Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: *Rhincodon typus* Smith 1828

Timothy D. Read[1], Robert A. Petit III[1,] Sandeep J. Joseph[1], Md. Tauqeer Alam[1], M. Ryan Weil[1**] , Maida Ahmad[1], Ravila Bhimani[1], Jocelyn S. Vuong[1], Chad P. Haase[1**], D. Harry Webb[2], Alistair D.M. Dove[2]*

[1] Department of Medicine, Division of Infectious Diseases & *Department of Human Genetics, Rollins School of Public Health, Emory University, Atlanta GA 30322*

[2]*Georgia Aquarium, 225 Baker Street Atlanta GA 30313*

*Corresponding author. *Georgia Aquarium, 225 Baker Street Atlanta GA 30313+1 (404) 581 4364. adove@georgiaaquarium.org*

** Current addresses: MRW: SRA International Inc, Atlanta Georgia; CPH: Otogenetics Corp, Atlanta Georgia

Abstract

The whale shark (*Rhincodon typus*) has by far the largest body size of any elasmobranch (shark or ray) species and is therefore also the largest extant species of the paraphyletic assemblage commonly referred to as "fishes". As both a phenotypic extreme and a member of the group basal to the remaining gnathostomes, which includes all tetrapods and therefore also humans, its genome is of substantial comparative interest. Whale sharks are also listed as a "vulnerable" species on the International Union for Conservation of Nature (IUCN)'s Red List of threatened species and are of growing popularity as both a target of ecotourism and as a charismatic conservation ambassador for the pelagic ecosystem. A genome map for this species would aid in defining effective conservation units and understanding global population structure. We characterised the nuclear genome of the whale shark using next generation sequencing (454, Illumina) and *de novo* assembly and annotation methods, based on material collected from the Georgia Aquarium. The data set consisted of 878,654,233 reads, which assembled into 11,347,816 contigs and 3,606,038 scaffolds. The estimated genome size was 3.44Gb. As expected, the proteome of the whale shark was most closely related to the only other complete genome of a cartilaginous fish, the Holocephali Elephant shark. The whale shark contained a novel Toll-like-receptor protein with sequence conservation to both the TLR4 and TLR13 proteins of mammals. The data are publicly available on a Galaxy bioinformatic server (http://whaleshark.georgiaaquarium.org). This represents the first shotgun elasmobranch genome and will aid studies of molecular systematics, biogeography, genetic differentiation, and conservation genetics in this and other shark species, as well as providing comparative data for studies of evolutionary biology and immunology across the jawed vertebrate lineages.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

INTRODUCTION

The Gnathostomata, or jawed vertebrates, arose roughly halfway through the Palaeozoic era, and radiated to produce many of the groups of animals most familiar to the general public: sharks, bony fishes, amphibians, reptiles, birds and mammals, including humans. The transition from jawless to jawed vertebrates included several important adaptations that have defined the success of vertebrate life, including the adoption of antibody-based immune systems (Venkatesh et al., 2014). The extant sister group to the gnathostomes is the Agnatha or jawless fishes, represented by the hagfish and lamprey, while the most basal group among the gnathostomes is the cartilaginous fishes, consisting of the holocephalans or ratfishes, and the elasmobranchs, including all the sharks and rays. As basal gnathostomes, cartilaginous fishes are important model species for comparative studies of human evolution, including anatomy, physiology and immunology. Venkatesh and co-authors (Venkatesh et al., 2007, 2014; Davies et al., 2009; Ravi et al., 2009; Inoue et al., 2010) have explored the genomic basis of some of these adaptations in the elephant fish, *Callorhinchus milii*, a cartilaginous fish from the Holocephali, however no elasmobranch species has had a complete nuclear genome published prior to this study.

Until relatively recently, little was known about the biology of the largest shark in the world, the circum-tropical, filter-feeding whale shark, *Rhincodon typus* Smith 1828 (Colman, 1997; Martin, 2007; Stevens, 2007; Rowat & Brooks, 2012) **(Figure 1)**. Advances in tagging technology, combined with the discovery of several reliable, seasonal, near-coastal aggregations in different parts of the world (Wilson, Taylor & Pearce, 2001; de la Parra Venegas et al., 2011; Rowat & Brooks, 2012) have spurred a rapid expansion in whale shark science since 2000. These efforts have been further enhanced by the three International Whale Shark Conferences (the most recent collected at https://peerj.com/collections/3-whale-shark-conference-2013/), which have served to promote collaboration on what is otherwise a fairly intractable species to study, due to its size and oceanic habits. The maintenance of a collection of whale sharks at Georgia Aquarium has provided research opportunities not previously available in the natural setting of whale sharks, including the ability to collect samples suitable for genome sequencing. *R. typus* is an excellent model for comparative genomic study because it is a member of the basal gnathostome lineage, because it represents the phenotypic extreme in body size among sharks and fishes generally, and because it is a charismatic subject of ecotourism, yet globally vulnerable to extinction.

