STEFAN ABREU FERNANDES

EVOLUTION OF IMMUNE GENES IN ANTARCTIC FISH

UNIVERSIDADE DO ALGARVE
Faculdade de Ciências e Tecnologia
2017/2018

STEFAN ABREU FERNANDES

EVOLUTION OF IMMUNE GENES IN ANTARCTIC FISH

Mestrado em Biologia Marinha

Trabalho efetuado sob a orientação de:
Professor Doutor Adelino V.M. Canário (UAlg/CCMAR)
Professor Doutor LiangBiao Chen (Shanghai Ocean University)

UNIVERSIDADE DO ALGARVE
Faculdade de Ciências e Tecnologia
2017/2018

EVOLUTION OF IMMUNE GENES IN ANTARCTIC FISH

Declaração de autoria de trabalho

Declaro ser o autor deste trabalho, que é original e inédito. Autores e trabalhos consultados estão devidamente citados no texto e constam da listagem de referências incluída.

_____

(Stefan Abreu Fernandes)

Agradecimentos

I would like to thank my supervisors, Professor Adelino V.M. Canário and Professor LiangBiao Chen, for their guidance and without whom I wouldn't have had the oportunity to do part of my thesis at the Shanghai Ocean University.

Obrigado, à professora Teresa Modesto pela sua disponibilidade e pelo seu apoio ao longo do meu mestrado, à professora Rita Castilho pela motivação e orientação que me permitiram finalizar a minha tese e à Doutora Regina Cunha pela a ajuda na elaboração dos meus resultados.

Obrigado, aos investigadores e técnicos do grupo de encdocrinologia comparativa e biologia integrativa do CCMAR que me acolheram num ambiente produtivo e com quem aprendi muito ao longo do meu percurso.

Um grande obrigado aos meus amigos, ao João pela sua honestidade e apoio que me permitiram melhorar a minha tese, à Ana que sempre que preciso está disposta a ajudar-me mesmo que isto signifique ficar horas ao telefone, à Raquel que com as palavras certas torna tudo mais fácil de concretizar, e ao Daniel sem quem a minha estadia em Xangai não teria sido tão aventurosa. Vocês foram os meus exemplos de insipiração e determinação que me ajudaram a concretizar a minha tese.

Aos meus pais, irmãos, cunhados e sobrinhos, obrigado por toda a força, amor, carinho e coragem que me deram mesmo a milhares de quilômetros de distância e por sempre me terem apoiado nas minhas escolhas que me permitiram alcançar este objetivo, por tudo isso e muito mais, obrigado.

Resumo

A superordem dos Notothenioidei inclui o maior número de representantes the peixes ósseos na plataforma continental da Antártida. As condicões abióticas e bióticas que dominam nesta região do mundo levaram à radiação e à especiação desta ordem. As baixas temperaturas que se deram durante o período do Eoceno Tardio levaram à flutuação da superfície ocupada pela calota glaciar no Oceano Antártico o que levou a uma redução do habitat disponível na plataforma continental. Por sua vez, a falta de habitat foi seguida por um declínio nas espécies de peixes ósseos e a uma alteração na relação predador-presa o que permitiu dispersão e diversificação das espécies que se adaptaram ao novo meio ambiente. Há 25 milhões de anos as condições ambientais tornaram-se mais estáveis criando um ambiente dominado por águas frias, ricas em oxygénio e nutrientes o que promoveu a adaptação radiativa dos Notothenoids. Os novos nichos ecológicos associados a condições ambientais estáveis providenciaram aos Notothenoids os requisitos para se tornarem a ordem de peixes ósseos domiante na plataforma continental da Antártida. Um dos pontos de interesse por esta ordem de peixes ósseos deve-se ao facto de ainda não se perceber qual será o impacto do aquecimento global nestas espécies. As diversas adaptações presentes nas várias espécies desta ordem também representam um fator de peso no que leva ao seu interesse científico. Estas vão desde elevada densidade de mitocôndrias e maior dimensão do miocárdio, à perda de resposta das proteínas de choque térmico, perda de hemoglobina, à evolução de proteínas anticongelantes - as adaptações observadas indicam quão bem se deu a adpatação desta ordem a um ambiente extremo. Entretanto, já foram desenvolvidos trabalhos que evidenciam que o sistema imune destas espécies também foi sujeito a adaptações promovidas pelo meio ambiente da Antártida. De modo geral, o sistema imune permite aos organismos superar perturbações que vão ao encontro da sua homeostase. O estudo do sistema imune em vertebrados, como os peixes, pode revelar aspetos importantes para o entendimento da evolução do sistema imune em vertebrados mais complexos. O objetivo deste trabalho é de estudar a evolução de genes que se enquadram no sistema imune de três espécies de peixes ósseos da região Antártida, *Eleginops maclovinus* que reside na região sub-polar da Antártida e duas espécies cuja a distribuição é limitada ao oceano da Antártida pela corrente circumpolar, *Notothenia coriiceps* e *Dissostichus mawsoni*. Para tal, foram escolhidas cinco famílias de genes que se relacionam com o sistema imune, estas foram os *toll-like receptors* (TLR), *immunoglobulin superfamily* (IgSf)*, phosphoinositide-3-kinase* (PIK3)*, AKT/protein kinase B* (AKT3) e as *semaphorins* (Sema). Os genomas e transcriptomas de 8 espécies de peixes foram obtidos de bases de dados de livre acesso enquanto os genomas e transcriptomas das espécies da Antártida foram proporcionados pelo laboratório do Professor

LiangBiao Chen da *Shanghai Ocean University*. Depois de identificadas as sequências proteícas destas famílias no peixe modelo *Danio rerio*, procedeu-se a uma pesquisa de similaridade por BLAST entre estas últimas e os transcriptomas das restantes espécies de forma a identificar as sequências potencialmente homólogas. Estas sequências por sua vez foram filtradas de modo a somente reter, para cada família, as sequências que apresentavam um rácio de identidade desejado. Após um alinhamento múltiplo de sequências (MSA), foram escolhidos para cada uma das famílias o melhor modelo evolutivo. Com os MSA de amino ácidos foi possível construir para cada família uma árvore filogenética na qual foi possível identificar genes ortólogos. Com os genes ortólogos foi possível construir uma árvore filogenética de espécies. De seguida, os MSA de amino ácidos foram convertidos para alinhamentos de codões para permitir a estimação da taxa de substituição de nucleótidos nas árvores filogenéticas das espécies, que é dada por ϖ= dN/dS onde dN equivale à taxa de substituição não-sinónima e dS a taxa de substituição sinónima. Com o valor de dS foi então possível resolver a equação $T=Ks/2r$, em que T representa o tempo de divergência a ser calculado, Ks é taxa de substituição sinónima e r é a taxa de substituição estimada obtida da bibliografia. Os resultados relativos ao número de sequências e a análise filogenética permitiram identificar variabilidade no número de genes encontrados em cada espécie tal como também foi possível observar que, quando presentes, os ortólogos das 11 espécies formavam uma árvore filogenética distincta. Estas observações levaram a estipular, tanto para os Notothenoids como para os restantes *taxa* analisados, que estas famílias de genes se enquadram num processo evolutivo denomiado de processo de nascimento e morte. As estimações dos tempos de divergência, obtidos nos nós para cada família de genes que representada num maior número de espécies resultaram em tempos de divergência similares ou superiores às estimativas dadas pelos registos fósseis. Por sua vez, os nós que apresentavam um menor número de epécies, indicaram tempos de divergência mais recentes do que os registos fósseis. As cinco famílias de genes nos Nototheniidae indicaram tempos de divergência recentes desde 7.1 milhões de anos (m.y.a) para Sema, 6.2 m.y.a para AKT3, 4.3 m.y.a para IgSf, 4 m.y.a para PIK3 e 2.5 m.y.a para TLR. As diferenças obtidas entre os tempos de divergência das cinco famílias de genes revelam uma possível relevância perante a adaptação dos Nototheniidae ao ambiente antártico, pois as famílias de genes que apresentam funções mais diversas também apresentam tempos de divergência mais antigos (Sema, AKT3, PIK3, IgSf) do que as famílias de genes com funções somente immunologicas (TLR). Finalmente, foi possível observar que estes tempos de divergência incluem-se dentro das estimativas cronológicas dadas para um fenómeno climatérico conhecido como transição climatérica do mioceno médio (MMCT) que

ocurreu entre os 25-5 m.y.a. Esta correlação levou a considerar que as adaptações do sistema imune dos nototheniidae sejam subsequentes ao MMCT.

Palavras-chave: Peixes da Antártida, adaptação, genes da imunidade, famílias de genes, fiolgenia.

Abstract

The Notothenioidei suborder has the largest representation of teleost fish in the Antarctic continental shelf. Their speciation and adaptive radiation was the result of particular abiotic and biotic conditions in the Southern Ocean. During the Late Eocene the cooler temperatures enlarged the ice cover leading to the loss of shelf habitat. This loss of natural environment resulted in a decline of fish diversity followed by the radiation and diversification of those who could adapt to the new conditions. Since then, the stable environment that has governed the Southern Ocean for the last 25 million years promoted the adaption of Notothenoids to cold, oxygen rich waters, allowing them to become the main teleost suborder in the Antarctic shelf habitats. Furthermore, previous work has shown that in Notothenoids immune related genes have undergone adaptations due to their exposure to the environmental conditions of the Antarctic Ocean. Their immune related adaptations give us the opportunity to study the phylogenetic diversification among Notothenioids and other teleosts that adapted to different environments. We studied the evolution of five immune related gene families in eight non-Antarctic vertebrates and three notothenoids, *Eleginops maclovinus*, *Notothenia coriiceps* and *Dissostichus mawsoni*, through phylogenetic analysis and divergence times estimation using nucleic and protein sequences. Genes for five different gene families were obtained from the genome and transcriptome of the investigated species. A possible birth-and-death process may have been identified for the five immune gene families. Furthermore, the divergence times estimated for the nototheniidae indicate that after the Middle Miocene climatic transition, those species relied on the gene families that presented a broader range of functions for their adaption to the Antarctic environment.

Keywords: Antarctic fish, adaptation, immune genes, gene families, phylogeny.

Table of contents

Index of figures

Figure 4.4: Phylogenetic Maximum-Likelihood gene tree for phosphatidylinositol 3-kinase of *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch). The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a VT+G+F substitution model. The bootstrap values are given in italic next to the nodes. The subfamilies containing the eleven orthologs pik3c3, pik3cd and pik3cg are highlighted in respectively green, orange and blue.

Figure 4.5: Phylogenetic Maximum-Likelihood gene tree for immunoglobulin superfamily between *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch). The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a VT+G+F substitution model. The bootstrap values are given in italic next to the nodes. The subfamilies containing the eleven orthologs IgSf3 and IgSf8 are highlighted in orange and blue respectively.

Figure 4.6: **Phylogenetic Maximum-Likelihood gene tree for semaphorin family of *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch).** The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a JTT+I+G+F substitution model. The bootstrap values are given in italic next to the nodes. The subfamilies containing the eleven orthologs sema3b, sema3bl and sema3c are highlighted in respectively green, orange and blue.

Figure 4.7: Phylogenetic Maximum-Likelihood tree for toll-like receptor based on 11 1:1 orthologous protein sequences from eleven fish, showing the relationships between Notothenioidei (*N. coriiceps*, *E. maclovinus*, *D. mawsoni*) and other fish species. The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumnae* was used as outgroup. The bootstrap values are given in italic next to the nodes.

Figure 4.8: Phylogenetic Maximum-Likelihood tree for AKT serine/threonine kinase 3 based on 11 1:1 orthologous protein sequences from eleven fish, showing the relationships between Notothenioidei (*N. coriiceps*, *E. maclovinus*, *D. mawsoni*) and other fish species. The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumna*e was set as outgroup. The bootstrap values are given in italic next to the nodes.

Figure 4.9: **Phylogenetic Maximum-Likelihood tree for phosphatidylinositol 3-kinase based on 33 1:1 orthologous protein sequences from eleven fish, showing the relationships between Notothenioidei (*N. coriiceps, E. maclovinus, D. mawsoni*) and other fish species.** The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumnae* was used as outgroup. The Bootstrap values are given in italic next to the nodes.

Index of tables

List of abbreviations

ACC - Antarctic circumpolar current

AFGP - antifreeze glycoprotein

AKT3 - AKT/ protein kinase B

dN - Non-synonymous (dN)

dS - Synonymous substitutions (dS)

IgSf - Immunoglobulin superfamily

MHC - Major histocompatibility complex

MRCA - Most recent common ancestor

MSA - Multiple sequence alignments

m.y.a - Million years ago

PIK3 - Phosphoinositide-3 kinase

SI - Supplementary information SI

Sema - Semaphorin

TLR - Toll-like receptor

TSWGD - Teleost-specific Whole Genome Duplication

WGD - Whole genome duplication

ϖ= dN/dS; - Non-synonymous to Synonymous substitutions ratio

1. Introduction

1.1 The Antarctic Ocean

The Antarctic or Southern Ocean has been a cold and stable environment for the last 20 million years (Dayton *et al.*, 1994) when the land bridges between East Antarctica and Australia (Tasmanian gateway) around 35.5 m.y.a (Stickley *et al.*, 2004) and between South America and the Antarctic Peninsula (Drake Passage) were interrupted allowing the circulation of a circumpolar current, the Antarctic circumpolar current (ACC) (Lyle *et al.,* 2007; Pfuhl & McCave, 2005).These events were caused by the displacement of tectonic plates, isolating the Antarctic continent and altering the atmospheric circulation leading to a cooling of this region (Cristini *et al.,* 2012) . The currently extended ice-sheet cover that expands over the Southern Ocean is believed to have started during the Eocene-Oligocene transition period when the low pCO_2 and the cooling of Antarctic sea water gave rise to a global "Ice-house" state (Sijp *et al.,* 2014) (Fig.1.1). Not only did this global cooling and expanding ice-sheet shaped the landscape of the southern pole, it has also changed the subaquatic landscape by occupying a great extent of the continental shelf (Clarke *et al.,* 2004). During the last glacial maxima the ice sheet extension reached as far as the continental shelf (MacKintosh *et al.,* 2011).

The ACC is the major current of the Antarctic Ocean and the largest of the earth (Mintenbeck, 2017) with current speeds of 173 Sv (1 Sv equals $10^6 m^3 s^{-1}$) (Donohue *et al.,* 2016) flowing from west to east connecting the various ocean basins while distributing heat and nutrients (Hassold *et al.,* 2009), acting as a boundary between the adjacent oceans and the Antarctic Ocean (Orsi *et al.,* 1995). The Antarctic ocean is characterized by oxygen rich waters (Lu *et al.,* 2016) with high nutrient concentrations (Dayton *et al.,* 1994). The water temperature around the Antarctic continent is permanently low varying between +2°C and -2°C (Mintenbeck, 2017). During the polar summer when the poles have a continuous solar exposure the increasing temperature of sea surface stratifies the water column reduces the mixed layer depth and stimulates the phytoplankton bloom (Llort *et al.,* 2015) which is one of the main starting points of the food web of this region (Constable *et al.,* 2014).

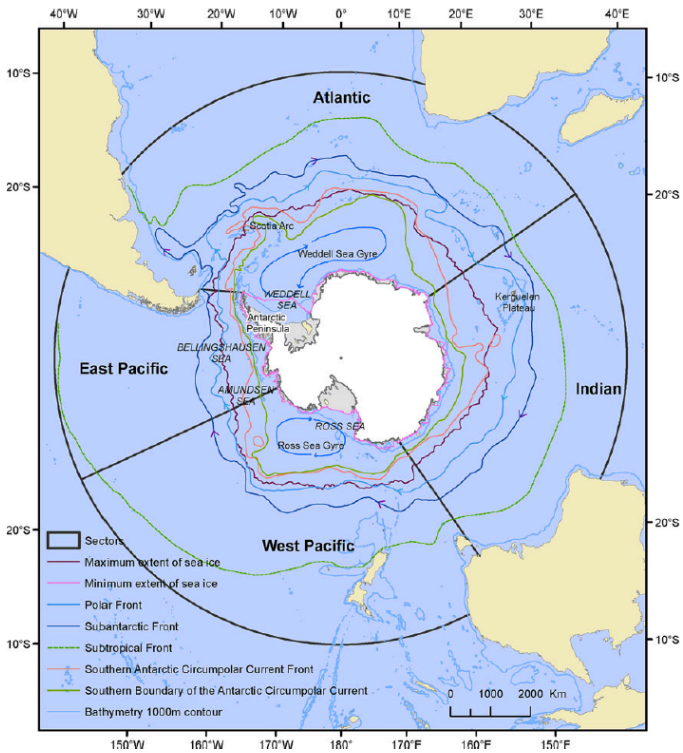

Figure 1.1: Representation of the Southern Ocean with minimum and maximum extent of sea ice and the Southern Boundary of the Antarctic Circumpolar Current, and the 1000 m countour. Adapted from Constable *et al.*, (2014).

The ongoing increase in greenhouse gases due to anthropic activity is causing a global warming that is threatening the south pole by increasing the sea water temperature, melting the ice cover, and increasing stratification (Gupta *et al.,* 2009). As a result, the ACC could start to slow down enhancing the mixing of the warmer waters of the adjacent oceans and increasing even more the sea water temperature of the Antarctic Ocean (Rusell *et al.,* 2006). When looking at the future of the Antarctic continent under the scope of global warming one is compelled to think about the ecosystem that was established under unchanging stable conditions for the last 20 million years which included the endemic Notothenoid teleosts.

1.2 Fish Immunology

The first evidence of an early immune system is attributed to the phagocytic activity of unicellular amebae comparable to the phagocytic activity of macrophages in higher organisms (Desjardins *et al.,* 2005). Nevertheless, a necessary ability to eliminate intruders or pathogens is the recognition of the self and non-self (Cooper, 2010). The non-specific innate immune system developed receptors (e.g toll-like receptors) that could identify features that were preserved on microbial pathogens like glycolipids of the cell-membrane and nucleic acids (Iwasaki & Medzhitov, 2015; Janeway, 1989; Takeda *et al.,* 2003). To these toll-like receptors

has also been attributed an important role in acquired immune defences since they activate antigen specific T cells (Fearon & Locksley, 1996Schnare *et al.,* 2001). It is to note that the bases of modern immunology reside in the phagocytosis theory presented by Metchnikoff who did his findings on starfish larvae (Tauber, 1992). Fish are considered an essential link to the study of the evolution of the vertebrate immune system because of their basal position and because they have the greatest number of species (Ahn *et al.,* 2016). The innate immune system is essential for fishes since they are exposed to the aquatic environment from early developmental stages (Rombout *et al.,* 2005). Besides, fish represent a significant contribution for the spread of the adaptive immunity in vertebrates since the adaptive immune system has its origin in primitive jawless fish (Litman *et al*.,2010).

The immune system allows animals to cope with everyday disturbances like injuries due to predation or infections by pathogens so as to maintain their homeostasis (Buchmann, 2014). Basically, when infected, the animal's immune system acts as a feedback mechanism to the non-self pathogen by producing defenses that will eliminate the intruder. As to better conceptualize the immune system and understand what are the processes involved in the different immune responses, the immune system was artificially divided into two components the innate and the adaptive immunity (Schultz & Grieder, 1987) (Fig.1.2).

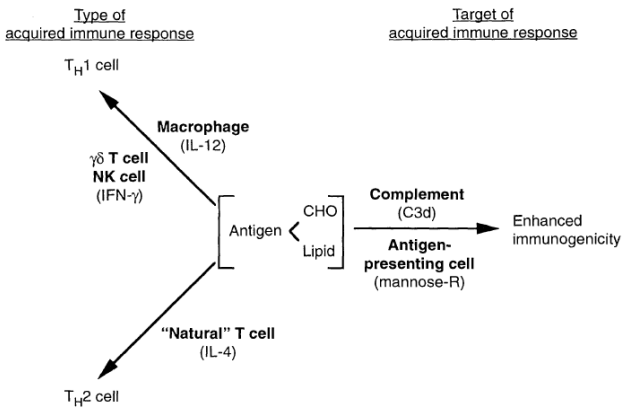

Figure 1.2: Representation of the two components of the immune system with the innate immunity represented in boldface. The innate system recognizes the pathogens and indicates to the adaptive system which are the complementary anitgens (From Fearon & Locksley (1996)).

They differ in their origin and function where the innate immunity relies on non-specific germ line encoded receptor proteins and the adaptive immunity relies on gene recombinations (V(D)J recombination) that produce specific antigen-receptors on the surface of B and T cells that allows them to recognize a large number of antigens (Thompson, 1995). Phylogenetically the innate immune system predates the adaptive immunity since it can be found in all multicellular

organisms whereas the adaptive immunity is only found in vertebrates (Janeway & Medzhitov, 2002). Furthermore, the study of teleost immunity can provide interesting knowledge about the possibilities that immune related genes could be responsible for the diversification of species (Eizaguirre *et al.,* 2009). For example, the cold adapted Atlantic cod (*Gadus morhua*) presents an interesting case study where immune related genes are thought to have brought forth speciation. Interestingly this species has lost the major histocompatibility complex (MHC) which is a key feature of the adaptive immune system (Star *et al.,* 2011). The loss of MHC II genes triggered the compensatory expansion of MHC I gene complex but authors did not clarify which one arose first and if they were related. More recently, Malmstrøm *et al.,* (2016) tried to unravel the relationship and order in which these events occurred so as to understand the role of immune related genes on speciation. As it turns out, the MHC II gene loss predates the MHC I gene expansion which could have triggered the gene expansion of MHC I genes, highlighting the role of MHC genes in teleost diversification.

1.3 The immune system of Antarctic fish

The Notothenoids are the most abundant and diverse teleost group in Antarctic waters and are a good example of a specific adaptive radiation known as species flock (Eastman & McCune, 2000). Notothenoids are mostly benthic fish specifically due to their lack of swim bladder but some species like *D. mawsoni* have developed a nearly pelagic lifestyle by reducing the density of their skeleton and by enhancing fat deposits (Eastman & Devries, (1981), Eastman, (2000)). Other adaptations that have contributed for their success in the extreme cold environment range from higher mitochondria density (O'Brien & Mueller, 2010), loss of the heat shock proteins response (Hofmann *et al.,* 2000), increased myocardium (Johnston *et al.,* 1983), loss of haemoglobin (Ruud, 1954) and evolution of anti-freeze proteins (Deng *et al.,* 2010). The evolution of anti-freeze proteins was a fundamental part of their success in the southern Ocean (Montgomery & Clements, 2000). In Notothenoids two kinds of antifreeze molecules can be found, the antifreeze glycoproteins and the antifreeze proteins (Evans & DeVries, 2017). Chen *et al.,* (1997) found that an antifreeze glycoprotein gene evolved from a pancreatic enzyme gene. In addition, Deng *et al.* (2010) also hinted that the origin of antifreeze proteins is linked to the neofunctionalization of an old sialic acid synthase gene. The work of Chen *et al.,* (2008) that focused on the transcriptomic and genomic evolution of the Notothenoid fish revealed an up-regulation of innate immunity related genes that are suggested to be responsible for the prevention of oxidative stress due to high oxygen exposure in these fish. Similarly, Bilyk & Cheng (2013) also pointed to the enhanced over expression of genes related to innate immune response in the Notothenoid *P.borchgrevinki*. Additionally, when

exposed to a different pathogen agonist, bacterial or viral, *N.coriiceps* presented different immune responses (Ahn *et al.,* 2016). While when exposed to bacterial pathogen, the immune response was based on antigen presentation, during a viral contamination the immune response induced the tumor necrosis factor (TNF) pathway. Ota *et al.,*(2003) showed in two Notothenoid species that an immunoglobulin gene (IgM) had undergone adaptive selection to prevent protein disfunction due to the cold environment or degradation by coevolving parasites. All these specific adaptations of the immune system in Notothenoids allow us to study and finally understand their diversification in a such extreme environment.

1.4 Fish genome evolution

Although the term "fish" doesn't refer to a monophyletic group in the tree of life it is widely used and refers mostly to the water dwelling animals belonging to the teleosts (coelacanth, lungfish, ray-finned fish), chondrichthyes (sharks, rays, chimeras) and jawless craniates (lampreys, hagfish) (Nelson, 2006). From those, the ray-finned fish or actinopterygian account for 95% of the fish species and represent half of the known species of vertebrates (Volff, 2005). The increasing availability of whole-genome sequencing has provided the evidence that vertebrates have gone through a series of whole-genome duplication (WGD) events (Dehal & Boore, 2005; Putnam *et al.,* 2008). In comparison with invertebrates, the genomes of vertebrates present a higher number of genes in each gene family (Meyer & Van de Peer, 2005). Common to all vertebrates are two episodes of whole genome duplication (1R and 2R) which are thought to have occurred at the earlier stages of their evolution (Dehal & Boore, 2005) whereas teleosts have undergone another specific whole genome duplication (3R) (Van de Peer *et al.,* 2003; Meyer & Van de Peer, 2005) between 225 and 333 million years ago (Hurley *et al.,* 2007; Near *et al.*, 2012) (Figure 1.3).

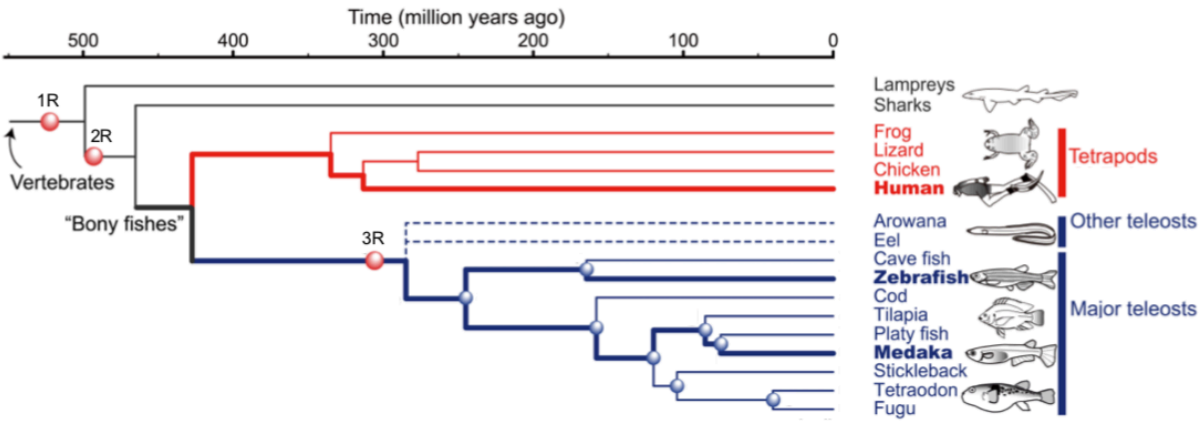

Figure 1.3: Species tree showing major vertebrate groups and their evolutionary relationship with the 3 rounds of whole genome duplication. 1R and 2R corresponds to the two whole genome replications in the vertebrate stem. 3R corresponds to the whole genome replication specific to teleost fish. Adapted from (OIST).

After genome duplication the resulting pairs of genes (paralogs) can have different fates. Immediately after duplication the daughter genes present similar functions, lessening the selective constrains to keep both of the duplicates resulting in the loss of one duplicate (non-functionalization), the gain of a new function by one of the paralogs (neo-functionalization), or each of the duplicated genes keeps a different subfunction of the ancestral gene that is complementary to the other duplicate and together work as one (subfunctionalization) and dosage selection where the duplicated changes are both kept presenting only few changes (Glasauer & Neuhauss, 2014). Although the teleost specific WGD cannot be considered as wholly responsible for the great diversity of fish (Hurley *et al.,* 2007; Santini *et al.,* 2009), authors have concluded that once a WGD has occurred it laid the foundations for posterior diversification (Berthelot *et al.,* 2014; Glasauer & Neuhauss, 2014).

1.5 Gene families

One of the first authors to classify genes into families was Tomoko Ohta, defined as "a group of genes or nucleotide sequences with the following characteristics: multiplicity, close linkage, sequence homology, and related or overlapping functions" (Ohta, 1980). In 2008, Ohta (Ohta, 2008) additionally differentiated the gene families in multigene families, which are groups of genes with sequence homology and related overlapping functions, and the superfamilies, which are groups of proteins or genes of common origin with nonoverlapping functions. Gene families are seen as a valuable characteristic to the organization of the genome presenting varying degrees of complexity and number of genes (Ohta, 2008). The organization of genes into gene families has been useful for the creation of databases that order the nucleotide or amino acid sequences into gene families such as Pfam (Finn *et al.,* 2018), Uniprot (Chen *et al.,* 2017) and InteproScan (Mitchell *et al.,* 2018).

Several sequence-based methods, that can be grouped into three categories, have been developed to identify members of a gene family (Frech & Chen, 2010). The first method groups genes into gene families by searching for similarities in sequence domains and motifs and is useful to identify gene function (Frech & Chen, 2010) as seen in the Pfam database (Finn *et al.,* 2018). The second method groups gene families by pairwise comparison of protein sequences using clustering techniques (Bernardes *et al,* 2015). The third method is the construction of a phylogenetic tree which is implemented when the scope of the research is focused on the evolutionary history of a gene family (Song *et al.*, 2017).

As presented by Hood *et al.,* (1975), evolution is thought to act in concert through whole gene families not only in individual genes. By studying whole gene families, researchers were

able to identify their role in diversification of several species. Ramasamy *et al.,* (2016) hypothesized that a change in the chemical-ecological environment caused the duplication of an olfactory gene family that led to the adaptation of *Drosophila suzukii* to the new conditions. Cortesi *et al.,* (2015) proposed that gene loss, pseudogenization, and gene duplication in the opsin gene family led to adaptation of percomorph fish to the diverse light conditions at which they are found today. These authors pointed out to the relevance that comparative studies of gene family to unravel their evolution, could finally lead to a better conception of speciation.

1.5.1 Toll-Like Receptors

The toll-like receptors (TLRs) comprise the largest family of the pattern recognition receptors (PRRs) which recognize pathogen associated molecular patterns (PAMPs) (Takeuchi & Akira, 2010). Their contribution to the immune system is significant since their correct recognition of the PAMP's initiates an adequate immune response (Kawai & Akira, 2010). TLRs are transmembrane proteins with the an extracellular domain composed of leucine-rich repeats (LRR) responsible for the recognition of the PAMPs, the transmembrane helical structure and the intracellular part known as the toll-interleukin receptor which mediates the signal transition for the immune response (Gay & Gangloff, 2007). TLRs play a pivotal role between innate and adaptive immune system (Werling *et al.,* 2009). As part of the PRRs they are an integral part of the innate immune response (Kawai & Akira, 2010) and they are also responsible for the activation of adaptive immune responses by triggering the release of T-cell stimulators (Schnare *et al.,* 2001).

