ratio of ancestral fish branch

Table S3: The likelihood values and likelihood ratio test statistics for branch-site test of positive selection with the whole teleost clade or the terminal teleost branches as foreground branches

a) Branch-site models applied to the teleost terminal branches			
Gene	**Null Model**	**Alternate Model**	**2ΔlnL***
atp1b2b	-10269.722557	-10269.722557	0
atoh1a	-10109.424091	-10109.424072	0.000038
apc	-79221.788615	-79180.351384	**82.874462**
htt	-93758.927026	-93770.876448	**23.898844**
hcn1	-21596.981253	-21594.058047	5.846412
gpr126	-30966.240082	-30951.689928	**29.100308**
foxd3	-7677.149695	-7676.937568	0.424254
fgfr1a	-25137.118486	-25137.118486	0
fgf10a	-7130.647777	-7130.637629	0.020296
llgl2	-33412.695475	-33411.619021	2.152908
lman2la	-14390.245705	-14390.245705	0
nsfa	-22955.3709	-22955.3709	0
pcsk5a	-64260.638514	-64260.638514	0
irx4a	-18215.184129	-18215.184129	0
kal1a	-28605.865408	-28605.865408	0
kbp	-19353.874455	-19353.874455	0
llgl1	-33054.836974	-33054.836974	0
tmie	-5344.711657	-5344.711657	0
atp2b1a	-38076.317471	-38076.317471	0
cav1	-4939.994098	-4939.994098	0
cdh41	-31249.463257	-31249.463257	0
cldnb	-6613.279727	-6610.98622	4.587014
col17a1b	-68944.506662	-68944.506662	0
cxcl12a	-4041.550913	-4040.456251	2.189324
cxcr4b	-16072.902208	-16072.902208	0
cxcr7b	-5549.261741	-5549.261741	0
dkk1b	-9666.897344	-9666.040032	1.714624
fgf3	-6295.807026	-6295.807026	0
eya4	-13893.268442	-13893.268164	0.000556
esr1	-18666.021676	-18664.318714	3.405924
erbb3b	-47470.453643	-47470.453643	0
erbb2	-43035.854976	-43035.854976	0
egr2b	-17181.694991	-17181.694991	0
kif1b	-42905.549247	-42905.549248	0.000002
b) Branch-site models applied to the whole teleost clade			
Gene	**Null model**	**Alternate Model**	**2ΔlnL***
apc	-78893.005695	-78893.005696	0
atoh1a	-10100.043905	-10100.043905	0
atp1b2b	-10237.142499	-10237.142499	0
atp2b2a	-37959.581101	-37959.581101	0
cav1	-4937.11817	-4937.11817	0
cdh4	-31232.491741	-31232.491741	0
cldnb	-6607.841895	-6607.841895	0
col17a1b	-68744.293501	-68744.293501	0
cxcl12a	-4034.504539	-4034.504539	0
cxcr4b	-16068.24404	-16068.24404	0

cxcr7b	-5546.190408	-5546.190408	0
dkk1b	-9641.616668	-9641.616668	0
egr2b	-17179.068246	-17179.068246	0
erbb2	-42955.872682	-42955.872682	0
erbb3b	-47321.776647	-47321.776647	0
esr1	-18609.759581	-18609.759581	0
eya4	-13878.760461	-13878.760461	0
fgf3	-6296.228302	-6296.228302	0
fgf10a	-7116.741129	-7116.741129	0
fgfr1a	-25104.242541	-25104.242541	0
foxd3	-7663.577546	-7663.577546	0
gpr126	-30945.566323	-30945.566323	0
hcn1	-21574.167754	-21574.167754	0
htt	-93618.209871	-93618.209871	0
irx4a	-18167.783011	-18167.783011	0
kal1a	-28581.260172	-28581.260172	0
kbp	-19327.852432	-19327.852432	0
kif1b	-42905.549247	-42905.549248	
llgl1	-33062.122641	-33062.122641	0
llgl2	-33361.326031	-33361.326031	0
lman2la	-14383.251145	-14383.251145	0
nsfa	-22906.896139	-22906.896139	0
pcsk5a	-64157.617609	-64157.617609	0
tmie	-5344.711657	-5344.711657	0

*When a LRT is significant at *p-value* < 0.01, it is indicated in boldface.

14

Appendix 4 (Supplementary materials for Chapter 6)

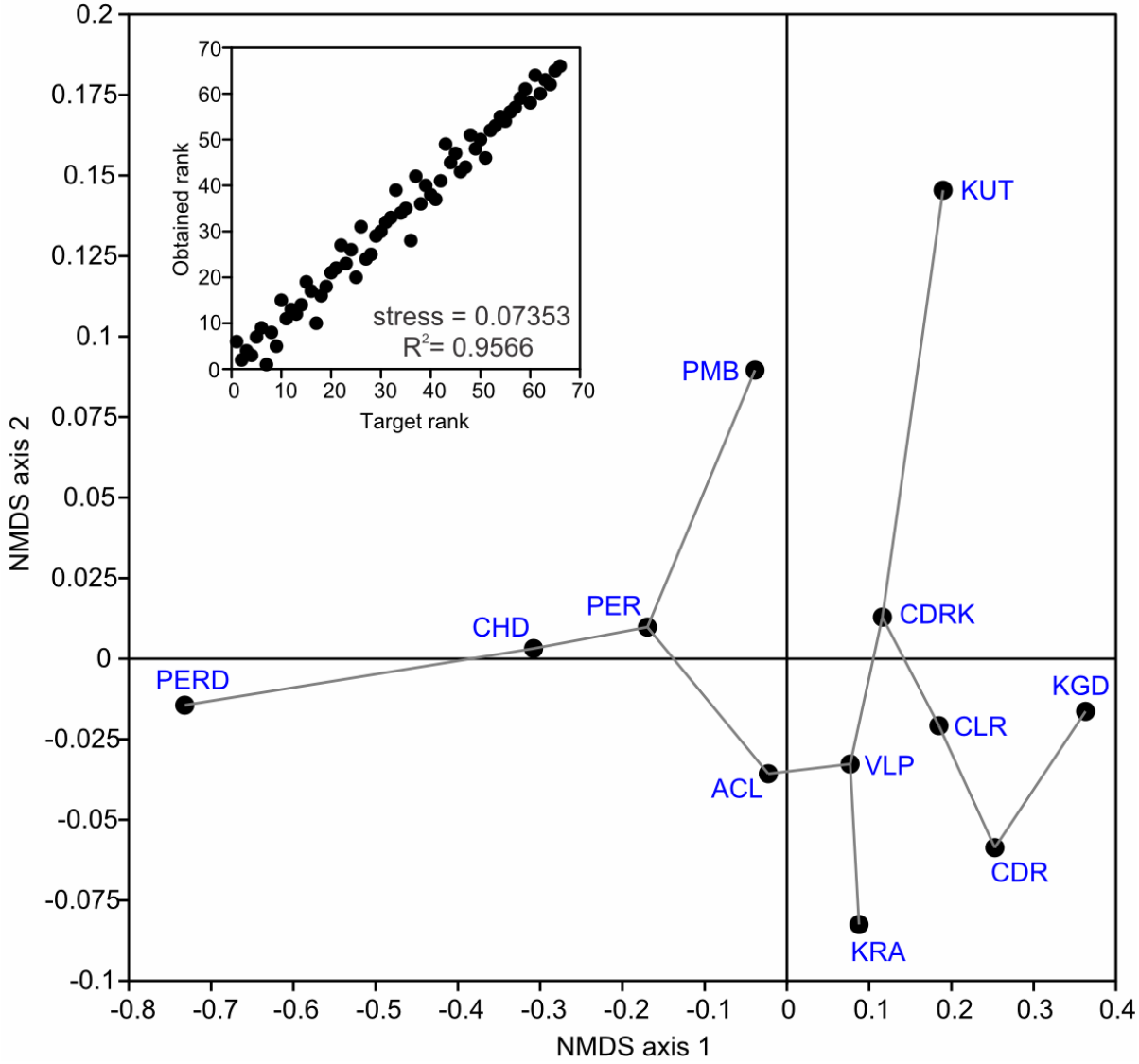


Figure 14.1.: Non-metric multidimensional scaling of DFA functions at the centroid using Euclidian distances. Connecting line is the minimum span tree. Shephard plot is shown in the inset.

