# The whale shark genome reveals patterns of vertebrate gene family evolution

Milton Tan[1]*, Anthony K. Redmond[2], Helen Dooley[3], Ryo Nozu[4], Keiichi Sato[4,5], Shigehiro Kuraku[6], Sergey Koren[7], Adam M. Phillippy[7], Alistair D.M. Dove[8], Timothy D. Read[9]

1. Illinois Natural History Survey at University of Illinois Urbana-Champaign, Champaign, IL, USA.
2. Smurfit Institute of Genetics, Trinity College Dublin, Dublin, Ireland.
3. University of Maryland School of Medicine, Institute of Marine & Environmental Technology, Baltimore, MD, USA.
4. Okinawa Churashima Research Center, Okinawa Churashima Foundation, Okinawa, Japan.
5. Okinawa Churaumi Aquarium, Motobu, Okinawa, Japan
6. RIKEN Center for Biosystems Dynamics Research (BDR), RIKEN, Kobe, Japan.
7. National Human Genome Research Institute, Bethesda, MD, USA.
8. Georgia Aquarium, Atlanta, GA, USA.
9. Department of Infectious Diseases, Emory University School of Medicine, Atlanta, GA, USA.

Keywords: fish, Elasmobranchii, Chondrichthyes, Cartilaginous Fishes, Gnathostomata, comparative genomics, innate immunity, gigantism

## Abstract

Due to their key phylogenetic position, cartilaginous fishes, which includes the largest fish species *Rhincodon typus* (whale shark), are an important vertebrate lineage for understanding the origin and evolution of vertebrates. However, until recently, this lineage has been understudied in vertebrate genomics. Using newly-generated long read sequences, we produced the best gapless cartilaginous fish genome assembly to date. The assembly has fewer missing ancestral genes than *Callorhinchus milii*, which has been widely-used for evolutionary studies up to now. We used the new assembly to study the evolution of gene families in the whale shark and other vertebrates, focusing on historical patterns of gene family origins and loss across early vertebrate evolution, innate immune receptor repertoire evolution, and dynamics of gene family evolution size in relation to gigantism. From inferring the pattern of origin of gene families across the most recent common ancestors of major vertebrate clades, we found that there were many shared gene families between the whale shark and bony vertebrates that were present in the most recent common ancestor of jawed vertebrates, with a large increase in novel genes at the origin of jawed vertebrates independent of whole genome duplication events. The innate immune system in the whale shark, which consisted of diverse pathogen recognition receptors (PRRs) including NOD-like receptors, RIG-like receptors, and Toll-like receptors. We discovered a unique complement of Toll-like receptors and triplication of

NOD1 in the whale shark genome. Further, we found diverse patterns of gene family evolution between PRRs within vertebrates demonstrating that the origin of adaptive immunity in jawed vertebrates is more complicated than simply replacing the need for a vast repertoire of germline encoded PRRs. We then studied rates of amino acid substitution and gene family size evolution across origins of vertebrate gigantism. While we found that cartilaginous fishes and giant vertebrates tended to have slower substitution rates than the background rate in vertebrates, the whale shark genome substitution rate was not significantly slower than *Callorhinchus*. Furthermore, rates of gene family size evolution varied among giants and the background, suggesting that differences in rate of substitution and gene family size evolution relative to gigantism are decoupled. We found that the gene families that have shifted in duplication rate in whale shark are enriched for genes related to driving cancer in humans, consistent with studies in other giant vertebrates than support the hypothesis that evolution of increased body size requires adaptations that result in reduction of per cell cancer rate.

# Introduction

Jawed vertebrates (Gnathostomata) comprise two extant major groups, the cartilaginous fishes (Chondrichthyes) and the bony vertebrates (Osteichthyes, including Tetrapoda)[1]. Comparison of genomes between these two groups not only provides insight into early jawed vertebrate evolution and the emergence of various biological features, but also enables inference of ancestral jawed vertebrate traits[2]. While the availability of sequence data from many species across phylogeny is key to the success of such analyses, until very recently genomic data from cartilaginous fishes was significantly underrepresented compared to other vertebrate linages. The first cartilaginous fish genome, that of *Callorhinchus milii* (known colloquially as ghost shark, elephant shark, or elephant fish), was used to study the early evolution of genes related to bone development and emergence of the adaptive immune system[2]. However, concerns were raised that *Callorhinchus*, being a member of the Holocephali (chimaeras, ratfishes) – one of the two major groups of cartilaginous fishes – and separated from the Elasmobranchii (sharks, rays, and skates) for approximately 420 million years[3] may not be representative of cartilaginous fishes as a group[4].

Until recently, only scant genetic resources have been available for the whale shark[5] specifically, and elasmobranchs in general. The first elasmobranch genome published was for a male whale shark ("Norton") by Read *et al.*[6]. Famously representing one of Earth's ocean giants, the whale shark is by far the largest of all fishes reaching a maximum confirmed length of nearly 19 meters in length[7]. Despite their large body size, endangered status, and appeal to both scientists and the general public, relatively little is known about their biology[8–10]. Due to its phylogenetic position among vertebrates, the scarcity of shark genomes, and its unique biology, the whale shark genome can be used to address questions related to vertebrate genome evolution[11,12], the relationship of gene evolution in sharks and unique shark traits[11,12], as well as the evolution of gigantism[13]. As expected, Read *et al.* found that whale shark genome was most closely related among sequenced fish genomes (at the time) to *Callorhinchus*. They also discovered a TLR similar to TLR21 in the whale shark thus suggesting that TLR21 was derived in the most recent common ancestor of jawed vertebrates. While this represented an important

step forward for elasmobranch genomics, the genome was fragmentary, and substantial improvements to the genome contiguity and annotation were expected from reassembling the genome using PacBio long-read sequences[6].

More recent work has focused upon further sequencing, assembling, and analyzing of the whale shark nuclear genome[6,11,13]. Hara *et al.*[11] generated novel elasmobranch genome assemblies for the brown-banded bamboo shark (*Chiloscyllium punctatum*) and cloudy catshark (*Scyliorhinus torazame*). They also reassembled the published Read *et al.* whale shark genome data, supplementing it with transcriptome sequence from blood cells sampled from a different male individual. Their comparative genomic study confirmed a reduced evolutionary rate in cartilaginous fish genomes[11], a conclusion previously drawn from only a handful of genes[14,15]. Park *et al.*[13] then reported on a whale shark genome assembly from a third male individual and compared genomic characteristics from animals to yeast to identify correlates of longevity and genome traits. They found that much of the increase in the gene length in the whale shark genome was due to relatively large introns that were enriched for CR1-like LINEs. Finally, in a white shark genome paper, Marra *et al.*[12] studied rates of positive selection across protein-coding genes in the white shark and whale shark genomes. Like Park *et al.*, they identified an abundance of LINEs in both the white shark and whale shark genomes, proposing this characteristic contributes to genomic stability. They also confirmed the reduced olfactory receptor diversity in cartilaginous fishes first reported by Hara *et al*.[11], and determined genes related to genome integrity and ubiquitination, as well some related to wound healing, were evolving under positive selection in both shark species.

Long read sequencing is an important factor in assembling longer contigs to resolve repetitive regions[16,17], which comprise the majority of vertebrate genomes. Herein we report on the best gapless assembly of the whale shark genome thus far, based on *de novo* assembly of long reads obtained with the PacBio single molecule real-time sequencing platform. We used this assembly and new annotation in a comparative genomic approach to investigate the origins and losses of gene families to identify patterns of gene family evolution associated with major early vertebrate evolution. We also focused on gene family evolution of innate immune pathogen recognition receptor (PRR) types. Immune protection in vertebrates is usually classified into two types; innate and adaptive. The innate system is evolutionary ancient and based upon germline-encoded receptors, known as pattern recognition receptors (PRRs). By contrast, the adaptive system, found in vertebrates alone, overlays the innate system and is based upon somatically rearranging receptors (variable lymphocyte receptors in jawless vertebrates; immunoglobulins and T cell receptors in jawed vertebrates). Greatly expanded PRR repertoires have been reported in several deuterostome invertebrate genomes compared to conserved repertoires in their jawed vertebrates homologs, leading to the proposal that the acquisition of adaptive immunity negated the need for vast PRR repertoires, or somehow constrained PRR expansion[18,19]. Despite cartilaginous fishes, including whale shark, providing a good model to test this by virtue of being the oldest extant jawed vertebrate lineage relative to mammals to possess both innate and adaptive systems[20,21], genomic investigations into shark immunity have been limited by the paucity of genomic resources[2,4]. Using the new whale shark genome assembly, we investigated the repertoires of three major PRR families: NOD-like receptors, RIG-like receptors, and Toll-like receptors. Finally, we compared the rates of

functional genomic evolution in multiple independent lineages of vertebrates that have experienced gigantism, including the whale shark, to test for relationships between gigantism and genomic evolution among vertebrates. Further, larger-bodied organisms tend to have lower cancer rates than expected given their increased numbers of cells relative to smaller-bodied organisms[22], suggesting genes involved in cancer suppression may evolve differently in vertebrate giants. Supporting this hypothesis, recent research in giant mammals such as elephants and whales have identified selection or duplication of various gene families that are related to causing cancer in humans[23–25], supporting the role of evolution in cancer-related genes in the evolution of gigantism. Hence, we studied whether gene families that have shifted in gene duplication rates were enriched for orthologs of known cancer genes.

# Results and Discussion

**Gapless Genome Assembly.** The new "Norton" whale shark genome assembly represents the best gapless assembly to date for the whale shark (Supplementary Note 1; Supplementary Table 1). The total length of the contigs of the new assembly was 2.93 Gbp. This number is smaller than a non-sequencing based estimate of the whale shark genome size of 3.73 Gbp by Hara *et al.*[11], which suggests that sections of the genome, potentially comprising primarily repetitive elements, are still missing. The new assembly had 57,333 contigs with a contig N50 of 144,422 bp. The assembly had fewer contigs than the number of scaffolds of previous assemblies, and a higher contig N50, representing a dramatic improvement in contiguity compared to the existing whale shark genome assemblies (Supplementary Table 1). This higher contiguity at the contig level (vs. scaffold level) was also better than the published *Callorhinchus*, brownbanded bamboo shark, cloudy catshark, and white shark genomes[2,11,12]. Although the BUSCO score of *Callorhinchus* is higher than in whale shark (which is expected as *Callorhinchus* was used in selecting the conserved ortholog set for BUSCO), when considering gene families across vertebrates rather than simply the set in BUSCO (See Ancestral Vertebrate Genome Evolution below and Supplementary Note 2), we infer more gene families are missing from *Callorhinchus* than the whale shark, suggesting that either the whale shark has a greater conservation of the gene families from the ancestral cartilaginous fish than *Callorhinchus*, or that the whale shark has fewer missing gene families in its assembly than *Callorhinchus*. When evaluating gene completeness using a rigid 1-to-1 ortholog core vertebrate gene set[26] (implemented in gVolante[27]), we found 85% core vertebrate genes were complete and found 97.4% core vertebrate genes included partial genes, which compares favorably to other sharks[11]. Thus, the gene content of the whale shark is quite complete for a cartilaginous fish genome and will be informative for questions regarding vertebrate gene evolution.

**Ancestral Vertebrate Genome Evolution.** We sought to use the new whale shark genome assembly to infer the evolution of gene families (i.e. orthogroups) of protein-coding genes across vertebrates. So-called orthogroups are a unit of comparison in comparative genomics defined as all genes descended from a single gene in the common ancestor of species considered[28]; hence, they are dependent on the phylogenetic breadth of species included. We inferred the origin of gene families in vertebrate evolution and compared the ancestral genomes

at major nodes in vertebrate phylogeny. This has relevance for the origin of biological functions in vertebrate evolution and informs the generality of genetic studies across taxa. Furthermore, inferred gene family losses may also relate to functional genomic evolution of lineages that inherit those losses[29]. We determined chordate-level gene families from proteomes for 24 chordates, including 22 representative vertebrates, sea squirt (*Ciona*), and lancelet (*Branchiostoma*), and inferred the history of gene family origin and loss by comparing the presence and absence of gene families across species. The lancelet *Branchiostoma* has some of the highest retention of gene families from both chordate and pre-animal ancestors[30,31], supporting its use as an outgroup for understanding the origins of novel genes within the vertebrate clade relative to other animals.

When comparing the number of gene families within the ancestral genomes of major chordate clades, we found a consistent increase in the total number of gene families from the most recent common ancestor (MRCA) of chordates to the MRCA of Gnathostomata (e.g. 10,255 gene families in the MRCA of Olfactores (Urochordata + Vertebrata) to 12,815 in the MRCA of Gnathostomata; Figure 1, Supplementary Note 2, Supplementary Figure 3). After the peak in the MRCA of Gnathostoma, there was a further increase in number of gene families to the ancestor of bony vertebrates and a decrease to the ancestor of cartilaginous fishes. In addition, there was an increase in the number of gene families conserved in all descendants of each ancestor; we inferred 3,291 gene families in the MRCA of Olfactores were conserved in all Olfactores studied, but 6,022 gene families in the MRCA of Gnathostomata were conserved in all gnathostomes studied. While marginally more gene families are conserved across all bony vertebrates (6,187 gene families), far more gene families are conserved between both cartilaginous fishes studied (10,631 gene families), which is likely a relatively high number due to only considering two species. We also found an increase in the number of novel gene families (989 in the MRCA of Olfactores to 2,106 in the MRCA of Gnathostomata). We found a variable number of novel gene families in vertebrate ancestors were retained in all descendants, with 294 retained from the MRCA of Olfactores, 259 retained from the MRCA of Vertebrata, and 379 retained from the MRCA of Gnathostomata; however, here, the Gnathostomata still had the largest number of conserved novel gene families. The MRCAs of Osteichthyes and Chondrichthyes both had larger numbers of conserved gene families and far lower numbers of novel gene families than earlier vertebrate ancestors (255 novel gene families in bony vertebrates, 124 novel gene families in cartilaginous fishes). Thus, the ancestor of jawed vertebrates had more novel gene families, and more novel gene families were retained from the jawed vertebrate ancestor in all descendants, relative to earlier ancestors in chordate history. Overall, this implies the origin of jawed vertebrates established a large proportion of novel gene families of both bony vertebrates and cartilaginous fishes.