There are few publications on the genetics and genomics of whale sharks. Some of the first efforts at discriminating substructure in the global population were based on microsatellite (Schmidt et al., 2009) or mitochondrial control loop (Castro et al., 2007) sequences and failed to detect as much global population structure as might be expected. In a recent review incorporating natural history data, (Sequeira et al., 2013) concluded that whale sharks are part of a single global metapopulation. These studies have been contradicted by a more recent paper that found distinct genetic differences between Atlantic and Indo-Pacific whale sharks (Vignaud et al., 2014) based on additional microsatellite loci. Alam et al (Alam et al., 2014) provided the first genomic exploration of the whale shark: the complete mitochondrial genome along with a phylogenomic comparison with representative members of the other major elasmobranch orders. The number of chromosomes in the whale shark genome has not yet been ascertained.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

In this report we present the preliminary whole genome shotgun sequencing (WGS) analysis of a *R. typus* male. We hope to eventually obtain enough raw data to present a more complete genome but the current data set will be of use to researchers studying whale shark biology and the evolution of the gnathostomata.

MATERIALS & METHODS

Sample collection

The genome sequence was derived from tissue samples collected in 2007 post mortem from a male whale shark of Taiwanese origin at Georgia Aquarium. The animal was originally collected near Hualien, Taiwan in 2004 as part of a pelagic trap fishery quota.

DNA extraction, library preparation & sequencing

The genomic DNA used for this study was isolated from liver and spleen tissues using the Qiagen Maxi Prep kit (Qiagen, Venlo, Netherlands). Over the course of the project 6 sequencing libraries were constructed from this input DNA and used for 13 separate sequencing experiments ("runs") (see Table 1).

De novo assembly & annotation

After low quality reads were filtered out, the remaining reads were assembled using SOAPdenovo (v. 2.04). Assemblies were created using k-mers 31-89. Statistics for each assembly were created using a script from the Assemblathon project. K-mer 63 was chosen as the "best" assembly. This was based on 63-mers (a) producing the largest contig (86,048 bp) and (b) having a NG50 very similar to the other top scores (63-mer: 3,358bp, 65-mer: 3,454bp, and 67-mer: 3,406bp).

Whale shark proteins were predicted *de novo* on the assembled contigs using AUGUSTUS (v. 3.0.3)(Stanke et al., 2006). The proteins were matched against the NCBI nr database using BLASTP (v. 2.2.26+)(Altschul et al., 1990) with a threshold cutoff E-value of 10^{-3}, and the INTERPRO profile database using InterProScan (v5) (Quevillon et al., 2005). BLAST2GO (v3.0.7) (Götz et al., 2008) was used for functional protein annotation based on the results of these analyses. KronaTools (v2.4) (Ondov, Bergman & Phillippy, 2011) was used to create taxonomic visualizations.

Ortholog analysis

The annotated complete predicted proteomes from 11 fish genomes (see results section below) was searched against itself (all vs all) using BLASTP (v.2.2.30) with a threshold cutoff E-value of 10^{-5}. The percent identity, E-value and alignment scores were parsed out from the BLASTP output in order to compute the percent match identity, which were utilized for identifying the orthologous sequences using the OrthoMCL algorithm(Li, Stoeckert & Roos, 2003). Core genes are defined as the protein-coding gene clusters that are shared by all fish genomes used in this study. Unique genes found in only one of the fish genomes was also identified in this analysis. MUSCLE (v. 3.6) (Edgar, 2004) was used with default settings to align the core genes, and each of the protein alignments was filtered by GBLOCKS (v0.91) (Talavera & Castresana, 2007) to remove gaps and highly divergent regions.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Zebrafish proteins with orthologs missing in the whale shark were tested for functional significance using WebGestalt (update 5/20/2014) (Duncan, Prodduturi & Zhang, 2010).

Phylogenetic reconstruction

The sequence alignment of all the amino acid sequences of the core genes were concatenated together to form a super alignment in order to perform phylogenetic analysis. Maximum likelihood (ML) based phylogenetic reconstruction was implemented using RAxML (v 7.2.8-ALPHA) (Stamatakis, 2014). The Jones-Taylor-Thornton (JTT) amino acid substitution model(Taylor & Jones, 1993) of rate heterogeneity with 4 discrete rate categories was used. To evaluate statistical support, a majority rule-consensus tree of 100 bootstrap replicates was computed.

Data accession and availability

Raw data from the project is available from the National Center for Biotechnology information short read archive under accession number SRP044374. We also created a Galaxy server instance (Afgan et al., 2010) to allow researchers download contigs and predicted proteins, and perform a limited set of bioinformatic analyses on the data (http://whaleshark.georgiaaquarium.org). Most of the results section is described by public histories on this website. Additional scripts and supporting information have been placed on a public GitHub site (https://github.com/Read-Lab-Confederation/whaleshark).

RESULTS & DISCUSSION

Sequencing and assembly

We generated *R. typus* sequencing libraries using 454 and Illumina of technologies, obtaining a total of ~ 278Gb bases of unassembled sequence (**Table 2** & Methods). Reference-free analysis of the quality filtered data using the *preqc* tool (v. 0.10.13) (Simpson, 2013) gave us an estimate of the genome size based on k-mer word frequency of 3.44Gb, within the range reported size of other chondrichthyes (Gregory, 2005; Gregory & Witt, 2008). We estimated that we had approximately 30-fold redundancy of in coverage of the genome. The DNA composition of the assembled contigs was 42% G+C.