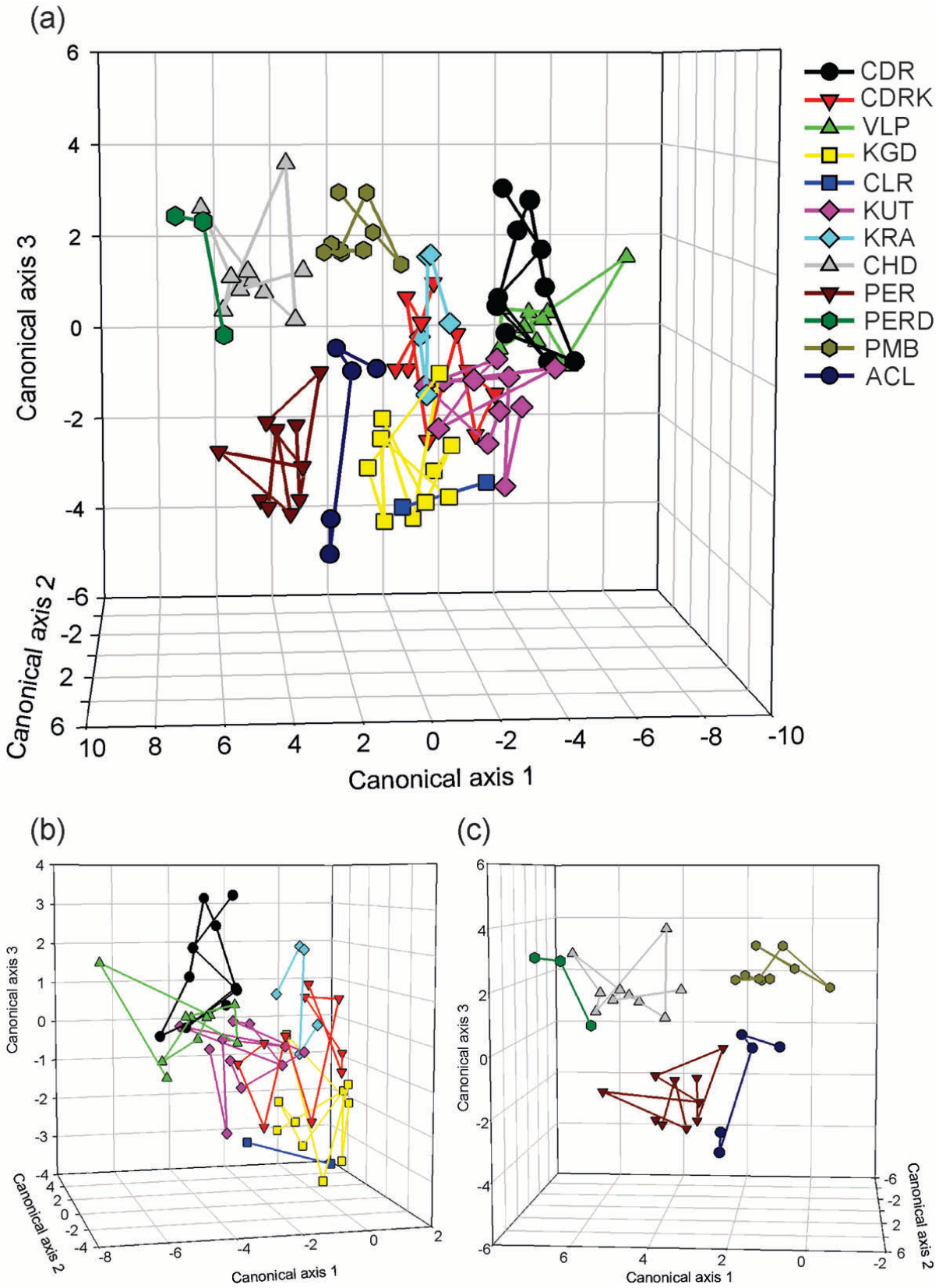


Figure 14.2.: MANOVA/CVA on the on the first three canonical axes. (a) Clusters of all 12 populations on the first three canonical axes, (b) clusters of populations north of Palghat gap and (c) clusters of populations south of Palghat gap. Points are connected by line just for eyeballing the clusters.

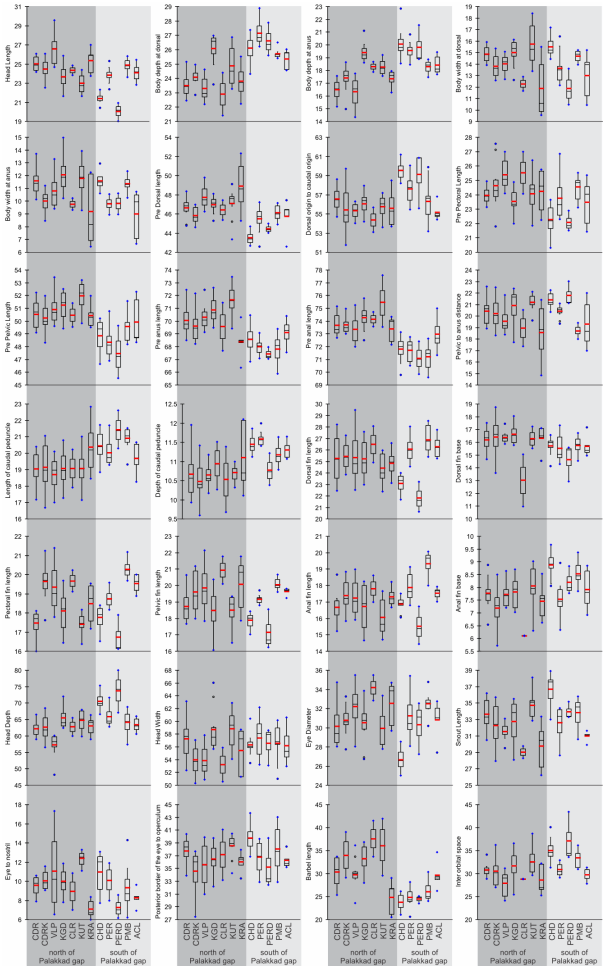


Figure 14.3.: Box plot of size adjusted morphometric characters. Redline is the mean.