Many studies have found evidence that two rounds (2R) of whole genome duplication occurred early in vertebrate evolution, resulting in gene duplicates referred to as ohnologs[32–35]. Gene copies arising from genome duplication should belong to the same gene family, but it is possible that the large number of novel gene families we found at the ancestor of jawed vertebrates was artefactually inflated by ohnologs split across multiple gene families. To estimate the extent of this potential oversplitting, we compared the 2,106 gene families inferred as novel at the base of jawed vertebrates to a strict set (1,381) and a relaxed set (2,642) of

ohnolog families previously determined using a synteny-aware method by Singh *et al.*[34]. We found that between 140 (strict) and 460 (relaxed) of the novel jawed vertebrate gene families included any human genes assigned to any ohnolog families identified by Singh *et al.*, indicating a minority of the gene families we identified corresponded to duplicated gene families. Furthermore, we found much agreement between how novel jawed vertebrate genes were assigned to gene families or to ohnologs; for the strict set, 110 ohnolog families matched to gene families we determined, while for the relaxed set, 288 ohnolog families matched to gene families we determined. There were relatively few discrepancies, but they tended to be ohnolog families split into multiple gene families. 13 strict ohnolog families corresponded to 30 gene families, and 67 relaxed ohnolog families corresponded to 153 gene families. Rarely, our gene families clustered multiple ohnologs as single gene families, although this is not necessarily erroneous, and may be due to our greater taxon sampling than used by Singh *et al*. In summary, this supports that only a small proportion of the gene families we identified are potentially spuriously split ohnologs. This reinforces the importance of the origin of jawed vertebrates for genomic novelty, not just for the vertebrate 2R whole genome duplication, but additional novel gene families as well.

We then tested whether gene families that were gained or lost during vertebrate evolution were enriched for certain GO, IPR, or Pfam annotations (Supplementary Note 2), which might indicate functional genomic shifts preceding the origin of these clades. We found an enrichment of connexin function in gene families that originated at the MRCA of Olfactores (Supplementary Table 5). These intercellular channels act as part of gap junctions[36], and this enrichment of novel connexin function in the MRCA of Olfactores is consistent with the identification of connexins in sea squirt that are orthologous to vertebrate connexins[37], their absence in lancelet[38], and the origin of intercellular signaling function in the MRCA of Olfactores. In both the MRCA of vertebrates and jawed vertebrates, we consistently found enrichment in novel genes for G-protein coupled receptors and immunoglobulin domains (Supplementary Tables 6–7). We also found enrichment for hormone activity but only in the MRCA of jawed vertebrates (Supplementary Tables 7). This indicates a relatively large gain of signaling molecules in the origins of these clades, implying increasing sophistication required increased regulation. Hence, not only did many hormones with a role in mammal homeostasis originate in the MRCA of jawed vertebrates – as previously shown in comparisons of shark genomes to other vertebrates[11] – but signaling was a dominant function of the novel gene families that originating in the MRCA of all vertebrates as well as jawed vertebrates. Additionally, in the MRCA of jawed vertebrates, we found enrichment for immune response and of various immune-related domains (Supplementary Table 7), consistent with the origin of mammalian-like adaptive immunity in jawed vertebrates.

Differences between bony vertebrates and cartilaginous fishes might arise through the function of gene families specific to each lineage. We found no enrichment for function or domain terms among the 124 gene families derived in the MRCA of cartilaginous fishes. By contrast, the 255 gene families derived in the MRCA of bony vertebrates were enriched for G-protein coupled receptor domains, immunoglobulin domains, and olfactory receptor domains (Supplementary Table 8), which may indicate further overrepresentation for novelty in gene families functioning in signaling in bony vertebrates, but not in cartilaginous fishes

(Supplementary Table 9). The addition of novel G-protein coupled receptor domain proteins in bony vertebrates is consistent with the relative paucity of these receptors noted in previous comparisons of cartilaginous fishes and bony vertebrates[12]. However, it should be noted that the lack of enrichment for functions among gene families that were novel in cartilaginous fishes may also be explained by the difference in the proportion of genes with annotations: only 86 of the 124 (69.4%) gene families that were novel to cartilaginous fishes were annotated, while 232 of the 255 (91.0%) gene families novel to bony vertebrates were annotated, indicating that the novel gene families in cartilaginous fishes were less likely to possess known protein domains or functions.

Biological differences between bony vertebrates and cartilaginous fishes might also be explained by differences in lost gene functions relative to the jawed vertebrate ancestor. We found no enrichment for functional terms for the 145 gene families lost in bony vertebrates, but the 729 gene families lost in cartilaginous fishes were enriched for a variety of domains. Even though we only included two cartilaginous fishes, we found olfactory receptor function was prominent among the gene families lost within cartilaginous fishes, which is also consistent with the low number of olfactory receptors noted in cartilaginous fishes previously[11,12,39]. We also found enrichment of losses of gene families associated with certain transposable elements including DDE superfamily endonuclease domains and L1 family transposase. The enrichment for a loss of L1 transposase domain in cartilaginous fishes is consistent with the absence of full L1 transposons that we noted in whale shark (Supplementary Note 1B) and noted previously for *Callorhinchus*[40]. Specific to the evolution of the whale shark, relative to the cartilaginous fish ancestor, we inferred 13 novel gene families and a loss of 554 gene families. Neither of these sets of gene families were enriched for any functional terms relative to jawed vertebrate genes overall. However, as sequencing effort and taxon sampling varies among these clades, it is possible that the lack of functional enrichment in losses may have been due to random loss due to variable levels of genome sequencing completeness. On the other hand, the lack of functional enrichment in groups of gene families lost between elasmobranchs, holocephalans, and bony vertebrates could also suggest that the biological differences evolved between these groups were not driven by differential gene family loss.

Thus, our analyses imply a dynamic history of gene family gain and loss across early vertebrate evolution. Of particular importance is the number of gene families gained in the MRCA of jawed vertebrates in establishing the gene families present in bony vertebrates and cartilaginous fishes. Although increasing numbers of gene families originated over time, many of these novel gene families among vertebrates, jawed vertebrates, and bony vertebrates were not conserved across all descendants, suggesting pervasive loss of novel genes as well. We also found enrichment for certain functions among these gene families, specifically G-protein coupled receptors, immunoglobulin domains, hormones, and olfactory receptors. These analyses demonstrate the ancestry of many of the major genomic functions in bony vertebrates preceded their origin, as illuminated by these gene families being shared with their cartilaginous fish sister group. The whale shark genome therefore provides an important additional resource to study the origins of gene families in vertebrates.

**Evolution of innate immunity in the whale shark**

Using the new whale shark genome assembly, we investigated the repertoires of three major PRR families: NOD-like receptors (NLRs), RIG-like receptors (RLRs), and Toll-like receptors (TLRs). Our analyses (see Methods; Supplementary Note 3) allowed us to explore ancestral vertebrate and jawed vertebrate PRR diversity, while searching for evidence of vertebrate PRR expansions. NOD-like receptors (NLRs) are intracellular receptors that detect a wide array of PAMPs (Pathogen associated molecular pattern) and DAMPs (Damage/Danger associated molecular pattern)[41–45]. Typified by the presence of a NACHT domain (although most also possess C-terminal leucine rich-repeats), NLRs are further divided into three subfamilies; the NODs, NLRPs (NALPs), and IPAF. We identified 43 NLR sequences, 23 of which contained a clearly-identifiable NACHT domain (Supplementary Table 11, Figure 2, Supplementary Figure 3). Unusually, we found three copies of NOD1 in whale shark (ultrafast bootstrap support, UFBOOT=100; Figure 2), all containing NACHT domains and occupying unique genomic locations. We hypothesize that this potentiates broader bacterial recognition or more nuanced responses to different pathogens. NLRPs are vital for the induction of inflammatory responses (through activation of a multi-protein complex called the inflammasome)[46], yet strikingly, we identified only a single NLRP-like sequence (lacking a detectable NACHT domain) in whale shark. Interestingly, we also identified a novel NLR gene family that appears to be closely related to the NLRPs (UFBOOT=67) in zebrafish and whale shark (UFBOOT=92), that has undergone significant expansion in the latter (whale shark clade UFBOOT=100; Figure 2, Supplementary Figure 3). It seems likely that this large repertoire of NLRP-related genes compensates for the paucity of true NLRPs in whale shark, thereby offering alternative routes for inflammasome activation. Interestingly, each of the vertebrate species examined have independently expanded a different NLR family, with NLRP genes expanded in human (clade UFBOOT=99; Supplementary Figure 3), and the 'fish-specific' FISNA in zebrafish (clade UFBOOT=86; Supplementary Figure 3), for which we surprisingly found a whale shark ortholog (UFBOOT=74). This suggests that NLR repertoire expansions may be as common in vertebrates as in deuterostome invertebrates, contradicting the idea that the presence of an adaptive immune system influences this.

RIG-like receptors (RLRs) are cytosolic PRRs that detect viral nucleic acid and initiate immune responses through Mitochondrial Antiviral Signaling Protein (MAVS)[47,48]. Bony vertebrates have three RLR proteins; RIG-1, MDA5, and LGP2. All contain a DEAD-helicase domain, a C-terminal RNA recognition domain, and an N-terminal CARD domain pair that mediate their interaction with MAVS[47,48]. We identified four candidate RLR genes in the whale shark; two non-overlapping, partial MDA5-like sequences, and one each for RIG-1 and LGP2 (Supplementary Table 11). Phylogenetic reconstruction of DEAD-Helicase domains of whale shark MDA5, RIG-1, and LGP2 (Figure 2; Supplementary Figure 4) support that all three RLRs had already diverged in the last common ancestor of extant jawed vertebrates. Further, and consistent with past findings, we found that MDA5 and LGP2 are the result of a vertebrate specific duplication, while RIG-1 split from these genes prior to the emergence of vertebrates[48]. CARD domain analyses designed to include MAVS confirmed the presence of this protein in whale shark, *Callorhinchus*, and despite difficulties identifying a sequence previously[49], coelacanth (UFBOOT=100; Supplementary Figure 4). Together, our results show that the mammalian RLR repertoire (and MAVS) was established prior to the emergence of extant jawed

vertebrates and, likely due to its importance in antiviral immunity, has been near ubiquitously conserved since.

Toll-like receptors (TLRs) are probably the best known of all innate immune PRRs. Past evolutionary studies suggest that the vertebrate TLR repertoire is highly conserved in vertebrates, with only small changes between species[35,49–53]. In contrast, large lineage-specific expansions have been observed in several invertebrates[18,19]. We identified 13 unique putative TLR sequences in whale shark (Supplementary Table 11), 11 of which are orthologous to TLR1/6/10, TLR2/28 (x2), TLR3, TLR7, TLR8, TLR9 (x2), TLR21, TLR22/23, and TLR27. The remaining two, along with a coelacanth sequence, represent a novel ancestral jawed vertebrate TLR gene family related to TLR21, for which we propose the name TLR29. This TLR repertoire is a unique combination compared to all other vertebrates previously studied, formed from a mix of mammalian, teleost, and 'living-fossil' TLRs (i.e. present in sharks, coelacanths, gars; TLR27, TLR29). Our analyses indicate that the ancestor of extant vertebrates possessed at least 15 TLRs, while the ancestor of jawed vertebrates possessed at least 19 TLRs (including three distinct TLR9 lineages), both larger repertoires than possessed by modern 2R species (Fig 2; Supplementary Figure 5). Unlike invertebrates where both loss and expansion of TLRs are extensive, our data suggest that the vertebrate TLR repertoire is probably primarily expanded through early genome duplication events (although a few new TLRs have arisen independently) followed by slow differential gene loss.

In summary, different levels of evolutionary constraint are apparent both within and between species and PRR families following the emergence of vertebrates. Such unique evolutionary histories suggest that PRR repertoire evolution is driven by specific functional needs on a case-by-case basis. Further, rather than a simple replacement scenario, where the acquisition of adaptive immunity supplanted the need for vast repertoires of PRRs in vertebrates, the ability to interact with the adaptive system, in a safe (i.e. limiting autoimmunity) and effective manner, is likely the major force restraining the proliferation of certain vertebrate PRRs.

**Rates of functional genomic evolution and gigantism.** Rates of genomic evolution vary considerably across vertebrates, either across clades or in relationship to other biological factors such as body size. We compared rates in two different aspects of genomic evolution with potential functional relationship to gigantism in the whale shark to other vertebrates: rates of amino acid substitution in protein-coding genes, and rates of gene family size evolution. For these analyses, we determined vertebrate-level gene families by inferring gene families for the representative vertebrates and excluding lancelet and sea squirt (see Methods).