We performed de novo assembly using the SOAPdenovo (Li et al., 2010) program. A range of different k-mer values were tried for the de Bruijn graph building step of the algorithm but we settled on using 63-mers because this setting yielded the fewest contigs with the largest N50 value. The assembly that we used for downstream analysis consisted of 11,347,816 contigs and 3,606,038 scaffolds.

The rather low N50 compared to other other recent vertebrate genome projects suggested that the assembly could benefit from more mate-pair and long read sequences, as well as deeper coverage of Illumina sequence to help correct sequence. The assembly incorporated an Illumina mate-pair library of approximately 3 kb. Attempts to construct larger insert mate-pair libraries resulted in failure. We are currently working on generating 5-10 fold genome coverage using the long-read Pacific Bioscience SMRT technology to complement the current assembly.

When we matched the CEGMA collection of ultra-conserved vertebrate proteins (CEGs)(Parra, Bradnam & Korf, 2007) against the contigs, we found that only 17% of the 248 CEGs had a complete

match, but ~57% had a partial hit. This result likely reflected the fragmented nature of the assembly at this stage.

Sequence contamination is an issue that has bedeviled WGS projects(Merchant, Wood & Salzberg, 2014). We therefore expected to see non-whale shark DNA originating from carryover from previous Illumina runs, and contamination from extrinsic laboratory sources during tissue preparation, the latter especially since the *R. typus* diet may contain unusually high levels of bacteria (Rohner et al., 2013). To determine the approximate extent of this issue, we used BLAST to compare the assemblies to the highly conserved bacterial 16S gene and found only four contigs with low sequence coverage (5-7 fold redundancy) had greater than 75% matches to the whole gene. Therefore, we concluded that bacterial contamination was present but not a major factor in this project.

Immediately prior to the release of these data (December 2014) there were only 110 nucleotide sequences in the NCBI database assigned a *R. typus* taxonomic origin. 109/110 of these sequences could be mapped to the contigs from the this project with a threshold match significance BLAST score of 10^{-5} or lower. The one sequence that did not match was a putative recombination activating protein 2 ortholog (NCBI gid:315571864) that turned out to have best matches only to other bony fishes and thus may have been misidentified in its origin.

Predicted proteins

We used the AUGUSTUS software (Stanke et al., 2006) for *de novo* prediction of 23,594 protein-coding genes on the WGS contigs. While the largest predicted protein was 4,709 amino acids in length, the majority of the proteins were less than 200 amino acids **(Figure 2)**. Of the predicted proteins, 16,413 (70%) of the proteins had a blastp match in the NCBI nr database. More than 99% of the protein best matches were to eukaryotes (**Figure 3**), providing further evidence that prokaryotic contamination in the project was limited. Within the eukaryotes, 87% of the matches were to Chordata with other fish species that have completed genomes predominant (**Figure 4**). The genome with the greater number of best matches (30% of Chordata) was the Elephant shark. These results therefore were in line with what would be expected of a novel Chondrichthyes genome sequence.

Ortholog analysis

In order to investigate orthologs patterns we compared the predicted *R. typus* proteome against proteomes from 10 other fishes and lamprey using BLASTP with a cutoff E-value of 10^{-5} and clustered into groups related by sequence similarity with the ORTHOMCL software pipeline. The predicted proteomes of atlantic cod (*Gadus morhua*) (Star et al., 2011), atlantic salmon (*Salmo salar*)(Davidson, 2013), coelacanth (*Latimeria chalumnae*)(Amemiya et al., 2013), fugu (*Takifugu ruprides*)(Aparicio et al., 2002), elephant shark (*Callorhinchus milii)(Venkatesh et al., 2014),* sea lamprey (*Petromyzon marinus*)(Smith et al., 2013), medaka (*Oryzias latipes*)(Kasahara et al., 2007), nile tilapia (*Oreochromis niloticus*)(Guyon et al., 2012), stickleback (*Gasterosteus aculeatus*)(Jones et al., 2012), green spotted pufferfish (*Tetraodon nigroviridis) (Jaillon et al., 2004)*, and zebrafish (*Danio rerio*)(Howe et al., 2013) were downloaded from the UCSC genome browser site (Karolchik, Hinrichs & Kent, 2009) in November 2014. We found that there was a 'core' set of 1846 ortholog groups with at

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

least one protein member present in each of the eleven genomes, representing the set of highly conserved functions. Of these genes, 155 orthologs were present with exactly one protein member in each of the groups. We concatenated and aligned these proteins and produced a maximum likelihood tree (**Figure 5**), which recapitulated the evolutionary relationship of the species: *R. typus* and *C. milii* forming a deep clade that diverged before the evolution of bony fish.