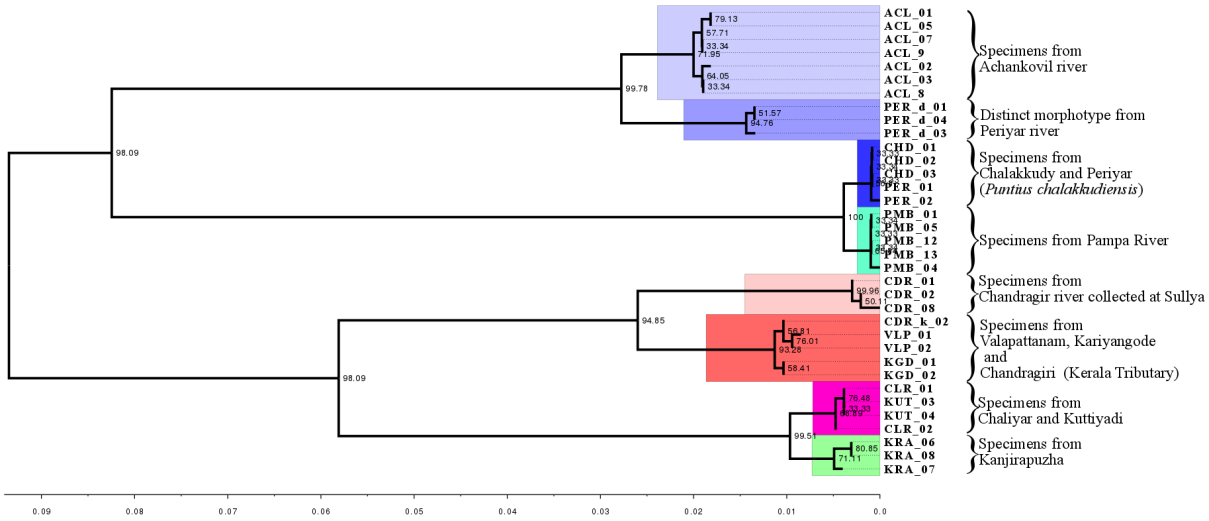


Figure 14.4.: Phylogenetic tree constructed using the concatenated alignment showing the relationships between the specimens collected from different river systems throughout their range, shLRT node support are shown, right side of the tree has each group labeled with their river of origin.

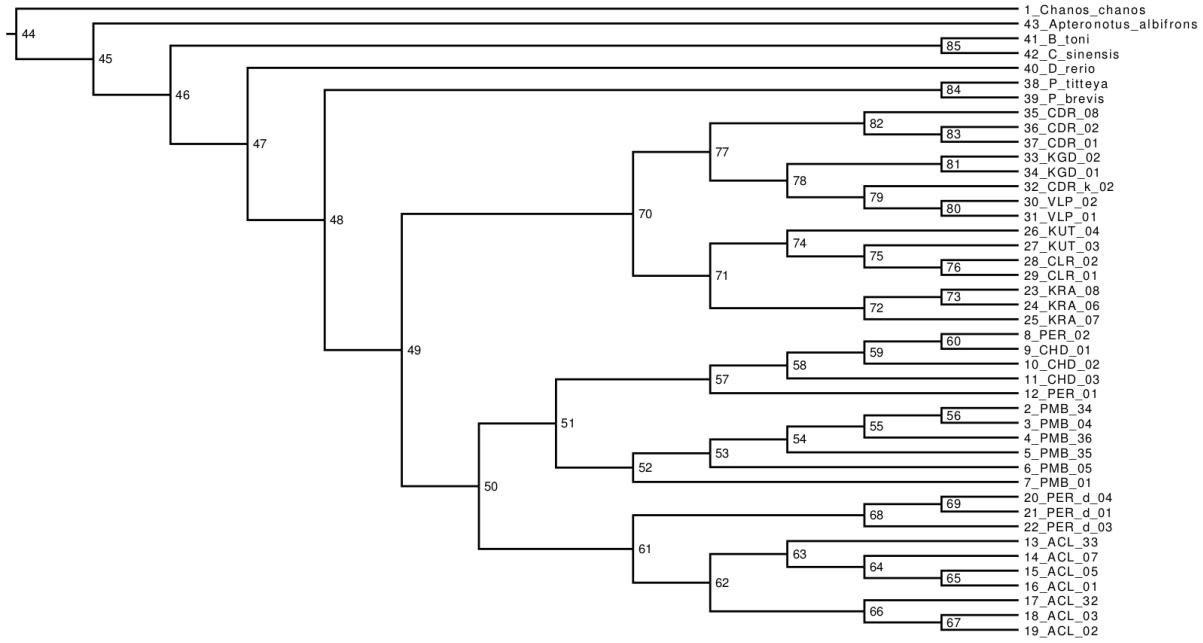


Figure 14.5.: Cladogram with corresponding node numbers for which the divergence times are presented in the table S5, tips have their numbers as the prefix followed by an underscore and the specimen name

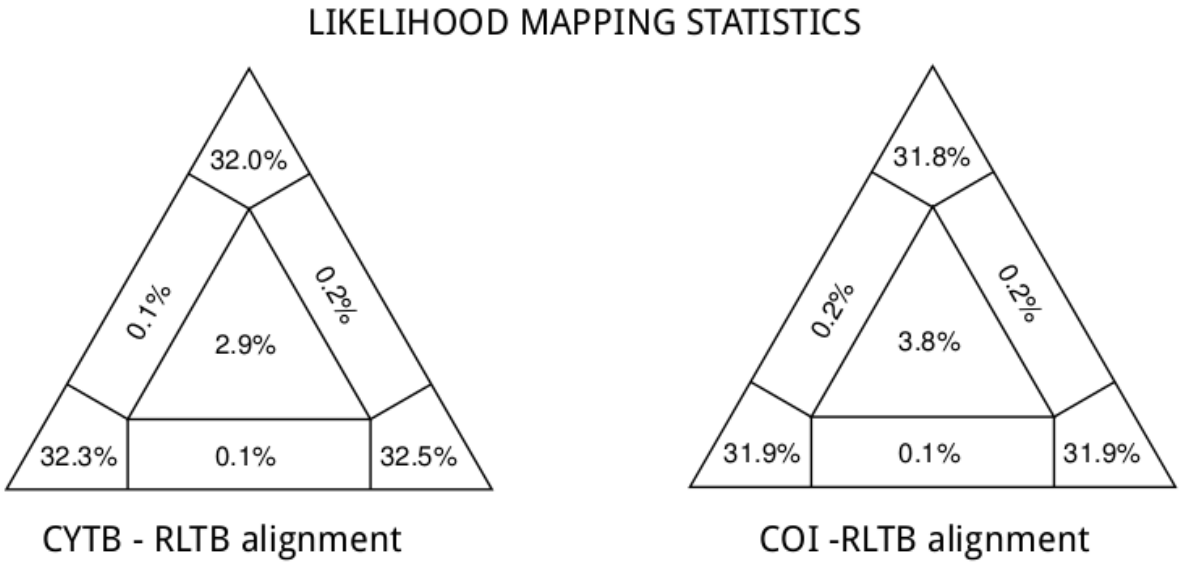


Figure 14.6.: Results of the likelihood mapping procedure for the CYTb and COI alignments used in this study, note that more than 90% of the quartets are resolved in both cases.

Table S1. Analysis of Variance of size adjusted characters.