Substitution rates across a set of single-copy orthologs varied across vertebrate genomes, and these rates were relatively low in the whale shark compared to most other vertebrates (Fig. 3). We used the two-cluster test to test for different rates of substitution among vertebrate clades[54]. Previous use of the two-cluster test, comparing the *Callorhinchus* genome to other vertebrates, supported that *Callorhinchus* has a slower substitution rate than coelacanth, teleosts, and tetrapods[2]. Although the whale shark was estimated to have a slower rate than *Callorhinchus*, this difference was not significant ($p = 0.7114$). Indeed, we found that cartilaginous fishes evolved more slowly than bony vertebrates overall, as well as more specific

lineages including gar, coelacanth, and tetrapods (all tests $p = 0.0004$). We found no evidence for the node-density artefact that can bias these analyses due to sparse taxon sampling[55]. The decreased rate of substitution in protein-coding genes in both *Callorhinchus* and whale shark relative to the bony vertebrates is concordant with an overall decreased rate of synonymous substitutions previously found in cartilaginous fishes[11].

We then tested whether rates of molecular substitution differed on branches leading to gigantism in vertebrates when compared to the background rate of molecular evolution in vertebrates. The origins of gigantism in elephants, whales, and whale sharks has previously been shown to correspond to shifts in the rate or mode of body size evolution[56–58]. We estimated time-varying rates of body size evolution in cartilaginous fishes, and consistent with previous research[58], found that gigantism in whale shark corresponds to a discrete shift in the rate of body size evolution to five times the background in cartilaginous fishes (Supplementary Note 4; Supplementary Figure 6)[58]. We thus compared rates of genomic evolution in vertebrate giants – represented by African elephant, minke whale, bowhead whale, and whale shark – to other vertebrates. We found that rates of amino acid substitution on branches leading to the whale shark, elephant, and baleen whales were significantly slower than the background rate of molecular evolution in vertebrates (log-likelihood ratio test $p < 0.0001$), consistent with earlier evidence that larger-bodied taxa have lower rates of protein evolution[59,60]. However, given that the whale shark genome does not appear to evolve significantly more slowly than the *Callorhinchus* genome (as noted above), or other small-bodied sharks as found previously when focusing on fourfold degenerate sites[11], there does not appear to be a compounded effect on substitution rates in the whale shark genome as both a vertebrate giant and a cartilaginous fish. This implies that substitution rates and body size may be decoupled within cartilaginous fishes, which are already overall slowly-evolving, unlike the pattern seen in other vertebrates.

Rates of change in gene family sizes, due to gain and loss of gene copies within gene families, can also vary across species[61]. Therefore, this represents another axis of functional genomic evolution potentially independent from rates of molecular substitution. We estimated rates for gene family size evolution separately for 6,898 gene families present in lamprey and jawed vertebrates and 6,181 gene families present in jawed vertebrates but absent in lamprey, because an assumption of the method used is that gene families are present in the MRCA of all taxa included[62]. Average global rates of gene gain and loss in vertebrates were estimated to be 0.0004–0.0005 gains/losses per million years (Figure 4). However, we found that the rate of gene family size evolution in giant vertebrates was significantly faster than in the remaining branches both for gene families present in lamprey ($p < 0.01$) and gene families absent in lamprey ($p = 0.02$). Additionally, when we estimated rates independently for each origin of gigantism, rates of gene gain and loss were estimated to be roughly an order of magnitude greater in baleen whales than the other giant lineages and the background, implying that baleen whales have a different pattern of gene family evolution than other vertebrates. That the rate of gene family size evolution in different vertebrate giants demonstrates that gigantism may not have the same relationship with overall patterns of gene family size evolution as it does on rates of amino acid substitution in gene families across the genome, and suggests that substitution rates and rates of gene duplication and loss may be decoupled.

Replicated shifts in rates of gene family size evolution for specific gene families across independent giant lineages might indicate the repeated effect of selection related to gigantism. We inferred that 2,375 gene families had a rate shift in gene family size evolution on at least one branch in the vertebrate phylogeny (Supplementary File 8, 9). For these gene families that had a rate shift, on average, ~5 independent rate shifts occurred among the vertebrate species considered, indicating the potential for independent rate shifts corresponding to independent origins of gigantism. We found only three gene families that independently shifted in gene family size evolution rate along the branches leading to the whale shark, elephant, and baleen whales: HIST1H4, HIST1H1, and a family formed by NF1A, NF1C, and NF1X. HIST1H4 is a known cancer-related genes listed by the Cancer Gene Census[63,64], supporting that evolution of cancer suppression may play a role in the evolution of gigantism in histone proteins. Although NF1A, NF1C, and NF1X are not listed as cancer genes, they are related to a cancer-related gene NF1B[63]. Though it is already known that elephants and baleen whales differ in the cancer-related gene families that have duplicated, the overlap of duplicated cancer-related gene families across multiple origins of gigantism is compelling. However, a caveat is that these three gene families have also shifted in rate along other branches among vertebrates, and thus gigantism is not required for a rate shift: HIST1H4 shifted in rate 18 branches in vertebrate evolution, HIST1H1 shifted along 13 branches, and NF1 shifted along seven branches, and so these rate shifts may occur for other adaptations besides gigantism.

We also looked at rate shifts in gene family size evolution restricted to the whale shark, as duplication in cancer-related gene families is not necessarily expected to occur in all independent origins of gigantism. For example, in elephants, the duplication of tumor suppressor protein TP53 has been implicated in reduced cancer rates in elephants relative to other mammals[23,24], while in baleen whales this gene is not duplicated. We confirmed a rate shift in gene family size evolution in TP53 in the lineage leading to elephant but did not find this gene family to have shifted in rate along the whale shark branch or the baleen whale branches. We also did not infer a rate shift in gene families previously suggested to be duplicated in cetaceans related to longevity or cancer suppression, such as PCNA, LAMTOR1, DLD, KCNMB1, and PDCD5[25,65]; this may be because while the elephant has twelve copies of TP53, the other duplications in cetaceans resulted in fewer extra copies, suggesting that focusing on gene families with a significant rate shift associated with a branch will result in a conservative identification of candidate gene families. We still found 188 gene families that have shifted in rate along the branch leading to the whale shark. This set of gene families was enriched for gene families that included human orthologs that are listed in the Cancer Gene Census[63,64]. 16 gene families included human orthologs that are related to cancer, representing a significant enrichment over all gene families tested (odds ratio = 2.022, $p = 0.0117$). These 16 gene families included 21 cancer-related human orthologs: AKAP9, CDH10/CDH11, CNTNAP2, COL1A1/COL2A1/COL3A1, CSMD3, CYSLTR2, DNMT3A, ELN, H3F3A/B, HLA-A, KMT2C/D, LRP1B, NBEA, PTPRD, SLC34A2, and SRGAP3. This supports that the evolution of cancer suppression or longevity in the whale shark may relate to gene duplication. However, these cancer genes are not exclusive to certain tumor types, tissue type, or roles in cancer. Additionally, these gene families were not enriched for any GO function, nor were any annotated for any cancer-relevant GO functions (e.g. regulation of cell cycle, apoptosis), suggesting that

shifts in gene family size evolution in the whale shark for cancer suppression genes were not directed at specific functions.

We acknowledge that attempting to identify genomic evolution specific to the whale shark lineage is hampered a lack of data for smaller-bodied elasmobranchs, because the lineage separating the whale shark and *Callorhinchus* also includes evolutionary history that is not specifically attributed to the whale shark's evolution of gigantism. Hence, identification of patterns common to the lineage leading to whale shark, elephant, and baleen whales is partially subdued by intervening divergences between elephant shark, whale shark, and other elasmobranchs. Future genome sequencing efforts including more elamosbranchs will help to better clarify genomic evolution specific to the whale shark.

## Conclusions

The whale shark is a large, charismatic species and is, by far, the largest species of fish. As a cartilaginous fish, a lineage for which only few genomes have thus far been sequenced, the whale shark genome represents an important resource for vertebrate comparative genomics. The new long read-based genome assembly reported in this paper provides the best gapless genome assembly thus far among cartilaginous fishes. Comparison of the whale shark to other vertebrates not only expands the number of shared gene families that were ancestral to jawed vertebrates, but implies a burst in novel genes at this time of early vertebrate evolution. These early gene families are involved in a diversity of functions including reproduction, metabolism, and development. Specifically, with respect to innate immunity genes, we found divergent patterns of gene gain and loss between NLRs, RIGs, and TLRs, and provide insight into their repertoires in the jawed vertebrate ancestor. These results rejected a scenario where the importance of PRRs is muted in vertebrates by the presence of adaptive immunity, rather indicating their necessity as part of a highly regulated, hybrid immune system. Finally, we demonstrated that the relationship between rates of gene family size evolution and rates of substitution to gigantism are decoupled, and explored the role of cancer-related gene evolution in gigantism in the whale shark and other vertebrate giants. The whale shark genome helps to build a foundation in shark and vertebrate comparative genomics useful to answer questions of broader vertebrate evolution and convergent evolution of distinctive traits. A continued increase in fully-sequenced elasmobranch genomes will continue to enhance research from finding unique, whale-shark specific evolutionary change to illuminating broader patterns of vertebrate evolution.

## Acknowledgements

The sequencing service was provided by the Norwegian Sequencing Centre (www.sequencing.uio.no), a national technology platform hosted by the University of Oslo and supported by the "Functional Genomics" and "Infrastructure" programs of the Research Council of Norway and the Southeastern Regional Health Authorities". We thank F. Thibaud-Nissen for assistance with genome annotation through NCBI RefSeq. We thank B. Morgan and the High

Performance Computing oversight committee for access and assistance with the Center for Advanced Science Innovation and Commerce (CASIC) supercomputer at Auburn University, the rest of the staff of Laboratory of Phyloinformatics in RIKEN BDR for transcriptome sequencing, and R. A. Petit III for assistance with computing at Emory University. Inference of gigantism in the whale shark was made possible by body size data kindly provided by C. Mull. We are thankful for funding provided from the Georgia Aquarium and the Emory School of Medicine Development. S. Koren and A.M. Phillippy were supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

## Author Contributions

M.T. contributed writing of the manuscript, experimental design, and implementing analyses of genome polishing, phylogenomics, orthology determination, gene family evolution, RNAseq mapping, and submission of genomic data to NCBI for accessioning and annotation by RefSeq. A.K.R. and H.D. performed analyses and wrote the results, supplementary text, and methods sections on innate immune genes. S.Koren and A.P. assembled the genome using Canu. R.N., K.S., and S.Kuraku generated whale shark blood transcriptome data. T.D.R. and A.D.M.D. initiated and managed the genome project. All authors provided feedback on the manuscript.

# Figures

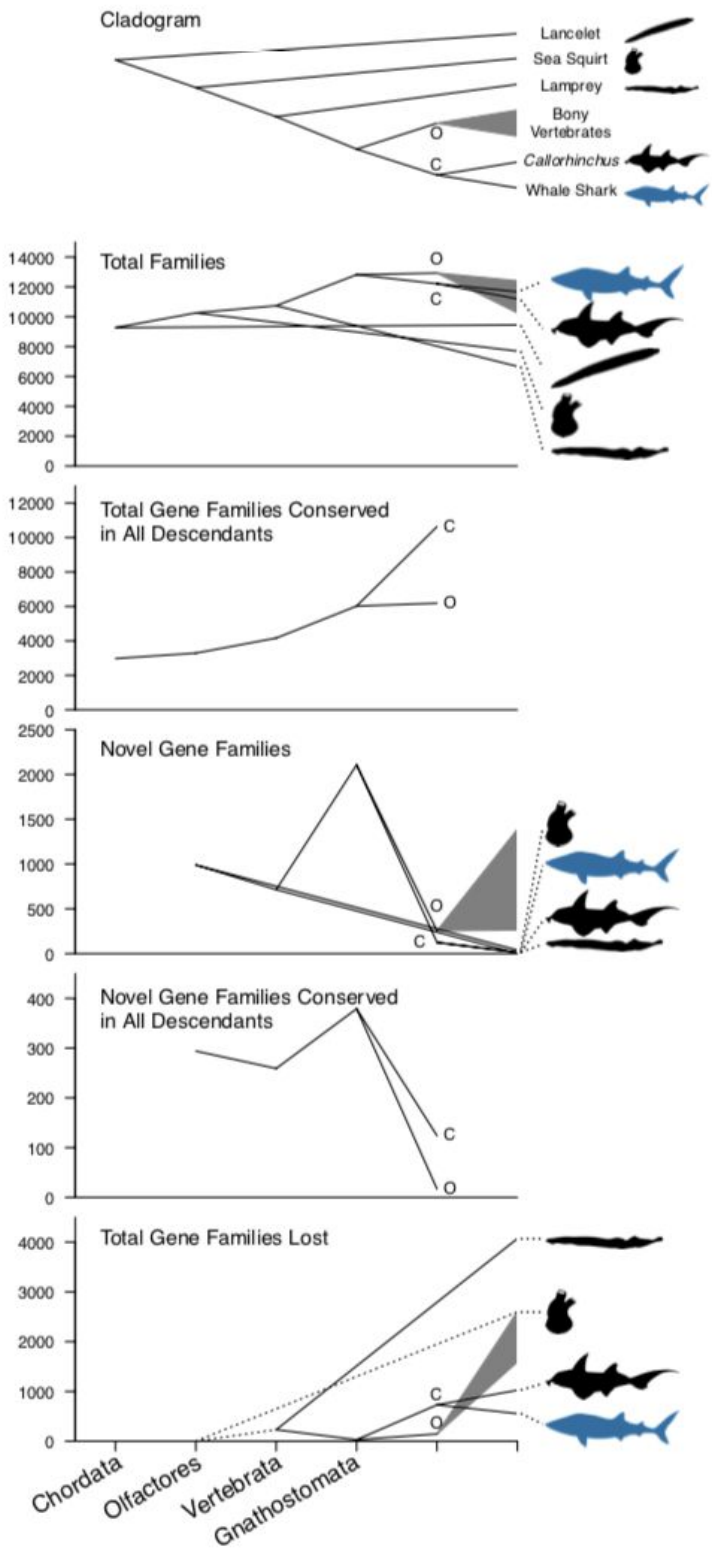

Cladogram
Lancelet
Sea Squirt
Lamprey
Bony Vertebrates
Callorhinchus
Whale Shark
O
C
Total Families
Total Gene Families Conserved in All Descendants
Novel Gene Families
Novel Gene Families Conserved in All Descendants
Total Gene Families Lost
Chordata
Olfactores
Vertebrata
Gnathostomata