In general, the number of genes belonging to the TLR gene family may vary, with mammals counting with 10-13 functional TLRs (Kawai & Akira, 2010) comprising TLR1-13 (Solbakken *et al.,* 2016) and bony fish with up to 17 (Rebl *et al.,* 2010) which may additionally include *TLR14-26* (Solbakken *et al.,* 2016). From the fish species where TLRs have been identified, zebrafish has 17 (Meijer *et al.,* 2004), pufferfish 11 (Oshiumi *et al.,* 2003) and cod 9 (Solbakken *et al.,* 2016). All the TLR genes are regrouped into six bigger families, *TLR1, TLR3, TLR4, TLR5, TLR7* and *TLR11* that are usually represented in different species by at least one ortholog (Roach *et al.,* 2005).

1.5.2 The Immunoglobulin superfamily

Genes encoding at least one immunoglobulin (Ig) domain are classified into the immunoglobulin superfamily and their proteins are relevant for the identification and elimination of exogenous entities (Garver *et al.,* 2008). The structure of the Ig domain with its

stable structure but high variability in amino acid sequence is essential for the function of those proteins giving them a high degree of diversity (Halaby & Mornon, 1998).

Their diversity and their occurrence in a high number of taxa have made these proteins one of the largest families (Natarajan *et al.,* 2015). The members of the immunoglobulin superfamily (IgSf) considered to be relevant to the immune response are identified by shared structural features that can be differentiated by function and size into two categories, a variable-domain (V-domain) and a constant-domain (C-domain) (Natarajan *et al.,* 2015). The different members of the IgSf present a broad variety of functions such as muscle proteins, surface antigen receptors, co-receptors of the immune system and cell ligand molecules (Natarajan *et al.,* 2015). Two of the most prominent IgSf-subfamilies are the T-cell receptors that act as antigen receptors and the antigen presenting molecules such as the MHC, Class I and II (Lefranc, 2014).

1.5.3 Semaphorins

The semaphorin family has more than 30 representatives divided into eight subfamilies and can be found in invertebrates (Classes 1 and 2), vertebrates (Classes 3 to 7) as well as in viruses (Class V) (Fig. 1.4). (Goodman *et al.,* 1999). They can be secreted or membrane-bound and are differentiated through sequence similarity and structural variation (Kikutani *et al.,* 2007), having in common a *sema* domain relevant for the binding of specific receptors, a PSI domain (plexins, semaphorins, and integrins) and a terminal domain C (Janssen *et al.,* 2010; Liu *et al.,* 2010; Nogi *et al.,* 2010).

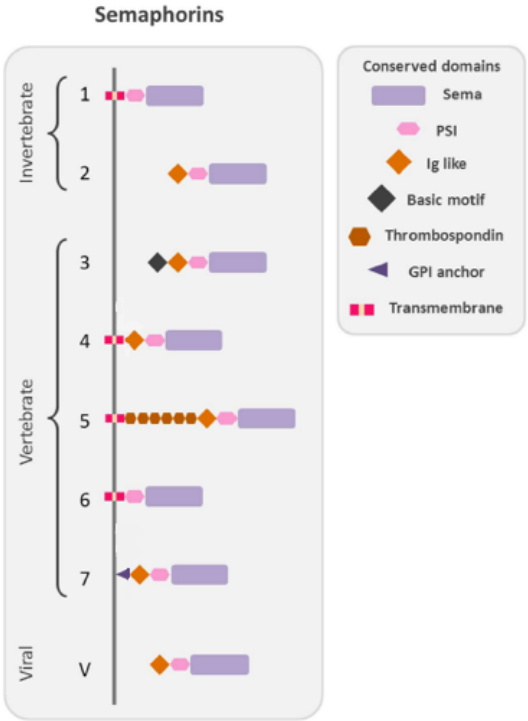

Figure 1.4: Schematic representation of the protein structure of semaphorin. Semaphorins are represented in their classification into eighth classes. Their conserved domains are drawn in different shapes and colors as indicated in the figure. Class 1 and 2 are found in invertebrates whereas class 3 to 7 are found in vertebrates and class V in viruses. Domains abbreviations: PSI (plexin semaphorin integrin); Ig-like (immunoglobulin like); GPI, (glycosylphosphatidylinositol) anchor. Adapted from Messina & Giacobini (2013).

Members of the semaphorin family (Sema) are involved in several biological processes such as immune, vascular development, endocrine system, cell migration, nervous system (Gu & Giraudo, 2013; Messina & Giacobini, 2013; Sun *et al.,* 2017). Several classes of the semaphorin family have been found to play a role in the immune regulation of some organisms. In class 4, Sema4D is highly expressed in resting T-cells of lymphoid organs such as lymph nodes, spleen and thymus (Kumanogoh & Kikutani, 2003) and is coupled to the regulation of B-cells (Kumanogoh *et al.,* 2000) whereas Sema4A is expressed in spleen, bone-marrow, dendritic cells and B cells and is involved in T-cell activation and proliferation (Kumanogoh *et al.,* 2002; Ito & Kumanogoh, 2016). In class 3, Sema3A and Sema3E have been associated with the regulation of immune cell trafficking (Choi *et al.,* 2008; Takamatsu *et al.,* 2010). Of further interest is Sema7A which is expressed in CD4+, CD8+ thymocytes and on activated T-cells (Mine *et al.,* 2000) and has been found to be involved in inflammatory immune response through stimulation of the production of macrophages and of cytokines in those macrophages and monocytes (Suzuki *et al.,* 2007; Kang *et al.,* 2014).

1.5.4 PIK3-AKT3

Phosphoinositide-3 kinase (PI3K or PIK3) are a family of lipid kinases present in all cells, producing phosphoinositides responsible for signalling pathways in several metabolic processes (Okkenhaug, 2013), including immune genes, as shown in the KEGG (Kyoto Encyclopedia of Genes and Genomes) annotated pathway (Fig.1.5). As mentioned in the extensive reviews of Koyasu (2003) and Okkenhaug & Vanhaesebroeck (2003) they are responsible for the regulation of TLRs, in lymphocyte development and in B- and T-cell regulation. Therefore, the PI3K-AKT-signaling pathway play an important role in the function of immune cells (Okkenhaug & Vanhaesebroeck, 2003). The *TLR2* and *TLR4* are a type of germline-encoded PRR important for transmembrane signalling pathways with a significant contribution to the immune system, since they enable the latter to recognize pathogenic particles initiating an adequate immune response (Kawai & Akira, 2010). Specifically, *TLR4* are specialized in bacterial lipopolysaccharide recognition (Kawasaki & Kawai, 2014) whereas *TLR2* recognize a large array of microbial components including of parasitical, viral, fungal and bacterial origins (Akira, *et al.,* 2006). Troutman *et al.,* (2012) supported that the interaction of TLR in the PI3K-AKT-signaling pathway is essential for the correct course of an immune-response. Moreover, as illustrated in Fig.1.5, AKT or protein kinase B is a key mediator in the PI3K-AKT-signaling pathway ensuring the proceeding of many of the metabolic steps (Lawlor & Alessi, 2001). The activity of the AKT protein is enhanced by its phosphorylation through the binding with a phosphoinositide (Li *et al.,* 2002). This phosphorylation is the beginning of

a cascade of phosphorylation of other proteins to regulate several metabolic ways in the cell (Lin et al., 2009).

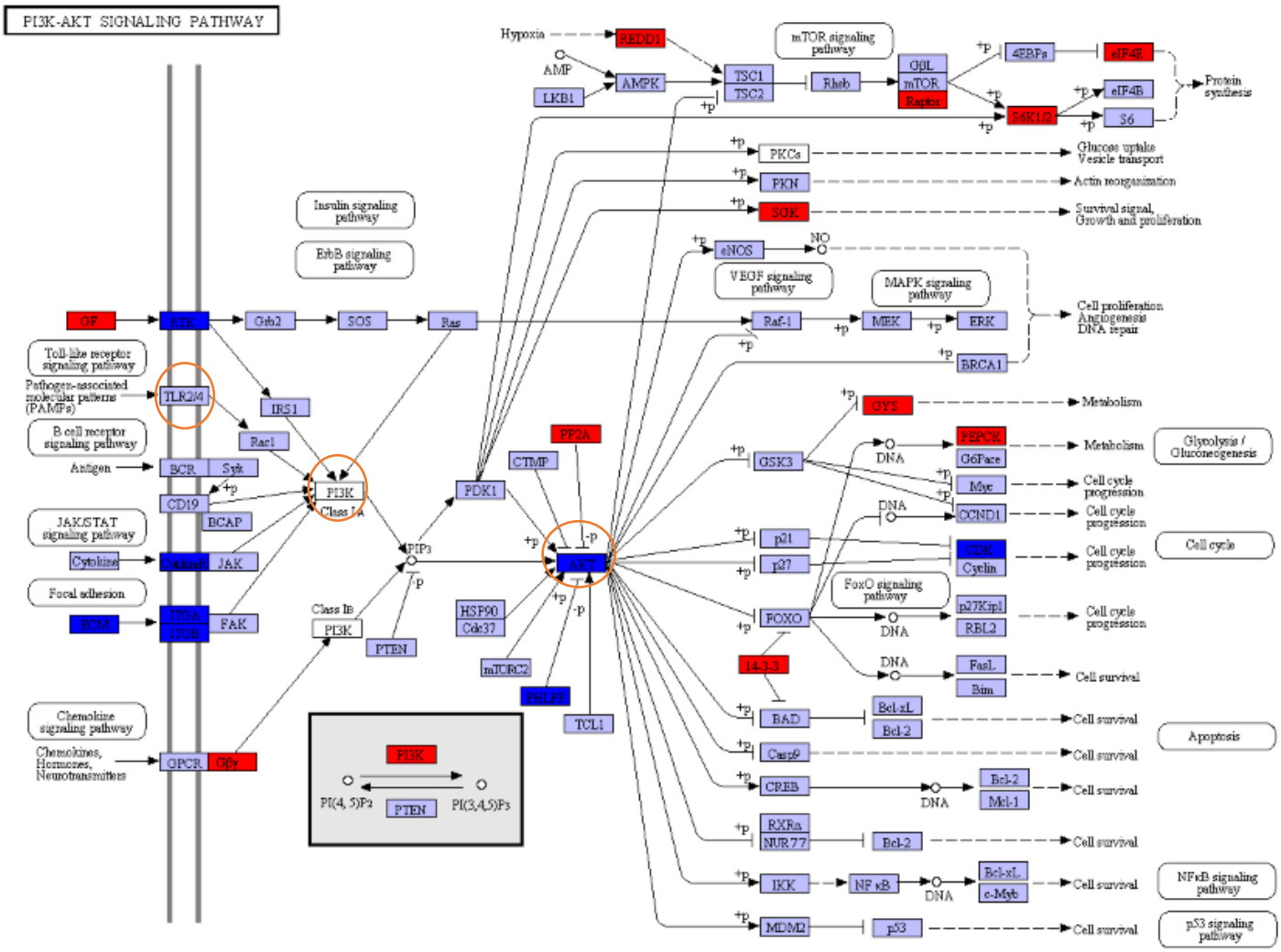

Figure 1.5: Representations of the PIK3-ATK-Signaling Pathway. The orange circles indicate the reviewed immune related genes. Adapted from Kanehisa *et al.,* (2016).

1.6 Homology

Richard Owen coined in 1843 for the first time the term homology as "*the same organ in different animals under every variety of form and function*" (Owen, 1848). At that time Owen didn't consider the source of this similarity and referred to it manly in relation to structures that were located in the same position in different organisms (Hall, 2013). It was only after the publication by Charles Darwin's of the Origin of Species (1859) and the postulation of the Theory of Evolution, which stated that all organisms descend from one common ancestor, that Ray Lankester (Lankester, 1870) introduced the notion that the similarity between structures was due to their shared ancestry.

Homology can be referred to as the similarity between two characters sharing a common ancestor and can be applied to different contexts such as structural, like the homology between tetrapod limbs (Amaral & Schneider, 2018), developmental as in the similarities between the processes that gave origin to a feature (Hall, 2013) and genetical as is the case between nucleotide sequences in DNA or amino acid sequences in proteins (Pearson, 2013).

It was Fitch (1970), based on the similarity concept of Owen, that applied the term homology to define homologous sequences as two or more sequences that present a high degree of similarity between them indicating a recent common ancestor. He also provided more specification by differentiating these homologous sequences as being orthologous or paralogous. As defined by Fitch (1970), when homologs originate due to a speciation event that occurred to the last common ancestor, they are called orthologs, whereas if two homologous sequences originate due to gene duplication then they are called paralogs. In most cases orthologues share similar functions and paralogs tend to diversify and specialize hence acquiring new functions (Koonin, 2005). Paralogs can subsequently be divided into two other subcategories based on the time of the duplication event. Paralogous sequences that arise from a lineage-specific duplication after a speciation event are termed as in-paralogs whereas if the duplication precedes the speciation event they are called out-paralogs (Koonin, 2005). The above mentioned distinctions have to be kept in mind for the downstream analysis of the phylogenetic trees, since it provides the means to understand if the relationships in the phylogenetic tree are due to speciation or duplication events (Salemi & Vandamme, 2003). As presented in (Pearson, 2013) the most used methodology to find homologous sequences is by applying a similarity search algorithm such as BLAST (Altschul *et al.* , 1990) that allows a rapid sequence similarity comparison between a reference species and a target species. A valuable strength of a sequence similarity search tools, are the statistics that are given for each match providing a mean to find the matches that are significantly similar, thus more likely to be homologous (Pearson, 2013). The Expect value (E) estimates the number of BLAST hits, presenting a similar score, which could occur by chance (Korf *et al.*, 2003). A small E-value tells us that the possibility of a BLAST hit resulting from chance are low, thus we can deduce that this match is probably due to a high degree of similarity between the two sequences. By applying a threshold based on the E-value it is easier to define which are the best alignments that should be kept for a specific investigation. Depending on the research one has to pay attention to the nature of the sequences that have to be aligned. To detect homology on a closer time range a DNA:DNA alignment might be satisfactory since the evolutionary look-back that the DNA provides doesn't extent further than 400 million years (Pearson, 2013). On the other hand, protein:protein alignments provide a more distant evolutionary look-back that can date to a last common ancestor shared 2.5 billion years ago (Pearson, 2013). Due to this difference,

the threshold E-value set between DNA:DNA alignments and protein:protein alignments has to differ (Pearson, 2013). A threshold of <0.001 is generally enough to assume that protein:protein alignments are homologous whereas for DNA:DNA alignments the value has to go as far as <10^-10 to assume homology (Pearson, 2013). Another way to look for possible homology between matched sequences is to check the percentage of shared identity (Pearson, 2013). This value represents the percentage of identical residues that are located in same position between two given sequences (amino acid or nucleotides) (Pearson, 2013; Fassler & Cooper, 2011). A minimum threshold of 30% identity can be considered sufficient to look for homologous sequences, when coupled with an appropriate E-value.

1.7 Sequence alignment and model selection

Multiple sequence alignments (MSA) are relevant in genomic and evolutionary studies, as they compare several protein or nucleotide sequences and identify their shared identical regions or homologous regions (Nuin *et al.,* 2006). A MSA is built on a sequential pairwise alignment where the order of the sequences is given by a phylogenetic tree (Edgar & Batzoglou, 2006). A MSA organizes the sequences in a matrix where each row represents a sequence and each column indicates homologous sites where insertions or deletions are denoted by gaps (Elias, 2006). With the identification of such homologous regions it is possible to deduce the function, the structure and the phylogeny between a set of sequences (Elias, 2006) .

A phylogenetic analysis has to go through statistical inferences to be validated (Posada & Buckley, 2004; Kelchner & Thomas, 2007), hence the need for a model that best fits the replacement rate of the amino acids or the substitution rate of nucleic acids (Posada & Crandall, 2001). These models provide a mean to estimate the probabilities of the different changes that could occur to a nucleotide or amino acid along the phylogeny (Posada & Crandall, 2001). These changes can be modelled with different methods ranging from distance methods to maximum likelihood, or maximum parsimony with each one presenting their own set of parameters and differing in their degree of complexity (Posada & Crandall, 2001).To assess the reliability and choose the best fitting model a statistical test can to be carried out that will compare all the available models that are given for a data set by means of the Akaike information criterion (AIC) (Akaike, 1974) and Bayesian information criterion (BIC) (Schwarz, 1978; Posada & Crandall, 2001).

1.8 Phylogenetic trees

In a phylogenetic analysis, a phylogenetic tree provides a mean to identify the evolutionary relationship between organisms (Vandamme, 2003). Such a tree is composed of nodes and branches where each node represents a unit (species or sequence) and each node is connected to another by only one branch where the pattern obtained by the branching is called the topology (Vandamme, 2003). The specific case of the terminal nodes or branch tips are termed as OTU (operational terminal units) which are the units for the construction of the tree (Vandamme, 2003). A phylogenetic tree may be unrooted, where a common ancestor isn't specified or rooted where one of the OTU is set as an outgroup to the other units of the tree, who then build the ingroup, resulting in an root node that represents the common ancestor to the ingroup and the outgroup (Vandamme, 2003; Horner & Pesole, 2004). Several methods, each presenting its sets of algorithms and assumptions, have been implemented to depict phylogenetic trees. These different types of approaches can be divided into two categories based on the method used to construct the tree, character-based or the distance based methods (Horner & Pesole, 2004). The distance based methods such as Neighbour-Joining or UPGMA rely on a matrix that compares the pairwise distances between the sequences and groups, or clusters the sequences by their level of similarity (Baldauf, 2003; Horner & Pesole, 2004). The character based methods such as Maximum-Likelihood, Maximum-Parsimony and Bayesian method compare the aligned sequences by looking for character substitution, where each position of the alignment is considered a character and the amino acid or nucleotide at this position is the state, to find the tree that best fits a given model of substitution (Vandamme, 2003; Horner & Pesole, 2004). Both kinds of methods have advantages and disadvantages; the distance based are faster to construct yielding only tree topology whereas the character-based can retrace the evolution of a specific site which in turn requires longer computational times (Baldauf, 2003). The choice of which method to use depends on the focus of the study but authors may consider comparing the trees obtained with different methods to confirm the phylogeny (Baldauf, 2003). Once a phylogenetic tree is built it is necessary to measure how accurate the dataset supports the tree (Baldauf, 2003; Horner & Pesole, 2004). Nowadays multiple methods can be used to estimate the reliability of a tree such as likelihood-based test, internal branch lengths, and bootstrapping (Z. Yang, 2014). The most commonly used is the bootstrap method where trees are randomly rebuild based on different subsamples of the original dataset and where the number of times a specific tree is built in each one of those subsamples is calculated (Baldauf, 2003; Horner & Pesole, 2004).

1.9 Nucleotide substitution and Divergence Time

A widespread method to analyse the evolutionary pressure exerted on protein-coding genes is the estimation of the ratio of non-synonymous (dN) to synonymous substitutions (dS) ω=dN/dS between a given set of sequences in a phylogeny (Mugal *et al.,* 2014). By calculating this ratio one can infer the type and the strength of selective pressure, where ω>1 indicates positive selection, ω=1 indicates neutral evolution and ω<1 indicates purifying selection (Gharib & Robinson-Rechavi, 2013). The differences between the sequences are due to changes caused by mutations in the DNA (Loewe & Hill, 2010). Such a mutation can result in the insertion or deletion of nucleotides in the DNA sequence or the replacement of a nucleotide with another nucleotide called a substitution (Li & Graur, 2002). In case of a substitution there are two possible outcomes; a transversion where a pyrimidine changes to a purine or vice-versa, and a transition where a pyrimidine changes to a pyrimidine or a purine to a purine (Li & Graur, 2002). The effect of the substitution on the translation of the codon into a protein can be synonymous, there is no effect on the translation, or non-synonymous where an amino acid is translated into a different amino acid (Li & Graur, 2002). The outcome of a mutation that occurred in a single organism depends on the evolutionary processes acting on the population (natural selection and genetic drift) which may spread the mutation through all the organisms leading to the fixation of the mutation in the population or it may lead to the loss of the mutation (Jeffares *et al.,* 2015). Depending on how it affects the fitness of an organism a mutation can either be advantageous by increasing fitness, deleterious by decreasing fitness or neutral when the effects of the mutation are so small that they don't affect selection (Loewe & Hill, 2010). Non-synonymous substitutions are usually linked to negative changes to the structure and function of proteins consequently they are deleterious while synonymous substitutions, as they don't change the amino acids, are neutral (Jeffares *et al.,* 2015). Their neutral nature makes the synonymous substitutions less prone to selective pressure leading them to accumulate with a linear rate which in turn can be used as an approximation to estimate the relative divergence time between two sequences (Huerta-Cepas & Gabaldón, 2011). Common limitations to synonymous substitution rates as estimates of the divergence time are associated to the species used for the analysis since too closely related species will not provide enough differences to show any significant changes in divergence time and too distantly related species will have had time to accumulate several mutations on the same site causing what is called mutational saturation (Wilke *et al.,* 2009). Nevertheless, the estimation of the divergence time between a set of sequences may be obtained from the number of synonymous substitutions seen between them as those are a result of the elapsed time since they separated from their last common

ancestor (Wilke *et al.,* 2009; Huerta-Cepas & Gabaldón, 2011). This approach is based on the assumptions of a molecular clock which considers that evolution takes place at a constant rate throughout lineages and that mutations are mostly neutral (Wilke *et al.,* 2009).

2. Objectives

The evolutionary study of whole gene families may be a valuable tool to understand how a given species adapted to its environment. This may be achieved through a phylogenetic analysis of gene and species phylogenies, and estimation of the genes divergence times. This work aims to study the evolution of five immune related gene families, in one Sub-Antarctic and two Antarctic teleosts, *Elegeniops maclovinus*, *Notothenia coriiceps* and *Dissostichus mawsoni*, respectively, by:

1) Identifying which was the evolutionary process that acted on the five gene families of the three target species through phylogenetic analysis.

2) Estimating the divergence time of five immune related gene families to understand if those estimations correlate to the adaptive radiation of notothenioids into the Antarctic Ocean.

3. Material and Methods

3.1 Fish species selection

The reference species used are one Sarcopterygian species the coelacanth (*Latimeria chalumnae*) and ten Actinopterygii, zebrafish (*Danio rerio*), Nile tilapia (*Oreochromis niloticus*); sea bass (*Dicentrarchus labrax*); stickleback (*Gasterosteus aculeatus*); medaka (*Oryzias latipes*); cod (*Gadus morhua*); bullhead notothen (*Notothenia coriisceps*); patagonian blenny (*Eleginops macluvinus*), antarctic toothfish (*Dissostichus mawsoni*) and spotted gar (*Lepisosteus oculatus*) These fish species have been chosen because a large amount of sequence data available that enable comparative studies of vertebrate evolution (Braasch *et al.,* 2015; Xiao *et al.,* 2015). The target species are two Antarctic species *Notothenia coriiceps* and *Dissostichus mawsoni* and one Subantarctic specie *Eleginops maclovinus.*

3.2 Sequence retrieval

The genome (coding sequences) and transcriptome of the eleven species were retrieved from the National Center for Biotechnology Information (NCBI) database using BLAST that functions as a search program that allows a rapid sequence similarity comparison between a reference species and a target species (Altschul *et al.,* 1990) or from the Ensembl genome database project. The genome (coding sequences) and transcriptome FASTA files of the majority of the species where retrieved from the ENSEMBL database (Zerbino *et al.*, 2018) or from the Reference sequence (RefSeq) database of NCBI (O'Leary *et al.*, 2016) using the *biomartr R* package (Drost & Paszkowski, 2017) in *Rstudio* (Racine, 2012). The package script provided was used without alterations except for the species name and directory specification. For further information about the retrieved genomes see Table *SI*.1. The genome (coding sequences) and transcriptome of *Eleginops maclovinus*, *Notohenia coriiceps* and *Dissostichus mawsoni* were sequenced and assembled (unpublished data) by a team of researchers from Shanghai Ocean University and provided by Professor Liangbiao Chen. With the help of an in-house script the query sequences of each gene family were retrieved from their respective transcriptome file.

Specifically, terms representing each family were introduced into the script that searched through the transcriptome FASTA file to retrieve the protein sequences whose headers included the terminology used. A gene family was considered a set of genes that presented an identical gene symbol or gene description, and as exemplified in the script 1 (*SI* Script1) each gene family was extracted based on those criteria. Once the transcript sequences

of the desired gene families were retrieved, another script (*SI* Script 2) was run to keep only the longest isoforms, and the resulting sequences, the candidate sequences, were kept for the downstream analysis. As presented in table 2.1 this process resulted in a varied number of sequences depending on the gene family, ranging from a minimum of 22 sequences for the AKT3 gene family to a maximum of 172 sequences for the semaphorin gene family.

3.3 Homology

The search for homologous sequences was performed using BLAST on a local server between the before mentioned candidate sequences of *D. rerio* and all the other reference species as well as target species based on an in-house script. The first step used BLASTP with an Expect value (e-value) threshold of 1e-5 followed by filtering so as to only keep the sequences that presented an identity ratio higher than 0.1. The final procedure comprises several sorting steps and an additional filtering of the retained sequences by applying another minimum identity ratio cut-off value of 0.3.

3.4 Sequence alignment and model selection

A combined file for each gene family containing the homologous sequences between the eleven species was created to proceed for a multiple sequence alignment generated using MUSCLE (v3.8.425) (Edgar, 2004). Once the homologous sequences aligned, the best amino acids replacement models were selected for each gene family with ModelTest-NG (v.0.1.0) (Posada & Crandall, 1998) based on Akaike information criteria (AIC) and Bayesian information criteria (BIC) (Table 2.1). The model selection was carried out against 114 protein replacement models. For the five gene families both AIC and BIC indicate the same model result. The amino acid alignments obtained with the above mentioned methodology were then submitted to the PAL2NAL (Suyama *et al.,* 2006) to be converted to codon alignments for further estimation of nucleotide substitution rate and divergence time analysis.

Table 2.1 Number of sequences retrieved for the five gene families. The model with the highest score was selected for the construction of the ML-Gene trees.

				BIC			**AIC**		
	Sequences	Sites	Patterns	Model	Score	Weight	Model	Score	Weight
TLR	124	4565	2207	VT+G+F	354792.308	1	VT+G+F	353089.372	1
AKT3	22	1039	587	JTT+G	17381.6187	0.7453	JTT+G	17173.8862	0.6766
PIK3	154	3697	2390	VT+G+F	327696.886	1	VT+G+F	325684.931	1
Igsf	94	3289	3014	VT+G+F	250629.314	1	VT+G+F	249379.155	1
Sema	172	1733	1508	JTT+I+G+F	243052.849	0.8916	JTT+I+G+F	241077.194	0.9921

The same procedure as the one mentioned above was applied to the orthologous sequences obtained for each of the gene families for the construction of the alignment of the species phylogenies. The best fitting model for each of the species phylogenies is presented in table 2.2.

Table 2.2 Number of sequences retrieved for 3 concatenated supergenes PIK3, IgSf, Sema and for 2 single gene orthologous sequence TLR and AKT3. The model with the highest score was selected for the construction of the ML-Species trees.

				BIC			AIC		
	Sequences	Sites	Patterns	Model	Score	Weight	Model	Score	Weight
TLR	11	989	737	JTT+I+G4	25314.9991	0.9927	JTT+I+G4+F	25144.1749	0.9982
AKT3	11	510	217	JTT-DCMUT+I+G4	6448.6407	0.5191	JTT-DCMUT+I+G4	6359.718	0.5046
PIK3	11	3005	1648	JTT+I+G4+F	53708.1844	0.9911	JTT+I+G4+F	53467.8631	0.9986
Igsf	11	2800	1585	JTT+I+G4	51232.046	0.9964	JTT+I+G4+F	51025.7785	0.9683
Sema	11	3443	1944	JTT+I+G4	54604.8443	0.9896	JTT+I+G4+F	54406.752	0.7743

3.5 Phylogenetic analysis

3.5.1 Gene Trees

To understand the evolutionary history of the genes, five gene trees, one for each gene family of toll-like receptors (*TLR*), immunoglobulin superfamily (*IgSf*), phosphatidylinositol-4,5-bisphosphate 3-kinase (*PIK3*), AKT serine/threonine kinase 3 (*AKT3*), semaphorins (*Sema*), were constructed based on the aligned amino acid sequences of the 11 fish species *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch). The sequence alignments and the gene family phylogenies were inferred by the Maximum-Likelihood method using RaxML v0.5.1 Beta (Kozlov *et al.,* 2018) selecting the corresponding model as shown in table 2.1, with a 1000 bootstrap estimates on the online CIPRES server (Miller *et al.,* 2010).

3.5.2 Species Trees

Based on subfamily trees that presented all the eleven orthologous sequences, five ML-Tree species trees, one for each subfamily were build. The evolution of the five gene families was mapped on a species tree based on concatenated orthologous amino acid sequences of the eleven studied species. In each gene family tree, the subfamilies that presented all the eleven

studied species with a reliable topology were selected to proceed to the concatenation of the orthologous sequences. The concatenated sequences were then aligned using MUSCLE (v3.8.425) (Edgar, 2004). The sequence alignments and the species phylogenies were inferred by Maximum-Likelihood method in RaxML v0.5.1 Beta (Kozlov *et al.,* 2018) selecting the corresponding model as shown in table 2.2, with 1000 bootstraps on the online available CIPRES server (Miller *et al.,* 2010).

3.6 Nucleotide substitution rate (ϖ = dN/dS)

The orthologous coding sequences obtained with PAL2NAL were concatenated before proceeding to the dN/dS estimation. The nucleotides substitution rate (ϖ), the non-synonymous substitution rate (dN) and the synonymous substitution rate (dS) were then estimated using CODEML (PAML 4 package (Z. Yang, 2007)) based on a "free-model" (model= 1, runmode= -2), which performs a pairwise analysis and allows branch-specific values for ϖ, dN and dS. A threshold to filter out the nodes presenting dS saturation was set at dS < 4.

3.7 Divergence time

Divergence time between gene families and species were calculated based on the synonymous substitution rate of the CODEML analysis. As presented by Wang *et al.*, (2015) the equation $T=Ks/2r$ can be applied to calculate the divergence time, where T is the divergence time to the most recent common ancestor, obtained from the Timetree database (Hedges *et al.,* 2006), Ks is the synonymous substitution rate recovered from the CODEML analysis and r is the estimated substitution rate given as mutations per site per year which has to be calculated by a prior conversion of the equation $r = (Ks/T)/2$. In this study the number of genes retrieved is not enough to obtain a reliable estimated substitution rate and the value obtained by Wang *et al.*, (2015) was used. In their case they calculated the estimated substitution rate by using a calibration point set at the divergence time between *D. rerio* (205−255 Mya), retrieved from the TimeTree database (Hedges *et al.,* 2006), and five other species (*O. latipes, G. aculteatus, Takifugu rubripes, Tetraodon nigroviridis* and *G. morhua*). From those six species five were studied in this thesis. With this method the authors obtained a substitution rate of $5.7 – 6.4 \times 10^{-9}$. This same substitution rate was then used here to estimate the divergence times.