Characters	$F_{11,83}$	*P*
As %SL		
Head Length	18.544	< 0.0001
Body depth at dorsal	16.223	< 0.0001
Body depth at anus	17.750	< 0.0001
Body width at anus	5.632	< 0.0001
Pre Dorsal length	9.243	< 0.0001
Dorsal origin to caudal origin	6.428	< 0.0001
Pre Pectoral Length	4.493	< 0.0001
Pre Pelvic Length	7.907	< 0.0001
Pre anus length	11.417	< 0.0001
Pre anal length	14.178	< 0.0001
Length of caudal peduncle	4.399	< 0.0001
Depth of caudal peduncle	5.999	< 0.0001
Dorsal fin length	5.323	< 0.0001
Pectoral fin length	12.549	< 0.0001
Pelvic fin length	5.789	< 0.0001
Anal fin length	7.655	< 0.0001
Anal fin base	6.048	< 0.0001
As %HL		
Head Depth	10.805	< 0.0001
Head Width	3.078	0.002
Eye Diameter	6.270	< 0.0001
Snout Length	7.279	< 0.0001
Eye to nostril	4.384	< 0.0001
Posterior border of the eye to operculum	2.723	0.005
Inter orbital space	6.244	< 0.0001

Table S2. MANOVA/CVA loadings for the first three canonical axes

Variable	Axis 1	Axis 2	Axis 3
Head Length	-1.0364	-0.3378	0.5413
Body depth at dorsal	0.4461	-0.4208	-0.3340
Body depth at anus	0.4143	0.3288	-0.6828
Body width at dorsal	-0.0005	-0.2239	-0.0034
Body width at anus	-0.4787	0.0597	0.3816
Dorsal origin to caudal origin	0.1388	0.0651	-0.0706
Pre Pectoral Length	-0.0115	-0.5478	-0.0668
Pre Pelvic Length	0.2428	0.3370	-0.1624
Pre anus length	-0.1227	-0.0760	-0.4543
Pre anal length	-0.5179	0.4469	0.1920
Length of caudal peduncle	0.0809	-0.2017	0.6069
Depth of caudal peduncle	0.7643	0.1929	-0.2190
Dorsal fin length	-0.0285	-0.0981	-0.4304
Pectoral fin length	0.0639	-0.8844	0.1344
Pelvic fin length	-0.2806	0.3896	0.2120
Anal fin length	0.7158	-0.4653	-0.1788
Anal fin base	-0.1213	0.2696	0.9329
Head Depth	0.0619	0.0274	0.1190
Head Width	-0.0654	-0.1287	-0.0967
Eye Diameter	-0.1283	-0.0016	-0.0499
Snout Length	-0.0750	0.0785	0.2402
Eye to nostril	-0.1422	-0.1402	-0.1485
Posterior border of the eye to operculum	-0.0285	0.0383	0.0150
Inter orbital space	-0.1294	-0.0948	0.1986

Table S3. Discriminant functions for the 12 populations

	Populations north of Palakkad gap							Populations south of Palakkad gap			
	CDR	CDRK	VLP	KGD	CLR	KUT	KRA	CHD	PER	PERD	PMB
Intercept	-6201.417	-6061.030	-6152.064	-6255.463	-6201.647	-6028.999	-6070.577	-5937.168	-6038.358	-5801.318	-6124.918
%SL											
Head Length	116.966	109.763	110.032	117.050	107.860	105.221	115.580	109.290	111.949	109.610	115.796
Body depth at dorsal	8.321	10.444	10.020	9.073	7.634	7.420	9.025	11.581	17.658	17.834	13.766
Body depth at anus	-18.599	-15.827	-11.040	-18.477	-13.905	-9.849	-15.127	-13.965	-14.779	-15.180	-19.939
Body width at dorsal	20.149	21.288	19.063	20.741	21.592	21.564	17.760	21.171	20.259	15.001	22.809
Body width at anus	9.591	6.393	7.776	9.896	6.685	5.073	8.743	4.399	2.658	3.568	6.887
Pre Dorsal length	12.236	11.450	13.039	12.493	14.091	12.927	12.404	10.279	9.611	10.057	8.655
Dorsal origin to caudal origin	26.672	25.995	26.669	26.561	26.624	26.628	26.178	28.112	28.348	28.288	26.565
Pre Pectoral Length	3.635	6.342	2.123	5.097	4.050	6.288	0.865	3.696	6.592	4.131	8.097
Pre Pelvic Length	-23.755	-23.002	-20.606	-24.742	-21.089	-21.232	-21.262	-21.101	-22.194	-19.730	-24.645
Pre anus length	34.842	34.622	35.154	36.618	36.098	36.600	31.326	32.546	33.906	29.768	33.541
Pre anal length	55.328	54.361	53.342	54.056	54.694	52.239	55.193	51.053	48.920	49.969	51.713
Pelvic to anus distance	-7.732	-6.470	-5.866	-7.977	-7.192	-6.518	-5.890	-5.252	-4.064	-0.310	-8.381
Length of caudal peduncle	18.109	18.300	15.944	17.492	16.862	16.661	19.075	18.905	17.870	21.191	22.018
Depth of caudal peduncle	9.541	5.572	11.587	8.075	10.065	6.552	10.066	15.336	12.804	6.802	9.616
Dorsal fin length	-7.998	-7.534	-7.303	-8.187	-6.995	-6.521	-9.911	-9.669	-7.328	-10.264	-9.016
Dorsal fin base	-24.671	-23.253	-22.832	-24.419	-24.185	-27.950	-23.020	-27.687	-28.325	-28.940	-28.126
Pectoral fin length	23.900	30.457	25.799	28.567	24.806	27.047	24.066	27.151	25.691	22.034	31.348
Pelvic fin length	-16.181	-16.838	-18.018	-16.522	-17.098	-15.964	-13.481	-16.619	-17.703	-12.778	-17.678
Anal fin length	-18.097	-16.276	-15.531	-19.047	-17.098	-14.204	-17.388	-12.102	-11.352	-16.695	-11.049
Anal fin base	3.583	-1.018	-0.017	2.786	0.232	-5.268	5.231	5.614	0.979	7.445	6.406
%HL											
Head Depth	12.149	12.121	11.756	11.723	11.678	11.607	12.400	13.029	12.628	13.485	12.583
Head Width	10.765	9.891	10.440	10.588	9.863	9.607	10.956	9.834	11.163	9.265	11.763
Eye Diameter	9.848	8.773	9.433	10.274	8.993	9.478	10.380	9.232	10.506	11.470	10.138
Snout Length	5.131	3.997	3.212	4.679	4.060	3.092	3.795	4.872	3.917	5.447	4.467
Eye to nostril	14.415	15.145	14.562	15.665	15.600	15.520	12.807	13.968	15.077	11.969	14.346
Posterior border of the eye to operculum	-4.104	-4.440	-4.169	-4.202	-4.019	-3.655	-4.102	-4.113	-4.539	-5.019	-4.621
Barbel length	-1.338	-0.317	-0.181	-1.296	0.398	0.674	-2.068	-2.303	-2.139	-2.374	-2.893
Inter orbital space	12.229	11.595	10.484	12.206	10.636	10.601	12.090	11.527	11.157	12.091	13.513

Table S4: Detailed results of the GMYC methods implemented for the *cytb* ultrametric tree and the Concatenated ultrametric tree, the species distinction made is displayed in the Figure 2b in the main text.

Result of GMYC species delimitation using the Concatenated ultrametric tree:

Method: single

Likelihood of null model:	171.6296
Maximum likelihood of GMYC model:	187.7243
Likelihood ratio:	32.18945
Result of LR test:	4.773749e-07***
Number of ML clusters:	**7**
Confidence interval:	7-8
Number of ML entities:	27
Confidence interval:	11-27
Threshold time:	-0.0001780219

Method: multiple

Likelihood of null model:	171.6296
Maximum likelihood of GMYC model:	189.5913
Likelihood ratio:	35.92345
Result of LR test:	9.838917e-07***
Number of ML clusters:	9 (clusters reported in figure 2d)
Confidence interval:	8-9
Number of ML entities:	15
Confidence interval:	12-19
Threshold time:	-0.03990777
	-0.01748675
	-0.004642341

Comparison of single and multiple threshold GMYC

Chi-square = 3.7340, df = 6, P = 0.7126

Result of GMYC species delimitation using the CYTb ultrametric tree:

Method: single

Likelihood of null model:	187.1918
Maximum likelihood of GMYC model:	209.7168
Likelihood ratio:	45.05
Result of LR test:	9.029038e-10***
Number of ML clusters:	8
Confidence interval:	8-8
Number of ML entities:	23
Confidence interval:	23-23
Threshold time:	-0.0002175

Method: multiple

Likelihood of null model:	187.1918
Maximum likelihood of GMYC model:	198.8942
Likelihood ratio:	23.40481
Result of LR test:	0.000282415***
Number of ML clusters:	6 (clusters reported in figure 2d)
Confidence interval:	6-6
Number of ML entities:	10
Confidence interval:	9-10
Threshold time:	-0.1688247
	-0.008929472
	-0.005099265

Comparison of single and multiple threshold GMYC

Chi-square = 21.6452, df = 6, P = 0.0014

Table S5: Table showing the divergence times for the RLTB's internal node number are in the first column which follows the Figure S5.