Figure 1. Phenograms visualizing the change in gene family numbers across vertebrate evolution. The first panel shows the cladogram for vertebrate nodes pictured in the phenograms. The cladogram is then projected in the subsequent plots such that the y-axis position of each node and tip corresponds to the number of gene families estimated. Ancestral nodes are labeled on the x-axis as they differ in relative time, except for the ancestral nodes for bony vertebrates (marked with "O") and cartilaginous fishes (marked with "C"). The range of counts inferred for bony vertebrates is indicated by a gray triangle, which includes nineteen bony vertebrate species. Panels indicate the total number of gene families at each ancestral node and each tip, the total number of gene families at an ancestral node that are conserved in all descendant taxa considered, the number of novel gene families that originated at each node, the number of novel gene families that originated at each node and are conserved in all descendant taxa considerd, and the total number of gene families lost at each node and tip (for the total number of gene families lost, we did not sample distant outgroups to determine how many gene families were lost between the origin of Chordata and Olfactores, hence the dotted lines). Note the large number of total gene families, novel gene families, novel gene families conserved in all descendants, and small number of lost gene families at the jawed vertebrate (Gnathostomata) ancestor. For more detail, see text, Supplementary Note 2, and Supplementary Figure 3. Silhouettes via Phylopic: lancelet CC BY-SA by Bennet McComish, photo by Hans Hillewaert; sea squirt CC BY-NC by Mali'o Kodis, photograph by Melissa Frey; lamprey CC BY by Gareth Monger; *Callorhinchus* CC BY-SA by Milton Tan, originally by Tony Ayling; whale shark CC BY-SA by Scarlet23 (vectorized by T. Michael Keesey).

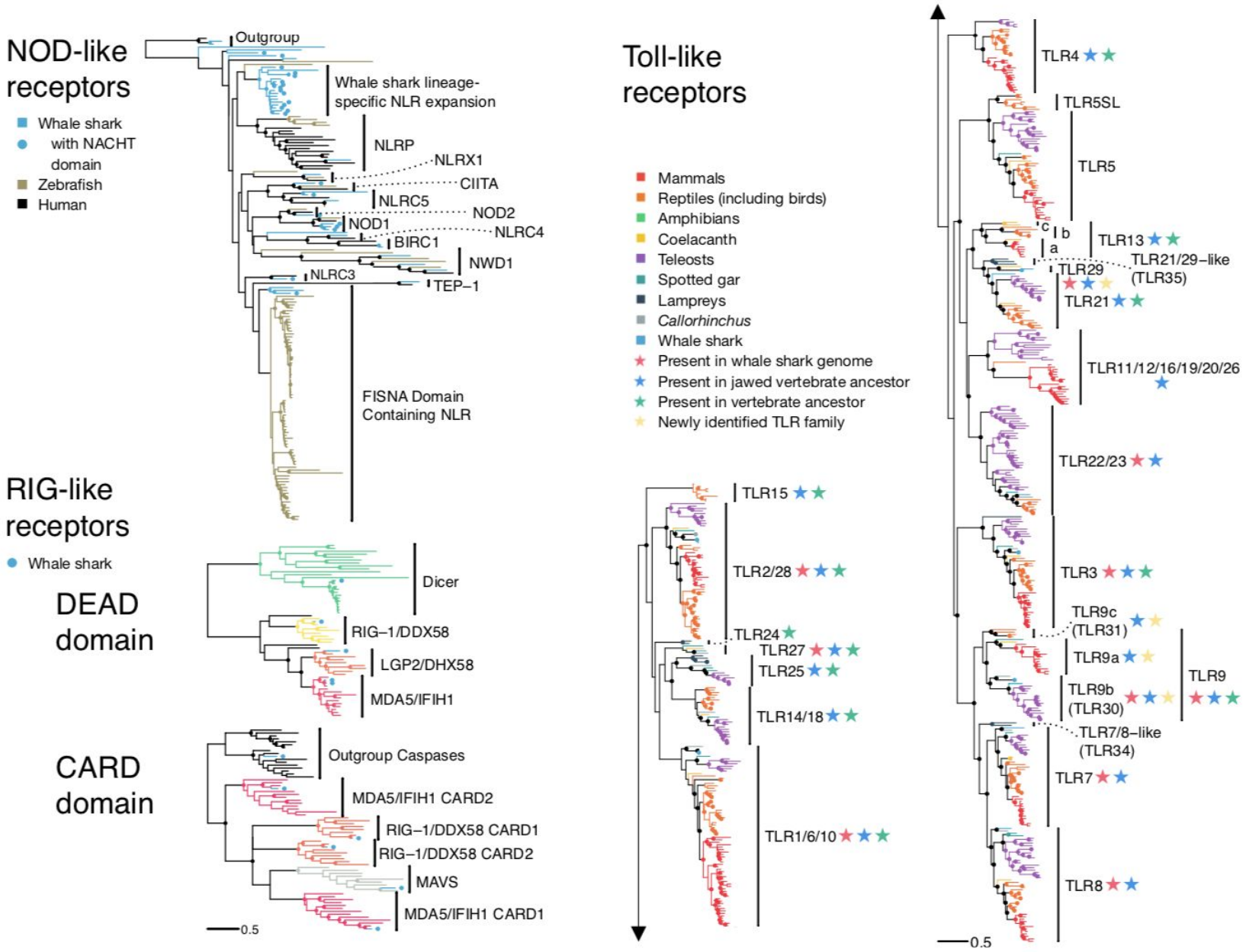

NOD-like receptors
Whale shark
with NACHT domain
Zebrafish
Human
Outgroup
Whale shark lineage-specific NLR expansion
NLRP
NLRX1
CIITA
NLRC5
NOD2
NOD1
NLRC4
BIRC1
NWD1
NLRC3
TEP–1
FISNA Domain Containing NLR
RIG-like receptors
Whale shark
DEAD domain
Dicer
RIG–1/DDX58
LGP2/DHX58
MDA5/IFIH1
CARD domain
Outgroup Caspases
MDA5/IFIH1 CARD2
RIG–1/DDX58 CARD1
RIG–1/DDX58 CARD2
MAVS
MDA5/IFIH1 CARD1
0.5
Toll-like receptors
Mammals
Reptiles (including birds)
Amphibians
Coelacanth
Teleosts
Spotted gar
Lampreys
Callorhinchus
Whale shark
Present in whale shark genome
Present in jawed vertebrate ancestor
Present in vertebrate ancestor
Newly identified TLR family
TLR15
TLR2/28
TLR24
TLR27
TLR25
TLR14/18
TLR1/6/10
TLR4
TLR5SL
TLR5
c
b
a
TLR13
TLR21/29–like (TLR35)
TLR29
TLR21
TLR11/12/16/19/20/26
TLR22/23
TLR3
TLR9c (TLR31)
TLR9a
TLR9b (TLR30)
TLR9
TLR7/8–like (TLR34)
TLR7
TLR8
0.5

Figure 2. PRR repertoires of the whale shark among vertebrates. Nodes supported ≥95% UFBOOT indicated with a dot. a) NLRs among whale shark, zebrafish, and human. NLRs in whale shark with a NACHT domain are indicated by a dot at the tip. b) RLR domain trees for DEAD (left) and CARD (right) domains among vertebrates. Branches are colored by gene, except for RLRs in whale shark which are colored distinctly and each labeled by a dot at each tip. c) TLRs among vertebrates. Each clade represents a separate TLR, families found within TLR13 labeled a (TLR13a), b (TLR32), and c (TLR33). TLR families are also labeled by stars indicating whether they were present in the whale shark genome, present in jawed vertebrate ancestor, present in the vertebrate ancestor, and novel to this study. For more detail, see Supplementary Figures 3–5.

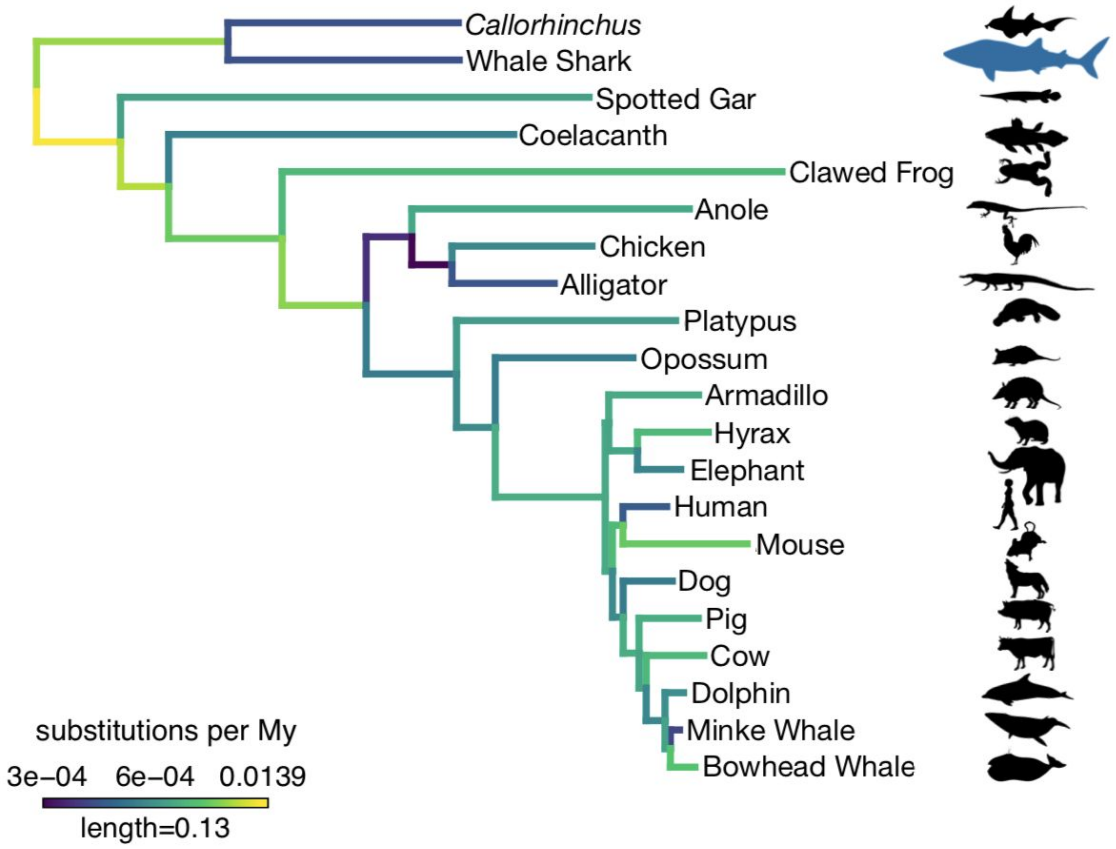



Figure 3. Amino acid substitution rate variation among jawed vertebrates. Branches are colored based on rates quantified by substitutions per site per million years of the maximum likelihood tree compared to a time-calibrated tree. Together, the whale shark and *Callorhinchus* have a significantly slower rate of molecular evolution than spotted gar and coelacanth, but not tetrapods. Furthermore, vertebrate giants – including the whale shark, elephant, and whales – have significantly lower rates of molecular evolution than other vertebrates. Note, color scale is on normalized reciprocal-transformed data, which emphasizes changes between smaller values of substitution per My. Silhouettes via Phylopic: *Callorhinchus*, CC BY-SA by Milton Tan,

originally by Tony Ayling; whale shark, CC BY-SA by Scarlet23 (vectorized by T. Michael Keesey); spotted gar, CC BY-NC-SA by Milton Tan; coelacanth, CC BY-NC-SA by Maija Karala; clawed frog, anole, platypus, opossum, elephant, CC BY by Sarah Werning; alligator, CC BY-NC-SA by Scott Hartman; mouse, CC BY-SA by David Liao; dolphin, CC BY-SA by Chris Huh; chicken, armadillo, hyrax, human, dog, pig, cow, minke whale, bowhead whale, public domain.