3.8 Fossil and biogeographic node ages estimates

Neopterygii MRCA *L. oculatus, D. rerio* - Broughton *et al.,* 2013 used the fossil record of the oldest extinct "semionotiform" *Acentrophorus varians* to infer the minimum node age for neopterygii at 260 m.y.a. The authors inferred a maximum node age of 386-375 m.y.a from two of the oldest stem-group actinopterans (Hurley *et al.* 2007).

Teleostei MRCA *D.rerio* (Otocephala), *G.morhua* (Acanthomorphata) - Benton *et al.,* (2015) inferred a minimum age of divergence between Otocephala and Euteleostei of 150 m.y.a with the age estimate of an early Orthogonikleithridae, *Leptolepides haerteisi*, based on the specimens retrieved by Arratia (1996) who identified it as part of the Euteleostei. To infer the maximum bound for the node age of Cupeocephala Benton *et al.,* (2015) estimated that the origin of crown Clupeocephala can be retraced to a time period that doesn't exceed the base of the Ladinian age which is 242 m.y.a. Based on this information a maximum age node for Teleostei of 242 m.y.a and a minimum age node of 150 m.y.a was set for this study.

Acanthomorphata MRCA *G.morhua* (Gadiformes), Percomorpha: Alfaro *et al.,* (2009) and Chen *et al.,* (2014) used the oldest fossil otoliths studied for the genus "*Acanthomorphorum*" estimated at 124-122 m.y.a, considered the first representative of the acanthomorphata (Nolf, 2004). Chen *et al.,* (2014) used this fossil record as a constraint for their molecular estimation of the appearance of Acanthomorphs and obtained a time interval of 136-166 m.y.a. The minimum node age could be set between 83 m.y.a and 96.9 m.y.a the former retrieved from the oldest fossil record found for Gadiformes and Zeiformes (*Cretazeus rinaldii*) (Chen *et al.,* 2014) and the latter found for the oldest fossil record for Percomorphs (Matschiner *et al,* 2011). Based on this information a maximum age node for Acanthomorpha of 150 m.y.a and a minimum age node of 96 m.y.a was set for this study.

Percomorpha MRCA Ovalentaria, Percomorpharia (=Eupercaria): The estimation of Percomorpha node age is derived from the interpretations of the fossil records presented in Matschiner *et al,* (2011) Benton *et al.,* (2015) and the references therein. A maximum node age may be set at 150.9-96.9 m.y.a corresponding to the age of the strata, where the fossil remnants of the oldest percomorph were found (Chen *et al.,* 1998 *in* Matschiner *et al,* 2011). Benton *et al.,* (2015) identified a minimum age estimate between Tetradodontiformes, a clade belonging to the Percomorpharia, and Ovalentaria of 69.71 m.y.a. The authors estimated this period based on the age of the layer from which the fossil of, *Cretatriacanthus guidotti* the youngest known tetradodontiform, was obtained. Based on this information a maximum age node for Percomorpha of 120 m.y.a and a minimum age node of 69 m.y.a was set for this study.

Perciformes: Betancur-R *et al.* (2013) pointed out to the necessity of clarifying the group Perciformes since as they stated in their work this group was for long considered as "the waste basket" in the teleost phylogeny including most of the modern teleost species but for which the monophyly was not established. They were able to identify Perciformes as a single monophyletic group containing a maximum of 71 families which diversified around 100 m.y.a.

Notothenioidei *E. maclovinus*, Nototheniidae: The first identified fossil belonging to the notothenoids is the one of *Proeleginops grandeastmanorum* described by Balushkin (1994), dating back to the late Eocene (~40m.y.a) during the La Meseta Formation on Seymour Island. Until now it was the only fossil record used as the calibration point for the radiation of notothenoids in the Antarctic Ocean (Near, 2004; Prisco *et al.,* 2007). Recently the fossil of *Mesetaichthys jerzmanskae* which is 10 m.y.a younger than *P. grandeastmanorum*, found by Balushkin (1994), has been described as a close relative to the Nototheniidae genus *Dissostichus* (Bieńkowska-Wasiluk *et al.,* 2013). Both *M. jerzmanskae* and *P. grandeastmanorum* could be used for the calibration of the Notothenioidei (Sub-Antarctic and Antarcitc species) clade. A maximum node age was set at 40 m.y.a corresponding to the first fossil record of suborder Notothenioidei and the minimum node age was set at 30 m.y.a with appearance of the first identified Nototheniidae *M.jerzmanskae*.

Nototheniidae *N. corriceps, D. mawsoni*: The maximum node limit for the Nototheniidae family was set at 30 m.y.a. which was obtained from the first fossil record of the closest relative to this family *M. jerzmanskae* (Bieńkowska-Wasiluk *et al.,* 2013). The radiation of the Nototheniidae in the Antarctic Ocean after the onset of the ACC which happened 20 m.y.a was used as the lower node limit.

4. Results

4.1 Sequence retrieval

The methodologies applied allowed retrieval of a variable number of genes and protein transcripts for each gene family in each of the species (Figure 3.1). Since the sequences retrieved from *D. rerio* were used as query for the search of homologs in the other species, the number of genes identified in those species was expected be lower than in *D. rerio*. The number of isoforms can be found in Table *SI*.2.

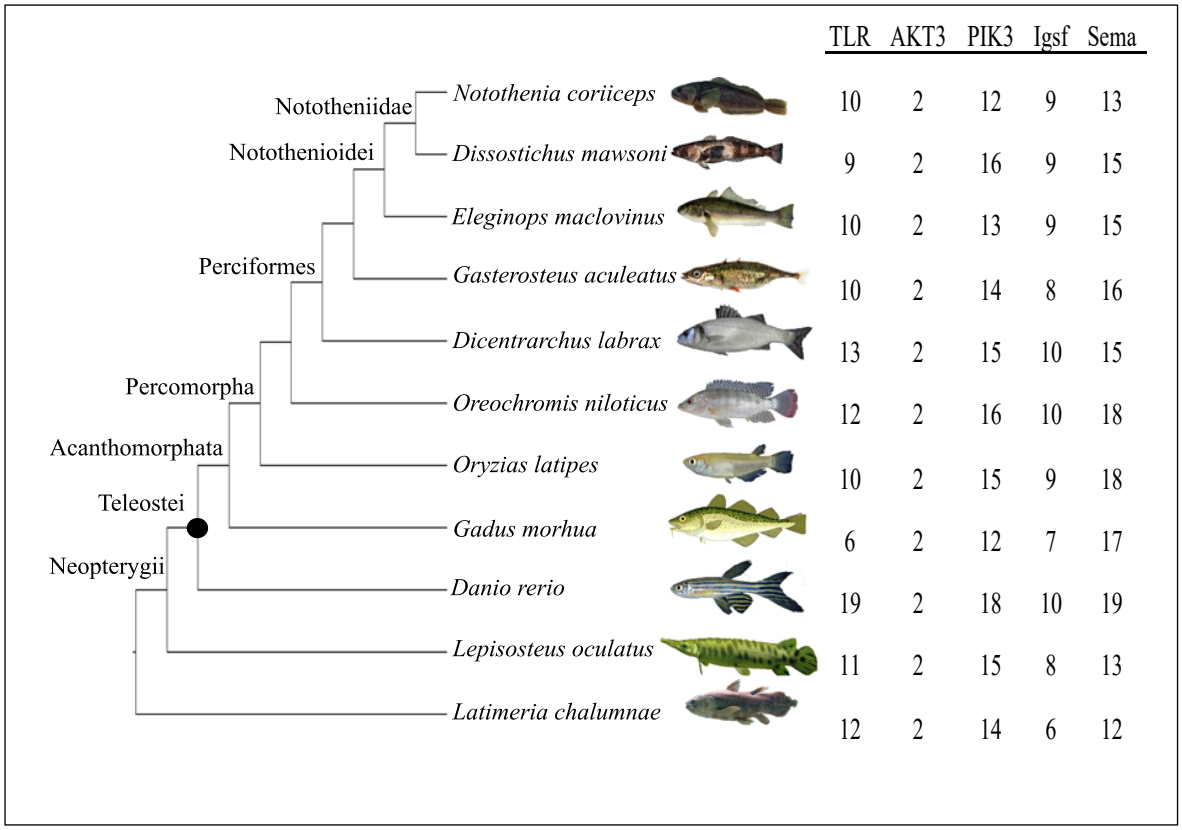

Figure 4.1: Cladogram of the used species with the number of genes retrieved in each gene family (*TLR, AKT3, PIK3, IgSf, Sema*). The black dot (●) indicates the teleost whole genome duplication. These numbers do not include the number of duplicates. The Accession numbers of each gene is given in Table *SI*.3 and the corresponding. References for the species images (Lecointre, 2004; Cada, 2005; Hillewaert, 2005; Strauss, 2006; Praxaysombath, 2008; Shandikov, 2013; O'Brien, 2011; smerikal, 2011; Ueda, 2013 ; Tamaki & Maeda, 2016a , 2016b)

.

Members for all the six TLR subfamilies were found in all species. In general, the TLR family had a higher variability in the number of retrieved genes between species than any other of the remaining families. The species with the lowest number of TLR genes was *G. morhua* with only 6 genes and the one with highest number, 19 TLR genes was *D. rerio* followed by *D. labrax* with 13 genes. Three of the TLR genes were only found in *D. rerio* (*tlr20.4*, *tlr4al*, *tlr4bb*) and two were only found between *D. rerio* and either *L.oculatus* (*tlr19*) or *L.chalumnae* (*tlr4ba*).

N. coriiceps and *G. morhua* were both lacking the tlr8b gene. The Notothenioidei were missing *tlr22* which was otherwise present from the Neopterygii on. *D. mawsoni* was the only member of the Percomorphs missing the *tlr21* gene. *E. maclivnus* and *D. mawsoni* were both lacking *tlr18*. The only TLR gene present in all the species was *tlr3*. The AKT3 was the smallest represented gene family with only 2 genes *AKT3a* and *AKT3b*. Although AKT3 had a small number of genes this number is conserved throughout all the species. The highest number of genes for the PIK3 family was found in *D. rerio* with a total of 18 genes, whereas the smallest number was found in *G. morhua* and *N. coriiceps* with twelve genes each. The *pik3r3a* was only found in *D. rerio* and *L. chalumnae*. The Nototheniidae were missing three PIK3 genes (*pik3c2b*, *pik3r4*, *pik3r6b*) whereas *N. coriiceps* was lacking two (*pik3c2a*, *pik3cb*) that were present in the other Notothenioidei. The IgSf presented a maximum of 10 genes in *D. rerio* and a minimum of 6 genes in *L. chalumnae*. The Notothenioidei share the same number of IgSf genes with *E. maclovinus* and *D. mawsoni*, both lacking *IgSf1*, whereas *N. coriiceps* was the only member of the Percomorphs missing *IgSf5a*. The gene family containing the highest number of genes was the semaphorin family with 19 genes found in *D. rerio*, however, those genes belonged either to *sema4* or *sema3*. *L. chalumnae* had the smallest number of semaphorin genes counting a total of 12 genes. In the semaphorin family the Notothenioidei lacked 3 identical genes with *sema4bb*, *sema3gb* and *sema3fa* missing. It should be noted that *sema4bb* and *sema3gb* were only present from the Neopterygii to the Percamorpha. In contrast, *sema3fa* was already present in *G. aculeatus*.

4.2 Phylogenetic analysis

4.2.1 Gene family Trees

In 4 out of the 5 gene family trees all the subfamilies that had a sequence for all the 11 investigated species, could be considered orthologs as they regrouped into their specific subtree (Ohta, 2008). A total of 10 subfamilies were found throughout the 5 gene families that comprised all the orthologous sequences. All the nodes had a strong bootstrap support with values comprised between 73 and 100 except for *pik3cd* for the Nototheniidae (bootsrap value

= 54) and the *pik3c3* for the Notothenioidei (bootsrap value = 64). From the 10 subfamilies, *akt3a* was the only one presenting a different configuration. The other 9 subfamilies presented a similar tree configuration consisting of *L. chalumnae* as the root species followed by the same order of nodes, Neopterygian, Teleosts, Percomorpha, Perciformes and ending with the crown node of Notothenioidei which also included the Nototheniidae.

The TLR family was built with 124 sequences and presented the longest amino-acid sequences after alignment with a total of 4565 sites. Only the *tlr3* gene subfamily presented 11 orthologous sequences. The AKT3 family contained the smallest number of sequences with only 22 sequences and had the shortest alignment from the five studied gene families with only 1039 sites. Both genes retrieved as members of the AKT3 family did not form a distinct pattern between the two subfamilies and only *akt3a* had all of the 11 orthologous sequences. Even though this subfamily did not regroup into a specific subfamily tree it was retained so as to continue the downstream analysis with 5 gene families. The PIK3 family comprised 154 sequences and had an alignment with a total length of 3697 sites. Three subfamilies *pik3cg*, *pik3c3* and *pik3cd* presented 11 orthologous sequences. The IgSf gene tree was built with 94 sequences with an alignment of containing 3289 sites. As in PIK3, 2 subfamilies *IgSf3* and *IgSf8* had all the 11 orthologous sequences. Lastly, the semaphorin gene tree with its 172 sequences counted the highest number of sequences but the alignment with a total of 1733 sites, only surpassed AKT3. The Sema family counted 3 subfamilies that totaled all 11 orthologous sequences of *sema3b*, *sema3bl* and *sema3c*.

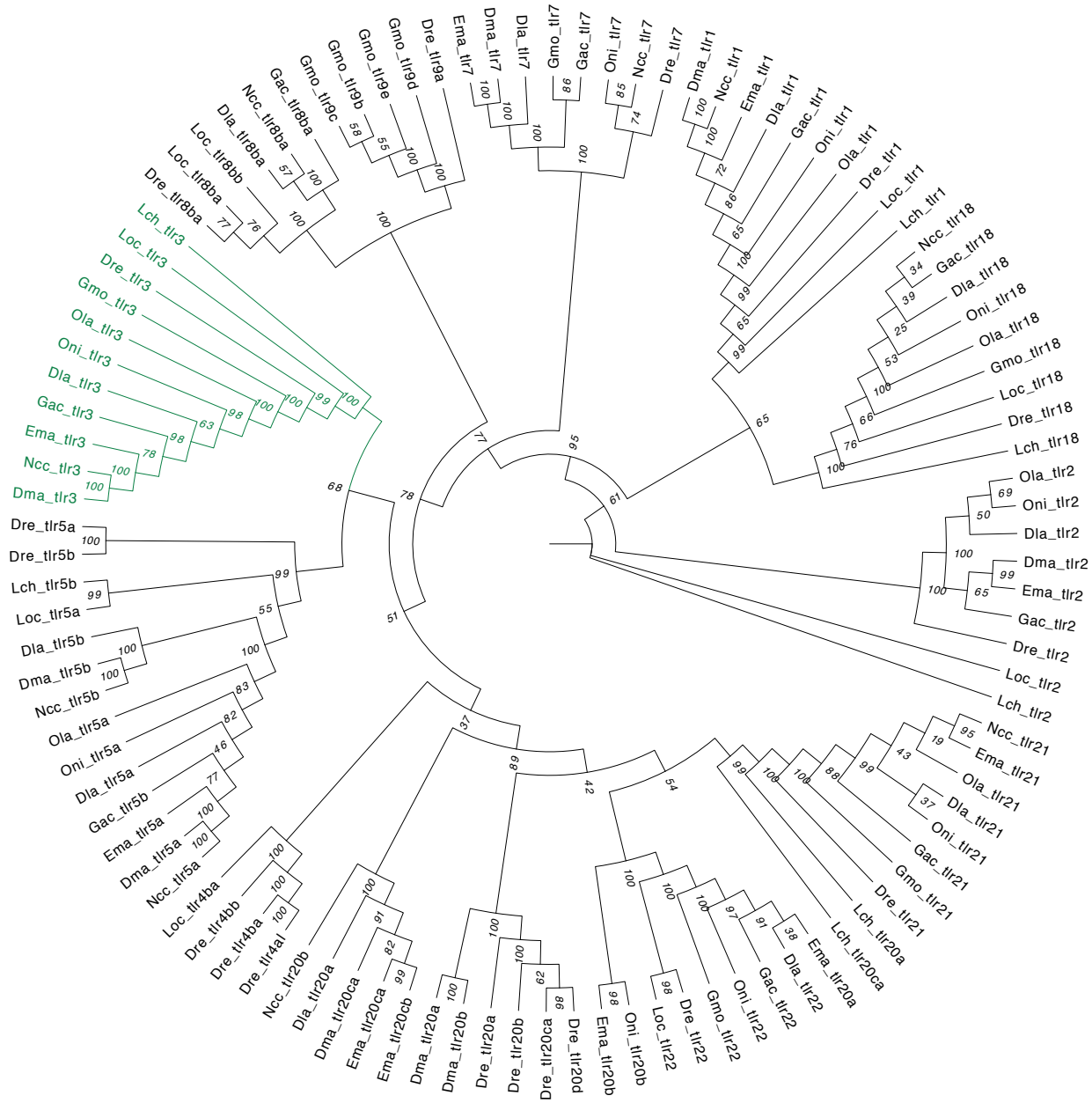

Figure 4.2: **Phylogenetic Maximum-Likelihood gene tree for Toll-Like Receptor family of *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch).** The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a VT+G+F substitution model. The bootstrap values are given in italic next to the nodes. The green highlighted *tlr3* subtree represents the only subtree composed by all the orthologs.

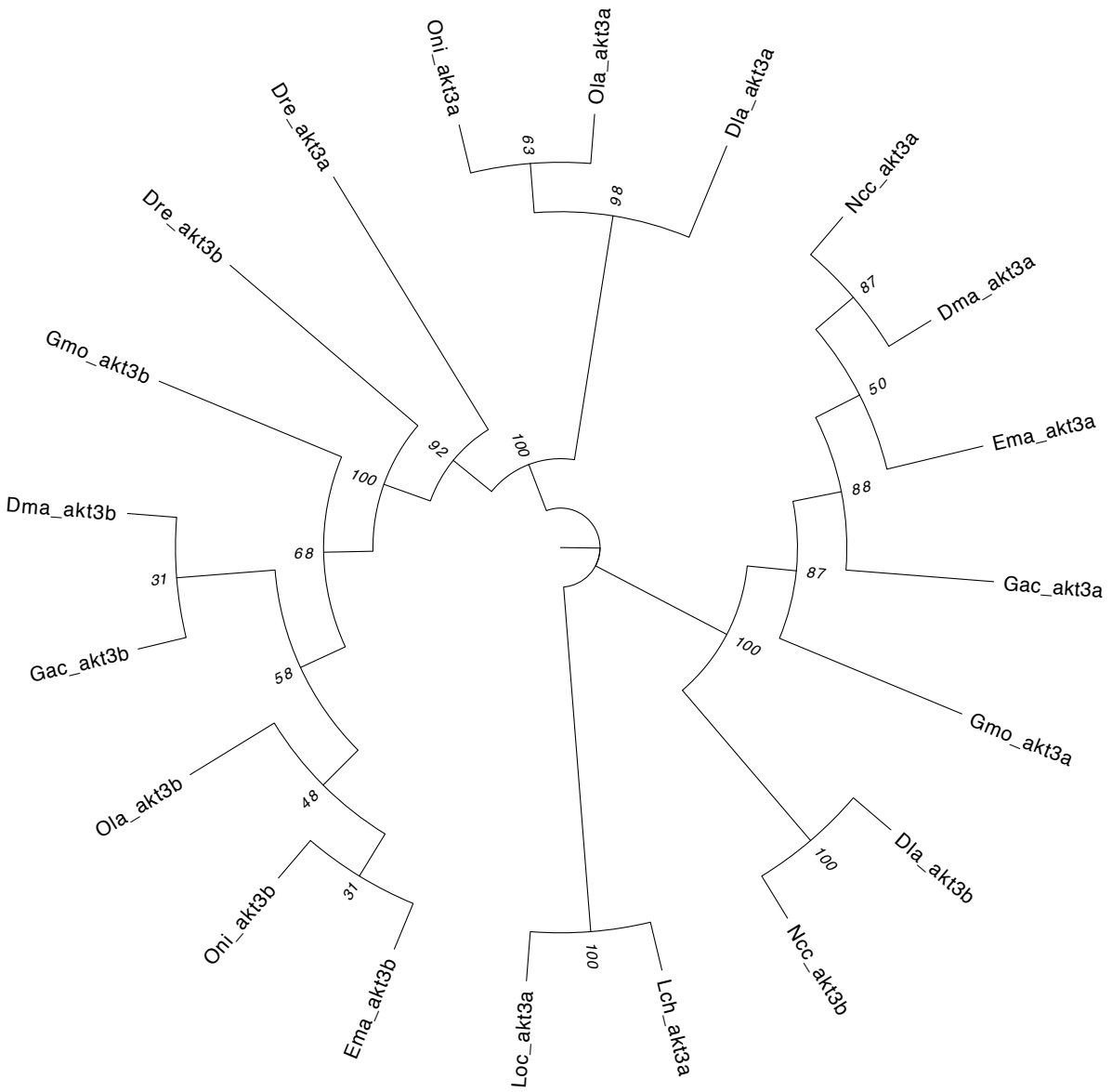

Figure 4.3: **Phylogenetic Maximum-Likelihood gene tree for AKT serine/threonine kinase 3 family of *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch).** The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a JTT+G substitution model. The bootstrap values are given in italic next to the nodes.

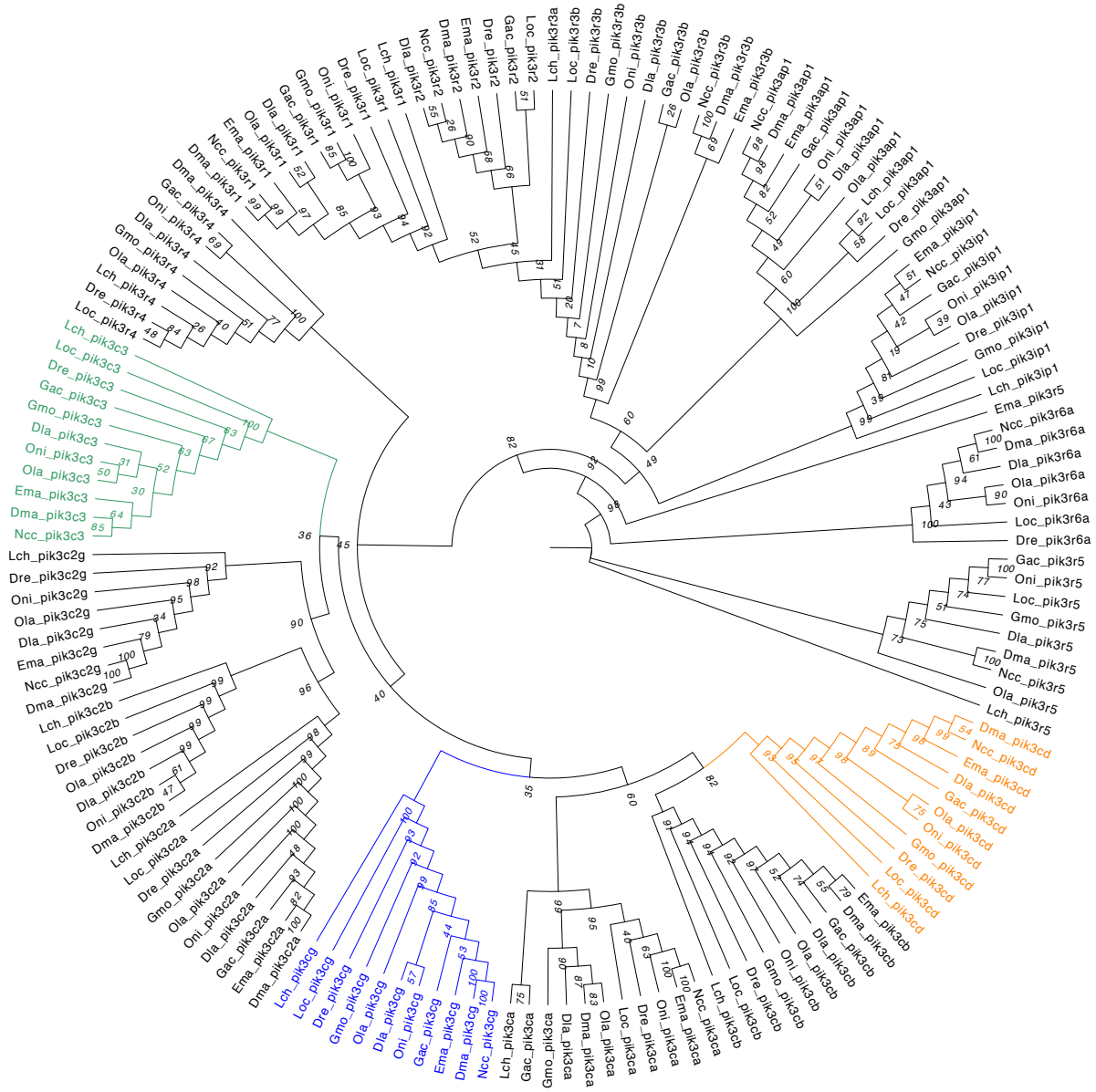

Figure 4.4: Phylogenetic Maximum-Likelihood gene tree for Phosphatidylinositol 3-kinase of *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch). The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a VT+G+F substitution model. The bootstrap values are given in italic next to the nodes. The subfamilies containing the eleven orthologs *pik3c3*, *pik3cd* and *pik3cg* are highlighted in respectively green, orange and blue.

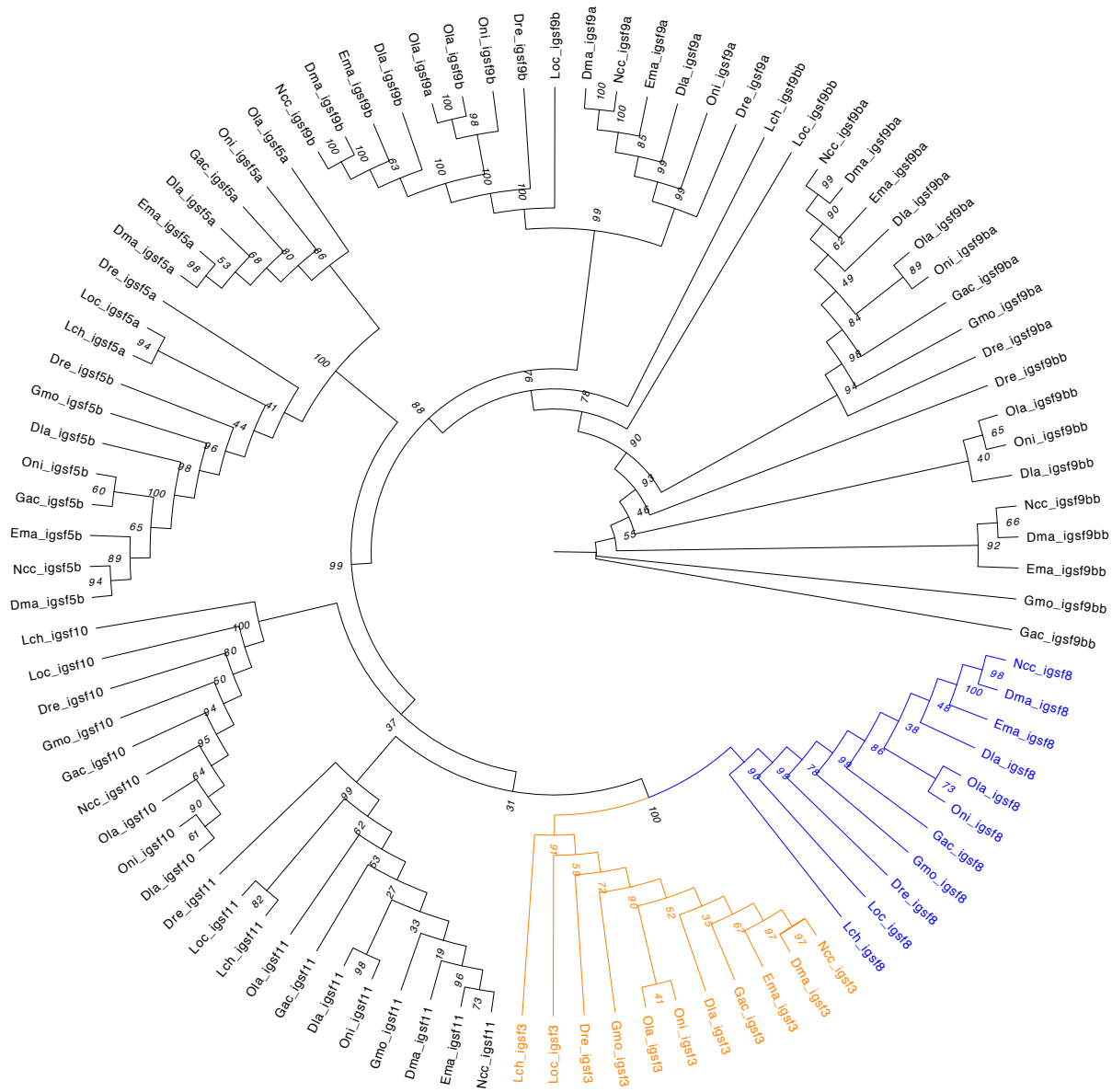

Figure 4.5: Phylogenetic Maximum-Likelihood gene tree for Immunoglobulin Superfamily between Notothenia coriiceps (Ncc), Eleginops maclovinus (Ema), Dissostichus mawsoni (Dma), Dicentrarchus labrax (Dla), Danio rerio (Dre), Lepisosteus oculatus (Loc), Oreochromis niloticus (Oni), Oryzias latipes (Ola), Gadus morhua (Gmo), Gasterosteus aculeatus (Gac) and Latimeria chalumnae (Lch). The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a VT+G+F substitution model. The bootstrap values are given in italic next to the nodes. The subfamilies containing the eleven orthologs *IgSf3* and *IgSf8* are highlighted in orange and blue respectively.