Node	95% confidence interval (x100 Ma)		mean (Ma)
n44	1.4385	–2.3455	177.78
n45	1.3785	–2.292	172.79
n46	1.0906	–1.9135	141.75
n47	0.8574	–1.5553	113.95
n48	0.6956	–1.2808	93.38
n49	0.4074	–0.8001	56.99
n50	0.2817	–0.5913	41.01
n51	0.0206	–0.0769	4.23
n52	0.0048	–0.0317	1.47
n53	0.003	–0.0228	1.02
n54	0.0018	–0.0165	0.71
n55	0.0009	–0.0121	0.47
n56	0.0003	–0.0089	0.29
n57	0.0035	–0.0271	1.2
n58	0.002	–0.0193	0.81
n59	0.001	–0.0139	0.54
n60	0.0004	–0.0099	0.33
n61	0.0632	–0.1767	10.82
n62	0.01	–0.0488	2.45
n63	0.0031	–0.0255	1.13
n64	0.0015	–0.018	0.74
n65	0.0001	–0.0076	0.21
n66	0.002	–0.0235	0.98
n67	0.0007	–0.0155	0.56
n68	0.0036	–0.0352	1.47
n69	0.0001	–0.0129	0.35
n70	0.2859	–0.6125	42.14
n71	0.0348	–0.1177	6.73
n72	0.0056	–0.0406	1.86
n73	0.0001	–0.0121	0.33
n74	0.003	–0.028	1.18
n75	0.0014	–0.0179	0.72
n76	0.0005	–0.0121	0.43
n77	0.1067	–0.2812	17.81
n78	0.009	–0.0485	2.38
n79	0.0032	–0.0272	1.2
n80	0.0006	–0.0145	0.5
n81	0.0001	–0.0172	0.47
n82	0.0055	–0.0421	1.9
n83	0.0009	–0.0199	0.7
n84	0.4691	–0.9559	67.87
n85	0.7776	–1.4534	105.25

Table S6. Confusion matrix for group identity based on discriminant functions. Populations in the row are original identities. Populations in the column are predicted identities. Diagonal elements indicate correct prediction of group identity. Off diagonal elements show wrong predictions

from \ to	CDR	CDRK	VLP	KGD	CLR	KUT	KRA	CHD	PER	PERD	PMB	ACL	Total	% correct
CDR	**10**	0	0	0	0	0	0	0	0	0	0	0	10	100.00
CDRK	0	**9**	0	0	***1***	0	0	0	0	0	0	0	10	90.00
VLP	0	0	**10**	0	0	0	0	0	0	0	0	0	10	100.00
KGD	0	0	0	**10**	0	0	0	0	0	0	0	0	10	100.00
CLR	0	0	0	0	**11**	0	0	0	0	0	0	0	11	100.00
KUT	0	0	0	0	0	**2**	0	0	0	0	0	0	2	100.00
KRA	0	0	0	0	0	0	**5**	0	0	0	0	0	5	100.00
CHD	0	0	0	0	0	0	0	**10**	0	0	0	0	10	100.00
PER	0	0	0	0	0	0	0	0	**10**	0	0	0	10	100.00
PERD	0	0	0	0	0	0	0	0	0	**3**	0	0	3	100.00
PMB	0	0	0	0	0	0	0	0	0	0	**9**	0	9	100.00
ACL	0	0	0	0	0	0	0	0	0	0	0	**5**	5	100.00
Total	10	9	10	10	12	2	5	10	10	3	9	5	95	98.95

Table S7

Samples procured from aquarium collectors, corresponding river systems and number of samples used

River System@	Collection Sites	n- MOR	n-MOL
Chandragiri/CDR[1]	Sullya	10	3
Chandragiri/CDRK[1]	Kottody, Nagapattinam	10	1
Karyangode/KGD[1]	Cherupuzha	10	2
Valapattanam/VLP[2]	Iritty	10	2
Bharatapuzha/KRA[3]	Kanjirapuzha	5	2

[1] Collected in 2005 and 2006
[2] Collected in 2010
[3] Collected in 2008
n-MOR: number of samples used for morphological analysis.
n-MOL: number of samples used for molecular analysis.

List of sampling sites from where we collected samples directly, corresponding river systems and number of samples used

River System@	Collection Sites	n- MOR	n-MOL
Kuttyadi/KUT[$, 1]	Chathangothunada	2	2(1)
Chaliyar/CLR[$, 2]	Chalipuzha, Pullooranpara	11(7)	2
Chalakudy/CHD[#, 3]	Athirapilly, Vettilapara	10(2)	3(1)
Periyar/PER[#, 3]	Paniyeli, Pooyamkutty	10(6)	2
Periyar/PERD[#, 3]	Pooyamkutty	3	3(2)
Pampa/PMB[#, 3]	Angel Valley, Azhutha, Koruthodu	9(2)	6(4)
Achankovil/ACL[#, 3]	Mukkada, Chuttipara, Kadakkola	5(3)	8(3)

[1]Collected in 2008
[2] Collected in 2007 and 2009
[3] Collected in 2010
[$]no permits were required
[#]permits were required

n-MOR: number of samples used for morphological analysis; numbers in parenthesis denotes number of fishes released back after measurements/sampling (*).
n-MOL: number of samples used for molecular analysis; numbers in parenthesis denotes number of fishes released back after measurements/sampling (*).

(*) Small and delicate fish could not be in many cases released back into the wild because manipulation and measurements of these animals eventually lead to their death. The impact of the loss of these small fishes will have minimal environmental effects.