The ortholog analysis revealed that there were 865 protein families present in the other 11 genomes, that were missing in the whale shark. This number was of the same order as the outgroup lamprey genome (764 missing orthologs) and higher than that seen in the other fishes (the elephant shark genome had only 108 missing protein sequences). Further, there were 543 proteins missing from both the whale shark and lamprey but represented in all the other 9 genomes. These absent proteins could be explained by the draft nature of the sequence data in this project, the preliminary de novo annotation and/or the evolutionary divergence of the whale shark and lamprey from the other species. We mapped the orthologs of the missing proteins in the well-annotated zebrafish genome and tested for enrichment of terms in the Gene Ontology or KEGG (Kyoto Encyclopedia of Genes and Genomes) databases using the WebGestalt GSAT analysis tool (Duncan et al., 2010). We found no specifically enriched terms or pathways in the missing protein set compared to the entire zebrafish proteome. This suggested that the absent genes were not overrepresented in any particular functional category, as might have happened through adaptive gene deletion.

The remaining 7,181 predicted proteins with no nr database match tended to be short (mean of 123 amino acids, compared to 175 for the protein dataset as a whole), suggesting that many were annotation overcalls, or fragments of proteins disrupted by contig gaps. The GC% of the 7,182 non-matching genes (44%) was similar to the gene set as a whole (41.8%). Several of these proteins are large enough that they are unlikely to be the result of spurious translation (19 were > 500 amino acids in length, the largest 1352 amino acids). These could represent novel *Chondrichthyes* functions although it is also formally possible that many of the proteins without a best match could be uncultivated microorganisms.

Preliminary comparison with Callorhinchus milii

The only other cartilaginous fish for which a complete genome has been assembled is the elephant fish *C. milii (Venkatesh et al., 2007, 2014; Davies et al., 2009)*, which is not an elasmobranch but a member of the Holocephali or ratfishes. There are striking differences between the genomes, most obviously in size. The whale shark genome, at 3.44Gb, is approximately 3.5x the size of the elephant fish genome at only 950Mb. The genomes were also diverged at the DNA level. In a discontiguous megablast alignment between the *C. milii* and whale shark scaffolds the combined length of matches with an E value of < 0.001 was only 42Mb of the Elephant Shark genome (71% nucleotide identity). Some of the features of the protein set of *R. typus* recapitulated discoveries made in *C. milii*. For example, homologs of the human SCP and SIBLING proline-glutamine families of bone-deposition proteins were missing from the whale shark genome (negative results of BLASTX alignment against the scaffolds); a result also seen in the other cartilaginous fish (Venkatesh et al., 2014). *C. milii* is reported to be missing a homolog of the important innate immunity protein TLR4 (Toll-Like Receptor 4), which detects lipopolysaccharide of infecting Gram negative bacteria (Venkatesh et al., 2014). We found that the human TLR protein had a significant match (BLASTP 1e-45) to a 925 amino acid protein

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

containing multiple leucine-rich repeat domains and a C-terminal TIR domain (Toll/Interleukin receptor) of the nucleotide-binding TLR2 superfamily. Interestingly, the was a much better match to the TLR13 protein from rodents (best match was to *Ictidomys tridecemli*, the North American thirteen-lined ground squirrel)(**Figure 6**). TLR13 is a recently discovered Toll-like receptor from rodents and bats that recognizes the 23S subunit of bacterial ribosomal RNA (Oldenburg et al., 2012). The whale shark protein is either an ancient homolog to the rodent TLR13 that was long ago lost from from other fish, or instead a diverged TLR with possibly novel pathogen-recognition function specific to elasmobranchs.

CONCLUSIONS

Given limited funding, we pursued a strategy of primarily using cost-effective Illumina short read sequencing to produce a preliminary *R. typus* genomic dataset. This allowed us to maximize coverage of the genome with high quality data and give estimates of the genome size and extent of bacterial contamination of the source DNA (both unknown at the start of the project), and to provide what we believe is a quite complete, if fragmented, draft of the genome. *De novo* gene prediction and comparisons with other fish genomes suggest the gene content and phylogenetic relationships of the proteins were what would generally have been expected of a cartilaginous fish. If we can obtain funding for the next stage of the work we aim to use a combination of up to 10x coverage with long reads using the Pacific Biosciences SMRT technology as well as deeper coverage with Illumina in a hybrid assembly to produce a genome with fewer contigs. We will also look to using RNA-seq data to aid gene annotation but we will likely have to rely on extraction from archived tissues due to the technical and ethical constraints on obtaining samples from live animals.

The genome sequence of an organism is now perhaps the single most important gateway to understanding its biology. We believe that despite the incomplete nature of the data, the draft sequence presented here will be a resource that can accelerate scientific investigation of the whale shark and of elasmobranchs in general. We have shown that the data encompasses almost all the current publicly-submitted whale shark nucleotide sequences although many genes are likely split over two or more contigs, and the large number of putatively 'missing' proteins probably reflects this reality in the draft sequence. Some caution should therefore be used when concluding that a protein homolog is "missing" from these data. Nevertheless, the current DNA sequence can be mined for new genotyping tools for populations genomics and the protein set can be compared intensively against known functions. The long term goals include understanding the genetic nature of the large body size of the whale shark, its metabolic adaptations to its planktonic diet and the evolution of its immune system in a comparative context within the gnathostomes.

This public data set is not only for research but can also be a teaching tool. We used an intramural version of the Galaxy server in a basic bioinformatics analysis course for undergraduates at Emory University (three of whom are on this author list). Students were inspired to improve their bioinformatic skills by the opportunity to explore the vast dataset of this wonderful organism. There are surely many important discoveries that will come from further careful analysis of the genome sequence.