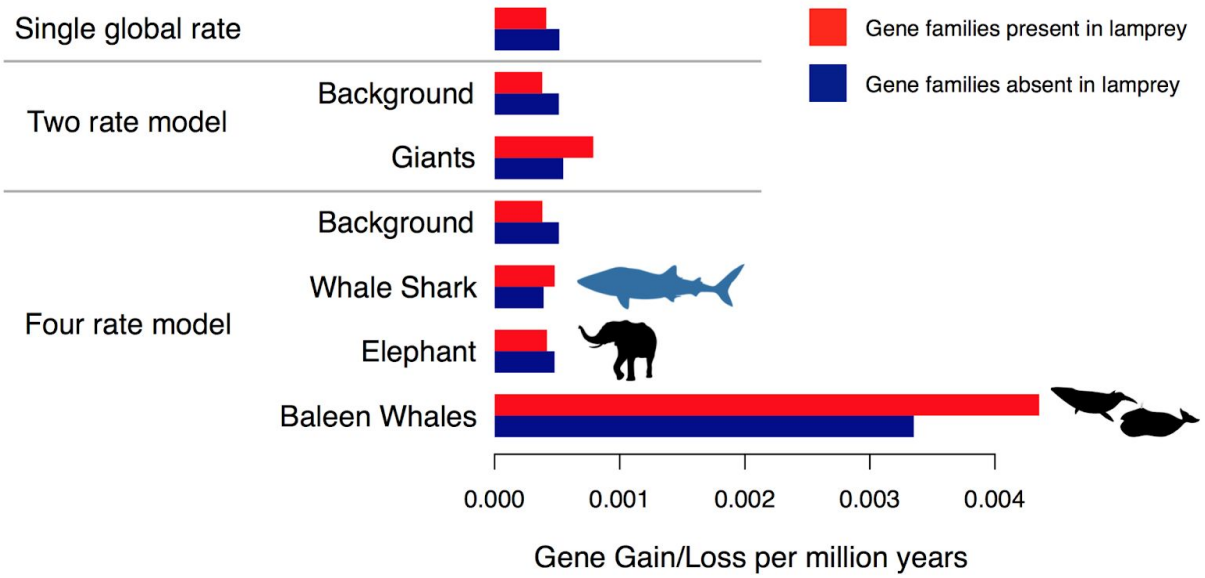



Figure 4. Estimated relative rates of gene family size evolution across giant vertebrates. Rates were estimated separately for gene families present in lamprey (n = 6898) and gene families absent in lamprey (n = 6181). Average global rates of gene gain and loss in vertebrates under a single global rate were estimated to be 0.0004 in gene families present in lamprey and 0.0005 gains/losses per million years in gene families absent in lamprey. Rates of gene gain and loss in giant vertebrates were significantly faster than in the remaining branches for gene families present ($p < 0.01$) and gene families absent in lamprey ($p = 0.02$) as estimated in the two-rate model. This greater rate in giants appears to be due primarily to a rate increase specific to baleen whales, as demonstrated in the four rate model where rates were estimated independently across giant lineages. Silhouttes via Phylopic: whale shark, CC BY-SA by Scarlet23 (vectorized by T. Michael Keesey); elephant, CC BY by Sarah Werning; whales, public domain.

# Methods

**Genome sequence assembly and assessment.** To improve on our earlier efforts to sequence and assemble the whale shark genome[6], we generated PacBio long read sequences from the same DNA sample. These sequences are available on NCBI SRA under the accession SRX3471980. This resulted in 61.8 Gbp of sequences, equivalent to ~20x fold coverage. The

initial assembly was performed using Canu 1.2[17] with adjusted parameters to account for the lower input coverage:

```
canu -p asm -d shark genomeSize=3.5g corMhapSensitivity=high corMinCoverage=2
errorRate=0.035
```

Illumina reads were aligned to the genome using BWA-MEM[66] v0.7.12-r1039 with default parameters and alignments were used as input into Pilon v1.18[67] to correct errors in the draft assembly. We performed assembly-free estimation of genome size, heterozygosity, and repeat content and also assessed gene completeness both with conserved orthologs and by mapping RNA-seq reads (Supplementary Note 1).

**Transcriptome sequencing.** Approximately 30 million short read pairs for whale shark transcripts were obtained with paired-end 127 cycles from blood cells of a male and a female by the Illumina HiSeq 1500 as describe previously[11]. Animal handling and sample collections at Okinawa Churaumi Aquarium were conducted by veterinary staff without restraining the individuals, in accordance with the Husbandry Guidelines approved by the Ethics and Welfare Committee of Japanese Association of Zoos and Aquariums. Downstream handling of nucleic acids were conducted in accordance with the Guideline of the Institutional Animal Care and Use Committee (IACUC) of RIKEN Kobe Branch (Approval ID: H16-11). Transcriptome sequence data are available at NCBI BioProject ID PRJDB8472 and DDBJ DRA ID DRA008572.

**Gene prediction.** Genes were predicted by RefSeq using their genome annotation pipeline[68]. This annotation included alignments of RNAseq data from grey bambooshark *Chiloscyllium griseum* kidney and spleen, nurse shark *Ginglymostoma cirratum* spleen and thymus, and brownbanded bambooshark *Chiloscyllium punctatum* retina, as well as protein alignments from Actinopterygii, and RefSeq protein sequences for Asian arowana *Scleropages formosus*, coelacanth, spotted gar, zebrafish, clawed frog, and human. Based on the RefSeq annotations, orthology, and synteny, we manually annotated the whale shark Hox gene clusters. After preliminary orthology determination, we determined additional genes absent in whale shark conserved among vertebrates, which we annotated by aligning protein sequences from representative vertebrates to whale shark using genBLAST[69,70] (Supplementary Note 1D, Supplementary File 3).

**Orthology inference.** We identified orthologs from the whale shark genome by comparison to publicly available chordate genomes. Chordate genomes were downloaded from RefSeq and Ensembl for 22 species representing major vertebrate clades, the sea squirt *Ciona intestinalis*, and lancelet *Branchiostoma floridae* (Supplementary Table 3). In selecting representative vertebrates, we specifically included the African elephant, as well as two baleen whale genomes, the minke whale and the bowhead whale, and the most closely-related genomes available for these taxa (rock hyrax and bottlenose dolphin, respectively). These ortholog clusters were used for the identification of origins of gene families in chordate evolution and genes that originated in the most recent common ancestor of jawed vertebrates, and studying enrichment or changes in functional annotation associated with these orthogroups (i.e. for analyses for the section titled "Ancestral Vertebrate Genome Evolution") (Supplementary File 5).

We also repeated orthologroup determination excluding sea squirt and lancelet to focus on vertebrate-level orthogroups (Supplementary File 7). These vertebrate-level orthogroups were used for phylogenomics, estimation of rates of molecular substitution, and estimation of rates of gene duplication and loss (i.e. for analyses for the section titled "Rates of functional genomic evolution and gigantism").

Ortholog clusters from proteomes were determined using OrthoFinder v1.1.8[28]. The OrthoFinder pipeline proceeds through several steps, first performing an all-by-all BLAST. We used DIAMOND for the all-to-all BLAST step[71]. With these hits, OrthoFinder determines reciprocal best hits while accounting for gene length bias and phylogenetic distance, then proceeds with clustering genes into orthogroups. Preliminary orthology determination suggested many missing orthologs in the chondrichthyan genomes. We thus performed orthology-based annotation using genBLAST[69,70], added newly identified proteins to the proteomes of whale shark and *Callorhinchus*, and reran the OrthoFinder pipeline including these proteins.

All proteins were then annotated for Gene Ontology (GO), InterPro (IPR), and Pfam terms using InterProScan 5[72], and representative annotations were assigned to each chordate orthogroup using KinFin[73] (Supplementary File 6). We also used KinFin to test for whether each gene family in giant vertebrates were significantly enriched or depleted relative to other vertebrates using the Mann-Whitney *U* test (see section titled "Rates of functional genomic evolution and gigantism"). P-values were adjusted for false discovery rate using the p.adjust method implemented in R to correct P-values for multiple hypothesis testing[74].

**Gene Family Origin and Loss**. To infer when gene families were gained and lost in vertebrate evolution, we mapped the origins and losses of gene families to the species tree parsimoniously, assuming that gene families have a single origin, but can be lost[73]. We were then able to count the number of gene families present at the MRCA of nodes, the number of novel gene families that originated along each branch, and the number of gene families lost along each branch (including gene families uniquely lost along each branch).

Since we recovered a large number of novel gene families at the base of jawed vertebrates, we wanted to confirm that this spike was not due to artefactual oversplitting of ohnologs (gene duplicates that arose from two rounds of whole genome duplication early in vertebrate evolution). Singh *et al.*[34] independently used a synteny-aware method to identify ohnologs in a subset of vertebrate genomes compared to our study. We compared our assignment of human orthologs to gene families (i.e. Chordate-level orthogroups) to their assignment of human orthologs to ohnolog families. Singh *et al.* present results at multiple stringencies. We selected the most conservative and the most relaxed ohnolog sets; the strict set of ohnologs includes 1,381 ohnologs, while the relaxed set includes 2,642 ohnologs. We then determined whether our gene families and Singh *et al.* ohnolog families matched in whether human orthologs were assigned to a single gene family or ohnolog, if orthologs that Singh *et al.* clustered as ohnologs were assigned to multiple gene families, or if orthologs that we clustered as gene families were assigned to multiple ohnologs. To find the common genes between the Ensembl protein IDs we clustered and Ensembl gene IDs provided by Singh *et al.*, we used biomaRt to translate identifiers[75,76].

Based on the representative annotations for each orthogroup determined above, we then determined whether groups of gene families that were gained or lost along branches in the vertebrate phylogeny were enriched for certain functions using a Fisher's Exact Test. Within each comparison, we adjusted the p-value to correct for multiple hypothesis testing by the Benjamin-Hochberg method using the p.adjust function in R[74,77]. Corrected p-values under the BH method can be interpreted at a significance threshold that is equivalent to the false discovery rate. We considered functions enriched with an adjusted p-value of 0.05 and false discovery rate of 0.05.

**Innate Immunity Analyses.**

*Homology identification*: Sequence similarity searches were performed using BLAST to identify putative homologs of TLRs, NLRs and RLRs[78]. An alternative approach using profile hidden Markov models, HMMER [version 3.1], was also tested for TLRs; the results obtained were identical, except that BLAST returned an additional putative TLR. Due to this HMMER results were not applied in subsequent analyses, and HMMER was not applied elsewhere[79] (http://hmmer.org/). Searches for TLR and RLR homologs were performed using all other sequences present in the TLR and RLR trees. Retention of sequences for further analyses was reliant on a reciprocal blast hit to a TLR or RLR in the Swissprot reviewed database or the NCBI non-redundant protein set[80].

For NLRs, detection is more complicated, as some NLRs do not contain computationally detectable NACHT domains (i.e. some family members, even in humans, are false negatives in domain-based search tools and databases), despite the NACHT domain being the defining feature of NLR family members. Further, some of these genes contain other domains and are also included in other gene families where the most members do not contain NACHT domains. As such, for the main analysis performed here, those sequences in the predicted proteome and translated transcriptome containing a predicted NACHT domain (according to the NCBI CD-search webserver[81]) are noted as such (and should be considered as the conservative set of whale shark NLR-like sequences). Additional sequences from the predicted protein set with a blast hit to known NLRs were also included to permit detection of potential orthologs of NLRs not found in the conservative set with definite/detectable NACHT domains. Proteins containing the closely related NB-ARC domain were also extracted from the whale shark proteome for use as an outgroup in NLR analyses, along with Human APAF-1 which also harbors an NB_ARC domain[82].

In cases where a transcript matches the genomic location of a predicted protein, the predicted protein is the sequence reported. Where multiple predicted proteins refer to the same genomic location, only a single sequence is retained for further analysis.

*Phylogenetic Datasets:* For the RLR datasets, members of each of the three vertebrate RLR families, some invertebrate RLRs, and a selection of DICER proteins sequences as an outgroup[48], were gathered to generate a phylogenetically informative dataset. Full length proteins were aligned for phylogenetic analysis of Dead-Helicase domains, and trimmed to the start and end of these domains based on the three human RLR sequences[48]. The same process was performed for the CARD domains[83].

For the TLR dataset, a large set of TLR nucleotide sequences were taken from a past study[53] and then subsampled to include fewer, but still phylogenetically representative species. A TLR sequence from grey bamboo shark (*Chiloscyllium griseum*) was also included[84]. Following trimming, the alignment consisted almost entirely of sites from the TIR domain, so TIR domains were not specifically extracted for this analysis.

For the NLR analysis the described set of human NLRs and NACHT domain containing proteins, as well as the closely related NB_ARCs as an outgroup[82], were downloaded from NCBI protein database. Zebrafish (where NLRs are massively expanded[85,86]) sequences were also included in this analysis but these were downloaded from the InterPro website (i.e. all *Danio rerio* proteins containing a NACHT domain)[87]. A very large number of zebrafish sequences were obtained, so to reduce the prevalence of pseudo-replicate sequences (that are likely to be uninformative in the context of understanding the whale shark NLR repertoire), cd-hit[88] was used to cluster zebrafish sequences with greater than 75% identity prior to phylogenetic analysis.