Table S8. Genbank details of the sequences used in the study

Sl.No	Trace file ID*	Sequence ID in Paper/voucher codes	Accession No. COI	Accession No. CytB	Source
1		CDR01	GQ247550	GQ247558	NCBI
2	023	CDR02	GQ247551 (NCBI)	JX470422	This study+NCBI
3	109	CDR03	JX462903	JX462890	This study
4	CDRK2	CDRK	JX462866	JX462876	This study
5		KGD01	GQ247554	GQ247559	NCBI
6	035	KGD02	JX470428	JX470423	This study
7		VLP01	GQ247555	GQ247561	NCBI
8	005	VLP02	JX470427	JX470421	This study
9		CLR01	GQ247552	GQ247560	NCBI
10	171	CLR02	GQ247553 (NCBI)	JX470426	This study+NCBI
11		CHD01	GQ247549	GQ247556	NCBI
12	043	CHD02	JX481180	JX470424	This study
13	49	CHD03	JX462904	JX462891	This study
14	052	PER01	JX481181	JX470425	This study
15	053	PER02	JX470429	GQ247557(NCBi)	This study+NCBI
16	215	PERD03	JX462905	JX462892	This study
17	288	PERD04	JX462906	JX462893	This study
18	289	PERD05	JX462907	JX462894	This study
19	301	PMB01	JX462908	JX462895	This study
20	305	PMB02	JX462909	JX462896	This study
21	306	PMB03	JX462910	JX462897	This study
22	ACL1	ACL01	JX462898	JX462898	This study
23	ACL2	ACL02	JX462899	JX462886	This study
24	ACL3	ACL03	JX462900	JX462887	This study
25	ACL5	ACL05	JX462901	JX462888	This study
26	ACL7	ACL07	JX462902	JX462889	This study
27	KRA6	ACL06	JX462867	JX462877	This study
28	KRA7	KRA07	JX462868	JX462878	This study
29	KRA8	KRA08	JX462869	JX462879	This study
30	KUT3	KUT03	JX462874	JX462883	This study
31	KUT4	KUT04	JX462875	JX462884	This study
32	MLA1	PMB11	JX481187	JX481182	This study
33	MLA2	PMB12	JX481185	JX481183	This study
34	MLA3	PMB13	JX481186	JX481184	This study
35	ACH2	ACL8	JX481188	JX470430	This study
36	ACH3	ACL9	JX481189	JX470431	This study

* the code (recognizable in the filename) for the sequence trace files uploaded at figshare
http://dx.doi.org/10.6084/m9.figshare.95635

15

Appendix 5 (Supplementary materials for Chapter 7)

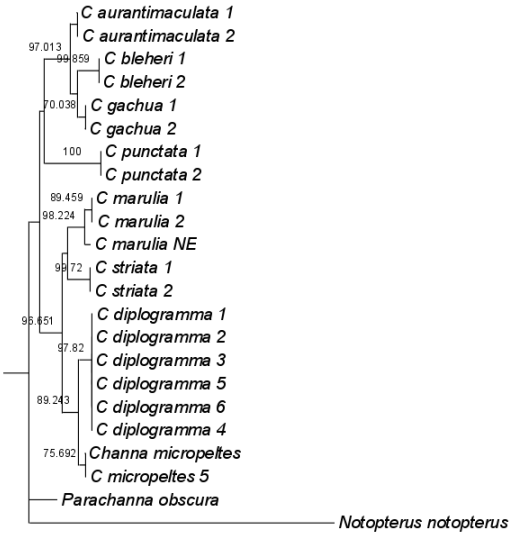


Figure 15.1.: Phylogenetic tree of the channid species used in the study with partial mitochondrial 16S rRNA gene sequences, rooted with Notopterus notopterus. Bootstrap values below 60 are not shown.

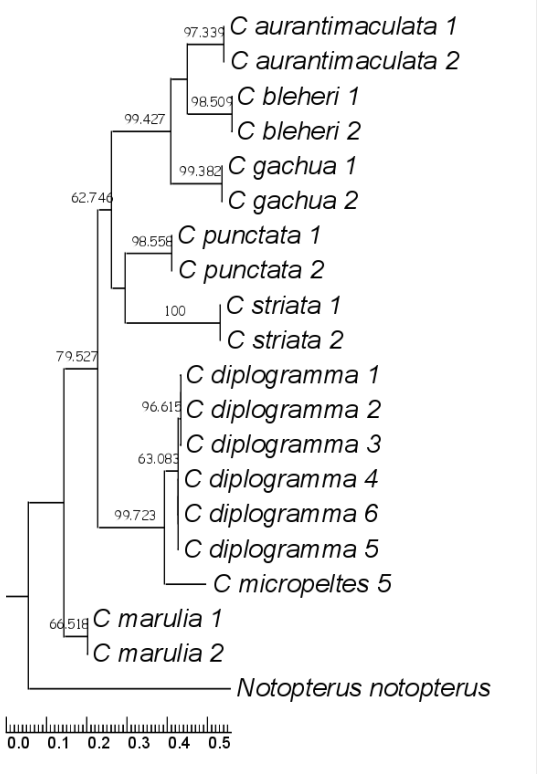


Figure 15.2.: Phylogenetic tree of the channid species used in the study with partial mitochondrial COI gene sequences, rooted with Notopterus notopterus. Bootstrap values below 60 are not shown.

Figure 15.3.: Photographs showing the gular scales of C.diplogramma (left) and C. micropeltes (right).

Table S1: Morphometric and meristic measurements of the specimen used in this study; the table contain measurements of the holotype of *C. diplogramma* (BMNH 1865.7.17.24 unique Holotype; NMW 73835, NMW 73838 and NMW 84220 Day's specimen) and Syntpes of *C. micropeltes* (RMNH D2318 Syntype; RMNH D1131&D1132 possible Syntypes); measurements are in millimeters.

	*UMT CM#1	UMT CM2	UMT CM3	UMT CM4	UMT CM5
	***Channa micropeltes* specimen measured in this study**				
Total length (mm)	433.35	477.03	607.24	654.93	338.93
Standard length (mm)	313.82	394.91	511.88	564.22	290.87
Body weight (g)	800	1200	2890	3300	355
Head Length (mm)	123.61	140.48	174.22	181.82	102.39
Pre dorsal length (mm)	117.91	131.31	165.1	172.1	95.01
Pre pectoral length (mm)	121.32	129.3	161.89	177.95	103.45
Pre pelvic length (mm)	131.72	135.37	185.06	204.12	106.05
Pre anal length (mm)	179.12	195.82	251.83	285.79	135.78
Body depth (mm)	76.37	89.01	136.05	143.18	66.74
Dorsal fin rays	43	43	44	44	43
Pectoral fin rays	16	16	17	17	17
Pelvic fin rays	6	6	6	6	6
Anal fin rays	28	29	27	28	28
Caudal fin rays	14	14	14	14	14
Lateral line scales	86	86	86	86	86
Scales below lateral line	16	16	16	16	16
Cheek scales	24	25	24	25	23
Gular scales	21	39	18	36	39
Total vertebrae	57	57	57	57	57

	*CRG-CHDIP# 20	CRG-CHDIP 21	CRG-CHDIP 22	CRG-CHDIP 23	CRG-CHDIP 24	CRG-CHDIP 25	CRG-CHDIP 26	CRG-CHDIP 27	CRG-CHDIP 28	CRG-CHDIP 29
	***Channa diplogramma* specimen measured in this study**									
Total length (mm)	445.56	328.21	488.88	589.19	525	148.56	189.72	129.75	107.24	172.39

Standard length (mm)	361.16	266.6	389.69	479.15	430.05	116.53	149.92	100.09	85.4	137.88
Body weight (g)	1200	400	1200	1700	160	234	541	149	78	405
Head Length (mm)	118	83.71	125.06	119.93	132.41	38.74	50	35.4	29.32	45.45
Pre dorsal length (mm)	118.72	91.11	132.66	150.8	138.92	43.59	53.64	38.42	33.09	48.53
Pre pectoral length (mm)	111.95	85.3	123.59	148.42	141.88	44.24	55.08	38.28	33.11	50.84
Pre pelvic length (mm)	125.65	89.64	133.14	152.76	151.75	46.8	57.23	42.2	34.17	53.82
Pre anal length (mm)	187.49	142.8	210.13	238.89	240.68	68.39	83.8	60.08	51.45	77.98
Body depth (mm)	92.48	60.29	97.79	84.63	91.44	20.57	26.25	15.86	12.09	24.04
Dorsal fin rays	43	43	43	43	44	44	43	43	43	43
Pectoral fin rays	17	17	17	17	17	17	17	17	17	17
Pelvic fin rays	6	6	6	6	6	6	6	6	6	6
Anal fin rays	28	28	26	28	27	27	27	28	28	28
Caudal fin rays	15	15	15	15	15	15	16	17	15	15
Lateral line scales	103	104	104	103	105	105	104	105	104	105
Scales below lateral line	15	15	15	15	15	15	15	15	15	15
Cheek scales	16	18	16	18	20	18	19	17	20	16
Gular scales	30	31	31	31	30	31	31	30	30	31
Total vertebrae	53	54	53	54	54	53	53	54	54	54