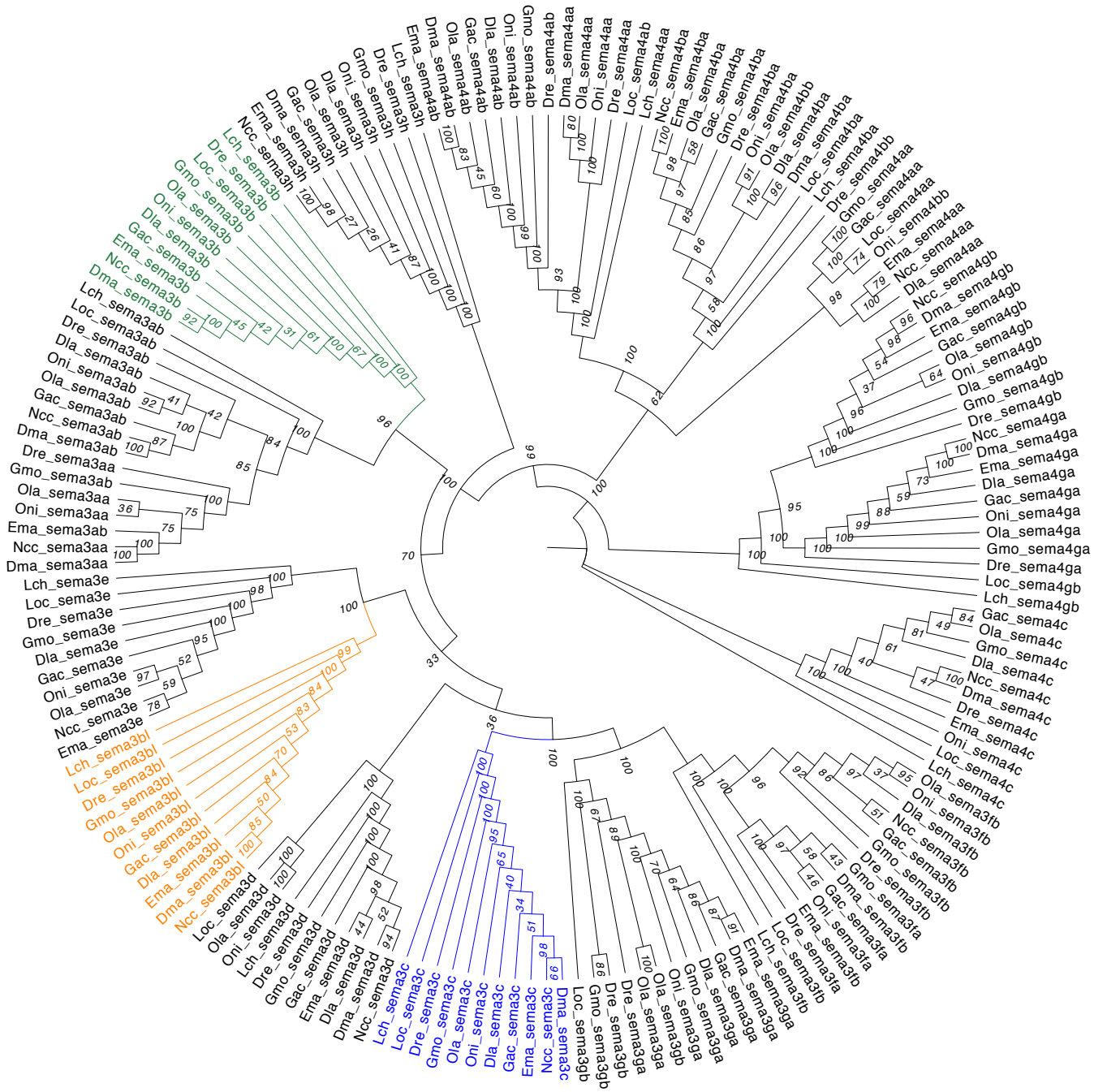

Figure 4.6: **Phylogenetic Maximum-Likelihood gene tree for Semaphorin family of *Notothenia coriiceps* (Ncc), *Eleginops maclovinus* (Ema), *Dissostichus mawsoni* (Dma), *Dicentrarchus labrax* (Dla), *Danio rerio* (Dre), *Lepisosteus oculatus* (Loc), *Oreochromis niloticus* (Oni), *Oryzias latipes* (Ola), *Gadus morhua* (Gmo), *Gasterosteus aculeatus* (Gac) and *Latimeria chalumnae* (Lch).** The tree was generated by RaxML (v0.5.1 Beta) with 1000 bootstrap estimates and a JTT+I+G+F substitution model. The bootstrap values are given in italic next to the nodes. The subfamilies containing the eleven orthologs *sema3b*, *sema3bl* and *sema3c* are highlighted in respectively green, orange and blue.

4.2.2 Species Trees

For each species tree, only those gene subfamilies represented in all 11 species were kept for further analysis. This step reduced considerably the number of genes used in the study. Specifically, Sema and PIK3 had 3 and IgSf had 2 subfamilies with 11 orthologous sequences each, this accounted for a total of 33 sequences for Sema and PIK3 and 22 for IgSf. For each gene family, those sequences were then concatenated into one supersequence per species yielding a total of eleven supersequences for the construction of the species tree. The TLR and AKT3 species tree were built only on 11 un-concatenated sequences since each counted one subfamily with eleven orthologs. The aim of this method was to identify how the species would cluster for each one of the gene families and to confirm if the topologies would be identical for the five gene families. This should allow to infer if the gene families are under selective pressure.

In the species phylogenies obtained for TLR, PIK3, IgSf and Sema the Neopterygii, Teleostei, Acanthomorphata, Notothenioidei and Nototheniidae nodes were conserved and were highly supported with bootstrap values of 100. The Percomorpha and Perciformes nodes were conserved in the TLR, IgSf and Sema species phylogeny ranging from minimum bootstrap values of 56 and 68 found in the IgSf and Sema Perciformes respectively and bootstrap values of 100 for all the Percomorpha nodes. The PIK3 species phylogeny had a minor difference with the above-mentioned species phylogenies by excluding *D. labrax* from the Perciformes. The AKT3 species phylogeny was characterized as the most irregular phylogeny when compared with the others. For the AKT3 species phylogeny the Notothenioidei was supported by 59 bootstrap values whereas Nototheniidae was supported by 80 bootstraps. The Teleostei were divided into two distinct clades with no confirmed phylogenies. One of the clades was composed of four of the five Perciformes and included *G. morhua* as a sixth member. *D. labrax* was integrated into the second clade with *D. rerio, O. latipes* and *O. niloticus*.

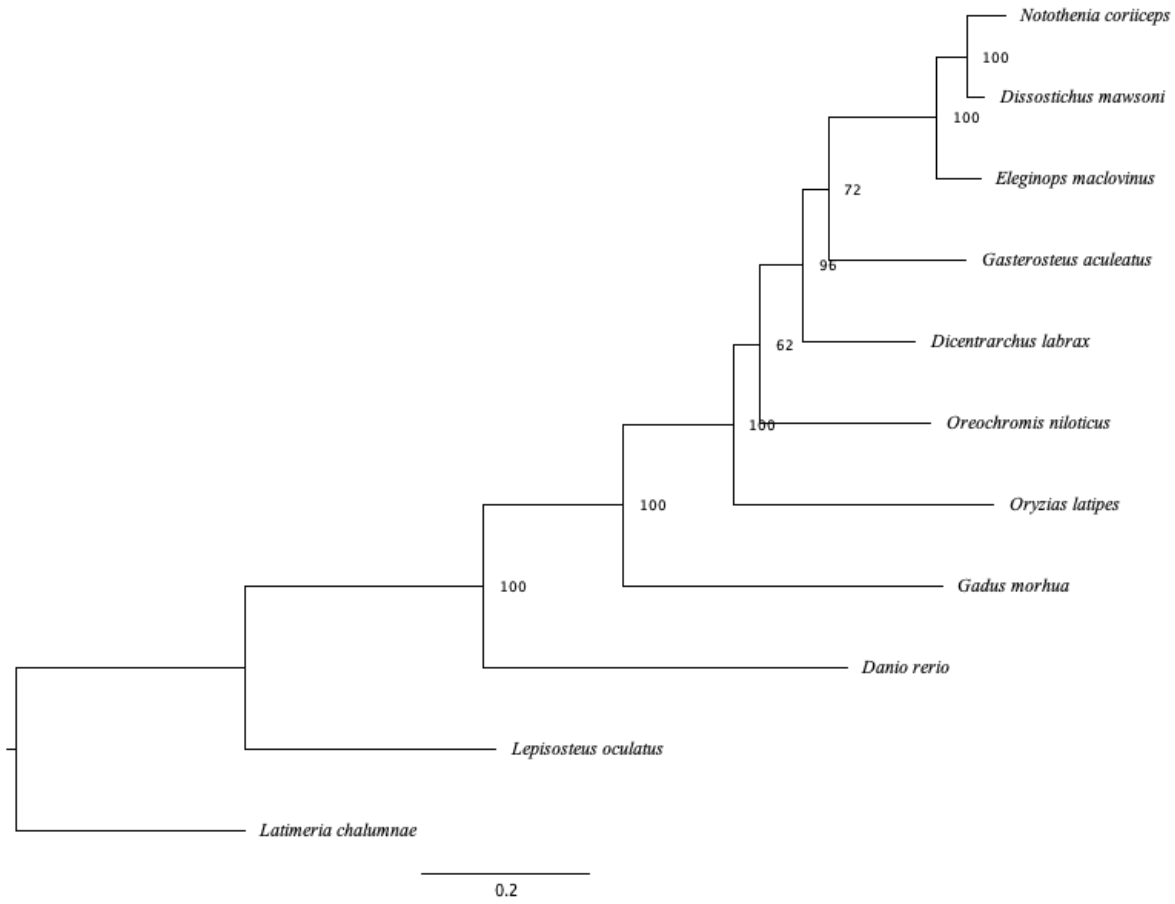

Figure 4.7: Phylogenetic Maximum-Likelihood tree for Toll-Like Receptor based on 11 1:1 orthologous protein sequences from eleven studied fish, showing the relationships between Notothenioidei (*N. coriiceps*, *E. maclovinus*, *D. mawsoni*) and other fish species. The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumnae* was used as outgroup. The bootstrap values are given in italic next to the nodes.

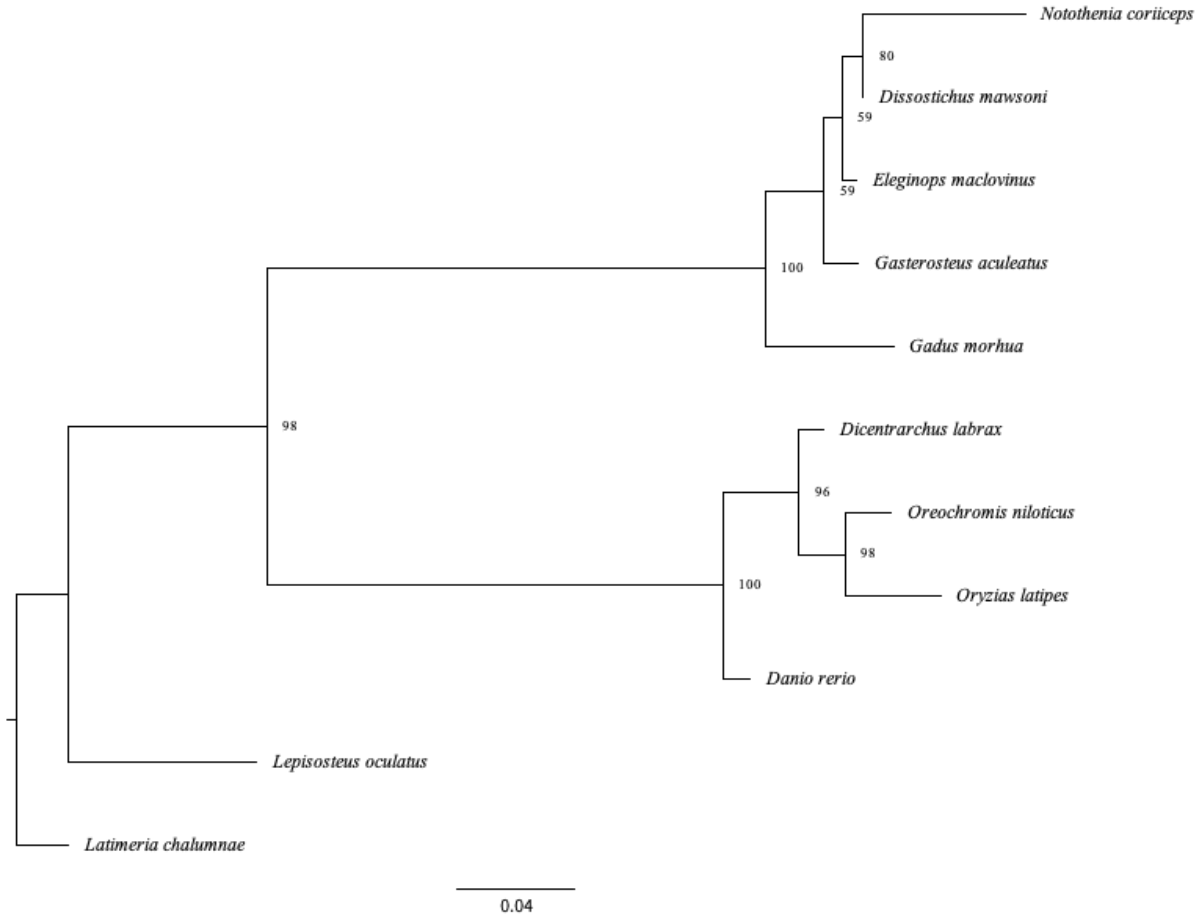

Figure 4.8: Phylogenetic Maximum-Likelihood tree for AKT serine/threonine kinase 3 based on 11 1:1 orthologous protein sequences from eleven studied fish, showing the relationships between Notothenioidei (*N. coriiceps*, *E. maclovinus*, *D. mawsoni*) and other fish species. The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumna*e was set as outgroup. The bootstrap values are given in italic next to the nodes.

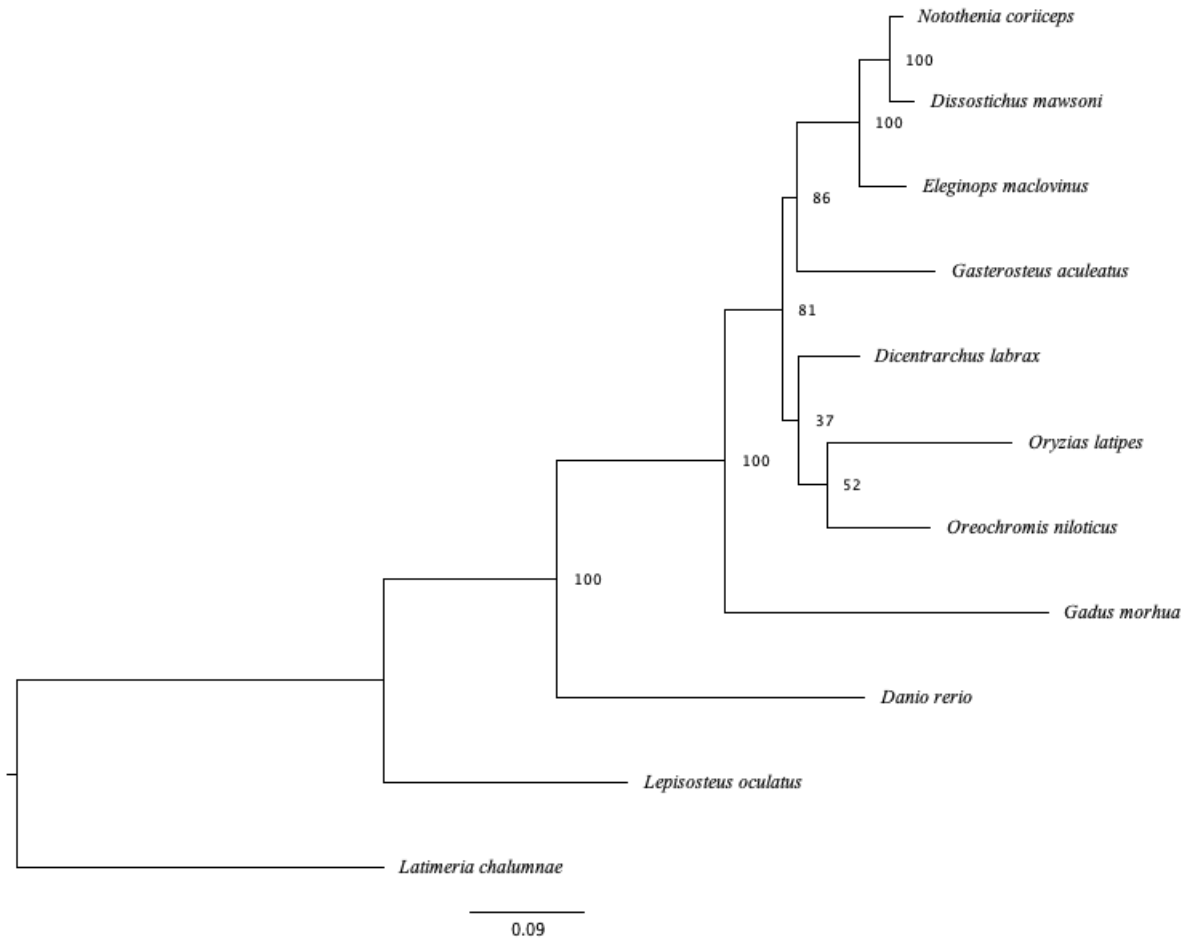

Figure 4.9: **Phylogenetic Maximum-Likelihood tree for Phosphatidylinositol 3-kinase based on 33 1:1 orthologous protein sequences from eleven studied fish, showing the relationships between Notothenioidei (*N. coriiceps, E. maclovinus, D. mawsoni*) and other fish species.** The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumnae* was used as outgroup. The bootstrap values are given in italic next to the nodes.

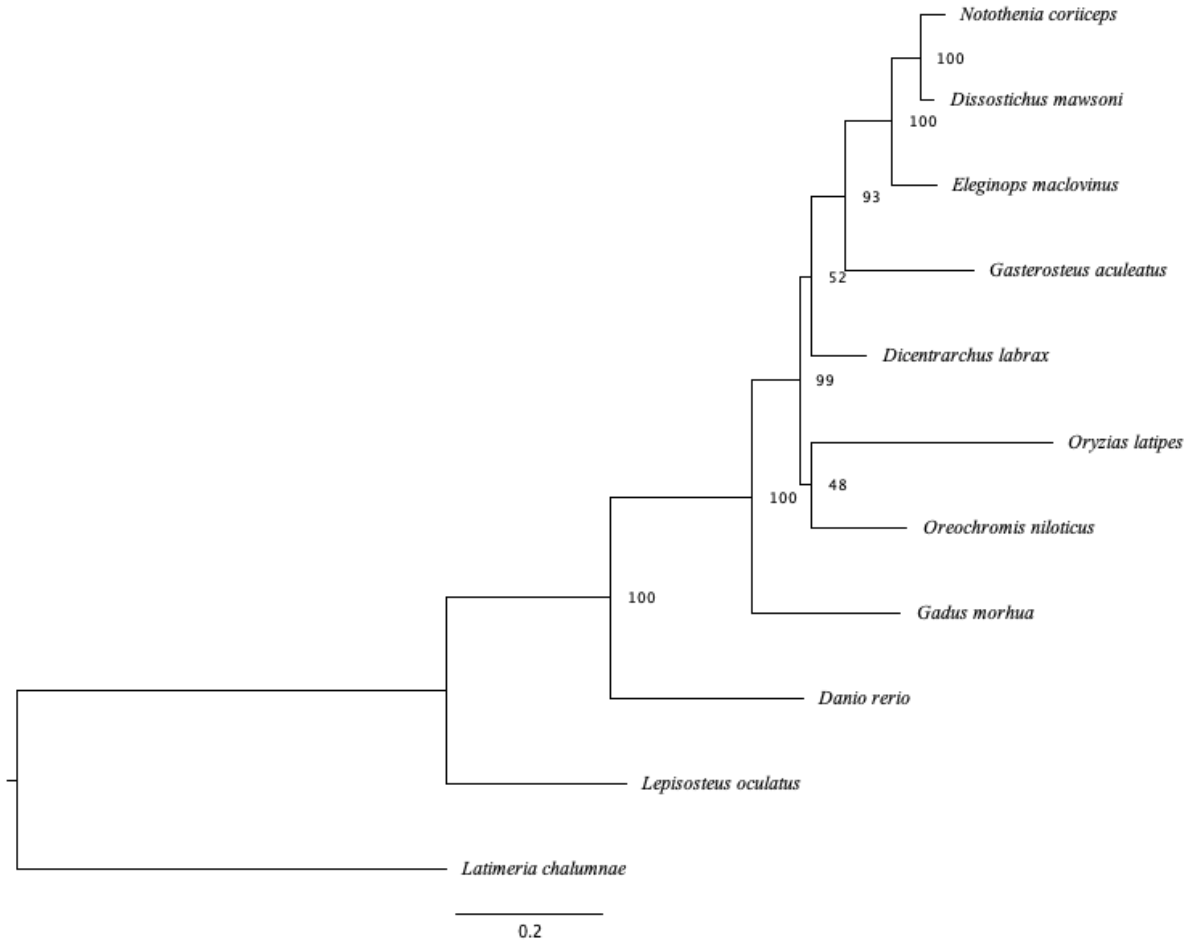

Figure 4.10: Phylogenetic Maximum-Likelihood tree for Immunoglobulin Superfamily based on 33 1:1 orthologous protein sequences from eleven studied fish, showing the relationships between Notothenioidei (*N. coriiceps, E. maclovinus, D. mawsoni*) and other fish species. The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumnae* was used as outgroup. The bootstrap values are given in italic next to the nodes.

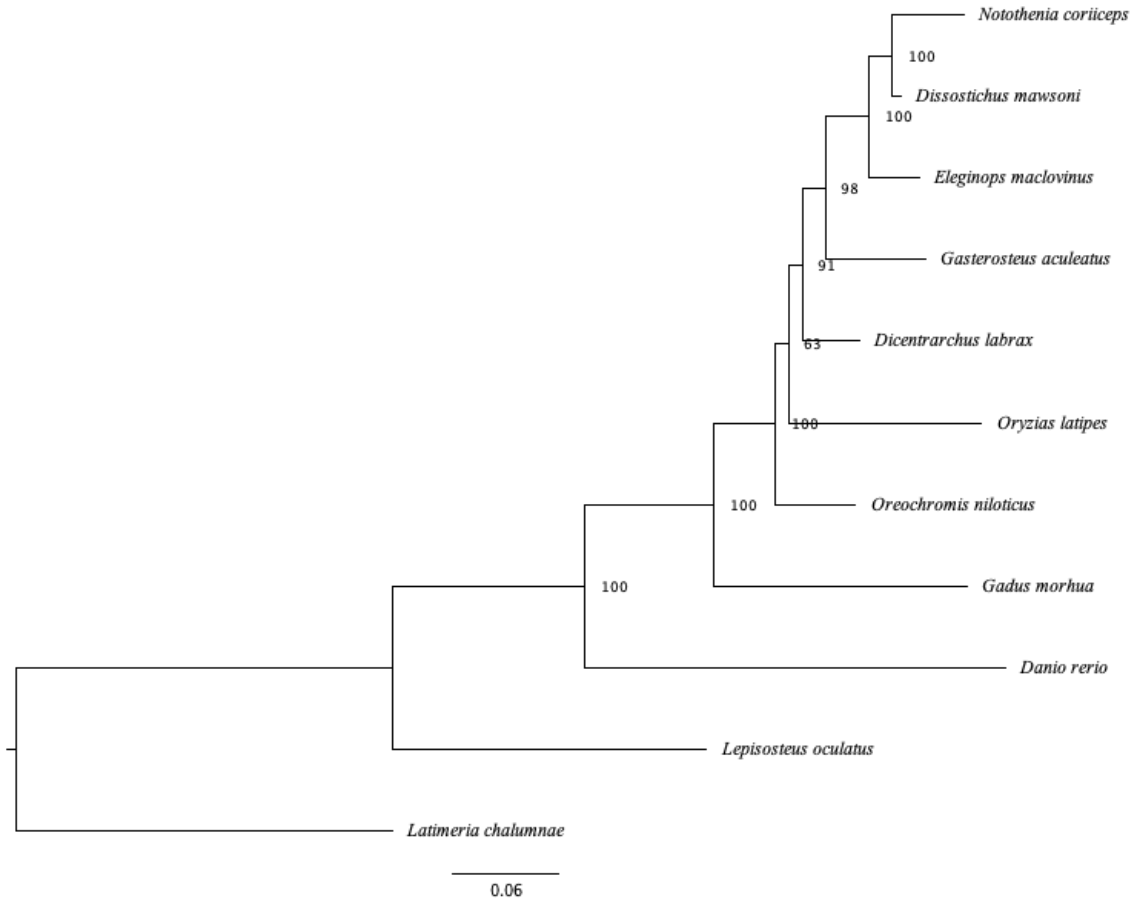

Figure 4.11: Phylogenetic Maximum-Likelihood tree for Semaphorin based on 33 1:1 orthologous protein sequences from eleven studied fish, showing the relationships between Notothenioidei (*N. coriiceps, E. maclovinus, D. mawsoni*) and other fish species. The tree was generated by RaxML (v0.5.1 Beta) with an VT+G+F substitution model. The actinopterygian *L. chalumnae* was used as outgroup. The bootstrap values are given in italic next to the nodes.

4.3 Divergence time estimate

Even though some discrepancies were observed between the gene families phylogenies, the same node order containing the same species as depicted in the cladogram of Figure 3.12, was kept for comparison. The differences in divergence time between the nodes should permit an overview of the evolution of the five gene families through the phylogeny and confirm if they are similar. The missing gene families in the individual nodes weren't caused by missing values but were caused by the threshold set for this specific analysis which eliminated all the pairwise values of synonymous substitution rate which yielded a value higher than four. For this same reason was *L. chalumnae* excluded from this analysis.

After the Acanthomorphata all nodes included divergence time estimates for the studied gene families. The Neopterygii node was only represented by the Semaphorin gene family whereas the Teleostei node was missing the TLR family. In Teleostei the mean age estimates with a 95% confidence interval of AKT3, PIK3 and IgSf are inside the age estimate obtained by the fossil record. For Acanthomorphata only TLR and AKT3 are comprised in the limits of the fossil record whereas the remaining gene families lie outside the upper limit of the fossil record. Percomorpha had their divergence time estimate for the five gene families inside the boundaries set by the fossil record. For Perciformes, Notothenioidei and Notothenoiidae the divergence time estimates of the five gene families are outside the lower limit of the fossil record.

Specifically for the Notothenoiidae, *N. coriiceps* and *D. mawsoni*, the divergence time estimates obtained for each gene family had noticeable variances with estimates varying between a maximum of 7.11 m.y.a for the semaphorin family followed by AKT3 with 6.17 mya, and PIK3 and IgSf, respectively, with 4 m.y.a and 4.43 m.y.a, to a minimum of 2.49 m.y.a for TLR (Table 3.2). The divergence time estimates from the Notothenoiidae to *E. maclovinus* yielded similar age estimates for TLR, PIK3 and IgSf whereas for Sema and AKT3, those estimates varied. The divergence time estimate for AKT3 between *N. coriiceps* and *E. maclovinus* was 11.54 m.y.a, whereas between *D. mawsoni* and *E. maclovinus* was 14.1 m.y.a.. With Sema these comparisons yielded, respectively, 14.12 m.y.a and 18.95 m.y.a (Table 3.2). In four of the five gene families, divergence times estimated between the tree Notothenioidei and their most recent common ancestors, the perciform *G. aculetaus,* yielded comparable results while the mean IgSf divergence time estimates indicated older divergence times than the others (Table *SI.*5).

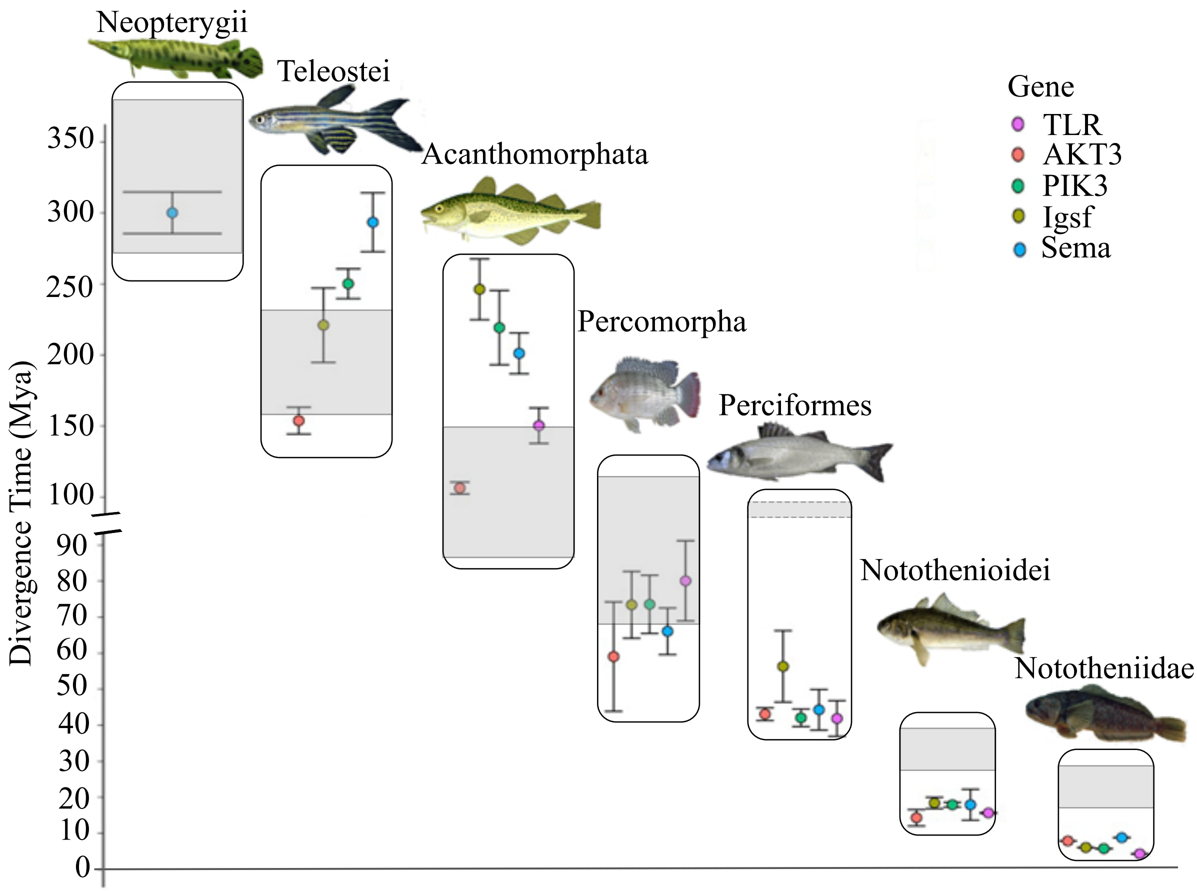

Figure 4.12: Divergence time estimates in millions of years for seven nodes, calculated with the synonymous substitution rate obtained for the five gene families, Toll-Like receptor (pink), AKT serine/threonine kinase 3 (red), Phosphatidylinositol 3-kinase (green), Immunoglobulin Superfamily (olive) and Semaphorins (blue) and fossil representation of the studied fish phylogeny. The circles represent the mean estimated divergence times and the whiskers mark the upper and lower limit of the 95% confidence interval for the age estimates. The gray boxes represent age estimate for the appearance of the node which were assigned with the fossil information retrieved from the literature (Materials and Methods). The mean values with their confidence intervals may be found in Table *SI*.5.