*Multiple sequence alignment and phylogenetic analyses:* Multiple sequence alignments were generated with MAFFT (version: 7.313)[89] using default parameters for the larger TLR and NLR datasets, but using the LINSi method for RLRs. trimAl (version: 1.2rev59)[90] was applied to remove gap rich sites, which are often poorly aligned, from the alignments using the “gappyout” algorithm. BMGE (version: 1.12)[91] was then used to help minimize the number of saturated sites in the remaining alignment (as identified using the BLOSUM30 matrix). The RLR analyses were not subjected to this BMGE analysis, as these were derived from conserved domains (meaning that alignments were based on relatively conserved sequence tracts and were already quite short). Phylogenetic analyses were performed in IQ-tree (version: omp-1.5.4)[92] using 1000 ultrafast bootstrap replicates[93] and the best-fitting model of amino acid substitution. Best-fitting substitution models were determined according to the Bayesian information criterion with ModelFinder from IQ-TREE[94], and ultrafast bootstrap support was computed to assess node support[95]. The following (best-fitting) models were applied for each dataset: LG+I+G for RLR CARD domains dataset, LG+I+F+G for RLR DEAD-Helicase domains dataset, JTT+I+F+G for the TLR dataset, and JTT+F+G for the NLR dataset.

**Phylogenomics.** Vertebrate-level orthogroups were filtered to single-copy orthologs for phylogenomic analyses. We determined orthologues from orthogroups by reconstructing orthogroup trees and used tree-based orthology determination using the UPhO pipeline [96]. The paMATRAX+ pipeline bundled with UPhO was used to perform alignment (mafft), mask gaps (trimAl), remove sequences containing too few unambiguous sites, and check that at the minimum number of taxa are present (custom script bundled with UPhO Al2Phylo), and then reconstruct phylogenies (FastTree)[89,90,97]. Next, we used UPhO to extract orthologs by identifying all maximum inclusive subtrees from orthogroups with at least five species, with the allowance for in-paralogs (paralogs that arose after all species divergences in the phylogeny, and thus do not affect relative relationships in the phylogeny), and retained the longest in-paralogous sequence for each species within each ortholog. For each single-copy ortholog, we aligned, trimmed, and sanitized sequences using the paMATRAX+ pipeline. Next, orthologs without lamprey, *Callorhinchus*, and whale shark were excluded.

To select the most reliable sequences for phylogenomics, we selected the most informative loci using MARE with default settings[98]. Of the remaining loci, we estimated a few metrics to further filter loci for more phylogenetically reliable sequences. First, we inferred maximum likelihood trees for each locus using IQ-TREE with ModelFinder to select the best model for each locus, and then inferred the maximum likelihood tree, assessing bootstrap support using the UFBoot2 rapid approximation with 1000 replicates[92,94,95]. Using these locus trees, we computed several metrics used for assessing the reliability of loci, including 1) average bootstrap support across nodes as a proxy for phylogenetic signal, 2) clock-likeness, a relative metric for the support of a strict clock model, by estimating the likelihood ratio of the JTT+G strict clock model compared to the JTT+G model where branches were free to vary in PAML 4[99], 3) LB score heterogeneity, a metric for the extremeness of long branches within a locus, implemented in TreSpeX[100], and 4) locus length. We filtered down the loci to retain sequences that had an average bootstrap support above 70, an alignment length over 1000 positions, clock-likeness below 1000 log-likelihood ratio between the free rate model and the strict clock model, and LB score heterogeneity below 100%. This resulted in 174 remaining loci representing a total of 267,860 positions. We concatenated the sequences and selected the best model of amino acid substitution and partitioning scheme using PartitionFinder2 with -raxml and the rcluster search, with -rcluster_max set to 100. A maximum likelihood phylogeny was then inferred using IQ-TREE with the best-fit partitioning scheme. The tree was rooted using the lamprey *Petromyzon*.

Numerous fossil-based node calibrations were identified from the literature. Most node ages were derived from age ranges published in the Fossil Calibration Database[3] and are listed in Supplementary Table 4. While previously the age of crown Chondrichthyes (here, the MRCA of Holocephali + Elasmobranchii) has been suggested to range from 333.56–422.4 Ma[3], the minimum age was recently pushed further back to 358 Ma based on multiple holocephalan fossils[101]. To assess the concordance of these fossil calibrations, we used treePL to estimate divergence times from the ML tree with each fossil calibration using penalized likelihood, then performed cross-validation and evaluated the concordance of the fossils to the time tree to identify and exclude outliers[102]. After excluding fossils that were discordant with the others, we estimated divergence times using treePL with only the remaining fossil calibrations.

**Evolution of Body Size in Cartilaginous Fishes.** We obtained a distribution of 500 time trees for chondrichthyans with divergence times estimated incorporating ten fossil calibrations[103], which are available online for download[104]. Body size data for chondrichthyans from a previous study[103] were kindly provided by Chris Mull. We used BAMM[105] to estimate time-varying rates of evolution of log-transformed body mass in Chondrichthyes. We then used BAMMtools to summarize results of the posterior distribution of results, including computing the marginal odds ratio of a shift in the rate of body size evolution leading to the whale shark, mean background rates of body size evolution in chondrichthyans, and mean rates of body size evolution in the branch leading to the whale shark. See Supplementary Note 4 for further details.

**Tests for Rates of Substitution.** We tested for differences in rates of molecular substitution between vertebrates by using the two-cluster test implemented in LINTRE[54], using amino acid

p-distances between taxa to estimate branch lengths. The two-cluster test is designed to test if the rates in two clades are significantly different by comparison to an outgroup. To focus on particular comparisons between two groups, we compared whale shark to *Callorhinchus*, cartilaginous fishes vs. teleosts (gar), cartilaginous fishes vs. coelacanth, and cartilaginous fishes vs. tetrapods. We used the full 1,249,191 residue alignment of single-copy vertebrate-level orthologs (i.e. the alignment prior to filtering for phylogenomic data), because focusing on the filtered genes for clock-likeness would likely underestimate rate heterogeneity across lineages. Tests for rates based on phylogenetic analysis may be biased because branch lengths may be underestimated in poorly sampled regions of the tree, but this node-density artefact can be tested for statistically[55,106,107]. We tested for the node-density artifact using an online tool provided by Pagel[108].

We also compared rates of genomic evolution of the three independent instances of vertebrate gigantism relative to the background rate of molecular evolution among vertebrates. To do this, we used PAML 4 to compute the likelihoods of the alignment of single-copy orthologs used for phylogenomics under two different models of molecular evolution[99]. We computed the likelihood of the data under a strict clock model (single-rate model) and under a local clock model (two-rate model) where the clock rate differed on branches leading to vertebrate giants.

**Rates of Gene Family Evolution.** We used the vertebrate-level orthogroups to estimate rates of gene duplication and loss. OrthoFinder output includes counts of the size of each orthogroup (i.e. gene families) for each species. We analyzed the evolution of gene family size under a birth-death process using CAFE 3[62], with the gene birth parameter specified by λ. We estimated species-by-species error rates in annotation with the caferror.py script.

To study genome-wide rates of gene duplication and loss in vertebrate giants and the whale shark, we estimated rates of gene duplication and loss across vertebrates (single λ model), and two multi-λ models: a two λ mode model where branches leading to gigantism had a second rate, and a four λ model where the rate categories were the background, a separate rate for each of the three independent origins of gigantism. We used a time-calibrated phylogeny of vertebrates for this analysis (see above). Because the birth-death model in CAFE 3 assumes gene family presence as the root state, vertebrate gene families were considered separately depending on whether or not they were present or absent in lamprey (6,898 and 6,181 gene families, respectively). Because we were specifically focused on the evolution of gene families including whale shark, we did not perform CAFE analysis on orthogroups that whale shark did not possess. To test for significance of the observed difference in likelihoods between the multi-λ models and the single λ model, we simulated gene family evolution with 100 replicates under these models and estimated the log-likelihood ratios from this null, simulated distribution (CAFE does not appear to provide a method to test for the fit between two multi-rate models). The p-value corresponds to the proportion of simulated replicates which had a smaller log-likelihood ratio than observed; however, nearly all observed log-likelihood ratios were below the smallest log-likelihood ratio found among the 100 simulated replicates, implying small p-values ($p < .01$). When fitting the λ model, CAFE 3 additionally computes rates of duplication and loss along each branch for each gene family, and tests whether significant rate shifts occur

along each branch (Supplementary File 7–8). Significance P-values < 0.05 indicate a significant rate shift in gene family size evolution rate.

In summary, we were able to quantify rates of gene family size evolution across vertebrates and estimate rates across different origins of gigantism; identify if any gene families differed in rates of gene duplication and loss differed from the overall rate variation in vertebrates; for gene families that differed in gene family size evolution rate from the background rate, determine gene families that had significant shifts in gene family size evolution rate along certain branches; and test whether or not gene families that shifted in gene family size evolution rate along branches associated with the evolution of gigantism were enriched for certain functions or cancer-related genes. Cancer-related gene families were determined by downloading the gene families from the COSMIC Cancer Gene Census[63,64] and determining which orthogroups included the human ortholog based on the Ensembl gene identifier provided by the CGC. Ensembl gene ENSG identifiers were matched to the Ensembl protein ENSP identifiers (which we used for orthogroup determination) using biomaRt[75,76].

# References


1. Nelson, J. S., Grande, T. C. & Wilson, M. V. H. *Fishes of the World*. (John Wiley & Sons, 2016).
2. Venkatesh, B. *et al.* Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).
3. Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaentologica Electronica* **18.1.1FC**, 1–106 (2014).
4. Redmond, A. K., Macqueen, D. J. & Dooley, H. Phylotranscriptomics suggests the jawed vertebrate ancestor could generate diverse helper and regulatory T cell subsets. *BMC Evol. Biol.* **18**, 169 (2018).
5. Alam, M. T., Petit, R. A., III, Read, T. D. & Dove, A. D. M. The complete mitochondrial genome sequence of the world's largest fish, the whale shark (Rhincodon typus), and its comparison with those of related shark species. *Gene* **539**, 44–49 (2014).
6. Read, T. D. *et al.* Draft sequencing and assembly of the genome of the world's largest fish,

the whale shark: *Rhincodon typus* Smith 1828. *BMC Genomics* **18**, 532 (2017).

7. McClain, C. R. *et al.* Sizing ocean giants: patterns of intraspecific size variation in marine megafauna. *PeerJ* **2**, e715 (2015).
8. Colman, J. G. A review of the biology and ecology of the whale shark. *J. Fish Biol.* **51**, 1219–1234 (1997).
9. Stevens, J. D. Whale shark (Rhincodon typus) biology and ecology: A review of the primary literature. *Fish. Res.* **84**, 4–9 (2007).
10. Rowat, D. & Brooks, K. S. A review of the biology, fisheries and conservation of the whale shark Rhincodon typus. *J. Fish Biol.* **80**, 1019–1056 (2012).
11. Hara, Y. *et al.* Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nat Ecol Evol* **2**, 1761–1771 (2018).
12. Marra, N. J. *et al.* White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability. *Proc. Natl. Acad. Sci. U. S. A.* 201819778 (2019).
13. Park, S. G. *et al.* The whale shark genome reveals how genomic and physiological properties scale with body size. *bioRxiv* 443036 (2018). doi:10.1101/443036
14. Martin, A. P., Naylor, G. J. & Palumbi, S. R. Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature* **357**, 153–155 (1992).
15. Martin, A. P. Substitution rates of organelle and nuclear genes in sharks: implicating metabolic rate (again). *Mol. Biol. Evol.* **16**, 996–1002 (1999).
16. Koren, S. *et al.* Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30**, 693–700 (2012).
17. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* (2017). doi:10.1101/gr.215087.116

18. Rast, J. P., Smith, L. C., Loza-Coll, M., Hibino, T. & Litman, G. W. Genomic insights into the immune system of the sea urchin. *Science* **314**, 952–956 (2006).
19. Huang, S. *et al.* Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res.* **18**, 1112–1126 (2008).
20. Flajnik, M. F. & Kasahara, M. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nat. Rev. Genet.* **11**, 47–59 (2010).
21. Dooley, H. Athena and the Evolution of Adaptive Immunity. *Immunobiology of the Shark* 29 (2014).
22. Peto, R., Roe, F. J., Lee, P. N., Levy, L. & Clack, J. Cancer and ageing in mice and men. *Br. J. Cancer* **32**, 411–426 (1975).
23. Abegglen, L. M. *et al.* Potential Mechanisms for Cancer Resistance in Elephants and Comparative Cellular Response to DNA Damage in Humans. *JAMA* **314**, 1850–1860 (2015).
24. Sulak, M. *et al.* TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants. *Elife* **5**, (2016).
25. Tollis, M. *et al.* Return to the sea, get huge, beat cancer: an analysis of cetacean genomes including an assembly for the humpback whale (Megaptera novaeangliae). *Mol. Biol. Evol.* (2019). doi:10.1093/molbev/msz099
26. Hara, Y. *et al.* Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. *BMC Genomics* **16**, 977 (2015).
27. Nishimura, O., Hara, Y. & Kuraku, S. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. *Bioinformatics* **33**, 3635–3637 (2017).
28. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16**, 157

(2015).