	***BMNH 1865.7.17.24**	***NMW 73835**	**NMW 73838**	**NMW 84220**
	***Channa diplogramma* type specimen measured in this study**			
Total length (mm)	97.1	424	275	459
Standard length (mm)	81.6	352	230	380
Body weight (g)	---	114	76.8	125
Head Length (mm)	28.4	109	77.5	120.5
Pre dorsal length (mm)	29.5	113.5	86.1	117.2
Pre pectoral length (mm)	29.1	120	114	131
Pre pelvic length (mm)	29.6	171.6	46.7	185

Pre anal length (mm)	43.8	65.7	---	68.8
Body depth (mm)	11.9	---	---	---
Dorsal fin rays	44	44	45	43
Pectoral fin rays	18	18	18	18
Pelvic fin rays	6	6	6	6
Anal fin rays	28	28	28	27
Caudal fin rays	14	14	14	14
Lateral line scales	---	106	107	106
Scales below lateral line	---	20	21	20
Cheek scales	21	20	22	19
Gular scales	---	---	---	---
Total vertebrae	55	---	---	---

	***RMNH D2318**	**RMNH D1131**	**RMNH D1132**
	***Channa micropeltes* type specimen measured in this study**		
Total length (mm)	710	261	301
Standard length (mm)	605	210	250
Body weight (g)	---	---	---
Head Length (mm)	185	59	---
Pre dorsal length (mm)	---	---	---
Pre pectoral length (mm)	---	---	---
Pre pelvic length (mm)	---	---	---
Pre anal length (mm)	---	---	---
Body depth (mm)	---	---	---
Dorsal fin rays	44	43	43
Pectoral fin rays	17	17	17
Pelvic fin rays	6	6	6
Anal fin rays	27	28	28
Caudal fin rays	14	14	14

Lateral line scales	---	---	---
Scales below lateral line	---	---	---
Cheek scales	---	---	---
Gular scales	---	---	---
Total vertebrae	---	---	---

***BMNH** – Natural History Museum, London, United Kingdom; ***RMNH** - Rijksmuseum van Natuurlijke Histoire RMNH/Naturalis, Leiden, The Netherlands; ***NHM** – Natural History Museum, Vienna, Austria; ***UMT** – Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia; ***CRG**- Conservation Research Group, Department of Aquaculture, St. Albert's College, Kochi, India.

#Voucher specimens of *C. diplogramma* examined in our study are currently deposited at the museum of CRG, Department of Aquaculture. St. Albert's College, Kochi, India (CRG-CHDIP-20-CRG-CHDIP- 29), while those of *C. micropeltes* at the Museum of the Institute of Tropical Aquaculture, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia (UMTCM1 to UMTCM5)

Table S2: List of fish species used for the study, their NCBI accession numbers, Voucher Numbers and the respective museums of deposition of the specimen

Species	NCBI-Accession Number	Gene region (partial)	Voucher Number	Museum	Remarks
Channa diplogramma	EU342210	COI	CD6	FGL-FMNC	Specimen H in figure 3
Channa diplogramma	EU342209	COI	CD5	FGL-FMNC	Specimen G in figure 3
Channa diplogramma	EU342208	COI	CD4	FGL-FMNC	Specimen E in figure 3
Channa diplogramma	EU342207	COI	CD3	FGL-FMNC	Specimen D in figure 3
Channa diplogramma	EU342206	COI	CD2	FGL-FMNC	Specimen C in figure 3
Channa diplogramma	EU342205	COI	CD1	FGL-FMNC	Specimen A in figure 3
Channa striata	EU342204	COI	CS2	FGL-FMNC	
Channa striata	EU342203	COI	CS1	FGL-FMNC	
Channa punctata	EU342202	COI	CP2	FGL-FMNC	
Channa punctata	EU342201	COI	CP1	FGL-FMNC	
Channa marulia	EU342200	COI	CM2	FGL-FMNC	
Channa marulia	EU342199	COI	CM1	FGL-FMNC	
Channa gachua	EU342198	COI	CG2	FGL-FMNC	
Channa gachua	EU342197	COI	CG1	FGL-FMNC	
Channa bleheri	EU342196	COI	CB2	FGL-FMNC	
Channa bleheri	EU342195	COI	CB1	FGL-FMNC	
Channa aurantimaculata	EU342194	COI	CA2	FGL-FMNC	
Channa aurantimaculata	EU342193	COI	CA1	FGL-FMNC	
Channa diplogramma	EU342192	16S	CD6	FGL-FMNC	Specimen H in figure 3
Channa diplogramma	EU342191	16S	CD5	FGL-FMNC	Specimen G in figure 3
Channa diplogramma	EU342190	16S	CD4	FGL-FMNC	Specimen E in figure 3
Channa diplogramma	EU342189	16S	CD3	FGL-FMNC	Specimen D in figure 3
Channa diplogramma	EU342188	16S	CD2	FGL-FMNC	Specimen C in figure 3

Channa diplogramma	EU342187	16S	CD1	FGL-FMNC	Specimen A in figure 3
Channa striata	EU342186	16S	CS2	FGL-FMNC	
Channa striata	EU342185	16S	CS1	FGL-FMNC	
Channa punctata	EU342184	16S	CP2	FGL-FMNC	
Channa punctata	EU342183	16S	CP1	FGL-FMNC	
Channa marulia	EU342182	16S	CM2	FGL-FMNC	
Channa marulia	EU342181	16S	CM1	FGL-FMNC	
Channa gachua	EU342180	16S	CG2	FGL-FMNC	
Channa gachua	EU342179	16S	CG1	FGL-FMNC	
Channa bleheri	EU342178	16S	CB2	FGL-FMNC	
Channa bleheri	EU342177	16S	CB1	FGL-FMNC	
Channa aurantimaculata	EU342176	16S	CA2	FGL-FMNC	
Channa aurantimaculata	EU342175	16S	CA1	FGL-FMNC	
Channa micropeltes	JF900369	COI	UMTCM5	UMT	
Channa micropeltes	JF900370	16S	UMTCM5	UMT	

UMT = Institute of Tropical Aquaculture, Universiti Malaysia Terengganu, Kuala Terengganu, Malaysia , **FGL-FMNC** = Fish Genetics Lab – Fatima Mata National College, Kollam, Kerala, India.