Table 3.1: Pairwise synonymous substitution value (dS-value) and the calculated divergence time (million years ago) from one-to-one concatenated genes between, *N. coriicpes* (Ncc), *D. mawsoni* (Dma), *E. maclovinus* (Ema) and *G. aculeatus* (Gac), using the estimated substitution rate at 5.7×10^{-9} mutations per site per year.

Species	Pairwise dS-value					Estimated Divergence Time (m.y.a)				
	TLR	AKT3	PIK3	IgSf	Sema	TLR	AKT3	PIK3	IgSf	Sema
Gac vs. Ncc	0.57	0.48	0.55	0.80	0.65	50.4	42.2	48.0	69.9	57.2
Gac vs. Dma	0.56	0.46	0.53	0.78	0.57	49.1	40.7	46.6	68.1	49.6
Gac vs. Ema	0.58	0.51	0.49	0.88	0.62	51.0	44.5	42.7	77.5	53.9
Ema vs. Ncc	0.16	0.16	0.19	0.20	0.22	14.3	14.1	16.9	17.9	18.9
Ema vs. Dma	0.16	0.13	0.18	0.18	0.16	14.1	11.5	16.2	16.2	14.1
Ncc vs. Dma	0.03	0.07	0.05	0.05	0.08	2.5	6.2	4.0	4.3	7.1

5. Discussion

In the present study, the number of genes for each species, the gene and species tree phylogenies and the clustering of orthologous genes allowed to idenfy a birth-and-death process for the five gene families in the investigated species. The divergence times estimated for the five gene families appeared to be dependend on the number of species used. The nodes containing more species yielded older divergence times whereas a reduced number of species yielded younger divergence times. Also, when looking specifically at the nototheniidae it appearded that the divergence time estimates were depended on gene functions. Furthermore, the recent divergence time estimates obtained for the five gene families in nototheniidae, seemed to indicate that the adaptation of the immune system in those species followed the Middle Miocene climatic transition.

5.1 Sequence retrieval & Phylogenetic analysis

Immune-related gene families have been shown to evolve in accordance to a process called birth-and-death (Nei *et al.,* 1997; Piontkivska & Nei, 2003). The birth-and-death process stipulates that the gene diversity seen between species is caused by species-specific genomic events like gene duplications resulting in genes that are maintained in the genome and others who become non-functional due to mutations or are lost (Nei & Rooney, 2005). Annilo *et al.,* (2006) established that the ATP-Binding cassette (ABC) multigene family, which encodes for transporter proteins, had a birth-and-death scenario due to the numerous gene losses and duplications found in sea squirt, zebrafish and chicken. Salaneck *et al.,* (2008) arrived at a

similar conclusion for the neuropeptide Y receptor (NPYR) gene family in which seven members which originated in jawed vertebrates, were subsequently lost in teleost and mammals whereas in basal actinopterygian they were kept. Pinhal *et al.,* 2011 has shown that in a fresh water stingray, the 5s rDNA gene family did not evolve under a concerted model as previously assumed for this gene family but that 5s rDNA duplicates have arisen by genome duplications and that purifying selection under birth-and-death evolution determined if they were kept or lost. Recently, Solbakken *et al.* (2016) identified in codfish an expansion of the TLR gene family following the loss of MHC II genes and concluded that this occurrence represented a birth-and-death scenario.

While the number of TLR genes may vary between species, it is now established to a certain degree that teleost fish have 20 TLR members (Rauta *et al.,* 2014). The difference between the total number of TLR found in *D. rerio* and other teleosts with *O. niloticus* and *L. chalumnae* both presenting the second highest count of TLR genes, is most probably due to *D. rerio* specific duplications (Palti, 2011; Rauta *et al.,* 2014). In this study, a maximum of 19 TLR genes have been found in *D. rerio* which is in concordance with the literature (Jault *et al.,* 2004). The TLR members could be classified into the six known TLR gene families (Fig3.2) (Roach *et al.,* 2005). The TLR1 family was represented by TLR-1, 2 and 18 but did not form a monophyletic group. The TLR4 family regrouped into one well supported clade, presenting three paralogs (*TLR4al*, *TLR4ba*, *TLR4bb*) identified in *D. rerio* and with only ortholog *TLR4ba* present in *L. oculatus*. For TLR5*,* two paralogs (*TLR5a*, *TLR5b*) were found and their clade was supported. *TLR19* was found only in *D. rerio* and *L. chalumnae*. *TLR20* with its four duplications and *TLR21-22* belong to the TLR11 family and formed a monophyletic clade. The *TLR3* gene was the only one who had an orthologous gene in all the species studied, which formed a monophyletic group supported by a high bootstrap value. The function of the *TLR3* gene is to recognize double-stranded RNA (Roach *et al.,* 2005) and might explain why it was kept in all the species. It has been found that in several teleost infected by dsRNA viruses *TLR3* expression increased (Su *et al.,* 2008; Sahoo *et al.,* 2015). Also, when exposed to Gram-negative bacteria up regulation of *TLR3* expression was observed in zebrafish (Phelan *et al.,* 2005), catfish (Bilodeau & Waldbieser, 2005) and catfish hybrids (Bilodeau *et al.,* 2006).

The immunoglobulin superfamily counts a large number of genes with various functions, grouped in several gene families (Ohta, 2008) and the method applied to retrieve the query genes yielded a vast number of different gene families in the different species (data not shown). The analysis found two gene families which had homologs in the eleven species investigated. These gene families were the immunoglobulin superfamily and the semaphorins. These results may be of use for the identification of at least one of those gene families as an integral part of the immune system. The first family retrieved belonging to the

immunoglobulins was the Immunoglobulin Superfamliy. The two genes selected for the IgSf species phylogeny were *IgSf3*, relevant for neuronal formation (Usardi *et al.,* 2017) and *IgSf8* (*EWI-2*) (Usardi *et al.,* 2017) responsible for recognition of viruses (Gordón-Alonso *et al.,* 2012). In the immunoglobulin superfamily, the grouping of the Sema3 and Sema4 sub-families among the IgSf in this study was due to the similarity of their immunoglobulin-like domain (Garver *et al.,* 2008; Messina & Giacobini, 2013). As indicated in the Introduction, the number of immune related studies integrating the semaphorin family is scarce.

Of the five vertebrate semaphorin subfamilies (Goodman *et al.,* 1999), Sema3 and Sema4 are considered as immune semaphorins (Kikutani *et al.,* 2007). Both subfamilies were greatly duplicated with 12 paralogs of Sema3 and 7 paralogs of Ssema4 (Table *SI.*2). These two subfamilies where distinctly separated into two well supported major trees (Fig.3.5). From all the five studied gene families the semaphorin gene family was represented with eight out of 19 genes, the highest count of conserved genes in the phylogeny. Nevertheless, from those eight only three regrouped into bootstrap supported clades (*Sema3b*, *Sema3bl*, *Sema3c*). *Sema3c* has been identified to mitigate cancer cell migration (Herman & Meadows, 2007) whereas *Sema3b* suppress cell growth and induces apoptosis in cancer cells (Potiron *et al.,* 2009). Other semaphorin genes retrieved in this study where identified to have more evident role in the immune system. Even though they did not form bootstrap supported clades their presence is nonetheless relevant. Such genes were *Sema3a*, which takes part in immune cell migrations (Takamatsu *et al.,* 2010), *Sema4a* which improves T-cell activation (Kumanogoh *et al.,* 2002) and *Sema4b* which is expressed in T- and B-cells and regulates the interaction between T-cells and basophils (Nakagawa *et al.,* 2011). Altogether there were 11 orthologs, indicating their relevance for the immune system. Interestingly, the number of Sema genes in *N. coriiceps* is the same as in *L. oculatus* (Table *SI.*2) who diverged before the TSWGD. And although the missing genes are not the same, one is tempted to suggest that the loss of some of those genes in *N. coriiceps* was due to its adaptation to the Antarctic environment. Those points may favor this gene family to be considered as a candidate for future immune related studies.

As observed for the previously discussed gene families, the PIK3 gene family had variable gene loss and duplications. In the 18 PIK3 genes found in *D. rerio* all belonged either to *PIK3a*, *c*, *i*, *r*. *Pik3a* and *PIK3i* had one member whereas *PIK3c* an *PIK3r* had each eight paralogs (Table *SI.*2). The genes that belong to the PIK3 family can be regrouped into four classes, class IA comprising *PIK3c-a*, *b*, *d* and *r1-r3*, class IB with *PIK3r-5*, *6a*, *6b* and *PIK3cg*, class II with *PIK3c2-a*, *b*, *g*, class III with *PIK3-c3* and *r4* (Li *et al.,* 2016; Okkenhaug, 2013). Except for *PIK3ap1* and *PIK3ip1* all the retrieved PIK3 sequences could be classified into one of the 4 classes. The genes selected to build the species phylogeny belonged to three of the four PIK3 gene classes, with class IA involved in antigen signaling (Okkenhaug & Vanhaesebroeck,

2003) represented by *PIK3cd*, class IB responsible for natural killer cells cytokine production (Orr *et al.,* 2009) represented by *PIK3cg* and class III who regulates the autophagosome activation pathway (Orhon *et al.,* 2015) represented by *PIK3c3*. It should be noted that the total count of PIK3 genes was identical in *G. morhua* and *N. coriiceps* and that both only had one class II gene, *PIK3c2a* in *G. morhua* and *PIK3c2g* in *N. coriiceps*, respectively. Considering the apparent loss of the class IA *PIK3r3a* gene in the Acanthomorphs, both species also lack an additional class IA gene.

AKT3 had two genes which are part of the protein kinase B (PKB) family. This family is composed of three conserved isoforms PKBα (AKT1), PKBβ (AKT2) and or PKBγ (AKT3) (Schultze *et al.,* 2011). Although the AKT3 gene family contained the lowest number of genes, the two genes present were highly conserved in all the studied species. This could be because of the essential role of AKT3 as mediator in the PIK3-AKT3 signalling pathway (Manning & Cantley, 2007) and the selective pressure to maintain the signalling cascade. Corroborating this, Schultze *et al.,* 2011 stated that all the AKT isoforms present a high degree of sequence conservation of their phosphorylation sites and that AKT substrates act simultaneously on several cellular functions (Manning & Cantley, 2007). Furthermore, the inaptitude to resolve the AKT3 gene family into two distinct subfamilies (Fig. 3.2) probably also results from the sequence conservation and similarity observed in the AKT/protein kinase B. The species phylogeny obtained with *AKT3a* (Fig.3.7) presented an atypical topology, regrouping *G. morhua* in a distinct clade with the perciform *G. aculeatus* and the Notothenioidei. Most of the investigations on AKT revealed that even though the isoforms have similar broad functions there is evidence that each one carries specific functions which are crucial for the cell survival and physiology (Manning & Cantley, 2007). This could indicate that the grouping of the Notothenioidei with *G. morhua* could to a certain degree be due to the similar environmental conditions to which they are exposed, raising the possibility this gene may be crucial for the survival in cold environments.

We had considered including the MHC class II in this study, but the analysis turned out to be incompatible with the aims of the research. Specifically, the MHC class II gene family was considered to be relevant in the process of adaptation of the cod fish to the arctic environment (Malmstrøm *et al.,* 2016). However, homology searches between *D. rerio* and the ten other investigated species showed an increased number of lost homologous sequences, which could not solely be caused by gene loss. This analysis implies that the evolutionary study of the number of genes, their respective copy number and duplications, retrieved for the five gene families should be taken with caution since the methodology applied may have biased the results, since this method did not allow us to identify the expanded MHC I reported by

Malmstrøm *et al.,* (2016) for *G. morhua* . Also, the query sequences for the homology searches used in this study were retrieved from *D. rerio* and therefore the comparison of the number of obtained genes for each gene and species is based on the similarity to those available from *D. rerio*.

The points discussed above which are, 1) that a variety of gain, and loss of genes can be identified in the different species, 2) the phylogenetic trees based on orthologous genes fit well with the accepted species phylogeny and, 3) in the five gene trees orthologous genes cluster together, allow us to understand how the five gene families evolved. The results are in agreement with what can be observed in gene families that are under birth-and-death regulation: 1) an increase in number of gene losses and/or expansion and 2), unlike what is stated in the birth-and-death opposing concerted evolution model of gene families, orthologs of those genes form inter species instead of intra species clusters (Nei *et al.,* 1997). This is valid not only for the Notothenoidei which is the aim of this study, but can be seen as a general evolutionary process taking place in the five studied gene families.

5.2 Divergence time analysis

The divergence time estimates did not present a pattern between the different nodes, which points to a node specific immune gene usage which in turn could depend of the various adaptations present in the different species contained in the node. A significant caveat in the estimation of the immune gene families divergence time in this study was the reduced number of genes and species used. Near *et al.,* (2012) demonstrated in their study, which estimated the divergence time of teleosts using multiple nuclear genes sequences, that when comparing with previously published studies using a single nuclear gene, the former presented age estimates similar to the fossil record whereas the latter underestimated the species divergence times. The number of species used in each node can also alter the divergence time estimation as verified by Schulte (2013) who concluded that the node ages from a South American lizard clade yielded younger estimates for under sampled nodes when compared to nodes with a higher number of taxa. In this study the number of genes used in each gene family did not seem to have a noticeable influence on the divergence time estimates. If this would be the case, then AKT3 and TLR had to generate the lowest divergence times in each node since both were estimated with only one orthologous gene each, while Sema and PIK3 were expected to have higher estimates since they had the highest number of concatenated orthologous sequences. While the number of genes does not offer a convincing explanation for the variation of the intra-nodal divergence time estimates the number of species used in each node may be the reason for this variation as discussed in the following paragraphs.

The nodes containing the highest numbers of species, Neopterygii, Teleostei, Acanthomorpha and Percomorpha yielded divergence times for each gene (Table *SI*.5) closer to the estimates given by the fossil record. It should be noted that by coincidence those are also the nodes where the fossil record spans a longer period of time. The apparent discrepancy between the fossil records and the divergence time obtained for the five Perciformes genes (Fig.3.12) could have resulted from an incomplete species sampling for this node.

Compared to the 182 species Betancur-R *et al.* (2013) identified in Perciformes which enabled them to estimate a divergence time of 100 m.y.a for this family, the five perciform species used here may not have been enough to achieve an acceptable time resolution. In our work, apart from the three Notothenioidiei, only *G. aculeatus* and *D. labrax* were included in the Perciformes. Additionally, both species diverged only recently in the order Perciformes with fossil record for *G. aculetaus* only dating back to a maximum of 13. m.y.a (Bell e*t al.,* 2009) whereas the Moronidae subfamily to which *D. labrax* belongs, are thought to have emerged around 25 m.y.a (Meynard *et al.,* 2012).

Similar to the Perciformes, the difference between the fossil record and the gene family divergence time estimates of the Notothenioidei and Nototheniidae (Fig.3.12), was due to the underrepresentation in the number of species included in those nodes. The current estimate of Notothenioidei species is 322 (Joseph, 2005). Using a larger set of Antarctic species Near *et al.,* (2012) estimated a divergence time of 34.6 m.y.a and 50.6 m.y.a. between Notothenioidei and Nototheniidae. In our study the Nototheioidei node included only 3 species, *E. maclovinus* and two Nototheniidae (*N. coriiceps* and *D. mawsoni*). *E. maclovinus* the sole member of the Eleginopsidae and three other non-Antarctic notothenoid families are thought to have diverged early in the Notothenioidei phylogeny (Papetti *et al.,* 2016). Near *et al.,* (2015), stipulated that *E. maclovinus* emerged in the Weddellian Province of East Gondwana where the seas were predominantly shallow, and the water temperatures were cool to temperate. The absence of AFGP in *E. maclovinus* further supports an early separation of this species from the other Antarctic notothenoids since the AFGPs are thought to have emerged in Antarctic notothenoids after the onset of the ACC (Cheng *et al.,* 2003). This could explain why the divergence time of the gene families did not indicate the same pattern between Notothenioidei and Nototheniidae (Fig 3.2) since their MRCA was not exposed to the same environmental pressures. Hence, the different divergence times estimated for the five gene families seem to have been conditioned by the adaption of Nototheniidae to the Antarctic because unlike the Nototehnioidei ancestor the Notothenoiidae ancestor most probably diverged into similar conditions to the ones seen today in the Antarctic Ocean.

Supporting this notion are the estimates for the radiation of the AFGP bearing Nototheniidae, including the Notothenoid and Dissostichus families, into the Antarctic ocean

which are comprised in a range of 25-5 m.y.a. Near (2004) by applying a molecular clock based on notothenoid partial gene mtDna, estimated that the radiation of AFGP bearing Antarctic Notothenoids occurred at the Oligocene-Miocene boundary around 24 ± 0.5 m.y.a.. Bargelloni & Lecointre (1998) also used mitochondrial DNA to estimate Antarctic Notothenoids radiation but yielded slightly different results inferring a range of 16-10 m.y.a.

Using notothenioid specific morphological characteristics, adaptations and a time calibrated phylogeny based on 49 species Colombo *et al.,* (2015) obtained an estimate of 13.4 m.y.a for the adaptive radiation of notothenioid. Interestingly, Chen *et al.* (1997), estimated the divergence time between *N. coriiceps* and *D. mawsoni* based on the differences of the AFGP gene sequences of 14-5 m.y.a. Shevenell *et al.,* (2004) determined for the Southern Ocean that the range between 14.2-13.8 m.y.a known as the Middle Miocene climatic transition (MMCT) was marked by environmental changes of the Southern Ocean with repeated fluctuations of sea-ice cover and the consequential drop of seawater temperature to polar norms. Near *et al.,* (2012, stipulated that these changes could have had to a lesser extent effects on the adaptation of Antarctic notothenioids.

Furthermore, the semaphorins yielded the oldest divergence time estimate (7 m.y.a.) in Nototheniidae followed closely by AKT3 (6.17 m.y.a). PIK3 and IgSf yielded similar divergence times of 4 m.y.a and 4.2 m.y.a respectively. As for AKT3 and Sema those gene families present several other functions apart from their implication in the immune system. The PIK3/AKT-pathway is essential for cell growth, survival and regulation of the cytoskeleton (Downward, 2004) whereas proteins encoded by the IgSf have functions such as muscle proteins, kinases and molecules with leucin-rich repeats (Natarajan *et al.,* 2015). The youngest divergence time estimated for a gene in Nototheniidae was the *TLR* gene (2.49 m.y.a) which was based on the single orthologous sequence of *TLR3*. In contrast to the other genes analyzed here, *TLR3* seems to present an exclusive immune related function. Roach *et al.,* (2005) stipulated that due to their importance in dsRNA recognition *TLR3* is under selective pressure to keep that specific function intact.

In summary, 1) the number of species used in each node influences the estimation of the gene divergence times, 2) the divergence time of the five genes in Notothenioidei and Nototheniidae is more recent than estimates from the fossil record and the literature, 3) the function of the individual genes (degree of conservation) helped to identify the pattern of gene divergence time in Nototheniidae. These observations also highlight the fact that the estimation of the divergence time is conditioned by a number of factors. Nevertheless, it is possible to hypothesize that during their adaptation to the Antarctic environment the Nototheniidae relied on genes with primarily vital metabolical cell functions, which are directly or indirectly relevant to the immune system. Finally, the recent divergence time estimates obtained for the

five gene families in nototheniidae, seemed to indicate that the adaptation of the immune system in those species followed the MMCT.

6 References

Ahn, D. H., Kang, S., & Park, H. (2016). Transcriptome analysis of immune response genes induced by pathogen agonists in the Antarctic bullhead notothen Notothenia coriiceps. *Fish and Shellfish Immunology*, *55*(June), 315–322. https://doi.org/10.1016/j.fsi.2016.06.004

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen recognition and innate immunity. *Cell*, *124*(4), 783–801. https://doi.org/10.1016/j.cell.2006.02.015

Alfaro, M. E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D. L., Harmon, L. J. (2009). Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences*, *106*(32), 13410–13414.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Amaral, D. B., & Schneider, I. (2018). Fins into limbs: Recent insights from sarcopterygian fish. *Genesis*, *56*(1), 8.

Annilo, T., Chen, Z. Q., Shulenin, S., Costantino, J., Thomas, L., Lou, H., Dean, M. (2006). Evolution of the vertebrate ABC gene family: Analysis of gene birth and death. *Genomics*, *88*(1), 1–11. https://doi.org/10.1016/j.ygeno.2006.03.001

Arratia, G. (1996). Basal teleosts and teleostean phylogeny.

Baldauf, S. L. (2003). Phylogeny for the faint of heart: A tutorial. *Trends in Genetics*, *19*(6), 345–351. https://doi.org/10.1016/S0168-9525(03)00112-4

Balushkin, A. V. (1994). Proeleginops grandeastmanorum gen. et sp. nov.(Perciformes, Notothenioidei, Eleginopsidae) from the Late Eocene of Seymour Island (Antarctica) is a fossil notothenioid, not a gadiform. *Journal of Ichthyology*, *34*(8), 10–23.

Bargelloni, L., & Lecointre, G. (1998). Four years in Notothenioid systematics: a molecular perspective. In *Fishes of Antarctica* (pp. 259–273). Springer.

Beg, M., Abdullah, N., Thowfeik, F. S., Altorki, N. K., & McGraw, T. E. (2017). Distinct Akt phosphorylation states are required for insulin regulated Glut4 and Glut1-mediated glucose uptake. *Elife*, *6*, e26896.

Bell, M. A., Stewart, J. D., & Park, P. J. (2009). The World's Oldest Fossil Threespine Stickleback Fish. *Copeia*, *2009*(2), 256–265. https://doi.org/10.1643/CG-08-059

Benton, M. J., Donoghue, P. C. J., Asher, R. J., Friedman, M., Near, T. J., & Vinther, J. (2015).

Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*, *18*(1), 1–106.

Bernardes, J. S., Vieira, F. R. J., Costa, L. M. M., & Zaverucha, G. (2015). Evaluation and improvements of clustering algorithms for detecting remote homologous protein families. *BMC Bioinformatics*, *16*(1), 34.

Berthelot, C., Brunet, F., Chalopin, D., Juanchich, A., Bernard, M., Noël, B., Guiguen, Y. (2014). The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. *Nature Communications*, *5*. https://doi.org/10.1038/ncomms4657

Betancur-R, R., Broughton, R. E., Wiley, E. O., Carpenter, K., López, J. A., Li, C., Cureton Ii, J. C. (2013). The tree of life and a new classification of bony fishes. *PLoS Currents*, *5*.

Bieńkowska-Wasiluk, M., Bonde, N., Møller, P. R., & Gaździcki, A. (2013). Eocene relatives of cod icefishes (perciformes: Notothenioidei) from Seymour Island, Antarctica. *Geological Quarterly*, *57*(4), 567–582. https://doi.org/10.7306/gq.1112

Bilodeau, A. L., Peterson, B. C., & Bosworth, B. G. (2006). Response of toll-like receptors, lysozyme, and IGF-I in back-cross hybrid (F1 male (blue× channel)× female channel) catfish challenged with virulent Edwardsiella ictaluri. *Fish & Shellfish Immunology*, *20*(1), 29–39.

Bilodeau, A. L., & Waldbieser, G. C. (2005). Activation of TLR3 and TLR5 in channel catfish exposed to virulent Edwardsiella ictaluri. *Developmental & Comparative Immunology*, *29*(8), 713–721.

Bilyk, K. T., & Cheng, C.-H. C. (2013). Model of gene expression in extreme cold - reference transcriptome for the high-Antarctic cryopelagic notothenioid fish Pagothenia borchgrevinki. *BMC Genomics*, *14*(1), 1–16. https://doi.org/10.1186/1471-2164-14-634

Braasch, I., Peterson, S. M., Desvignes, T., McCluskey, B. M., Batzel, P., & Postlethwait, J. H. (2015). A new model army: Emerging fish models to study the genomics of vertebrate Evo-Devo. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *324*(4), 316–341.

Buchmann, K. (2014). Evolution of innate immunity: Clues from invertebrates via fish to mammals. *Frontiers in Immunology*, *5*(SEP), 1–8. https://doi.org/10.3389/fimmu.2014.00459

Cada, R. N. (2005). Gadus morhua. Retrieved January 28, 2019, from https://upload.wikimedia.org/wikipedia/commons/6/66/Gamor_u0.gif

Chen, C., Huang, H., & Wu, C. H. (2017). Protein bioinformatics databases and resources. In *Protein Bioinformatics* (pp. 3–39). Springer.

Chen, L., DeVries, A. L., & Cheng, C.-H. C. (1997). Evolution of antifreeze glycoprotein gene

from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences* , *94*(8), 3811–3816. Retrieved from http://www.pnas.org/content/94/8/3811.abstract

Chen, W.-J., Bonillo, C., & Lecointre, G. (1998). Phylogeny of the Channichthyidae (Notothenioidei, Teleostei) based on two mitochondrial genes. In *Fishes of Antarctica* (pp. 287–298). Springer.

Chen, W.-J., Santini, F., Carnevale, G., Chen, J.-N., Liu, S.-H., LavouÃ©, S., & Mayden, R. L. (2014). New insights on early evolution of spiny-rayed fishes (Teleostei: Acanthomorpha). *Frontiers in Marine Science*, *1*(October), 1–17. https://doi.org/10.3389/fmars.2014.00053

Chen, Z., Cheng, C.-H. C., Zhang, J., Cao, L., Chen, L., Zhou, L., Chen, L. (2008). Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences*, *105*(35), 12944–12949. https://doi.org/10.1073/pnas.0802432105

Cheng, C. H. C., Chen, L., Near, T. J., & Jin, Y. (2003). Functional Antifreeze Glycoprotein Genes in Temperate-Water New Zealand Nototheniid Fish Infer an Antarctic Evolutionary Origin. *Molecular Biology and Evolution*, *20*(11), 1897–1908. https://doi.org/10.1093/molbev/msg208

Choi, Y. I., Duke-Cohan, J. S., Ahmed, W. B., Handley, M. A., Mann, F., Epstein, J. A., Reinherz, E. L. (2008). PlexinD1 glycoprotein controls migration of positively selected thymocytes into the medulla. *Immunity*, *29*(6), 888–898.

Clarke, A., Aronson, R. B., Crame, J. A., Gili, J.-M., & Blake, D. B. (2004). Evolution and diversity of the benthic fauna of the Southern Ocean continental shelf. *Antarctic Science*, *16*(4), 559–568.

Colombo, M., Damerau, M., Hanel, R., Salzburger, W., & Matschiner, M. (2015). Diversity and disparity through time in the adaptive radiation of Antarctic notothenioid fishes. *Journal of Evolutionary Biology*, *28*(2), 376–394. https://doi.org/10.1111/jeb.12570

Constable, A. J., Melbourne-Thomas, J., Corney, S. P., Arrigo, K. R., Barbraud, C., Barnes, D. K. A., Ziegler, P. (2014). Climate change and Southern Ocean ecosystems I: How changes in physical habitats directly affect marine biota. *Global Change Biology*, *20*(10), 3004–3025. https://doi.org/10.1111/gcb.12623

Cooper, E. L. (2010). Evolution of immune systems from self/not self to danger to artificial immune systems (AIS). *Physics of Life Reviews*, *7*(1), 55–78. https://doi.org/10.1016/j.plrev.2009.12.001

Cortesi, F., Musilová, Z., Stieb, S. M., Hart, N. S., Siebeck, U. E., Malmstrøm, M., Salzburger, W. (2015). Ancestral duplications and highly dynamic opsin gene evolution in

percomorph fishes. *Proceedings of the National Academy of Sciences*, *112*(5), 1493–1498. https://doi.org/10.1073/pnas.1417803112

Cristini, L., Grosfeld, K., Butzin, M., & Lohmann, G. (2012). Influence of the opening of the Drake Passage on the Cenozoic Antarctic Ice Sheet: A modeling approach. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *339–341*, 66–73. https://doi.org/10.1016/j.palaeo.2012.04.023

Dayton, P. K., Mordida, B. J., & Bacon, F. (1994). Polar Marine Communities1. *American Zoologist*, *34*(1), 90–99. https://doi.org/10.1093/icb/34.1.90

Dehal, P., & Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biology*, *3*(10), e314.

Deng, C., Cheng, C.-H. C., Ye, H., He, X., & Chen, L. (2010). Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(50), 21593–21598. https://doi.org/10.1073/pnas.1007883107

Desjardins, M., Houde, M., & Gagnon, E. (2005). Phagocytosis: The convoluted way from nutrition to adaptive immunity. *Immunological Reviews*, *207*, 158–165. https://doi.org/10.1111/j.0105-2896.2005.00319.x

di Prisco, G., Eastman, J. T., Giordano, D., Parisi, E., & Verde, C. (2007). Biogeography and adaptation of Notothenioid fish: Hemoglobin function and globin-gene evolution. *Gene*, *398*(1–2 SPEC. ISS.), 143–155. https://doi.org/10.1016/j.gene.2007.02.047

Donohue, K. A., Tracey, K. L., Watts, D. R., Chidichimo, M. P., & Chereskin, T. K. (2016). Mean Antarctic Circumpolar Current transport measured in Drake Passage. *Geophysical Research Letters*, *43*(22), 11,760-11,767. https://doi.org/10.1002/2016GL070319

Downward, J. (2004). PI 3-kinase, Akt and cell survival. *Seminars in Cell and Developmental Biology*, *15*(2), 177–182. https://doi.org/10.1016/j.semcdb.2004.01.002

Drost, H. G., & Paszkowski, J. (2017). Biomartr: Genomic data retrieval with R. *Bioinformatics*, *33*(8), 1216–1217. https://doi.org/10.1093/bioinformatics/btw821

Eastman, J. T. (2000). Antarctic notothenioid fishes as subjects for research in evolutionary biology. *Antarctic Science*, *12*(03), 276–287. https://doi.org/10.1017/S0954102000000341

Eastman, J. T. (2005). The nature of the diversity of Antarctic fishes. *Polar Biology*, *28*(2), 93–107. https://doi.org/10.1007/s00300-004-0667-4

Eastman, J. T., & Devries, A. L. (1981). Buoyancy Adaptations in a Swim-Bladderless Antarctic Fish. *Journal of Morphology*, (167), 91–102.