29. Shen, X.-X. *et al.* Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. *Cell* **0**, (2018).
30. Louis, A., Roest Crollius, H. & Robinson-Rechavi, M. How much does the amphioxus genome represent the ancestor of chordates? *Brief. Funct. Genomics* **11**, 89–95 (2012).
31. Richter, D. J., Fozouni, P., Eisen, M. B. & King, N. Gene family innovation, conservation and loss on the animal stem lineage. *Elife* **7**, (2018).
32. Ohno, S., Wolf, U. & Atkin, N. B. Evolution from fish to mammals by gene duplication. *Hereditas* **59**, 169–187 (1968).
33. Dehal, P. & Boore, J. L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol.* **3**, e314 (2005).
34. Singh, P. P., Arora, J. & Isambert, H. Identification of Ohnolog Genes Originating from Whole Genome Duplication in Early Vertebrates, Based on Synteny Comparison across Multiple Genomes. *PLoS Comput. Biol.* **11**, e1004394 (2015).
35. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* **48**, 427–437 (2016).
36. White, T. W. & Paul, D. L. Genetic diseases and gene knockouts reveal diverse connexin functions. *Annu. Rev. Physiol.* **61**, 283–310 (1999).
37. Okamura, Y. *et al.* Comprehensive analysis of the ascidian genome reveals novel insights into the molecular evolution of ion channel genes. *Physiol. Genomics* **22**, 269–282 (2005).
38. Alexopoulos, H. *et al.* Evolution of gap junctions: the missing link? *Curr. Biol.* **14**, R879–80 (2004).
39. Niimura, Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* **1**, 34–44 (2009).

40. Ivancevic, A. M., Kortschak, R. D., Bertozzi, T. & Adelson, D. L. LINEs between Species: Evolutionary Dynamics of LINE-1 Retrotransposons across the Eukaryotic Tree of Life. *Genome Biol. Evol.* **8**, 3301–3322 (2016).

41. Proell, M., Riedl, S. J., Fritz, J. H., Rojas, A. M. & Schwarzenbacher, R. The Nod-like receptor (NLR) family: a tale of similarities and differences. *PLoS One* **3**, e2119 (2008).

42. Ting, J. P.-Y. *et al.* The NLR gene family: a standard nomenclature. *Immunity* **28**, 285–287 (2008).

43. Caruso, R., Warner, N., Inohara, N. & Núñez, G. NOD1 and NOD2: signaling, host defense, and inflammatory disease. *Immunity* **41**, 898–908 (2014).

44. Fritz, J. H. & Kufer, T. A. Editorial: NLR-Protein Functions in Immunity. *Front. Immunol.* **6**, 306 (2015).

45. Keestra-Gounder, A. M. & Tsolis, R. M. NOD1 and NOD2: Beyond Peptidoglycan Sensing. *Trends Immunol.* **38**, 758–767 (2017).

46. Schroder, K. & Tschopp, J. The inflammasomes. *Cell* **140**, 821–832 (2010).

47. Loo, Y.-M. & Gale, M., Jr. Immune signaling by RIG-I-like receptors. *Immunity* **34**, 680–692 (2011).

48. Mukherjee, K., Korithoski, B. & Kolaczkowski, B. Ancient origins of vertebrate-specific innate antiviral immunity. *Mol. Biol. Evol.* **31**, 140–153 (2014).

49. Boudinot, P. *et al.* A tetrapod-like repertoire of innate immune receptors and effectors for coelacanths. *J. Exp. Zool. B Mol. Dev. Evol.* **322**, 415–437 (2014).

50. Roach, J. C. *et al.* The evolution of vertebrate Toll-like receptors. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9577–9582 (2005).

51. Star, B. *et al.* The genome sequence of Atlantic cod reveals a unique immune system. *Nature* **477**, 207–210 (2011).

52. Malmstrøm, M. *et al.* Evolution of the immune system influences speciation rates in teleost fishes. *Nat. Genet.* (2016). doi:10.1038/ng.3645

53. Wang, J., Zhang, Z., Liu, J., Zhao, J. & Yin, D. Ectodomain Architecture Affects Sequence and Functional Evolution of Vertebrate Toll-like Receptors. *Sci. Rep.* **6**, 26705 (2016).

54. Takezaki, N., Rzhetsky, A. & Nei, M. Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12**, 823–833 (1995).

55. Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst. Biol.* **55**, 637–643 (2006).

56. Puttick, M. N. & Thomas, G. H. Fossils and living taxa agree on patterns of body mass evolution: a case study with Afrotheria. *Proc. Biol. Sci.* **282**, 20152023 (2015).

57. Slater, G. J., Goldbogen, J. A. & Pyenson, N. D. Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proc. Biol. Sci.* **284**, (2017).

58. Pimiento, C., Cantalpiedra, J. L., Shimada, K., Field, D. J. & Smaers, J. B. Evolutionary pathways toward gigantism in sharks and rays. *Evolution* (2019). doi:10.1111/evo.13680

59. Martin, A. P. & Palumbi, S. R. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U. S. A.* **90**, 4087–4091 (1993).

60. Bromham, L., Rambaut, A. & Harvey, P. H. Determinants of rate variation in mammalian DNA sequence evolution. *J. Mol. Evol.* **43**, 610–621 (1996).

61. Hahn, M. W., Demuth, J. P. & Han, S.-G. Accelerated rate of gene gain and loss in primates. *Genetics* **177**, 1941–1949 (2007).

62. Han, M. V., Thomas, G. W. C., Lugo-Martinez, J. & Hahn, M. W. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Mol. Biol. Evol.* **30**, 1987–1997 (2013).

63. Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction

across all human cancers. *Nat. Rev. Cancer* **18**, 696–705 (2018).

64. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
65. Keane, M. *et al.* Insights into the Evolution of Longevity from the Bowhead Whale Genome. *Cell Rep.* **10**, 112–122 (2015).
66. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997* (2013).
67. Walker, B. J. *et al.* Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
68. Rhincodon typus Annotation Report. Available at: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Rhincodon_typus/100/. (Accessed: 17th December 2018)
69. She, R., Chu, J. S.-C., Wang, K., Pei, J. & Chen, N. GenBlastA: enabling BLAST to identify homologous gene sequences. *Genome Res.* **19**, 143–149 (2009).
70. She, R. *et al.* genBlastG: using BLAST searches to build homologous gene models. *Bioinformatics* **27**, 2141–2143 (2011).
71. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
72. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
73. Laetsch, D. R. & Blaxter, M. L. KinFin: Software for Taxon-Aware Analysis of Clustered Protein Sequences. *G3* **7**, 3349–3357 (2017).
74. Wright, S. P. Adjusted p-values for simultaneous inference. *Biometrics* **48**, 1005 (1992).
75. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases

and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).

76. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.* **4**, 1184–1191 (2009).
77. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
78. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
79. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
80. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
81. Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res.* **43**, D222–6 (2015).
82. Urbach, J. M. & Ausubel, F. M. The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 1063–1068 (2017).
83. Korithoski, B. *et al.* Evolution of a Novel Antiviral Immune-Signaling Interaction by Partial-Gene Duplication. *PLoS One* **10**, e0137276 (2015).
84. Krishnaswamy Gopalan, T. *et al.* Transcriptome profiling reveals higher vertebrate orthologous of intra-cytoplasmic pattern recognition receptors in grey bamboo shark. *PLoS One* **9**, e100018 (2014).
85. Laing, K. J., Purcell, M. K., Winton, J. R. & Hansen, J. D. A genomic view of the NOD-like receptor family in teleost fish: identification of a novel NLR subfamily in zebrafish. *BMC Evol. Biol.* **8**, 42 (2008).

86. Howe, K. *et al.* The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).

87. Hunter, S. *et al.* InterPro: the integrative protein signature database. *Nucleic Acids Res.* **37**, D211–5 (2009).

88. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).

89. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

90. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

91. Criscuolo, A. & Gribaldo, S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).

92. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

93. Minh, B. Q., Nguyen, M. A. T. & von Haeseler, A. Ultrafast approximation for phylogenetic bootstrap. *Mol. Biol. Evol.* **30**, 1188–1195 (2013).

94. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

95. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Le, S. V. UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.* (2017). doi:10.1093/molbev/msx281

96. Ballesteros, J. A. & Hormiga, G. A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology. *Mol. Biol. Evol.* **33**, 2481 (2016).

97. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

98. Misof, B. *et al.* Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. *BMC Bioinformatics* **14**, 348 (2013).

99. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

100. Struck, T. H. TreSpEx-Detection of Misleading Signal in Phylogenetic Reconstructions Based on Tree Information. *Evol. Bioinform. Online* **10**, 51–67 (2014).

101. Coates, M. I., Gess, R. W., Finarelli, J. A., Criswell, K. E. & Tietjen, K. A symmoriiform chondrichthyan braincase and the origin of chimaeroid fishes. *Nature* **541**, 208–211 (2017).

102. Near, T. J., Meylan, P. A. & Shaffer, H. B. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. *Am. Nat.* **165**, 137–146 (2005).

103. Stein, R. W. *et al.* Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. *Nat Ecol Evol* **2**, 288–298 (2018).

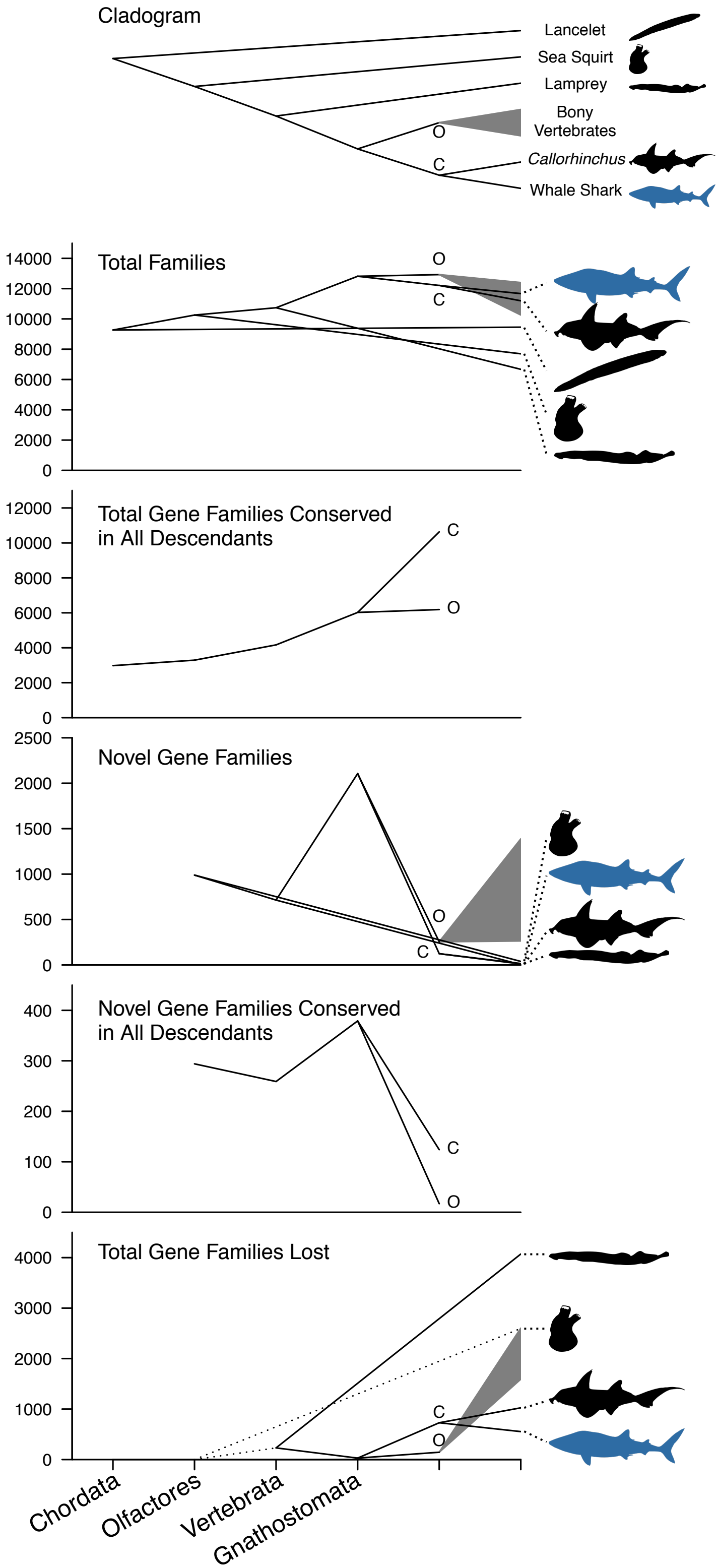
104. A Global Phylogeny of Sharks. *VertLife* (2018). Available at: http://vertlife.org/sharktree/. (Accessed: 25th February 2018)