FUNDING

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

The major funding from this project came from the Georgia Aquarium, with additional resources provided by Division of Infectious Diseases development funds to TDR. Coca Cola Inc. contributed towards establishing the Galaxy web server. Funding for equipment used at the Emory Genome Center was provided by the Georgia Research Alliance, Emory School of Medicine, Department of Human Genetics and the Atlanta Clinical and Translational Sciences Institute.

The funders played no role in the scientific direction of the study or writing of the manuscript.

ACKNOWLEDGMENTS

We gratefully acknowledge the help and support of current and past members of the zoological operations and veterinary care teams at Georgia Aquarium, particularly Dr. Tim Mullican, Dr. Greg Bossart, Chris Coco, Chris Schreiber, Dr. Tonya Clauss, Helen Ellis, Tim Binder, Ray Davis and Dr. Bruce Carlson. We also wish to acknowledge the valuable input of Jessica Peterson, Megan Cole, and Karin Fredrikson. Special thanks to Lex Nederbragt for serving as unofficial peer-reviewer.

Sequence data was generated at the Emory Genome Center, Hudson Alpha Institute and 454 Inc.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

TABLES

Table 1. Sequencing runs and libraries used in this study

SRA ID	Tissue	Library ID	Technology	Type*	Ave insert size (std dev)	Sequence length (bp)	Number of reads	Total bp
SRR1521182	Spleen	1	LS454	SE	na	40-1304	1,268,373	728,329,555
SRR1521183	Spleen	1	LS454	SE	na	40-1323	1,323,602	718,846,097
SRR1521184	Spleen	1	LS454	SE	na	40-1328	1,279,760	680,625,037
SRR1521191	Spleen	2	Illumina	PE	293(101)	100	210,821,824	21,082,182,400
SRR1521192	Spleen	2	Illumina	PE	300(91)	100	585,821,484	58,582,148,400
SRR1521195	Spleen	2	Illumina	PE	328(90)	100	585,054,464	58,505,446,400
SRR1521197	Spleen	2	Illumina	PE	286(100)	100	224,670,734	22,467,073,400
SRR1521198	Spleen	3	Illumina	MP	7161(755)	100	571,738,680	57,173,868,000
SRR1521199	Spleen	2	Illumina	PE	290(100)	100	300,519,032	30,051,903,200
SRR1521200	Spleen	4	Illumina	SE	na	51	108,403,623	5,420,181,150
SRR1521201	Spleen	5	Illumina	PE	274(54)	100	34,239,020	3,423,902,000
SRR1521204	Spleen	5	Illumina	PE	236(46)	100	90,708,094	9,070,809,400
SRR1521190	Liver	6	Illumina	PE	215(43)	100	99,078,844	9,907,874,400

- SE - single end; PE paired end; MP mate pair

Table 2. Features of the assembled whale shark genome

Total size of contigs	3,501,951,163
Number of contigs	11,347,816
Total size of scaffolds	2,934,639,008
Number of scaffolds	3,606,038
N50 scaffold sizes	4,554
%G+C nucleotides in contigs	42
Number of predicted proteins	23,594
Mean length proteins	175
Median length proteins	135

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Afgan E, Baker D, Coraor N, Chapman B, Nekrutenko A, Taylor J. 2010. Galaxy CloudMan: delivering cloud compute clusters. *BMC bioinformatics* 11 Suppl 12:S4.

Alam MT, Petit RA 3rd, Read TD, Dove ADM. 2014. 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.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of molecular biology* 215:403–410.

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. *Nature* 496:311–316.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, Dehal P, Christoffels A, Rash S, Hoon S, Smit A et al. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297:1301–1310.

Castro ALF, Stewart BS, Wilson SG, Hueter RE, Meekan MG, Motta PJ, Bowen BW, Karl SA. 2007. Population genetic structure of Earth's largest fish, the whale shark (Rhincodon typus). *Molecular ecology* 16:5183–5192.

Colman JG. 1997. A review of the biology and ecology of the whale shark. *Journal of fish biology*.

Davidson WS. 2013. Understanding salmonid biology from the Atlantic salmon genome. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* 56:548–550.

Davies WL, Carvalho LS, Tay B-H, Brenner S, Hunt DM, Venkatesh B. 2009. Into the blue: gene duplication and loss underlie color vision adaptations in a deep-sea chimaera, the elephant shark Callorhinchus milii. *Genome research* 19:415–426.

Duncan D, Prodduturi N, Zhang B. 2010. WebGestalt2: an updated and expanded version of the

Web-based Gene Set Analysis Toolkit. *BMC bioinformatics* 11:P10.

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

Gregory TR. 2005. Genome size evolution in animals. *The evolution of the genome*.

Gregory TR, Witt JDS. 2008. Population size and genome size in fishes: a closer look. *Genome / National Research Council Canada = Genome / Conseil national de recherches Canada* 51:309–313.

Guyon R, Rakotomanga M, Azzouzi N, Coutanceau JP, Bonillo C, D'Cotta H, Pepey E, Soler L, Rodier-Goud M, D'Hont A et al. 2012. A high-resolution map of the Nile tilapia genome: a resource for studying cichlids and other percomorphs. *BMC genomics* 13:222.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 36:3420–3435.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L et al. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496:498–503.