Eastman, J. T., & McCune, A. R. (2000). Fishes on the Antarctic continental shelf: Evolution of a marine species flock? *Journal of Fish Biology*, *57*(SUPPL. A), 84–102.

https://doi.org/10.1006/jfbi.2000.1604

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Edgar, R. C., & Batzoglou, S. (2006). Multiple sequence alignment. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *16*(3), 368–373.

Eizaguirre, C., Lenz, T. L., Traulsen, A., & Milinski, M. (2009). Speciation accelerated and stabilized by pleiotropic major histocompatibility complex immunogenes. *Ecology Letters*, *12*(1), 5–12.

Elias, I. (2006). Settling the intractability of multiple alignment. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *13*(7), 1323–1339. https://doi.org/10.1089/cmb.2006.13.1323

Evans, C. W., & DeVries, A. L. (2017). Coping with Ice: Freeze Avoidance in the Antarctic Silverfish (Pleuragramma antarctica) from Egg to Adult. In M. Vacchi, E. Pisano, & L. Ghigliotti (Eds.), *The Antarctic Silverfish: a Keystone Species in a Changing Ecosystem* (pp. 27–46). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-55893-6_2

Fassler, J., & Cooper, P. (2011). *BLAST Glossary*. *Bethesda Md: National Center for Biotechnology Information*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK62051/

Fearon, D. T., & Locksley, R. M. (1996). The Instructive Role of Innate Immunity in the Acquired Immune Response. *Science*, *272*(5258), 50–54. https://doi.org/10.1126/science.272.5258.50

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Bateman, A. (2018). The Pfam protein families database in 2019. *Nucleic Acids Res.*, 1–6. https://doi.org/10.1093/nar/gky995 The

Frech, C., & Chen, N. (2010). Genome-wide comparative gene family classification. *PLoS ONE*, *5*(10). https://doi.org/10.1371/journal.pone.0013409

Garver, L. S., Xi, Z., & Dimopoulos, G. (2008). Immunoglobulin superfamily members play an important role in the mosquito immune system. *Developmental and Comparative Immunology*, *32*(5), 519–531. https://doi.org/10.1016/j.dci.2007.09.007

Gay, N. J., & Gangloff, M. (2007). Structure and Function of Toll Receptors and Their Ligands. *Annual Review of Biochemistry*, *76*(1), 141–165. https://doi.org/10.1146/annurev.biochem.76.060305.151318

Gharib, W. H., & Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and

variation in GC. *Molecular Biology and Evolution*, *30*(7), 1675–1686. https://doi.org/10.1093/molbev/mst062

Glasauer, S. M. K., & Neuhauss, S. C. F. (2014). Whole-genome duplication in teleost fishes and its evolutionary consequences. *Molecular Genetics and Genomics*, *289*(6), 1045–1060. https://doi.org/10.1007/s00438-014-0889-2

Goodman, C. S., Kolodkin, A. L., Luo, Y., Püschel, A. W., & Raper, J. A. (1999). Unified nomenclature for the semaphorins/collapsins. *Cell*, *97*(5), 551–552.

Gordón-Alonso, M., Sala-Valdés, M., Rocha-Perugini, V., Pérez-Hernández, D., López-Martín, S., Ursa, A., Sánchez-Madrid, F. (2012). EWI-2 association with α-actinin regulates T cell immune synapses and HIV viral infection. *The Journal of Immunology*, 1103708.

Gu, C., & Giraudo, E. (2013). The role of semaphorins and their receptors in vascular development and cancer. *Experimental Cell Research*, *319*(9), 1306–1316.

Gupta, A. Sen, Santoso, A., Taschetto, A. S., Ummenhofer, C. C., Trevena, J., & England, M. H. (2009). Projected changes to the Southern Hemisphere ocean and sea ice in the IPCC AR4 climate models. *Journal of Climate*, *22*(11), 3047–3078. https://doi.org/10.1175/2008JCLI2827.1

Halaby, D. M., & Mornon, J. P. E. (1998). The immunoglobulin superfamily: An insight on its tissular, species, and functional diversity. *Journal of Molecular Evolution*, *46*(4), 389–400. https://doi.org/10.1007/PL00006318

Hall, B. K. (2013). Homology, homoplasy, novelty, and behavior. *Developmental Psychobiology*, *55*(1), 4–12. https://doi.org/10.1002/dev.21039

Hassold, N. J. C., Rea, D. K., van der Pluijm, B. A., & Parés, J. M. (2009). A physical record of the Antarctic Circumpolar Current: Late Miocene to recent slowing of abyssal circulation. *Palaeogeography, Palaeoclimatology, Palaeoecology*, *275*(1–4), 28–36. https://doi.org/10.1016/j.palaeo.2009.01.011

Hedges, S. B., Dudley, J., & Kumar, S. (2006). TimeTree: A public knowledge-base of divergence times among organisms. *Bioinformatics*, *22*(23), 2971–2972. https://doi.org/10.1093/bioinformatics/btl505

Herman, J. G., & Meadows, G. G. (2007). Increased class 3 semaphorin expression modulates the invasive and adhesive properties of prostate cancer cells. *International Journal of Oncology*, *30*(5), 1231–1238.

Hillewaert, H. (2005). Dicentrarchus labrax. Retrieved January 28, 2019, from https://commons.wikimedia.org/wiki/File:Dicentrarchus_labrax_(Belgium).jpg#/media/File:Dicentrarchus_labrax_(Belgium).jpg

Hofmann, G. E., Buckley, B. a, Airaksinen, S., Keen, J. E., & Somero, G. N. (2000). Heat-

shock protein expression is absent in the antarctic fish Trematomus bernacchii (family Nototheniidae). *The Journal of Experimental Biology*, *203*(Pt 15), 2331–2339. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10887071

Hood, L., Campbell, J. H., & Elgin, S. C. R. (1975). The organization, expression, and evolution of antibody genes and other multigene families. *Annual Review of Genetics*, *9*(1), 305–353.

Horner, D. S., & Pesole, G. (2004). Phylogenetic analyses: A brief introduction to methods and their application. *Expert Review of Molecular Diagnostics*, *4*(3), 339–350. https://doi.org/10.1586/14737159.4.3.339

Huerta-Cepas, J., & Gabaldón, T. (2011). Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics*, *27*(1), 38–45. https://doi.org/10.1093/bioinformatics/btq609

Hurley, I. A., Mueller, R. L., Dunn, K. A., Schmidt, E. J., Friedman, M., Ho, R. K., Coates, M. I. (2007). A new time-scale for ray-finned fish evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1609), 489–498.

Ito, D., & Kumanogoh, A. (2015). mTOR Complex Signaling through the SEMA4A–Plexin B2 Axis Is Required for Optimal Activation and Differentiation of CD8+ T Cells. *Journal of Immunology*, *195*(3), 934.943.

Iwasaki, A., & Medzhitov, R. (2015). Control of adaptive immunity by the innate immune system. *Inmunologia Básica*, *16*(4), 1–7. https://doi.org/10.1038/ni.3123.Control

Janeway, C. A. (1989). Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harbor Symposia on Quantitative Biology*, *54*(1), 1–13. https://doi.org/10.1101/SQB.1989.054.01.003

Janeway, C. A., & Medzhitov, R. (2002). Innate Immune Recognition. *Annual Review of Immunology*, *20*(1), 197–216. https://doi.org/10.1146/annurev.immunol.20.083001.084359

Janssen, B. J. C., Robinson, R. A., Pérez-Brangulí, F., Bell, C. H., Mitchell, K. J., Siebold, C., & Jones, E. Y. (2010). Structural basis of semaphorin–plexin signalling. *Nature*, *467*(7319), 1118.

Jault, C., Pichon, L., & Chluba, J. (2004). Toll-like receptor gene family and TIR-domain adapters in Danio rerio. *Molecular Immunology*, *40*(11), 759–771. https://doi.org/10.1016/j.molimm.2003.10.001

Jeffares, D. C., Tomiczek, B., Sojo, V., & dos Reis, M. (2015). A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In *Parasite Genomics Protocols* (pp. 65–90). Springer.

Johnston, I. A., Fitch, N., Zummo, G., Wood, R. E., Harrison, P., & Tota, B. (1983).

Morphometric and ultrastructural features of the ventricular myocardium of the haemoglobin-less icefish Chaenocephalus aceratus. *Comparative Biochemistry and Physiology Part A: Physiology*, *76*(3), 475–480.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2016). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361.

Kang, H.-R., Lee, H. S., Park, D. E., Lee, J. W., Song, W.-J., Park, H.-W., Min, K.-U. (2014). The Role Of Semaphorin 7A In Alternatively Activation Of Macrophages. *Journal of Allergy and Clinical Immunology*, *133*(2), AB247.

Kawai, T., & Akira, S. (2010). The role of pattern-recognition receptors in innate immunity: Update on toll-like receptors. *Nature Immunology*, *11*(5), 373–384. https://doi.org/10.1038/ni.1863

Kawasaki, T., & Kawai, T. (2014). Toll-like receptor signaling pathways. *Frontiers in Immunology*, *5*(SEP), 1–8. https://doi.org/10.3389/fimmu.2014.00461

Kelchner, S. A., & Thomas, M. A. (2007). Model use in phylogenetics: nine key questions. *Trends in Ecology and Evolution*, *22*(2), 87–94. https://doi.org/10.1016/j.tree.2006.10.004

Kikutani, H., Suzuki, K., & Kumanogoh, A. (2007). Immune semaphorins: increasing members and their diverse roles. *Advances in Immunology*, *93*, 121–143.

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, *39*(1), 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Korf, I., Yandell, M., & Bedell, J. (2003). BLAST. In *BLAST: An Essential Guide to the Basic Local Alignment Search Tool* (pp. 55–71). https://doi.org/10.1177/0049124103253373

Koyasu, S. (2003). The role of PI3K in immune cells. *Nature Immunology*, *4*(4), 313–319.

Kozlov, A., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2018). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *BioRxiv*, 1–5. https://doi.org/10.1101/447110

Kumanogoh, A., & Kikutani, H. (2003). Immune semaphorins: a new area of semaphorin research. *Journal of Cell Science*, *116*(17), 3463–3470. https://doi.org/10.1242/jcs.00674

Kumanogoh, A., Marukawa, S., Suzuki, K., Takegahara, N., Watanabe, C., Ch'ng, E. S., Kikutani, H. (2002). Class iv semaphorin sema4a enhances t-cell activation and interacts with tim-2. *Nature*, *419*(6907), 629–633. https://doi.org/10.1038/nature01037

Kumanogoh, A., Watanabe, C., Lee, I., Wang, X., Shi, W., Araki, H., Yasui, T. (2000). Identification of CD72 as a lymphocyte receptor for the class IV semaphorin CD100: a novel mechanism for regulating B cell signaling. *Immunity*, *13*(5), 621–631.

Lankester, E. R. (1870). II.—On the use of the term homology in modern zoology, and the

distinction between homogenetic and homoplastic agreements. *Annals and Magazine of Natural History*, *6*(31), 34–43.

Lawlor, M. A., & Alessi, D. R. (2001). PKB/Akt: a key mediator of cell proliferation, survival and insulin responses? *Journal of Cell Science*, *114*(Pt 16), 2903–2910. https://doi.org/10.1042/bst0290001

Lecointre, G. (2004). Eleginops maclovinus. Retrieved January 28, 2019, from http://glecointre.mnhn.fr/Collections.html%0D%0A%0D%0A

Lefranc, M. P. (2014). Immunoglobulin and T cell receptor genes: IMGT®and the birth and rise of immunoinformatics. *Frontiers in Immunology*, *5*(FEB), 1–22. https://doi.org/10.3389/fimmu.2014.00022

Li, W.-H., & Graur, D. (2002). *Fundamentals of Molecular Evolution*. (Andrew D. Sinauer, Ed.) (Second Edi). Sunderland, Massachusetts: SINAUER ASSOCIATES, INC.,.

Li, Y., Dowbenko, D., & Lasky, L. A. (2002). AKT/PKB phosphorylation of p21Cip/WAF1 enhances protein stability of p21Cip/WAF1 and promotes cell survival. *Journal of Biological Chemistry*, *277*(13), 11352–11361.

Li, Z., Yao, J., Xie, Y., Geng, X., & Liu, Z. (2016). Phosphoinositide 3-kinase family in channel catfish and their regulated expression after bacterial infection. *Fish and Shellfish Immunology*, *49*, 364–373. https://doi.org/10.1016/j.fsi.2016.01.002

Lin, H.-K., Wang, G., Chen, Z., Teruya-Feldstein, J., Liu, Y., Chan, C.-H., Nimer, S. (2009). Phosphorylation-dependent regulation of cytosolic localization and oncogenic function of Skp2 by Akt/PKB. *Nature Cell Biology*, *11*(4), 420–432.

Litman, G. W., Rast, J. P., & Fugmann, S. D. (2010). The origins of vertebrate adaptive immunity. *Nature Reviews Immunology*, *10*(8), 543–553. https://doi.org/10.1038/nri2807

Liu, H., Juo, Z. S., Shim, A. H.-R., Focia, P. J., Chen, X., Garcia, K. C., & He, X. (2010). Structural basis of semaphorin-plexin recognition and viral mimicry from Sema7A and A39R complexes with PlexinC1. *Cell*, *142*(5), 749–761.

Llort, J., Lévy, M., Sallée, J.-B., & Tagliabue, A. (2015). Onset, intensification, and decline of phytoplankton blooms in the Southern Ocean. *ICES Journal of Marine Science*, *72*(6), 1971–1984. https://doi.org/10.1093/icesjms/fsv053

Loewe, L., & Hill, W. G. (2010). The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1544), 1153–1167. https://doi.org/10.1098/rstb.2009.0317

Lu, Z., Hoogakker, B. A. A., Hillenbrand, C.-D., Zhou, X., Thomas, E., Gutchess, K. M., Rickaby, R. E. M. (2016). Oxygen depletion recorded in upper waters of the glacial Southern Ocean. *Nature Communications*, *7*, 11146. Retrieved from http://dx.doi.org/10.1038/ncomms11146

Lyle, M., Gibbs, S., Moore, T. C., & Rea, D. K. (2007). Late Oligocene initiation of the Antarctic circumpolar current: Evidence from the South Pacific. *Geology*, *35*(8), 691–694. https://doi.org/10.1130/G23806A.1

MacKintosh, A., Golledge, N., Domack, E., Dunbar, R., Leventer, A., White, D., Lavoie, C. (2011). Retreat of the East Antarctic ice sheet during the last glacial termination. *Nature Geoscience*, *4*(3), 195–202. https://doi.org/10.1038/ngeo1061

Malmstrøm, M., Matschiner, M., Tørresen, O. K., Star, B., Snipen, L. G., Hansen, T. F., Jentoft, S. (2016). Evolution of the immune system influences speciation rates in teleost fishes. *Nature Genetics*, *48*(10), 1204–1210. https://doi.org/10.1038/ng.3645

Manning, B. D., & Cantley, L. C. (2007). AKT/PKB Signaling: Navigating Downstream. *Cell*, *129*(7), 1261–1274. https://doi.org/10.1016/j.cell.2007.06.009

Matschiner, M., Hanel, R., & Salzburger, W. (2011). On the origin and trigger of the notothenioid adaptive radiation. *PLoS ONE*, *6*(4). https://doi.org/10.1371/journal.pone.0018911

Meijer, A. H., Gabby Krens, S. F., Medina Rodriguez, I. A., He, S., Bitter, W., Snaar-Jagalska, B. E., & Spaink, H. P. (2004). Expression analysis of the Toll-like receptor and TIR domain adaptor families of zebrafish. *Molecular Immunology*, *40*(11), 773–783. https://doi.org/10.1016/j.molimm.2003.10.003

Messina, A., & Giacobini, P. (2013). Semaphorin signaling in the development and function of the gonadotropin hormone-releasing hormone system. *Frontiers in Endocrinology*, *4*, 133.

Meyer, A., & Van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, *27*(9), 937–945.

Meynard, C. N., Mouillot, D., Mouquet, N., & Douzery, E. J. P. (2012). A phylogenetic perspective on the evolution of Mediterranean teleost fishes. *PLoS ONE*, *7*(5). https://doi.org/10.1371/journal.pone.0036443

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In *Gateway Computing Environments Workshop (GCE), 2010* (pp. 1–8). Ieee.

Mine, T., Harada, K., Matsumoto, T., Yamana, H., Shirouzu, K., Itoh, K., & Yamada, A. (2000). CDw108 expression during T-cell development. *Tissue Antigens*, *55*(5), 429–436.

Mintenbeck, K. (2017). Impacts of Climate Change on the Southern Ocean. In *Climate Change Impacts on Fisheries and Aquaculture* (pp. 663–701). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119154051.ch20

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Fraser, M. I. (2018). InterPro in 2019: improving coverage, classification and access to protein

sequence annotations. *Nucleic Acids Research*.

Montgomery, J., & Clements, K. (2000). Disaptation and recovery in the evolution of Antarctic fishes. *Trends in Ecology and Evolution*. https://doi.org/10.1016/S0169-5347(00)01896-6

Mugal, C. F., Wolf, J. B. W., & Kaj, I. (2014). Why time matters: Codon evolution and the temporal dynamics of dN/dS. *Molecular Biology and Evolution*, *31*(1), 212–231. https://doi.org/10.1093/molbev/mst192

Nakagawa, Y., Takamatsu, H., Okuno, T., Kang, S., Nojima, S., Kimura, T., Katayama, I. (2011). Identification of semaphorin 4B as a negative regulator of basophil-mediated immune responses. *The Journal of Immunology*, 1003485.

Natarajan, K., Mage, M. G., & Margulies, D. H. (2015). Immunoglobulin Superfamily. *ELS*, 1–7. https://doi.org/10.1002/9780470015902.a0000926.pub2

Near, T. J. (2004). Estimating divergence times of notothenioid fishes using a fossil-calibrated molecular clock. *Antarctic Science*, *16*(1), 37–44. https://doi.org/10.1017/S0954102004001798

Near, T. J., Dornburg, A., Harrington, R. C., Oliveira, C., Pietsch, T. W., Thacker, C. E., Beaulieu, J. M. (2015). Identification of the notothenioid sister lineage illuminates the biogeographic history of an Antarctic adaptive radiation. *BMC Evolutionary Biology*, *15*(1), 1–14. https://doi.org/10.1186/s12862-015-0362-9

Near, T. J., Dornburg, A., Kuhn, K. L., Eastman, J. T., Pennington, J. N., Patarnello, T., Jones, C. D. (2012). Ancient climate change, antifreeze, and the evolutionary diversification of Antarctic fishes. *Proceedings of the National Academy of Sciences*, *109*(9), 3434–3439. https://doi.org/10.1073/pnas.1115169109

Near, T. J., Eytan, R. I., Dornburg, A., Kuhn, K. L., Moore, J. A., Davis, M. P., Smith, W. L. (2012). Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences*, *109*(34), 13698–13703. https://doi.org/10.1073/pnas.1206625109

Nei, M., Gu, X., & Sitnikova, T. (1997). Evolution by the birth-and-death process in multigene families of the vertebrate immune system. *Proceedings of the National Academy of Sciences*, *94*(15), 7799–7806. https://doi.org/10.1073/pnas.94.15.7799

Nei, M., & Rooney, A. P. (2005). Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.*, *39*, 121–152.

Nelson, J. S. (2006). *Fishes of the World* (Fourth). John Wiley & Sons.

Nogi, T., Yasui, N., Mihara, E., Matsunaga, Y., Noda, M., Yamashita, N., Kumanogoh, A. (2010). Structural basis for semaphorin signalling through the plexin receptor. *Nature*, *467*(7319), 1123.

Nolf, D. (2004). Otolithes de poissons aptiens du Maestrazgo (province de Castellon, Espagne orientale). *Bulletin van Het Koninklijk Belgisch Instituut Voor Natuurwetenschappen. Aardwetenschappen= Bulletin de l'Institut Royal Des Sciences Naturelles de Belgique. Sciences de La Terre*.

Nuin, P. A. S., Wang, Z., & Tillier, E. R. M. (2006). The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, *7*(February). https://doi.org/10.1186/1471-2105-7-471

O'Brien, K. (2011). Notothenia coriiceps. Retrieved from https://www.polartrec.com/expeditions/biology-of-antarctic-fishes/photos

O'Brien, K. M., & Mueller, I. A. (2010). The unique mitochondrial form and function of Antarctic channichthyid icefishes. *Integrative and Comparative Biology*, *50*(6), 993–1008.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Ohta, T. (1980). Evolution and variation of multigene families. Lecture notes in biomathematics. Vol. 37. Springer-Verlag, New York.

Ohta, T. (2008). Gene Families: Multigene Families and Superfamilies. *Encyclopedia of Life Sciences*, (March 2008). https://doi.org/10.1038/npg.els.0005126

OIST, O. I. of S. and T. G. U. (n.d.). Gene loss pattern after teleost-specific whole genome duplication. Okinawa, Japan 904-0495: Okinawa Institute of Science and Technology Graduate University. Retrieved from https://www.oist.jp/news-center/photos/gene-loss-pattern-after-teleost-specific-whole-genome-duplication

Okkenhaug, K. (2013). Signalling by the phosphoinositide 3-kinase family in immune cells. *Annual Review of Immunology*, *31*(1), 675–704. https://doi.org/10.1146/annurev-immunol-032712-095946.Signalling

Okkenhaug, K., & Vanhaesebroeck, B. (2003). PI3K in lymphocyte development, differentiation and activation. *Nature Reviews Immunology*, *3*(4), 317–330. https://doi.org/10.1038/nri1056

Orhon, I., Dupont, N., Pampliega, O., Cuervo, A. M., & Codogno, P. (2015). Autophagy and regulation of cilia function and assembly. *Cell Death and Differentiation*, *22*(3), 389–397. https://doi.org/10.1038/cdd.2014.171

Orr, S. J., Quigley, L., & McVicar, D. W. (2009). In vivo expression of signaling proteins in reconstituted NK cells. *Journal of Immunological Methods*, *340*(2), 158–163.

Orsi, A. H., Whitworth, T., & Nowlin, W. D. (1995). On the meridional extent and fronts of

the Antarctic Circumpolar Current. *Deep-Sea Research Part I*, *42*(5), 641–673. https://doi.org/10.1016/0967-0637(95)00021-W

Oshiumi, H., Tsujita, T., Shida, K., Matsumoto, M., Ikeo, K., & Seya, T. (2003). Prediction of the prototype of the human Toll-like receptor gene family from the pufferfish, *Fugu rubripes*, genome. *Immunogenetics*, *54*, 791–800. https://doi.org/10.1007/s00251-002-0519-8

Ota, T., Nguyen, T.-A., Huang, E., Detrich, H. W., & Amemiya, C. T. (2003). Positive Darwinian selection operating on the immunoglobulin heavy chain of Antarctic fishes. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, *295*(August 2002), 45–58. https://doi.org/10.1002/jez.b.00004

Owen, R. (1848). *On the archetype and homologies of the vertebrate skeleton*. author.

Palti, Y. (2011). Toll-like receptors in bony fish: From genomics to function. *Developmental and Comparative Immunology*, *35*(12), 1263–1272. https://doi.org/10.1016/j.dci.2011.03.006

Papetti, C., Windisch, H. S., La Mesa, M., Lucassen, M., Marshall, C., & Lamare, M. D. (2016). Non-Antarctic notothenioids: Past phylogenetic history and contemporary phylogeographic implications in the face of environmental changes. *Marine Genomics*, *25*, 1–9. https://doi.org/10.1016/j.margen.2015.11.007

Pearson, R. W. (2013). An Introduction to Sequence Similarity ("Homology") Searching Curr. *Current Protocols in Bioinformatics*, *1*(10), 1286–1292. https://doi.org/10.1002/0471250953.bi0301s42.An

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics/Editoral Board, Andreas D. Baxevanis...[et Al.]*.

Pfuhl, H. A., & McCave, I. N. (2005). Evidence for late Oligocene establishment of the Antarctic Circumpolar Current. *Earth and Planetary Science Letters*, *235*(3–4), 715–728. https://doi.org/10.1016/j.epsl.2005.04.025

Phelan, P. E., Mellon, M. T., & Kim, C. H. (2005). Functional characterization of full-length TLR3, IRAK-4, and TRAF6 in zebrafish (Danio rerio). *Molecular Immunology*, *42*(9), 1057–1071.

Pinhal, D., Yoshimura, T. S., Araki, C. S., & Martins, C. (2011). The 5S rDNA family evolves through concerted and birth-and-death evolution in fish genomes: An example from freshwater stingrays. *BMC Evolutionary Biology*, *11*(1). https://doi.org/10.1186/1471-2148-11-151

Piontkivska, H., & Nei, M. (2003). Birth-and-death evolution in primate MHC class I genes: Divergence time estimates. *Molecular Biology and Evolution*, *20*(4), 601–609. https://doi.org/10.1093/molbev/msg064

Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, *53*(5), 793–808. https://doi.org/10.1080/10635150490522304

Posada, D., & Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics*, *14*(9), 817–818. https://doi.org/10.1093/bioinformatics/14.9.817

Posada, D., & Crandall, K. A. (2001). Selecting the Best-Fit Model of Nucleotide Substitution. *Systematic Biology*, *50*(4), 580–601. https://doi.org/10.1080/106351501750435121

Potiron, V. A., Roche, J., & Drabkin, H. A. (2009). Semaphorins and their receptors in lung cancer. *Cancer Letters*, *273*(1), 1–14.

Praxaysombath, B. (2008). Oreochromis niloticus. Retrieved January 28, 2019, from http://ffish.asia/photos/sp/409_Oreochromis_niloticus_NUOL-P01749.jpg

Putnam, N. H., Butts, T., Ferrier, D. E. K., Furlong, R. F., Hellsten, U., Kawashima, T., Yu, J.-K. (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature*, *453*(7198), 1064.

Racine, J. S. (2012). RStudio: A Platform-Independent IDE for R and Sweave. *Journal of Applied Econometrics*, *27*(1), 167–172.

Ramasamy, S., Ometto, L., Crava, C. M., Revadi, S., Kaur, R., Horner, D. S., Rota-Stabelli, O. (2016). The evolution of olfactory gene families in Drosophila and the genomic basis of chemical-ecological adaptation in Drosophila suzukii. *Genome Biology and Evolution*, *8*(8), 2297–2311. https://doi.org/10.1093/gbe/evw160

Rauta, P. R., Samanta, M., Dash, H. R., Nayak, B., & Das, S. (2014). Toll-like receptors (TLRs) in aquatic animals: Signaling pathways, expressions and immune responses. *Immunology Letters*, *158*(1–2), 14–24. https://doi.org/10.1016/j.imlet.2013.11.013

Rebl, A., Goldammer, T., & Seyfert, H. M. (2010). Toll-like receptor signaling in bony fish. *Veterinary Immunology and Immunopathology*, *134*(3–4), 139–150. https://doi.org/10.1016/j.vetimm.2009.09.021

Roach, J. C., Glusman, G., Rowen, L., Kaur, A., Purcell, M. K., Smith, K. D., Aderem, A. (2005). The evolution of vertebrate Toll-like receptors. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(27), 9577–9582. https://doi.org/10.1073/pnas.0502272102

Rombout, J. H. W. M., Huttenhuis, H. B. T., Picchietti, S., & Scapigliati, G. (2005). Phylogeny and ontogeny of fish leucocytes. *Fish and Shellfish Immunology*, *19*(5 SPEC. ISS.), 441–455. https://doi.org/10.1016/j.fsi.2005.03.007

Rusell, J. L., Stouffer, R. J., & Dixon, K. W. (2006). Intercomparison of the Southern Ocean circulations in IPCC coupled model control simulations. *Journal of Climate*, *19*(18), 4560–4575. https://doi.org/10.1175/JCLI3869.1

Ruud, J. T. (1954). Vertebrates without erythrocytes and blood pigment. *Nature*, *173*(4410), 848–850.

Sahoo, B. R., Dikhit, M. R., Bhoi, G. K., Maharana, J., Lenka, S. K., Dubey, P. K., & Tiwari, D. K. (2015). Understanding the distinguishable structural and functional features in zebrafish TLR3 and TLR22, and their binding modes with fish dsRNA viruses: an exploratory structural model analysis. *Amino Acids*, *47*(2), 381–400.

Salaneck, E., Larsson, T. A., Larson, E. T., & Larhammar, D. (2008). Birth and death of neuropeptide Y receptor genes in relation to the teleost fish tetraploidization. *Gene*, *409*(1–2), 61–71. https://doi.org/10.1016/j.gene.2007.11.011

Salemi, M., & Vandamme, A.-M. (2003). *The phylogenetic handbook: a practical approach to DNA and protein phylogeny*. Cambridge University Press.

Santini, F., Harmon, L. J., Carnevale, G., & Alfaro, M. E. (2009). Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evolutionary Biology*, *9*(1), 194.

Schnare, M., Barton, G. M., Holt, A. C., Takeda, K., Akira, S., & Medzhitov, R. (2001). Toll-like receptors control activation of adaptive immune responses. *Nature Immunology*, *2*(10), 947–950. https://doi.org/10.1038/ni712

Schulte, J. A. (2013). Undersampling Taxa Will Underestimate Molecular Divergence Dates: An Example from the South American Lizard Clade Liolaemini. *International Journal of Evolutionary Biology*, *2013*, 1–12. https://doi.org/10.1155/2013/628467

Schultz, K. T., & Grieder, F. (1987). Structure and function of the immune system. *Toxicologic Pathology*, *15*(3), 262–264. https://doi.org/10.1177/019262338701500301

Schultze, S. M., Jensen, J., Hemmings, B. A., Tschopp, O., & Niessen, M. (2011). Promiscuous affairs of PKB/AKT isoforms in metabolism. *Archives of Physiology and Biochemistry*, *117*(2), 70–77. https://doi.org/10.3109/13813455.2010.539236

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Shandikov, G. A. (2013). Dissostichus mawsoni. Retrieved January 28, 2019, from https://commons.wikimedia.org/wiki/File:Dissostichus_mawsoni_lateral.jpg

Shevenell, A. E., Kennett, J. P., & Lea, D. W. (2004). Middle Miocene Southern Ocean Cooling and Antarctic Cryosphere Expansion. *Science*, *305*(September), 1766–1770.