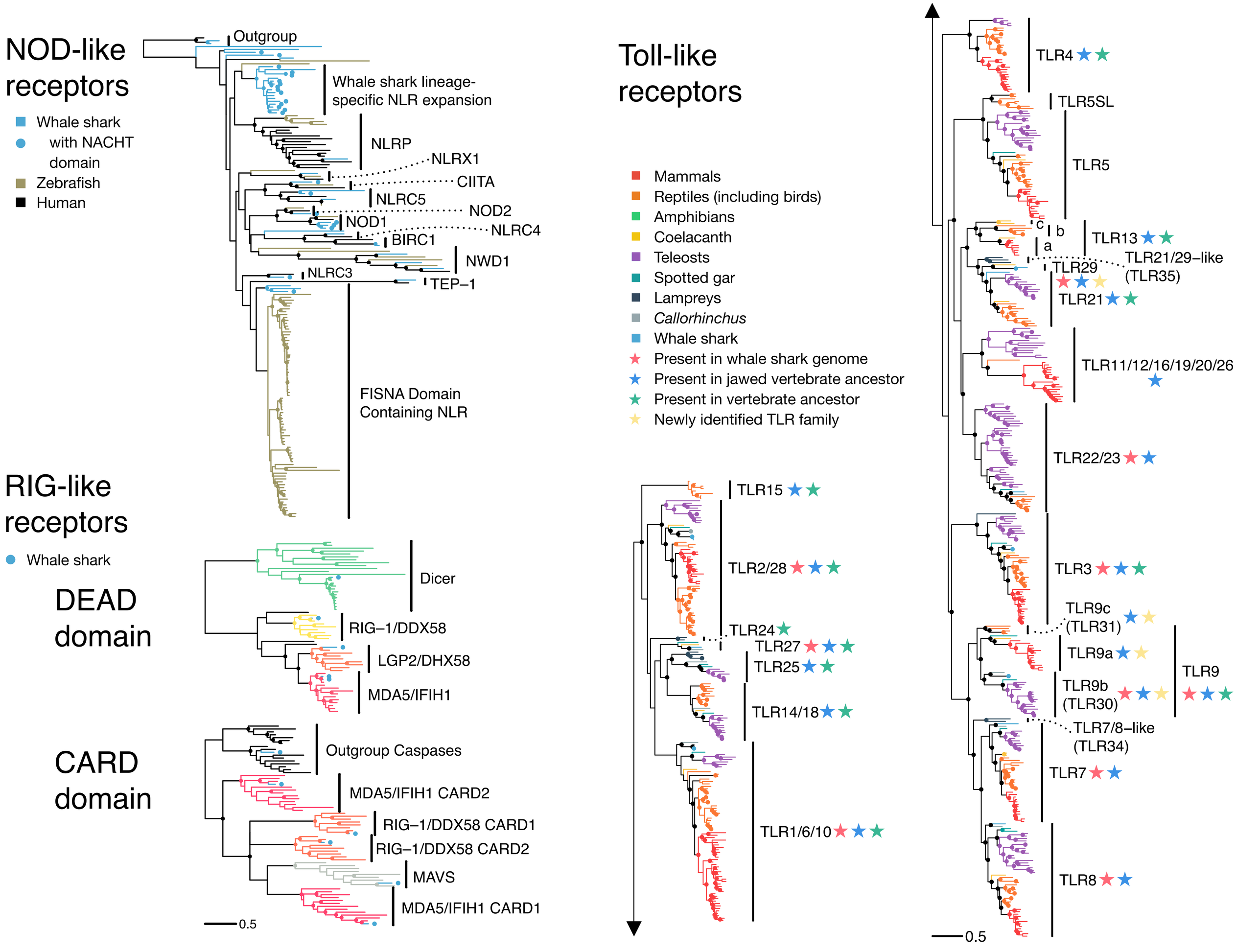
105. Rabosky, D. L. *et al.* Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**, 1958 (2013).

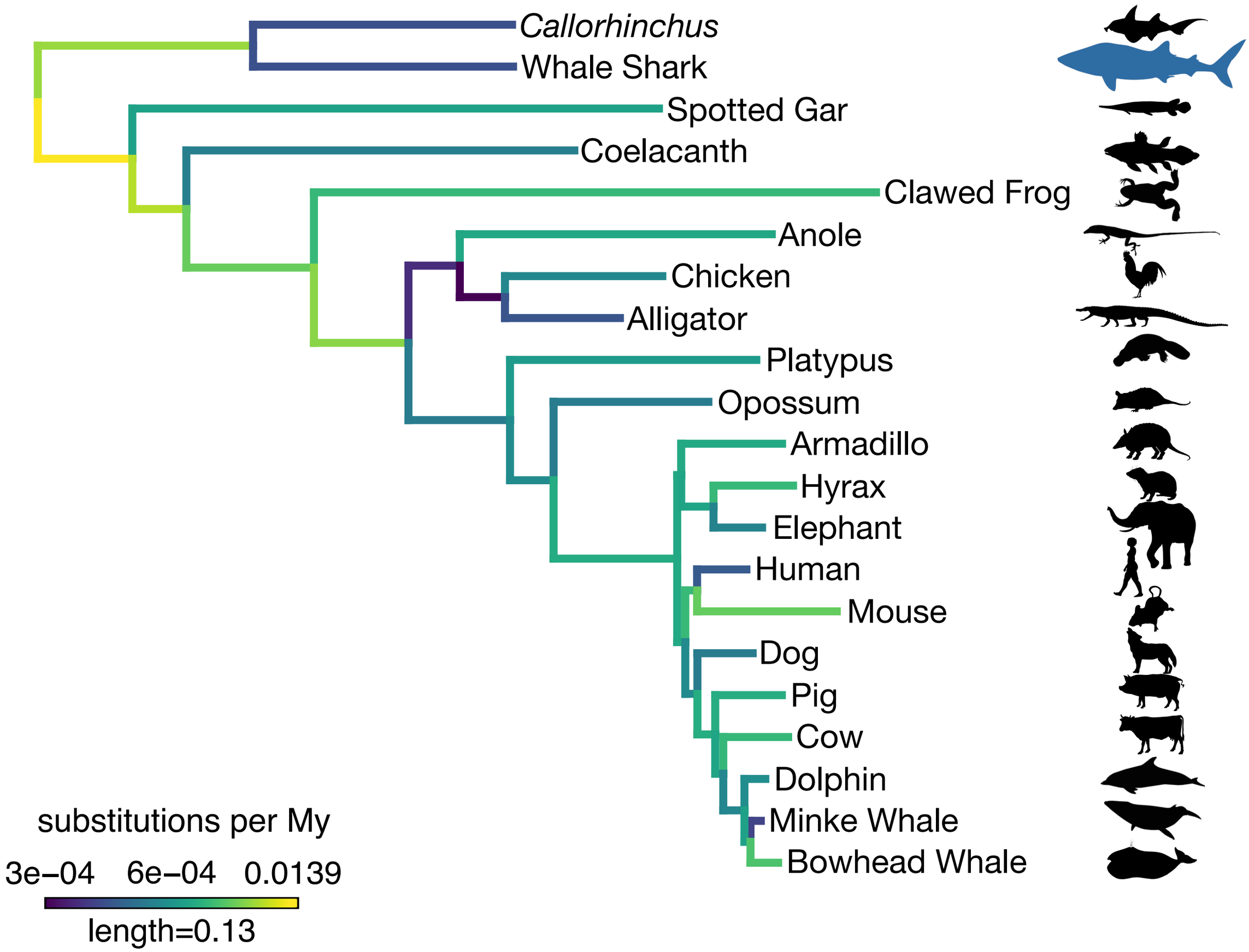
106. Webster, A. J., Payne, R. J. H. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478 (2003).

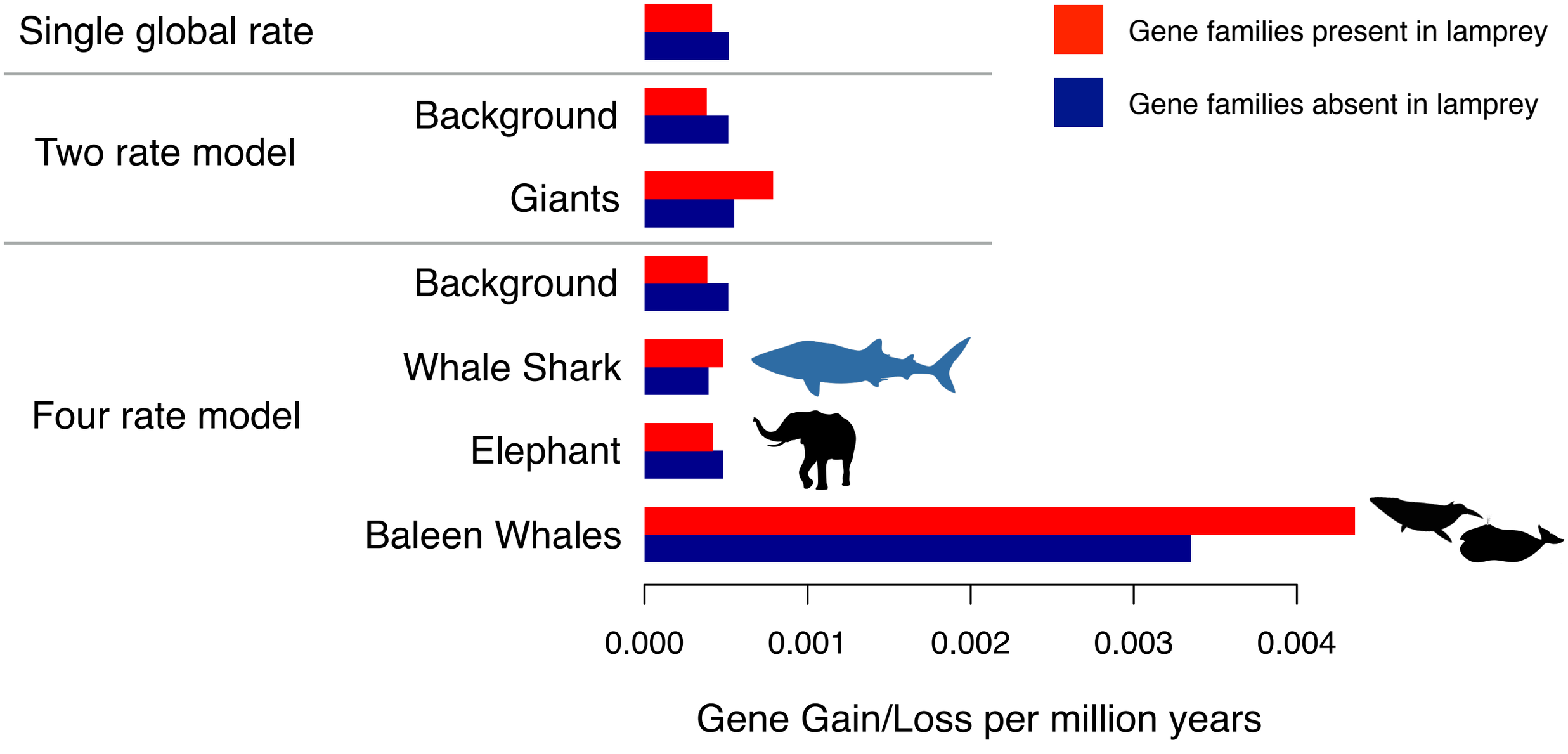
107. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).

108. Pagel, M. Test for Punctuational Evolution and the Node-Density Artifact. Available at: http://www.evolution.reading.ac.uk/pe/index.html.


Cladogram
Lancelet
Sea Squirt
Lamprey
Bony Vertebrates
Callorhinchus
Whale Shark
O
C
Total Families
14000
12000
10000
8000
6000
4000
2000
0
Total Gene Families Conserved in All Descendants
12000
10000
8000
6000
4000
2000
0
Novel Gene Families
2500
2000
1500
1000
500
0
Novel Gene Families Conserved in All Descendants
400
300
200
100
0
Total Gene Families Lost
4000
3000
2000
1000
0
Chordata
Olfactores
Vertebrata
Gnathostomata


NOD-like receptors
Whale shark
with NACHT domain
Zebrafish
Human
Outgroup
Whale shark lineage-specific NLR expansion
NLRP
NLRX1
CIITA
NLRC5
NOD2
NOD1
NLRC4
BIRC1
NWD1
NLRC3
TEP–1
FISNA Domain Containing NLR
RIG-like receptors
Whale shark
DEAD domain
Dicer
RIG–1/DDX58
LGP2/DHX58
MDA5/IFIH1
CARD domain
Outgroup Caspases
MDA5/IFIH1 CARD2
RIG–1/DDX58 CARD1
RIG–1/DDX58 CARD2
MAVS
MDA5/IFIH1 CARD1
0.5
Toll-like receptors
Mammals
Reptiles (including birds)
Amphibians
Coelacanth
Teleosts
Spotted gar
Lampreys
Callorhinchus
Whale shark
Present in whale shark genome
Present in jawed vertebrate ancestor
Present in vertebrate ancestor
Newly identified TLR family
TLR4
TLR5SL
TLR5
c
b
a
TLR13
TLR29
TLR21/29–like (TLR35)
TLR21
TLR11/12/16/19/20/26
TLR22/23
TLR15
TLR2/28
TLR24
TLR27
TLR25
TLR14/18
TLR1/6/10
TLR3
TLR9c (TLR31)
TLR9a
TLR9b (TLR30)
TLR9
TLR7/8–like (TLR34)
TLR7
TLR8
0.5


Callorhinchus
Whale Shark
Spotted Gar
Coelacanth
Clawed Frog
Anole
Chicken
Alligator
Platypus
Opossum
Armadillo
Hyrax
Elephant
Human
Mouse
Dog
Pig
Cow
Dolphin
Minke Whale
Bowhead Whale
substitutions per My
3e−04
6e−04
0.0139
length=0.13


Single global rate
Two rate model
Background
Giants
Four rate model
Background
Whale Shark
Elephant
Baleen Whales
Gene families present in lamprey
Gene families absent in lamprey
0.000
0.001
0.002
0.003
0.004
Gene Gain/Loss per million years