Inoue JG, Miya M, Lam K, Tay B-H, Danks JA, Bell J, Walker TI, Venkatesh B. 2010. Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. *Molecular biology and evolution* 27:2576–2586.

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

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

484:55–61.

Karolchik D, Hinrichs AS, Kent WJ. 2009. The UCSC Genome Browser. *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* Chapter 1:Unit1.4.

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

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13:2178–2189.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K et al. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research* 20:265–272.

Martin RA. 2007. A review of behavioural ecology of whale sharks (Rhincodon typus). *Fisheries research* 84:10–16.

Merchant S, Wood DE, Salzberg SL. 2014. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2:e675.

Oldenburg M, Krüger A, Ferstl R, Kaufmann A, Nees G, Sigmund A, Bathke B, Lauterbach H, Suter M, Dreher S et al. 2012. TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. *Science* 337:1111–1115.

Ondov BD, Bergman NH, Phillippy AM. 2011. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics* 12:385.

Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067.

De la Parra Venegas R, Hueter R, González Cano J, Tyminski J, Gregorio Remolina J, Maslanka M, Ormos A, Weigt L, Carlson B, Dove A. 2011. An unprecedented aggregation of whale sharks,

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Rhincodon typus, in Mexican coastal waters of the Caribbean Sea. *PloS one* 6:e18994.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. *Nucleic acids research* 33:W116–20.

Ravi V, Lam K, Tay B-H, Tay A, Brenner S, Venkatesh B. 2009. Elephant shark (Callorhinchus milii) provides insights into the evolution of Hox gene clusters in gnathostomes. *Proceedings of the National Academy of Sciences of the United States of America* 106:16327–16332.

Rohner CA, Couturier L, Richardson AJ, Pierce SJ, Prebble C, Gibbons MJ, Nichols PD. 2013. Diet of whale sharks Rhincodon typus inferred from stomach content and signature fatty acid analyses. *Marine ecology progress series* 493:219–235.

Rowat D, Brooks KS. 2012. A review of the biology, fisheries and conservation of the whale shark Rhincodon typus. *Journal of fish biology* 80:1019–1056.

Schmidt JV, Schmidt CL, Ozer F, Ernst RE, Feldheim KA, Ashley MV, Levine M. 2009. Low genetic differentiation across three major ocean populations of the whale shark, Rhincodon typus. *PloS one* 4:e4988.

Sequeira AMM, Mellin C, Meekan MG, Sims DW, Bradshaw CJA. 2013. Inferred global connectivity of whale shark Rhincodon typus populations. *Journal of fish biology* 82:367–389.

Simpson JT. 2013. Exploring Genome Characteristics and Sequence Quality Without a Reference. *arXiv [q-bio.GN]*.

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE et al. 2013. Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nature genetics* 45:415–21, 421e1–2.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio

prediction of alternative transcripts. *Nucleic acids research* 34:W435–9.

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. *Nature* 477:207–210.

Stevens JD. 2007. Whale shark (Rhincodon typus) biology and ecology: A review of the primary literature. *Fisheries research* 84:4–9.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic biology* 56:564–577.

Taylor WR, Jones DT. 1993. Deriving an amino acid distance matrix. *Journal of theoretical biology* 164:65–83.

Venkatesh B, Kirkness EF, Loh Y-H, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC et al. 2007. Survey sequencing and comparative analysis of the elephant shark (Callorhinchus milii) genome. *PLoS biology* 5:e101.

Venkatesh B, Lee AP, Ravi V, Maurya AK, Lian MM, Swann JB, Ohta Y, Flajnik MF, Sutoh Y, Kasahara M et al. 2014. Elephant shark genome provides unique insights into gnathostome evolution. *Nature* 505:174–179.

Vignaud TM, Maynard JA, Leblois R, Meekan MG, Vázquez-Juárez R, Ramírez-Macías D, Pierce SJ, Rowat D, Berumen ML, Beeravolu C et al. 2014. Genetic structure of populations of whale sharks among ocean basins and evidence for their historic rise and recent decline. *Molecular ecology* 23:2590–2601.