Sijp, W. P., von der Heydt, A. S., Dijkstra, H. A., Flögel, S., Douglas, P. M. J., & Bijl, P. K. (2014). The role of ocean gateways on cooling climate on long time scales. *Global and Planetary Change*, *119*, 1–22. https://doi.org/10.1016/j.gloplacha.2014.04.004

smerikal. (2011). Latimeria chalumnae. Retrieved January 28, 2019, from https://www.flickr.com/photos/smerikal/6227540054

Solbakken, M. H., Tørresen, O. K., Nederbragt, A. J., Seppola, M., Gregers, T. F., Jakobsen, K. S., & Jentoft, S. (2016). Evolutionary redesign of the Atlantic cod (Gadus morhua L.) Toll-like receptor repertoire by gene losses and expansions. *Scientific Reports*, *6*(January), 1–14. https://doi.org/10.1038/srep25211

Song, J., Zheng, S., Nguyen, N., Wang, Y., Zhou, Y., & Lin, K. (2017). Integrated pipeline for inferring the evolutionary history of a gene family embedded in the species tree: A case study on the STIMATE gene family. *BMC Bioinformatics*, *18*(1), 1–8. https://doi.org/10.1186/s12859-017-1850-2

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T. F., Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, *477*(7363), 207–210. https://doi.org/10.1038/nature10342

Stickley, C. E., Brinkhuis, H., Schellenberg, S. A., Sluijs, A., Röhl, U., Fuller, M., Williams, G. L. (2004). Timing and nature of the deepening of the Tasmanian Gateway. *Paleoceanography*, *19*(4), 1–18. https://doi.org/10.1029/2004PA001022

Strauss, T. (2006). Lepisosteus oculatus. Retrieved January 28, 2019, from https://commons.wikimedia.org/wiki/File:Lepisosteus_oculatus_03.jpg#/media/File:Lepisosteus_oculatus_03.jpg

Su, J., Zhu, Z., Wang, Y., Zou, J., & Hu, W. (2008). Toll-like receptor 3 regulates Mx expression in rare minnow Gobiocypris rarus after viral infection. *Immunogenetics*, *60*(3–4), 195–205.

Sun, T., Yang, L., Kaur, H., Pestel, J., Looso, M., Nolte, H., Santoni, M.-J. (2017). A reverse signaling pathway downstream of Sema4A controls cell migration via Scrib. *J Cell Biol*, *216*(1), 199–215.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.), 609–612. https://doi.org/10.1093/nar/gkl315

Suzuki, K., Okuno, T., Yamamoto, M., Pasterkamp, R. J., Takegahara, N., Takamatsu, H., Kolodkin, A. L. (2007). Semaphorin 7A initiates T-cell-mediated inflammatory responses through α1β1 integrin. *Nature*, *446*(7136), 680.

Takamatsu, H., Takegahara, N., Nakagawa, Y., Tomura, M., Taniguchi, M., Friedel, R. H., Okuno, T. (2010). Semaphorins guide the entry of dendritic cells into the lymphatics by activating myosin II. *Nature Immunology*, *11*(7), 594.

Takeda, K., Kaisho, T., & Akira, S. (2003). TOLL-LIKE RECEPTORS. *Annual Review of Immunology*, *21*(1), 335–376. https://doi.org/10.1146/annurev.immunol.21.120601.141126

Takeuchi, O., & Akira, S. (2010). Pattern Recognition Receptors and Inflammation. *Cell*,

140(6), 805–820. https://doi.org/10.1016/j.cell.2010.01.022

Tamaki, Y., & Maeda, K. (2016a). Danio rerio. Retrieved January 28, 2019, from https://oist-prod-www.s3-ap-northeast-1.amazonaws.com/s3fs-public/photos/20160912-Inoue-Prize-Zebrafish.jpg

Tamaki, Y., & Maeda, K. (2016b). Oryzias latipes. Retrieved January 28, 2019, from https://oist-prod-www.s3-ap-northeast-1.amazonaws.com/s3fs-public/photos/20160912-Inoue-Prize-Zebrafish.jpg

Tauber, A. I. (1992). The birth of immunology: III. The fate of the phagocytosis theory. *Cellular Immunology*, *139*(2), 505–530. https://doi.org/https://doi.org/10.1016/0008-8749(92)90089-8

Thompson, C. B. (1995). New insights into V(D)J recombination and its role in the evolution of the immune system. *Immunity*, *3*(5), 531–539. https://doi.org/10.1016/1074-7613(95)90124-8

Troutman, T. D., Bazan, J. F., & Pasare, C. (2012). Toll-like receptors, signaling adapters and regulation of the pro-inflammatory response by PI3K. *Cell Cycle*, *11*(19), 3559–3567. https://doi.org/10.4161/cc.21572

Ueda, K. (2013). Gasterosteus aculeatus. Retrieved January 28, 2019, from https://www.flickr.com/photos/ken-ichi/9467936453/in/photostream/

Usardi, A., Iyer, K., Sigoillot, S. M., Dusonchet, A., & Selimi, F. (2017). The immunoglobulin-like superfamily member IGSF3 is a developmentally regulated protein that controls neuronal morphogenesis. *Developmental Neurobiology*, *77*(1), 75–92.

Van de Peer, Y., Taylor, J. S., & Meyer, A. (2003). Are all fishes ancient polyploids? In *Genome Evolution* (pp. 65–73). Springer.

Vandamme, A.-M. (2003). Basic concepts of molecular evolution. In *The Phylogenic Handbook-A practical approach to DNA and protein phylogeny* (pp. 1–23). Cambridge University Press.

Volff, J. N. (2005). Genome evolution and biodiversity in teleost fish. *Heredity*, *94*(3), 280–294. https://doi.org/10.1038/sj.hdy.6800635

Werling, D., Jann, O. C., Offord, V., Glass, E. J., & Coffey, T. J. (2009). Variation matters: TLR structure and species-specific pathogen recognition. *Trends in Immunology*, *30*(3), 124–130. https://doi.org/10.1016/j.it.2008.12.001

Wilke, T., Schultheiß, R., & Albrecht, C. (2009). As Time Goes by: A Simple Fool's Guide to Molecular Clock Approaches in Invertebrates [*]. *American Malacological Bulletin*, *27*(1–2), 25–45. https://doi.org/10.4003/006.027.0203

Xiao, J., Zhong, H., Liu, Z., Yu, F., Luo, Y., Gan, X., & Zhou, Y. (2015). Transcriptome analysis revealed positive selection of immune-related genes in tilapia. *Fish and Shellfish*

Immunology, *44*(1), 60–65. https://doi.org/10.1016/j.fsi.2015.01.022

Yang, F., Shi, L., Liang, T., Ji, L., Zhang, G., Shen, Y., Xu, L. (2017). Anti-tumor effect of evodiamine by inducing Akt-mediated apoptosis in hepatocellular carcinoma. *Biochemical and Biophysical Research Communications*, *485*(1), 54–61.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591.

Yang, Z. (2014). *Molecular evolution: a statistical approach*. Oxford University Press.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098

7 Supplementary information

Table *SI*.1: Collections of sequence data for 8 of the species studied

Organism name	Database	Download date (2018)	Assembly name/ Biosample	Assembly date	Genebuild last geneset update	Assembly accession	Genebuild initial release date
Latimeria chalumnae	ensembl	Wed Mar 28	LatCha1	2011-09	2012-11	GCA_000225785.1	2011-10
Lepisosteus oculatus	ensembl	Mon Apr 2	LepOcu1	2011-12	2016-10	GCA_000242695.1	2013-12
Danio rerio	ensembl	Mon Mar 12	GRCz10	2014-09	2017-06	GCA_000002035.3	2015-05
Gadus morhua	ensembl	Wed Mar 28	gadMor1	2010-01	2011-08	N/A	2011-08
Oreochromis niloticus	ensembl	Sun Mar 11	Orenil1.0	2011-01	2016-10	GCA_000188235.1	2012-03
Oryzias latipes	ensembl	Sun Mar 11	HdrR	2005-10	2013-04	N/A	2006-10
Dicentrarchus labrax	http://seabass.mpipz.mpg.de	WedMar28	dicLab v1.0c	2012-06	2015-05	N/A	N/A
Gasterosteus aculeatus	ensembl	Wed Mar 28	BROADS1	2006-02	2010-05	N/A	2006-08

Table *SI.2*: Comparison of gene copy numbers for the five gene families among, *Notothenia coriiceps*, *Eleginops maclovinus*, *Dissostichus mawsoni, Dicentrarchus labrax, Danio rerio, Lepisosteus oculatus*, *Oreochromis niloticus, Oryzias latipes, Gadus morhua*, *Gasterosteus aculeatus and Latimeria chalumnae.*

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
Toll-Like receptor											
tlr1	1	1	2	N/A	1	1	1	1	1	1	1
tlr2	1	1	2	N/A	1	1	1	1	1	1	N/A
tlr3	1	1	1	1	1	1	1	1	1	1	1
tlr4al	N/A	N/A	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr4ba	N/A	1	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr4bb	N/A	N/A	1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr5a	1	1	3	N/A	1	1	1	N/A	1	1	1
tlr5b	1	1	2	N/A	1	1	1	1	N/A	1	1
tlr7	N/A	1	2	2	1	2	2	2	2	2	2
tlr8b	1	2	1	N/A	1	1	1	1	N/A	N/A	1
tlr9	1	1	1	5	1	1	1	1	1	N/A	1
tlr18	1	1	3	1	1	1	1	1	N/A	N/A	1
tlr19	1	N/A	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr20a/tlr20.1	1	N/A	2	N/A	N/A		1	N/A	1	1	N/A
tlr20b/tlr20.2	1	N/A	2	N/A	N/A	1	N/A	N/A	1	1	1
tlr20c/tlr20.3	1	N/A	2	N/A	N/A	N/A	1	N/A	2	1	N/A
tlr20d/tlr20.4	N/A	N/A	2	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr21	N/A	N/A	1	1	1	1	1	1	1	N/A	1
tlr22	N/A	1	1	1	N/A	1	1	1	N/A	N/A	N/A
Phosphatidyl -inositol 3-kinase											

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
pik3ap1	1	1	2	1	1	1	1	1	1	1	1
pik3c2a	1	1	2	1	1	1	1	1	1	1	N/A
pik3c2b	1	1	1	N/A	1	1	1	N/A	N/A	1	N/A
pik3c2g	1	N/A	1	N/A	1	1	1	N/A	1	1	1
pik3c3	1	1	3	1	1	1	1	1	1	1	1
pik3ca	2	2	2	3	2	3	3	3	2	2	2
pik3cb	1	1	2	1	1	1	1	1	1	1	N/A
pik3cd	1	1	1	1	1	1	1	1	1	1	1
pik3cg	1	1	1	1	1	1	1	1	1	1	1
pik3ip1	1	1	1	1	1	1	N/A	1	1	N/A	1
pik3r1	1	1	3	2	2	2	1	2	2	1	1
pik3r2	N/A	1	3	N/A	N/A	N/A	1	1	1	1	1
pik3r3a	1	N/A	3	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pik3r3b	N/A	1	2	1	1	2	1	2	1	1	1
pik3r4	1	1	2	1	1	1	1	1	N/A	1	
pik3r5	1	1	1	1	2	2	1	2	1	1	1
pik3r6a	N/A	1	1	N/A	1	1	1	N/A	N/A	1	1
pik3r6b	N/A	N/A	2	N/A	N/A	1	N/A	1	N/A	1	N/A
Immunoglobulin Superfamily											
igsf3	1	1	1	1	1	1	1	1	1	1	1
igsf5a	1	1	1	N/A	1	1	1	1	1	1	N/A
igsf5b	N/A	N/A	1	1	N/A	1	1	1	1	1	1
igsf8	1	1	1	1	1	1	1	1	1	1	1
igsf9a	N/A	N/A	1	N/A	1	1	1	N/A	1	1	1
igsf9b	N/A	1	1	N/A	1	1	1	N/A	1	1	1
igsf9ba	N/A	1	1	1	1	1	1	1	1	1	1

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
igsf9bb	1	1	1	1	1	1	1	1	1	2	1
igsf1	1	1	1	1	1	1	1	1	N/A	N/A	1
igsf11	1	1	1	1	1	1	1	1	1	1	1
Semaphorins											
sema3aa	N/A	N/A	1	N/A	1	1	N/A	N/A	N/A	1	1
sema3ab	1	1	1	2	1	1	2	2	2	1	1
sema3b	1	1	1	1	1	1	1	1	1	1	1
sema3bl	1	1	1	1	1	1	1	1	1	1	1
sema3c	1	1	1	1	1	1	1	1	1	1	N/A
sema3d	2	2	1	1	2	2	2	1	1	1	1
sema3e	1	1	1	1	1	1	1	1	1	N/A	1
sema3fa	N/A	N/A	1	1	N/A	1	N/A	1	N/A	N/A	N/A
sema3fb	1	1	1	1	1	1	2	1	2	2	1
sema3ga	N/A	N/A	1	1	1	1	1	1	1	1	
sema3gb	N/A	1	1	1	1	N/A	N/A	N/A	N/A	N/A	N/A
sema3h	1	N/A	1	1	1	1	1	1	1	1	1
sema4aa	1	1	1	1	1	1	1	1	1	1	1
sema4ab	N/A	1	1	1	1	1	1	1	1	1	
sema4ba	1	1	1	2	2	2	2	2	1	2	1
sema4bb	N/A	N/A	1	N/A	1	1	N/A	N/A	N/A	N/A	N/A
sema4c	1	1	1	1	1	1	1	1	1	1	1
sema4ga	N/A	N/A	1	1	1	1	1	1	1	1	1
sema4gb	1	1	1	1	1	1	1	1	1	1	1
AKT serine/ threonine kinase 3											

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
akt3a	3	3	2	3	5	4	3	3	1	2	3
akt3b	1	1	1	1	1	1	2	2	1	1	1

Table *SI*.3: Gene accession number for the longest sequence obtained for each species.

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
Toll-Like receptor											
tlr1	ENSLACG00000010038.1	ENSLOCG00000012910.1	ENSDARG00000100649.1	N/A	ENSORLG00000004420.1	NC_031966.1	dicLab1_genemodels_DLAgn_00109390_1	ENSGACG00000017958.1	evm.model.scaffold273.16	Dissostichus_mawsoni_GLEAN_10003375	gene12881
tlr2	ENSLACG00000012590.1	ENSLOCG00000018220.1	ENSDARG00000037758.5	N/A	ENSORLG00000002540.2	ENSONIG00000014114.1	dicLab1_genemodels_DLAgn_00214290_1	ENSGACG00000018669.1	evm.model.scaffold1613.5	Dissostichus_mawsoni_GLEAN_10002497	N/A
tlr3	ENSLACG00000011410.1	ENSLOCG00000013826.1	ENSDARG00000016065.10	ENSGMOG00000000786.1	ENSORLG00000008184.1	ENSONIG00000010172.1	dicLab1_genemodels_DLAgn_00183840_1	ENSGACG00000016874.1	evm.model.scaffold69.39	Dissostichus_mawsoni_GLEAN_10021765	gene14033
tlr4al	N/A	N/A	ENSDARG00000075671.4	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr4ba	N/A	ENSLOCG00000003751.1	ENSDARG00000019742.9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr4bb	N/A	N/A	ENSDARG00000022048.5	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr5a	ENSLACG00000000352.1	ENSLOCG00000018000.1	ENSDARG00000044415.9	N/A	ENSORLG00000016221.1	ENSONIG00000001333.1	dicLab1_genemodels_DLAgn_00066790_1	N/A	evm.model.scaffold229.27	Dissostichus_mawsoni_GLEAN_10009026	gene15521
tlr5b	ENSLACG00000015379.1	ENSLOCG00000018000.1	ENSDARG00000052322.8	N/A	NC_019882.2	ENSONIG00000001333.1	dicLab1_genemodels_DLAgn_00069890_1	ENSGACG00000004381.1	N/A	Dissostichus_mawsoni_GLEAN_10014118	gene22462
tlr7	N/A	N/A	ENSDARG00000075685.3	ENSGMOG00000007185.1	N/A	ENSONIG00000016050.1	dicLab1_genemodels_DLAgn_00139190_1	ENSGACG00000015050.1	evm.model.scaffold71.45	Dissostichus_mawsoni_GLEAN_10017339	gene9014

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
tlr8b	ENSLACG00000016240.1	ENSLOCG00000009982.1	ENSDARG00000073675.3	N/A	ENSORLG00000014255.1	ENSONIG00000019220.1	dicLab1_genemodels_DLAgn_00051580_1	ENSGACG00000003992.1	N/A	N/A	gene9016
tlr9a	N/A	N/A	ENSDARG00000044490.7	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr9b	N/A	N/A	N/A	ENSGMOG00000003222.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr9c	N/A	N/A	N/A	ENSGMOG00000003269.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr9d	N/A	N/A	N/A	ENSGMOG00000011244.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr9e	N/A	N/A	N/A	ENSGMOG00000011256.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A
tlr18	ENSLACG00000017699.1	ENSLOCG00000007992.1	ENSDARG00000040249.8	ENSGMOG00000003793.1	ENSORLG00000015704.1	ENSONIG00000006684.1	dicLab1_genemodels_DLAgn_00055750_1	ENSGACG00000001745.1	N/A	N/A	gene5762
tlr20a/tlr20.1	ENSLACG00000001078.1	N/A	ENSDARG00000092668.2	N/A	N/A	N/A	dicLab1_genemodels_DLAgn_00100740_1	N/A	evm.model.scaffold147.1	Dissostichus_mawsoni_GLEAN_10000475	N/A
tlr20b/tlr20.2	ENSLACG00000001078	N/A	ENSDARG00000094411.2	N/A	N/A	ENSONIG00000011671.1	N/A	N/A	evm.model.scaffold1495.1	Dissostichus_mawsoni_GLEAN_10000764	gene7986

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
tlr20c/tlr20.3	ENSLACG00000002657.1	N/A	ENSDARG00000041164.7	N/A	N/A	N/A	dicLab1_genemodels_DLAgn_00100740_1	N/A	evm.model.scaffold364.21	Dissostichus_mawsoni_GLEAN_10018235	N/A
tlr20d/tlr20.4	N/A	N/A	ENSDARG00000088701.3	N/A	N/A	N/A	N/A	N/A	evm.model.scaffold364.22	N/A	N/A
tlr21	N/A	N/A	ENSDARG00000058045.7	ENSGMOG00000018200.1	ENSORLG00000013437.1	ENSONIG00000020525.1	dicLab1_genemodels_DLAgn_00058280_1	ENSGACG00000009364.1	evm.model.scaffold93.53	N/A	gene11519
tlr22	N/A	ENSLOCG00000018316.1	ENSDARG00000104045.1	ENSGMOG00000010841.1	N/A	ENSONIG00000006496.1	dicLab1_genemodels_DLAgn_00078370_1	ENSGACG00000005449.1	N/A	N/A	N/A
Phosphatid-ylinositol 3-kinase											
pik3ap1	ENSLACG00000008915.1	ENSLOCG00000011598.1	ENSDARG00000078285.4	ENSGMOG00000006202.1	ENSORLG00000014121.1	ENSONIG00000010335.1	dicLab1_genemodels_DLAgn_00010760_1	ENSGACG00000003216.1	evm.model.scaffold9.22	Dissostichus_mawsoni_GLEAN_10007944	gene13803
pik3c2a	ENSLACG00000017040.2	ENSLOCG00000002963.1	ENSDARG00000060841.6	ENSGMOG00000010481.1	101155105	ENSONIG00000015560.1	dicLab1_genemodels_DLAgn_00171600_1	ENSGACG00000006738.1	evm.model.scaffold1027.4	Dissostichus_mawsoni_GLEAN_10011998	N/A
pik3c2b	ENSLACG00000017407.1	ENSLOCG00000012102.1	ENSDARG00000086927.3	N/A	ENSORLG00000010776.1	ENSONIG00000000290.1	dicLab1_genemodels_DLAgn_00214900_1	N/A	N/A	Dissostichus_mawsoni_GLEAN_10001230	N/A
pik3c2g	ENSLACG00000004021.1	N/A	ENSDARG00000099803.2	N/A	105354205	ENSONIG00000008497.1	dicLab1_genemodels_DLAgn_00169860_1	N/A	evm.model.scaffold120.12	Dissostichus_mawsoni_GLEAN_10005289	gene471

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
pik3c3	ENSLACG00000016070.1	ENSLOCG00000009397.1	ENSDARG00000054829.9	ENSGMOG00000017662.1	ENSORLG00000007721.1	ENSONIG00000012919.1	dicLab1_genemodels_DLAgn_00144560_1	ENSGACG00000019111.1	evm.model.scaffold232.8	Dissostichus_mawsoni_GLEAN_10005899	gene18828
pik3ca	ENSLACG00000011888.1	ENSLOCG00000001119.1	ENSDARG00000075456.4	ENSGMOG00000003009.1	ENSORLG00000008938.1	ENSONIG00000008379.1	dicLab1_genemodels_DLAgn_00003050_1	ENSGACG00000001192.1	evm.model.scaffold47.51	Dissostichus_mawsoni_GLEAN_10001047	gene21394
pik3cb	ENSLACG00000013175.1	ENSLOCG00000004421.1	ENSDARG00000075253.4	ENSGMOG00000013996.1	ENSORLG00000015529.1	ENSONIG00000007924.1	dicLab1_genemodels_DLAgn_00029270_1	ENSGACG00000005645.1	evm.model.scaffold469.11	Dissostichus_mawsoni_GLEAN_10009163	N/A
pik3cd	ENSLACG00000008900.1	ENSLOCG00000006919.1	ENSDARG00000003250.9	ENSGMOG00000000722.1	ENSORLG00000005422.1	ENSONIG00000002388.1	dicLab1_genemodels_DLAgn_00132120_1	ENSGACG00000007313.1	evm.model.scaffold115.59	Dissostichus_mawsoni_GLEAN_10013699	gene17170
pik3cg	ENSLACG00000006986.1	ENSLOCG00000015829.1	ENSDARG00000017757.8	ENSGMOG00000009314.1	ENSORLG00000009458.1	ENSONIG00000011855.1	dicLab1_genemodels_DLAgn_00206460_1	ENSGACG00000019186.1	evm.model.scaffold646.15	Dissostichus_mawsoni_GLEAN_10020819	gene27107
pik3ip1	ENSLACG00000016745.1	ENSLOCG00000004841.1	ENSDARG00000003281.6	ENSGMOG00000011675.1	ENSORLG00000006788.1	ENSONIG00000013396.1	N/A	ENSGACG00000008705.1	evm.model.scaffold101.61	N/A	gene11466
pik3r1	ENSLACG00000013626.2	ENSLOCG00000010510.1	ENSDARG00000038524.8	ENSGMOG00000003321.1	ENSORLG00000002670.1	ENSONIG00000001749.1	dicLab1_genemodels_DLAgn_00081990_1	ENSGACG00000001062.1	evm.model.scaffold1025.5	Dissostichus_mawsoni_GLEAN_10012840	gene20774
pik3r2	N/A	ENSLOCG00000001519.1	ENSDARG00000018060.11	N/A	N/A	N/A	dicLab1_genemodels_DLAgn_00003950_1	ENSGACG00000008201.1	evm.model.scaffold25.18	Dissostichus_mawsoni_GLEAN_10021203	gene6595

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
pik3r3a	ENSLACG00000007678.1	N/A	ENSDARG00000103038.1	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
pik3r3b	N/A	ENSLOCG00000005759.1	ENSDARG00000034409.6	ENSGMOG00000000803.1	ENSORLG00000013341.1	ENSONIG00000007853.1	dicLab1_genemodels_DLAgn_00145130_1	ENSGACG00000005180.1	evm.model.scaffold69.11	Dissostichus_mawsoni_GLEAN_10011059	gene18971
pik3r4	ENSLACG00000010358.1	ENSLOCG00000001622.1	ENSDARG00000060469.6	ENSGMOG00000003459.1	101173091	ENSONIG00000002040.1	dicLab1_genemodels_DLAgn_00215370_1	ENSGACG00000002977.1	N/A	Dissostichus_mawsoni_GLEAN_10001744	N/A
pik3r5	ENSLACG00000003800.1	ENSLOCG00000013254.1	ENSDARG00000102762.1	ENSGMOG00000017849.1	ENSORLG00000003122.1	ENSONIG00000008249.1	dicLab1_genemodels_DLAgn_00180100_1	ENSGACG00000006852.1	evm.model.scaffold46.46	Dissostichus_mawsoni_GLEAN_10021904	gene24931
pik3r6a	N/A	ENSLOCG00000013248.1	ENSDARG00000100336.1	N/A	ENSORLG00000003102.1	ENSONIG00000018161.1	dicLab1_genemodels_DLAgn_00180090_1	N/A	N/A	Dissostichus_mawsoni_GLEAN_10021903	gene24929
pik3r6b	N/A	N/A	ENSDARG00000091140.3	N/A	N/A	ENSONIG00000008242.1	N/A	ENSGACG00000006848.1	N/A	Dissostichus_mawsoni_GLEAN_10013057	N/A
Immuno-globulin Superfamily											
igsf3	ENSLACG00000013643.2	ENSLOCG00000011156.1	ENSDARG00000077002.3	ENSGMOG00000000627.1	ENSORLG00000017870.1	ENSONIG00000016010.1	dicLab1_genemodels_DLAgn_00051750_1	ENSGACG00000003812.1	evm.model.scaffold54.55	Dissostichus_mawsoni_GLEAN_10015527	gene27465
igsf5a	ENSLACG00000001489.1	XP_015196887.1	ENSDARG00000087983.3	N/A	ENSORLG00000018654.1	XP_005454662.2	dicLab1_genemodels_DLAgn_00039170_1	ENSGACG00000020189.1	evm.model.scaffold20.65	Dissostichus_mawsoni_GLEAN_10013291	N/A

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
igsf5b	N/A	N/A	ENSDARG00000090953.2	ENSGMOG00000008789.1	N/A	XP_005474614.1	dicLab1_genemodels_DLAgn_00030580_1	ENSGACG00000006903.1	evm.model.scaffold160.31	Dissostichus_mawsoni_GLEAN_10013064	gene17536
igsf8	ENSLACG00000015060.1	ENSLOCG00000005749.1	ENSDARG00000038467.6	ENSGMOG00000001253.1	ENSORLG00000005224.1	ENSONIG00000005111.1	dicLab1_genemodels_DLAgn_00095260_1	ENSGACG00000006633.1	evm.model.scaffold23.15	Dissostichus_mawsoni_GLEAN_10019441	gene27045
igsf9a	N/A	N/A	ENSDARG00000075864.4	N/A	XP_011477675.1	ENSONIG00000013054.1	dicLab1_genemodels_DLAgn_00086800_1	N/A	evm.model.scaffold878.1	Dissostichus_mawsoni_GLEAN_10013587	gene15055
igsf9b	N/A	ENSLOCG00000016092.1	ENSDARG00000010408.10	N/A	ENSORLG00000012192.1	ENSONIG00000010698.1	dicLab1_genemodels_DLAgn_00115840_1	N/A	evm.model.scaffold202.15	Dissostichus_mawsoni_GLEAN_10020178	gene22095
igsf9ba	N/A	N/A	ENSDARG00000033845.9	ENSGMOG00000010464.1	ENSORLG00000005174.1	ENSONIG00000011122.1	dicLab1_genemodels_DLAgn_00033960_1	ENSGACG00000010675.1	evm.model.scaffold102.21	Dissostichus_mawsoni_GLEAN_10010237	gene7018
igsf9bb	ENSLACG00000004883.1	ENSLOCG00000004847.1	ENSDARG00000069467.5	ENSGMOG00000011998.1	XP_023818071.1	ENSONIG00000015795.1	dicLab1_genemodels_DLAgn_00038810_1	ENSGACG00000020185.1	evm.model.scaffold20.20	Dissostichus_mawsoni_GLEAN_10001500	gene5428
igsf10	ENSLACG00000009276.1	XP_015216099.1	ENSDARG00000077497.4	ENSGMOG00000013742.1	ENSORLG00000005965.1	ENSONIG00000003578.1	dicLab1_genemodels_DLAgn_00033430_1	ENSGACG00000010238.1	N/A	N/A	gene15396
igsf11	ENSLACG00000007575.1	ENSLOCG00000002662.1	ENSDARG00000017217.9	ENSGMOG00000013368.1	ENSORLG00000013761.1	ENSONIG00000016695.1	dicLab1_genemodels_DLAgn_00029830_1	ENSGACG00000006157.1	evm.model.scaffold32.70	Dissostichus_mawsoni_GLEAN_10019777	gene8051
Semaphorins											

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
sema3aa	N/A	N/A	ENSDARG00000019235.10	N/A	ENSORLG00000015034.1	ENSONIG00000011531.1	N/A	N/A	N/A	Dissostichus_mawsoni_GLEAN_10016131	gene15947
sema3ab	ENSLACG00000013866.1	ENSLOCG00000016397.1	ENSDARG00000042210.7	ENSGMOG00000018773.1	ENSORLG00000009475.1	ENSONIG00000014966.1	dicLab1_genemodels_DLAgn_00167340_1	ENSGACG00000012703.1	evm.model.scaffold13.66	Dissostichus_mawsoni_GLEAN_10006936	gene16003
sema3b	ENSLACG00000016211.1	ENSLOCG00000014325.1	ENSDARG00000011672.11	ENSGMOG00000017239.1	ENSORLG00000003270.1	ENSONIG00000001792.1	dicLab1_genemodels_DLAgn_00133350_1	ENSGACG00000005848.2	evm.model.scaffold16.38	Dissostichus_mawsoni_GLEAN_10010796	gene21409
sema3bl	ENSLACG00000015321.1	ENSLOCG00000014307.1	ENSDARG00000007560.11	ENSGMOG00000016820.1	ENSORLG00000003333.1	ENSONIG00000001819.1	dicLab1_genemodels_DLAgn_00133330_1	ENSGACG00000005862.1	evm.model.scaffold16.39	Dissostichus_mawsoni_GLEAN_10010797	gene21410
sema3c	ENSLACG00000008248.1	ENSLOCG00000016023.1	ENSDARG00000034300.8	ENSGMOG00000011696.1	ENSORLG00000016082.1	ENSONIG00000000089.1	dicLab1_genemodels_DLAgn_00209160_1	ENSGACG00000019956.1	evm.model.scaffold21.60	Dissostichus_mawsoni_GLEAN_10016199	gene10776
sema3d	ENSLACG00000005273.1	ENSLOCG00000013481.1	ENSDARG00000017369.10	ENSGMOG00000019387.1	ENSORLG00000002866.1	ENSONIG00000004442.1	dicLab1_genemodels_DLAgn_00167350_1	ENSGACG00000012711.1	evm.model.scaffold83.41	Dissostichus_mawsoni_GLEAN_10006937	gene16002
sema3e	ENSLACG00000015836.1	ENSLOCG00000016395.1	ENSDARG00000036571.6	ENSGMOG00000019371.1	ENSORLG00000009509.1	ENSONIG00000014961.1	dicLab1_genemodels_DLAgn_00167330_1	ENSGACG00000012696.1	evm.model.scaffold83.39	N/A	gene16004
sema3fa	N/A	N/A	ENSDARG00000011163.10	ENSGMOG00000016462.1	N/A	ENSONIG00000004064.1	N/A	ENSGACG00000010605.1	N/A	N/A	N/A