Table S3: Genetic distance values calculated for the partial mitochondrial COI sequences of different Channa species used in the study

	CA1	CA2	CB1	CB2	CG1	CG2	CM1	CM2	CP1	CP2	CS1	CS2	Cmi4	CD1	CD2	CD3	CD4	CD5	CD6	NN
CA1	0.000																			
CA2	0.002	0.000																		
CB1	0.304	0.314	0.000																	
CB2	0.304	0.314	0.000	0.000																
CG1	0.327	0.337	0.397	0.397	0.000															
CG2	0.327	0.337	0.397	0.397	0.000	0.000														
CM1	0.533	0.548	0.662	0.662	0.576	0.576	0.000													
CM2	0.533	0.548	0.662	0.662	0.576	0.576	0.000	0.000												
CP1	0.507	0.521	0.538	0.538	0.554	0.554	0.403	0.403	0.000											
CP2	0.490	0.504	0.538	0.538	0.554	0.554	0.403	0.403	0.004	0.000										
CS1	0.699	0.717	0.849	0.849	0.633	0.633	0.512	0.512	0.457	0.473	0.000									
CS2	0.710	0.728	0.862	0.862	0.643	0.643	0.504	0.504	0.465	0.480	0.002	0.000								
Cmi4	0.673	0.690	0.762	0.762	0.809	0.809	0.554	0.554	0.466	0.466	0.662	0.653	0.000							
CD1	0.594	0.609	0.799	0.799	0.643	0.643	0.446	0.446	0.451	0.451	0.569	0.560	**0.210**	0.000						
CD2	0.594	0.609	0.799	0.799	0.643	0.643	0.446	0.446	0.451	0.451	0.569	0.560	**0.210**	0.000	0.000					
CD3	0.594	0.609	0.799	0.799	0.643	0.643	0.446	0.446	0.451	0.451	0.569	0.560	**0.210**	0.000	0.000	0.000				
CD4	0.555	0.570	0.749	0.749	0.601	0.601	0.415	0.415	0.438	0.438	0.562	0.554	**0.187**	0.007	0.007	0.007	0.000			
CD5	0.555	0.570	0.749	0.749	0.601	0.601	0.415	0.415	0.438	0.438	0.562	0.554	**0.187**	0.007	0.007	0.007	0.000	0.000		
CD6	0.555	0.570	0.749	0.749	0.601	0.601	0.415	0.415	0.438	0.438	0.562	0.554	**0.187**	0.007	0.007	0.007	0.000	0.000	0.000	
NN	1.341	1.373	1.234	1.234	1.101	1.101	0.878	0.878	0.959	0.972	0.939	0.939	0.983	0.928	0.928	0.928	0.871	0.871	0.871	0.000

Table S4: Genetic distance values calculated for the partial 16s rRNA gene sequences of different Channa species used in the study

	CA1	CA2	CB1	CB2	CG1	CG2	CM1	CM2	CP1	CP2	CS1	CS2	Cmi	Cmi4	CD6	CD5	CD4	CD3	CD2	CD1	PCO	NN	CMNE
CA1	0.000																						
CA2	0.002	0.000																					
CB1	0.040	0.044	0.000																				
CB2	0.037	0.040	0.002	0.000																			
CG1	0.033	0.036	0.040	0.037	0.000																		
CG2	0.031	0.034	0.042	0.039	0.002	0.000																	
CM1	0.132	0.127	0.162	0.156	0.137	0.133	0.000																
CM2	0.137	0.131	0.162	0.156	0.137	0.133	0.002	0.000															
CP1	0.171	0.165	0.205	0.197	0.155	0.159	0.185	0.190	0.000														
CP2	0.171	0.165	0.205	0.197	0.155	0.159	0.185	0.190	0.000	0.000													
CS1	0.140	0.134	0.144	0.138	0.138	0.134	0.089	0.089	0.214	0.214	0.000												
CS2	0.141	0.135	0.145	0.139	0.139	0.135	0.089	0.089	0.202	0.202	0.004	0.000											
Cmi	0.154	0.148	0.173	0.167	0.155	0.151	0.078	0.082	0.183	0.183	0.083	0.083	0.000										
Cmi4	0.161	0.154	0.180	0.173	0.161	0.157	0.082	0.086	0.190	0.190	0.087	0.088	0.002	0.000									
CD6	0.169	0.163	0.180	0.174	0.172	0.167	0.088	0.092	0.214	0.214	0.087	0.088	**0.024**	**0.027**	0.000								
CD5	0.169	0.163	0.180	0.174	0.172	0.167	0.088	0.092	0.214	0.214	0.087	0.088	**0.024**	**0.027**	0.000	0.000							
CD4	0.169	0.163	0.180	0.174	0.172	0.167	0.088	0.092	0.214	0.214	0.087	0.088	**0.024**	**0.027**	0.000	0.000	0.000						
CD3	0.169	0.163	0.180	0.174	0.172	0.167	0.088	0.092	0.214	0.214	0.087	0.088	**0.024**	**0.027**	0.000	0.000	0.000	0.000					
CD2	0.169	0.163	0.180	0.174	0.172	0.167	0.088	0.092	0.214	0.214	0.087	0.088	**0.024**	**0.027**	0.000	0.000	0.000	0.000	0.000				
CD1	0.176	0.169	0.187	0.180	0.178	0.174	0.092	0.096	0.222	0.222	0.092	0.093	**0.027**	**0.030**	0.002	0.002	0.002	0.002	0.002	0.000			
PCO	0.135	0.130	0.181	0.174	0.155	0.151	0.152	0.157	0.194	0.194	0.176	0.177	0.152	0.158	0.179	0.179	0.179	0.179	0.179	0.186	0.000		
NN	0.673	0.654	0.779	0.757	0.692	0.682	0.644	0.655	0.688	0.688	0.666	0.663	0.708	0.716	0.694	0.694	0.694	0.694	0.694	0.714	0.620	0.000	
CMNE	0.138	0.133	0.164	0.158	0.139	0.135	0.021	0.023	0.194	0.194	0.080	0.080	0.077	0.081	0.087	0.087	0.087	0.087	0.087	0.091	0.150	0.600	0.000

CA = Channa aurantimaculata, CB = Channa bleheri, CG = Channa gachua, CM = Channa marulia,
CP = Channa punctata, CS = Channa Striata, Cmi = Channa micropeltes, CD = Channa diplogramma,
PCO = Parachanna obscura, NN = Notopterus notopterus, CMNE = Channa marulia from North East India

Methods: The number of base substitutions per site from between sequences are shown in the tables above. Analyses were conducted using the Maximum Composite Likelihood model [1]. The rate variation among sites was modeled with a gamma distribution (shape parameter = 0.2238 for COI and 0.2424 for 16s). The differences in the composition bias among sequences were considered in evolutionary comparisons [2]. Substitution pattern and rates were estimated under the General Time Reversible model (+G) [4]. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories, [+G]). Mean evolutionary rates in these categories were 0.00, 0.02, 0.15, 0.74, 4.09 substitutions per site for COI and 0.00, 0.02, 0.18, 0.79, 4.01 for 16s rRNA gene. The nucleotide frequencies are A = 23.76%, T/U = 28.75%, C = 30.48%, and G = 17.01% for partial mitochondrial COI sequence and A = 29.82%, T/U = 21.54%, C = 25.60%, and G = 23.04% for the partial 16S rRNA gene sequence. The maximum Log likelihood for this computation was -2367.503 for COI and -1842.909 for 16s datasets. The analysis involved 21 nucleotide sequences for COI and 23 nucleotide sequences for 16s. Codon positions included were 1st+2nd+3rd+Noncoding for COI. All positions containing gaps and missing data were eliminated (complete deletion). There were a total of 477 positions for COI and 470 for 16s in the final dataset after deletion of the gapped sites. Evolutionary analyses were conducted in MEGA5 [3].

1. Tamura K., Nei M., and Kumar S. (2004). Prospects for inferring very large phylogenies by using the neighbor-joining method. Proceedings of the National Academy of Sciences (USA) 101:11030-11035.

2. Tamura K. and Kumar S. (2002). Evolutionary distance estimation under heterogeneous substitution pattern among lineages Molecular Biology and Evolution 19:1727-1736.

3. Tamura K., Dudley J., Nei M., and Kumar S. (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24:1596-1599.

4. Nei M. and Kumar S. (2000). Molecular Evolution and Phylogenetics. Oxford University Press, New York.