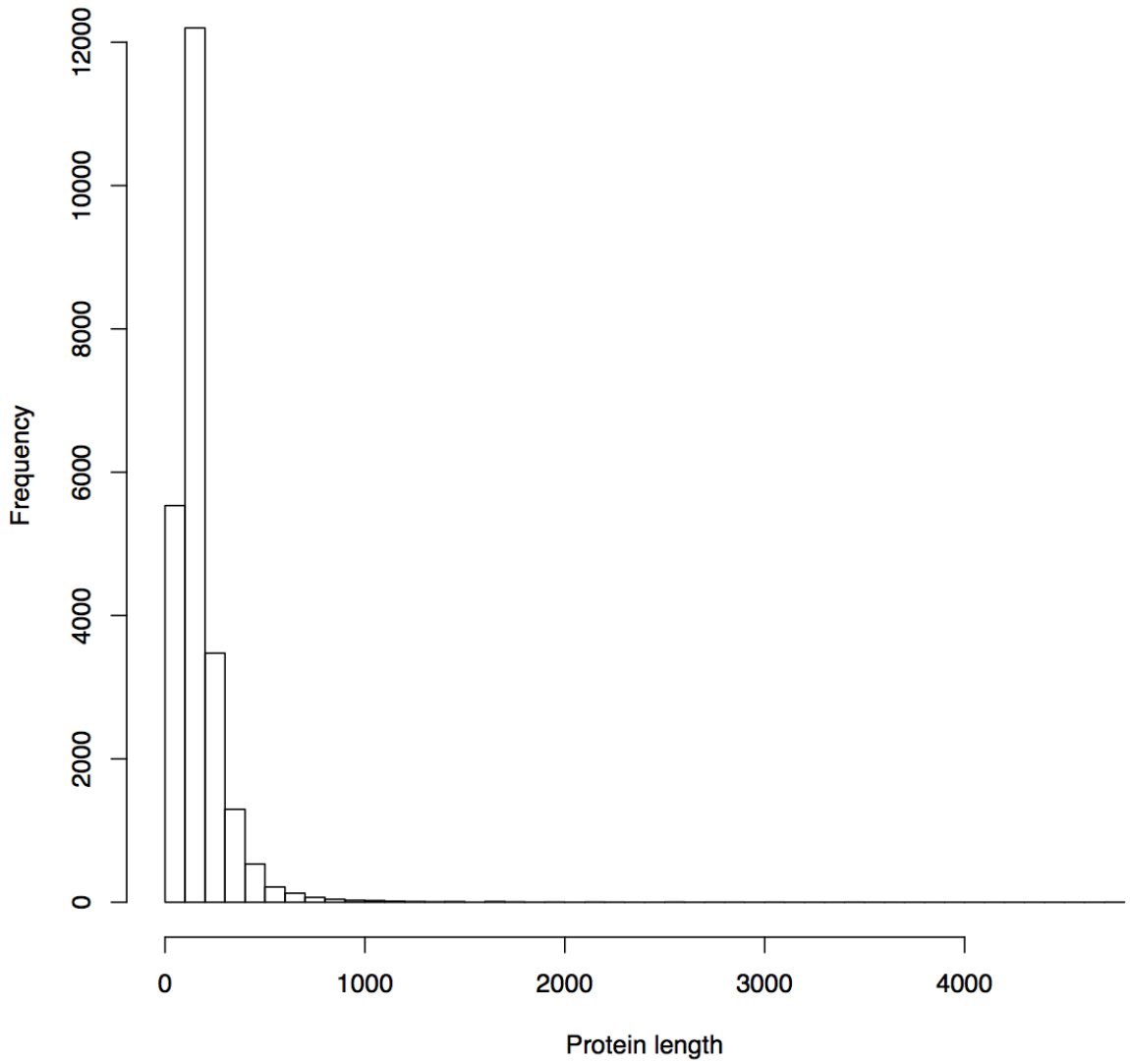
Wilson SG, Taylor JG, Pearce AF. 2001. The seasonal aggregation of whale sharks at Ningaloo Reef, Western Australia: currents, migrations and the El Nino/Southern Oscillation. *Environmental biology of fishes*.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Figure 1 Whale shark (*Rhincodon typus*) in the Gulf of Mexico with a *Homo sapiens* for size comparison (Photo credit: Georgia Aquarium. Rights free use permitted).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Figure 2. Histogram of predicted protein sizes



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Figure 3. Overview of taxonomy of whale shark protein best matches to the nr database. Figure was constructed from best BLAST matches to the nr database using the Krona (Ondov et al., 2011)tool.