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
sema3fb	ENSLACG00000005125.1	ENSLOCG00000010911.1	ENSDARG00000055373.7	ENSGMOG00000002430.1	ENSORLG00000014370.1	ENSONIG00000020280.1	dicLab1_genemodels_DLAgn_00089240_1	ENSGACG00000011989.1	evm.model.scaffold188.46	Dissostichus_mawsoni_GLEAN_10008090	gene19956
sema3ga	N/A	N/A	ENSDARG00000042545.5	ENSGMOG00000019678.1	ENSORLG00000014173.1	ENSONIG00000016871.1	dicLab1_genemodels_DLAgn_00127620_1	ENSGACG00000012228.1	evm.model.scaffold1552.1	Dissostichus_mawsoni_GLEAN_10005944	N/A
sema3gb	N/A	ENSLOCG00000014436.1	ENSDARG00000013607.9	ENSGMOG00000001255.1	XP_011475770.1	N/A	N/A	N/A	N/A	N/A	N/A
sema3h	ENSLACG00000018034.1	N/A	ENSDARG00000042616.5	ENSGMOG00000002408.1	ENSORLG00000002802.1	ENSONIG00000001675.1	dicLab1_genemodels_DLAgn_00133580_1	ENSGACG00000005697.1	evm.model.scaffold16.14	Dissostichus_mawsoni_GLEAN_10010776	gene3960
sema4aa	ENSLACG00000013812.1	ENSLOCG00000006579.1	ENSDARG00000077103.4	ENSGMOG00000012258.1	ENSORLG00000016896.1	ENSONIG00000002132.1	dicLab1_genemodels_DLAgn_00083690_1	ENSGACG00000015326.1	evm.model.scaffold88.27	Dissostichus_mawsoni_GLEAN_10001854	gene2977
sema4ab	N/A	ENSLOCG00000000333.1	ENSDARG00000062352.6	ENSGMOG00000018516.1	XP_023815848.1	ENSONIG00000010347.1	dicLab1_genemodels_DLAgn_00197820_1	ENSGACG00000001284.1	evm.model.scaffold360.2	Dissostichus_mawsoni_GLEAN_10005076	N/A
sema4ba	ENSLACG00000018085.2	ENSLOCG00000014333.1	ENSDARG00000074414.5	ENSGMOG00000013124.1	ENSORLG00000008247.1	ENSONIG00000002592.1	dicLab1_genemodels_DLAgn_00156190_1	ENSGACG00000010807.1	evm.model.scaffold248.5	Dissostichus_mawsoni_GLEAN_10015343	gene24172
sema4bb	N/A	N/A	ENSDARG00000076104.5	N/A	XP_011491188.1	ENSONIG00000002563.1	N/A	N/A	N/A	N/A	N/A

Gene	*Latimeria chalumnae*	*Lepisosteus oculatus*	*Danio rerio*	*Gadus morhua*	*Oryzias latipes*	*Oreochromis niloticus*	*Dicentrarchus labrax*	*Gasterosteus aculeatus*	*Eleginops maclovinus*	*Dissostichus mawsoni*	*Notothenia coriiceps*
sema4c	ENSLACG00000002282.1	ENSLOCG00000015594.1	ENSDARG00000079611.4	ENSGMOG00000008050.1	ENSORLG00000015302.1	ENSONIG00000015832.1	dicLab1_genemodels_DLAgn_00123460_1	ENSGACG00000013062.1	evm.model.scaffold504.5	Dissostichus_mawsoni_GLEAN_10016674	gene10355
sema4ga	N/A	N/A	ENSDARG00000076595.6	ENSGMOG00000013766.1	ENSORLG00000007299.1	ENSONIG00000018453.1	dicLab1_genemodels_DLAgn_00013790_1	ENSGACG00000008034.1	evm.model.scaffold72.40	Dissostichus_mawsoni_GLEAN_10020493	gene5009
sema4gb	ENSLACG00000006411.1	ENSLOCG00000012636.1	ENSDARG00000088143.2	ENSGMOG00000009878.1	ENSORLG00000011518.1	ENSONIG00000007902.1	dicLab1_genemodels_DLAgn_00104470_1	ENSGACG00000003493.1	evm.model.scaffold225.9	Dissostichus_mawsoni_GLEAN_10016615	gene11901
AKT serine/threo-nine kinase 3											
akt3a	ENSLACG00000004447.1	ENSLOCG00000012762.1	ENSDARG00000104810.1	ENSGMOG00000001153.1	ENSORLG00000004813.1	ENSONIG00000000863.1	dicLab1_genemodels_DLAgn_00015480_1	ENSGACG00000006296.1	evm.model.scaffold32.53	Dissostichus_mawsoni_GLEAN_10019793	gene10588
akt3b	ENSLACG00000003568.1	ENSLOCG00000005077.1	ENSDARG00000087205.2	ENSGMOG00000004973.1	ENSORLG00000001445.1	ENSONIG00000000038.1	dicLab1_genemodels_DLAgn_00038440_1	ENSGACG00000019752.1	evm.model.scaffold1716.2	Dissostichus_mawsoni_GLEAN_10003410	gene25711

Table *SI*.4: Results of the pairwise Codeml analysis. Non-synonymous substitution (dn), synonymous substitution (dS), their ratio (dn/ds) and the calculated divergence time (million years) from one-to-one concatenated genes between, ten of the investigated species with the gene name and the corresponding node given, using the estimated substitution rate of 5.7×10^{-9} mutations per site per year.

Node	Gene	Species 1	Species 2	dnds	dn	ds	Divergence Time
Neopterygii	Sema	*L.oculatus*	*D.labrax*	0.0489	0.1413	2.8927	253.75
Neopterygii	Sema	*L.oculatus*	*D.mawsoni*	0.0419	0.1305	3.1149	273.24
Neopterygii	Sema	*L.oculatus*	*D.rerio*	0.044	0.1622	3.6906	323.74
Neopterygii	Sema	*L.oculatus*	*O.niloticus*	0.0386	0.1376	3.5666	312.86
Neopterygii	Sema	*L.oculatus*	*G.morhua*	0.0415	0.1437	3.4621	303.69
Neopterygii	Sema	*L.oculatus*	*E.maclovinus*	0.0378	0.1357	3.593	315.18
Neopterygii	Sema	*L.oculatus*	*G.aculeatus*	0.0353	0.1282	3.6321	318.61
Acanthomorphata	AKT3	*G.morhua*	*D.mawsoni*	0.0218	0.0257	1.1751	103.08
Acanthomorphata	AKT3	*G.morhua*	*G.aculeatus*	0.0228	0.0266	1.1659	102.27
Acanthomorphata	AKT3	*G.morhua*	*N.coriiceps*	0.0278	0.0337	1.2095	106.10
Acanthomorphata	AKT3	*G.morhua*	*E.maclovinus*	0.0192	0.025	1.3011	114.13
Acanthomorphata	Igsf	*G.morhua*	*G.aculeatus*	0.0377	0.082	2.1781	191.06
Acanthomorphata	Igsf	*G.morhua*	*D.labrax*	0.0324	0.0778	2.399	210.44
Acanthomorphata	Igsf	*G.morhua*	*O.niloticus*	0.0358	0.0963	2.6932	236.25
Acanthomorphata	Igsf	*G.morhua*	*D.mawsoni*	0.0312	0.0852	2.7297	239.45
Acanthomorphata	Igsf	*G.morhua*	*E.maclovinus*	0.0289	0.0881	3.0455	267.15
Acanthomorphata	Igsf	*G.morhua*	*N.coriiceps*	0.0282	0.0898	3.1809	279.03
Acanthomorphata	Igsf	*G.morhua*	*O.latipes*	0.0325	0.1114	3.4284	300.74
Acanthomorphata	PIK3	*G.morhua*	*D.mawsoni*	0.0561	0.1162	2.0708	181.65
Acanthomorphata	PIK3	*G.morhua*	*N.coriiceps*	0.055	0.1161	2.1121	185.27
Acanthomorphata	PIK3	*G.morhua*	*G.aculeatus*	0.0492	0.1106	2.2477	197.17
Acanthomorphata	PIK3	*G.morhua*	*D.labrax*	0.0436	0.1045	2.3982	210.37
Acanthomorphata	PIK3	*G.morhua*	*O.niloticus*	0.0459	0.1115	2.4285	213.03
Acanthomorphata	PIK3	*G.morhua*	*E.maclovinus*	0.0445	0.1154	2.5967	227.78
Acanthomorphata	PIK3	*G.morhua*	*O.latipes*	0.0343	0.125	3.6459	319.82
Acanthomorphata	Sema	*G.morhua*	*D.labrax*	0.0383	0.0785	2.0515	179.96
Acanthomorphata	Sema	*G.morhua*	*D.mawsoni*	0.0353	0.0729	2.0647	181.11
Acanthomorphata	Sema	*G.morhua*	*E.maclovinus*	0.033	0.0782	2.3682	207.74
Acanthomorphata	Sema	*G.morhua*	*G.aculeatus*	0.0389	0.0728	1.872	164.21
Acanthomorphata	Sema	*G.morhua*	*O.niloticus*	0.0289	0.0735	2.5384	222.67
Acanthomorphata	Sema	*G.morhua*	*N.coriiceps*	0.0362	0.0929	2.5652	225.02
Acanthomorphata	Sema	*G.morhua*	*O.latipes*	0.0345	0.0897	2.5998	228.05
Acanthomorphata	TLR	*G.morhua*	*D.labrax*	0.1988	0.2722	1.3688	120.07
Acanthomorphata	TLR	*G.morhua*	*E.maclovinus*	0.2013	0.3009	1.495	131.14
Acanthomorphata	TLR	*G.morhua*	*D.mawsoni*	0.182	0.295	1.6208	142.18
Acanthomorphata	TLR	*G.morhua*	*N.coriiceps*	0.1756	0.2981	1.6975	148.90
Acanthomorphata	TLR	*G.morhua*	*G.aculeatus*	0.1507	0.26	1.7257	151.38
Acanthomorphata	TLR	*G.morhua*	*O.latipes*	0.1583	0.3222	2.0353	178.54

Node	Gene	Species 1	Species 2	dnds	dn	ds	Divergence Time
Acanthomorphata	TLR	*G.morhua*	*O.niloticus*	0.1334	0.2736	2.0508	179.89
Teleostei	AKT3	*D.rerio*	*D.labrax*	0.0114	0.0185	1.6232	142.39
Teleostei	AKT3	*D.rerio*	*O.latipes*	0.0183	0.0329	1.7982	157.74
Teleostei	AKT3	*D.rerio*	*O.niloticus*	0.0185	0.034	1.8393	161.34
Teleostei	Igsf	*D.rerio*	*N.coriiceps*	0.0715	0.1414	1.9766	173.39
Teleostei	Igsf	*D.rerio*	*D.mawsoni*	0.0663	0.1351	2.0394	178.89
Teleostei	Igsf	*D.rerio*	*E.maclovinus*	0.0613	0.1362	2.2213	194.85
Teleostei	Igsf	*D.rerio*	*D.labrax*	0.0548	0.1285	2.3435	205.57
Teleostei	Igsf	*D.rerio*	*O.niloticus*	0.0576	0.1478	2.5664	225.12
Teleostei	Igsf	*D.rerio*	*G.aculeatus*	0.0436	0.1355	3.1044	272.32
Teleostei	Igsf	*D.rerio*	*O.latipes*	0.0474	0.1604	3.3854	296.96
Teleostei	PIK3	*D.rerio*	*D.labrax*	0.0457	0.1181	2.5858	226.82
Teleostei	PIK3	*D.rerio*	*O.niloticus*	0.0475	0.1258	2.6473	232.22
Teleostei	PIK3	*D.rerio*	*E.maclovinus*	0.0463	0.123	2.656	232.98
Teleostei	PIK3	*D.rerio*	*O.latipes*	0.0471	0.1383	2.9378	257.70
Teleostei	PIK3	*D.rerio*	*G.aculeatus*	0.0418	0.1256	3.001	263.25
Teleostei	PIK3	*D.rerio*	*D.mawsoni*	0.0425	0.1285	3.0254	265.39
Teleostei	PIK3	*D.rerio*	*N.coriiceps*	0.0415	0.1294	3.1136	273.12
Teleostei	Sema	*D.rerio*	*D.labrax*	0.0543	0.1519	2.7972	245.37
Teleostei	Sema	*D.rerio*	*E.maclovinus*	0.0509	0.1485	2.92	256.14
Teleostei	Sema	*D.rerio*	*D.mawsoni*	0.0453	0.1425	3.1471	276.06
Teleostei	Sema	*D.rerio*	*O.latipes*	0.0488	0.1589	3.2524	285.30
Teleostei	Sema	*D.rerio*	*O.niloticus*	0.0395	0.1456	3.6841	323.17
Teleostei	Sema	*D.rerio*	*N.coriiceps*	0.0451	0.1676	3.7181	326.15
Teleostei	Sema	*D.rerio*	*G.aculeatus*	0.0364	0.1419	3.9007	342.17
Percomorpha	AKT3	*O.niloticus*	*D.labrax*	0.0551	0.0251	0.4555	39.96
Percomorpha	AKT3	*O.latipes*	*D.labrax*	0.0338	0.0252	0.746	65.44
Percomorpha	AKT3	*O.niloticus*	*O.latipes*	0.0324	0.0263	0.8129	71.31
Percomorpha	Igsf	*O.niloticus*	*O.latipes*	0.0916	0.0803	0.8766	76.89
Percomorpha	Igsf	*O.niloticus*	*D.labrax*	0.0949	0.0482	0.5076	44.53
Percomorpha	Igsf	*O.niloticus*	*D.mawsoni*	0.1012	0.0598	0.5914	51.88
Percomorpha	Igsf	*O.niloticus*	*N.coriiceps*	0.1102	0.0659	0.5981	52.46
Percomorpha	Igsf	*O.niloticus*	*E.maclovinus*	0.0994	0.0596	0.6001	52.64
Percomorpha	Igsf	*O.niloticus*	*G.aculeatus*	0.0618	0.0568	0.9193	80.64
Percomorpha	Igsf	*O.latipes*	*D.labrax*	0.077	0.0663	0.861	75.53
Percomorpha	Igsf	*O.latipes*	*D.mawsoni*	0.0833	0.0784	0.9408	82.53
Percomorpha	Igsf	*O.latipes*	*N.coriiceps*	0.09	0.0861	0.9561	83.87
Percomorpha	Igsf	*O.latipes*	*E.maclovinus*	0.078	0.0799	1.0243	89.85
Percomorpha	Igsf	*O.latipes*	*G.aculeatus*	0.0576	0.0787	1.3672	119.93
Percomorpha	PIK3	*O.niloticus*	*G.aculeatus*	0.0912	0.0568	0.623	54.65
Percomorpha	PIK3	*O.niloticus*	*O.latipes*	0.0712	0.07	0.9831	86.24
Percomorpha	PIK3	*O.niloticus*	*D.labrax*	0.0798	0.0409	0.5134	45.04
Percomorpha	PIK3	*O.niloticus*	*E.maclovinus*	0.0898	0.0547	0.6093	53.45

Node	Gene	Species 1	Species 2	dnds	dn	ds	Divergence Time
Percomorpha	PIK3	*O.niloticus*	*D.mawsoni*	0.086	0.0583	0.6779	59.46
Percomorpha	PIK3	*O.niloticus*	*N.coriiceps*	0.0854	0.0584	0.6831	59.92
Percomorpha	PIK3	*O.latipes*	*D.labrax*	0.0673	0.0631	0.9376	82.25
Percomorpha	PIK3	*O.latipes*	*E.maclovinus*	0.0708	0.0733	1.035	90.79
Percomorpha	PIK3	*O.latipes*	*G.aculeatus*	0.0738	0.0774	1.0483	91.96
Percomorpha	PIK3	*O.latipes*	*D.mawsoni*	0.0725	0.0765	1.0552	92.56
Percomorpha	PIK3	*O.latipes*	*N.coriiceps*	0.0699	0.0765	1.0946	96.02
Percomorpha	Sema	*O.niloticus*	*D.labrax*	0.0726	0.0351	0.4834	42.40
Percomorpha	Sema	*O.niloticus*	*D.mawsoni*	0.0668	0.0365	0.5466	47.95
Percomorpha	Sema	*O.niloticus*	*E.maclovinus*	0.0729	0.0416	0.5705	50.04
Percomorpha	Sema	*O.niloticus*	*N.coriiceps*	0.0973	0.0608	0.6246	54.79
Percomorpha	Sema	*O.niloticus*	*G.aculeatus*	0.0529	0.0418	0.7904	69.33
Percomorpha	Sema	*O.niloticus*	*O.latipes*	0.0624	0.0522	0.8367	73.39
Percomorpha	Sema	*O.latipes*	*D.labrax*	0.0701	0.0539	0.7691	67.46
Percomorpha	Sema	*O.latipes*	*E.maclovinus*	0.0626	0.0517	0.8259	72.45
Percomorpha	Sema	*O.latipes*	*D.mawsoni*	0.0599	0.052	0.8679	76.13
Percomorpha	Sema	*O.latipes*	*N.coriiceps*	0.0809	0.0766	0.9473	83.10
Percomorpha	Sema	*O.latipes*	*G.aculeatus*	0.0545	0.0562	1.0322	90.54
Percomorpha	TLR	*O.latipes*	*D.labrax*	0.2033	0.216	1.0623	93.18
Percomorpha	TLR	*O.latipes*	*D.mawsoni*	0.2116	0.2345	1.1085	97.24
Percomorpha	TLR	*O.latipes*	*N.coriiceps*	0.2174	0.2421	1.1134	97.67
Percomorpha	TLR	*O.latipes*	*E.maclovinus*	0.2052	0.2309	1.1249	98.68
Percomorpha	TLR	*O.latipes*	*G.aculeatus*	0.155	0.2291	1.4781	129.66
Percomorpha	TLR	*O.niloticus*	*O.latipes*	0.2161	0.2231	1.0323	90.55
Percomorpha	TLR	*O.niloticus*	*D.labrax*	0.3533	0.1695	0.4797	42.08
Percomorpha	TLR	*O.niloticus*	*D.mawsoni*	0.3053	0.1914	0.6267	54.97
Percomorpha	TLR	*O.niloticus*	*N.coriiceps*	0.3139	0.1987	0.6329	55.52
Percomorpha	TLR	*O.niloticus*	*E.maclovinus*	0.3055	0.1961	0.6419	56.31
Percomorpha	TLR	*O.niloticus*	*G.aculeatus*	0.2338	0.1878	0.8033	70.46
Perciformes	AKT3	*G.aculeatus*	*D.mawsoni*	0.0132	0.0061	0.4634	40.65
Perciformes	AKT3	*G.aculeatus*	*N.coriiceps*	0.0297	0.0143	0.4807	42.17
Perciformes	AKT3	*G.aculeatus*	*E.maclovinus*	0.0142	0.0072	0.5073	44.50
Perciformes	Igsf	*D.labrax*	*G.aculeatus*	0.0544	0.0415	0.7621	66.85
Perciformes	Igsf	*D.labrax*	*D.mawsoni*	0.0905	0.0363	0.4012	35.19
Perciformes	Igsf	*D.labrax*	*N.coriiceps*	0.1063	0.0446	0.419	36.75
Perciformes	Igsf	*D.labrax*	*E.maclovinus*	0.0877	0.0383	0.4364	38.28
Perciformes	Igsf	*G.aculeatus*	*D.mawsoni*	0.0598	0.0464	0.7768	68.14
Perciformes	Igsf	*G.aculeatus*	*N.coriiceps*	0.068	0.0542	0.797	69.91
Perciformes	Igsf	*G.aculeatus*	*E.maclovinus*	0.0524	0.0463	0.8834	77.49
Perciformes	PIK3	*D.labrax*	*G.aculeatus*	0.0983	0.0429	0.4361	38.25
Perciformes	PIK3	*D.labrax*	*E.maclovinus*	0.0951	0.0388	0.4085	35.83
Perciformes	PIK3	*D.labrax*	*D.mawsoni*	0.0945	0.0419	0.4434	38.89
Perciformes	PIK3	*D.labrax*	*N.coriiceps*	0.0941	0.0425	0.4515	39.61

Node	Gene	Species 1	Species 2	dnds	dn	ds	Divergence Time
Perciformes	PIK3	*G.aculeatus*	*E.maclovinus*	0.0921	0.0448	0.4863	42.66
Perciformes	PIK3	*G.aculeatus*	*D.mawsoni*	0.0911	0.0484	0.531	46.58
Perciformes	PIK3	*G.aculeatus*	*N.coriiceps*	0.087	0.0475	0.5466	47.95
Perciformes	Sema	*D.labrax*	*E.maclovinus*	0.0916	0.0355	0.3871	33.96
Perciformes	Sema	*D.labrax*	*N.coriiceps*	0.1358	0.0557	0.4101	35.97
Perciformes	Sema	*D.labrax*	*D.mawsoni*	0.0908	0.0314	0.3455	30.31
Perciformes	Sema	*D.labrax*	*G.aculeatus*	0.0741	0.0378	0.51	44.74
Perciformes	Sema	*G.aculeatus*	*D.mawsoni*	0.0497	0.0281	0.5659	49.64
Perciformes	Sema	*G.aculeatus*	*E.maclovinus*	0.0555	0.0342	0.6152	53.96
Perciformes	Sema	*G.aculeatus*	*N.coriiceps*	0.0803	0.0524	0.6523	57.22
Perciformes	TLR	*D.labrax*	*D.mawsoni*	0.4301	0.1512	0.3514	30.82
Perciformes	TLR	*D.labrax*	*N.coriiceps*	0.4279	0.1564	0.3654	32.05
Perciformes	TLR	*D.labrax*	*E.maclovinus*	0.401	0.1502	0.3745	32.85
Perciformes	TLR	*D.labrax*	*G.aculeatus*	0.2812	0.135	0.48	42.11
Perciformes	TLR	*G.aculeatus*	*D.mawsoni*	0.2717	0.1522	0.5602	49.14
Perciformes	TLR	*G.aculeatus*	*N.coriiceps*	0.2717	0.1561	0.5745	50.39
Perciformes	TLR	*G.aculeatus*	*E.maclovinus*	0.2626	0.1527	0.5816	51.02
Notothenioidei	AKT3	*E.maclovinus*	*D.mawsoni*	0.0241	0.0032	0.1316	11.54
Notothenioidei	AKT3	*E.maclovinus*	*N.coriiceps*	0.0726	0.0117	0.1607	14.10
Notothenioidei	Igsf	*E.maclovinus*	*D.mawsoni*	0.0932	0.0172	0.1843	16.17
Notothenioidei	Igsf	*E.maclovinus*	*N.coriiceps*	0.1239	0.0253	0.2044	17.93
Notothenioidei	PIK3	*E.maclovinus*	*D.mawsoni*	0.1175	0.0216	0.1842	16.16
Notothenioidei	PIK3	*E.maclovinus*	*N.coriiceps*	0.1026	0.0198	0.1927	16.90
Notothenioidei	Sema	*E.maclovinus*	*D.mawsoni*	0.0948	0.0153	0.161	14.12
Notothenioidei	Sema	*E.maclovinus*	*N.coriiceps*	0.1816	0.0392	0.2157	18.92
Notothenioidei	TLR	*E.maclovinus*	*D.mawsoni*	0.3028	0.0487	0.1608	14.11
Notothenioidei	TLR	*E.maclovinus*	*N.coriiceps*	0.3405	0.0554	0.1626	14.26
Nototheniidae	AKT3	*N.coriiceps*	*D.mawsoni*	0.1166	0.0082	0.0703	6.17
Nototheniidae	Igsf	*N.coriiceps*	*D.mawsoni*	0.2413	0.0119	0.0492	4.32
Nototheniidae	PIK3	*N.coriiceps*	*D.mawsoni*	0.1886	0.0086	0.0456	4.00
Nototheniidae	Sema	*N.coriiceps*	*D.mawsoni*	0.34	0.0275	0.081	7.11
Nototheniidae	TLR	*N.coriiceps*	*D.mawsoni*	1.0631	0.0302	0.0284	2.49

Table *SI*.5: Mean divergence time estimates, upper and lower confidence limits in millions of years for seven nodes, calculated for the five gene families, Toll-Like receptor (TLR), AKT serine/threonine kinase 3 (AKT3), Phosphatidylinositol 3-kinase (PIK3), Immunoglobulin Superfamily (Igsf) and Semaphorins (Sema).

Node	Gene	Mean	Upper confidence limit	Lower confidence limit
Neopterygii	Sema	300.15	314.78	285.52
Acanthomorphata	AKT3	106.39	110.59	102.20
Acanthomorphata	Igsf	246.30	267.73	224.87
Acanthomorphata	PIK3	219.30	245.47	193.12
Acanthomorphata	Sema	201.25	215.60	186.90
Acanthomorphata	TLR	150.30	162.75	137.85
Teleostei	AKT3	153.82	163.27	144.37
Teleostei	Igsf	221.02	247.21	194.82
Teleostei	PIK3	250.21	260.71	239.71
Teleostei	Sema	293.48	314.23	272.72
Percomorpha	AKT3	58.90	74.54	43.26
Percomorpha	Igsf	73.70	83.25	64.16
Percomorpha	PIK3	73.85	82.17	65.53
Percomorpha	Sema	66.15	72.81	59.48
Percomorpha	TLR	80.57	92.04	69.11
Perciformes	AKT3	42.44	44.26	40.62
Perciformes	Igsf	56.09	66.31	45.87
Perciformes	PIK3	41.40	43.89	38.90
Perciformes	Sema	43.69	49.50	37.87
Perciformes	TLR	41.20	46.29	36.10
Notothenioidei	AKT3	12.82	15.17	10.48
Notothenioidei	Igsf	17.05	18.67	15.43
Notothenioidei	PIK3	16.53	17.22	15.85
Notothenioidei	Sema	16.52	20.93	12.11
Notothenioidei	TLR	14.18	14.33	14.04
Nototheniidae	AKT3	6.17	6.17	6.17
Nototheniidae	Igsf	4.32	4.32	4.32
Nototheniidae	PIK3	4.00	4.00	4.00
Nototheniidae	Sema	7.11	7.11	7.11
Nototheniidae	TLR	2.49	2.49	2.49

Script 1

```
from Bio import SeqIO
import re

seqin = open()
TLRout = open('', 'w')
PIK3out = open('','w')
AKT3out = open('', 'w')
MHCout = open('', 'w')
Igout = open('','w')

record = re.findall(r"(>.+\n[ABCDEFGHIJKLMNOPQRSTUVWXYZ\n]+)",seqin.read())

for seq in record[:]:
    seq2 = seq.split("\n")
    seqid = seq2[0]
    sequence = "".join(seq2[1:])
    print(seqid)
    print(sequence)
    if "tlr" in seq.split("\n")[0].lower() or "toll-like receptor" in seq.split("\n")[0].lower():
        if not "related" in seq.split("\n")[0].lower():
            # print(seq.split("\n")[0].lower())

            TLRout.write(seqid + '\n' + sequence + '\n')

    if "pik3" in seq.split("\n")[0].lower() or "phosphatidylinositol 3-kinase" in seq.split("\n")[
        0].lower() or "phosphoinositide 3-kinase" in seq.split("\n")[0].lower():
        if not "related" in seq.split("\n")[0].lower():
            # print(seq.split("\n")[0].lower())

            PIK3out.write(seqid + '\n' + sequence + '\n')

    if "akt3" in seq.split("\n")[0].lower() or "RAC-gamma" in seq.split(
            "\n")[0] or "v-akt murine thymoma viral oncogene homolog 3" in seq.split("\n")[0]:
        if not "related" in seq.split("\n")[0]:
            # print(seq.split("\n")[0].lower())

            AKT3out.write(seqid + '\n' + sequence + '\n')

    if "mhc" in seq.split("\n")[0].lower() or "major histocompatibility complex" in seq.split("\n")[0].lower():
        if not ("vmhc" in seq.split("\n")[0].lower() or "ventricular myosin heavy chain-like" in seq.split("\n")[0].lower()):
            if not "related" in seq.split("\n")[0].lower():
                # print(seq.split("\n")[0].lower())

                MHCout.write(seqid + '\n' + sequence + '\n')
```

```
    if "immunoglobulin" in seq.split("\n")[0].lower():
        if not "related" in seq.split("\n")[0].lower():
            Igout.write(seqid + '\n' + sequence + '\n')
```

Script 2

```
from Bio import SeqIO
import re
import os,csv


###First concatenate all the fasta files of the various genes into one file for each species
DreDir= ()
list_of_files=[DreDir + fasta for fasta in os.listdir(DreDir) if fasta.endswith("_Dre_Ensembl.fa") ]

with open('Prot_Comb_Dre.fasta', 'w') as w_file:
    for filen in list_of_files:
        with open(filen, 'rU') as o_file:
            seq_records = SeqIO.parse(o_file, 'fasta')
            for seq_record in seq_records:
                w_file.write(">"+ str(seq_record.description) + "\n"+str(seq_record.seq)+"\n")
            #SeqIO.write(seq_records, w_file, 'fasta')


### From here starts the real cleaning of the fasta file

seqin = open(‘’)
fin = re.findall(r'(>.+\n.+\n)', seqin.read())

### Create empty Dictionary
d = {}
### Start of iteration
for records in fin:
    record = records.split('\n')
    seqid = record[0]
    sequence = record[1]

### Next genID and length are added to dictionary
    gene = str(re.search(r'gene:(\w+\.+\w+)',seqid).group(1))
    seqlen = len(sequence)

    putindict = [seqid,str(seqlen),sequence]
    if not gene in d:
        d[gene] = putindict
    else:
        if int(d[gene][1]) < seqlen:
            d[gene] = putindict


print(len(d))
```

```
fout = open('Trscript_Dre.fa','w')

for seq in d:

    x = d[seq][0]+'\n'+d[seq][2]+'\n'
    fout.write(x)

### Delete similar gene symbols###


seqin = open('')
fin = re.findall(r'(>.+\n.+\n)', seqin.read())

### Create empty Dictionary
d = {}
### Start of iteration
for records in fin:
    record = records.split('\n')
    seqid = record[0]
    sequence = record[1]

    ### Next genID and length are added to dictionary
    gene = str(re.search(r'gene_symbol:([\s\S]+) \[',seqid).group(1))
    seqlen = len(sequence)

    putindict = [seqid,str(seqlen),sequence]
    if not gene in d:
        d[gene] = putindict
    else:
        if int(d[gene][1]) < seqlen:
            d[gene] = putindict


print(len(d))
fout = open('Proteome_Dre.fa','w')

for seq in d:

    x = d[seq][0]+'\n'+d[seq][2]+'\n'
    fout.write(x)
```