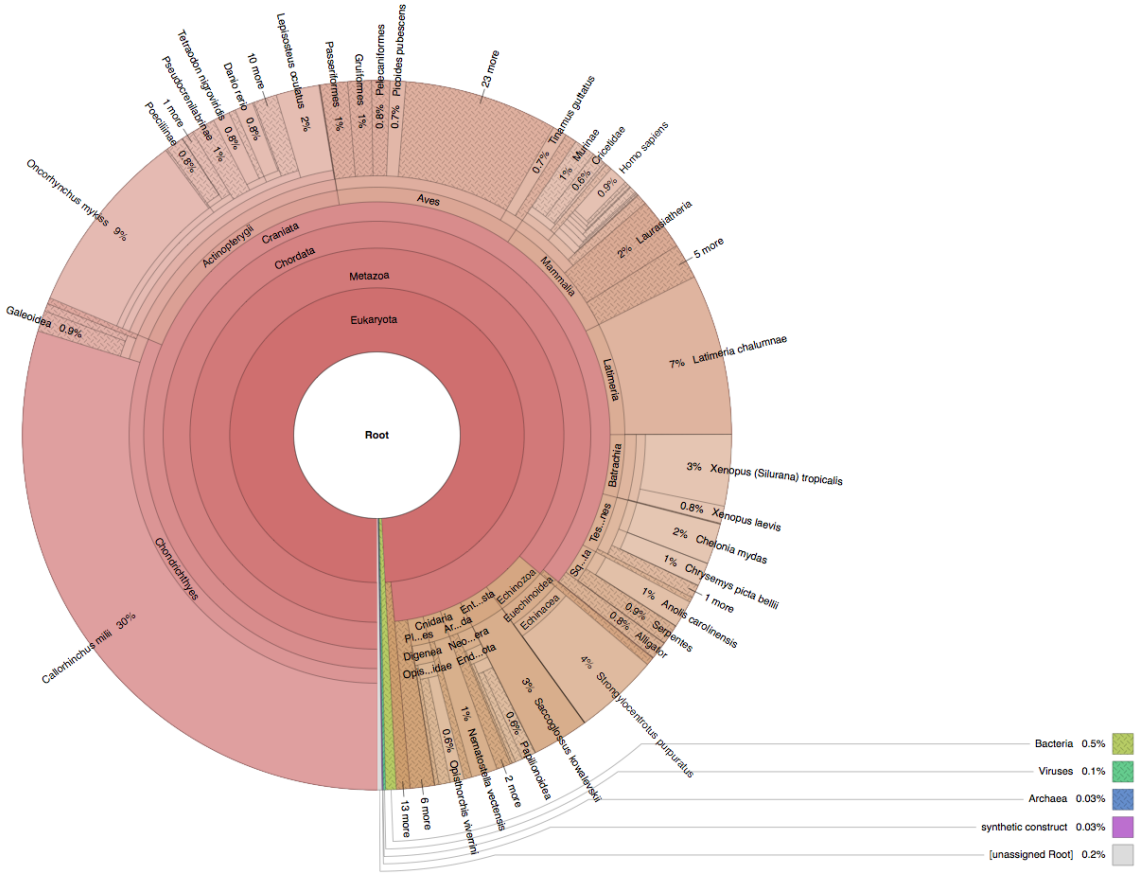
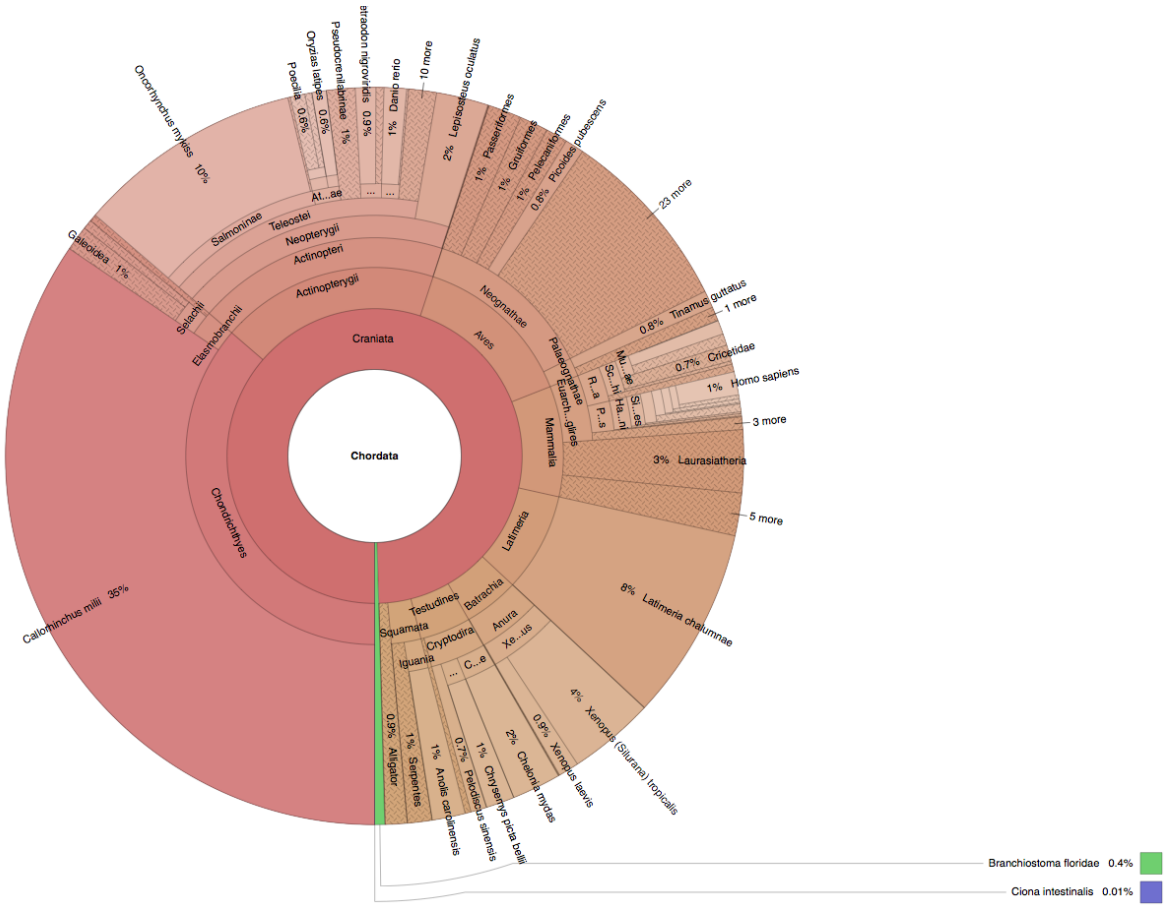
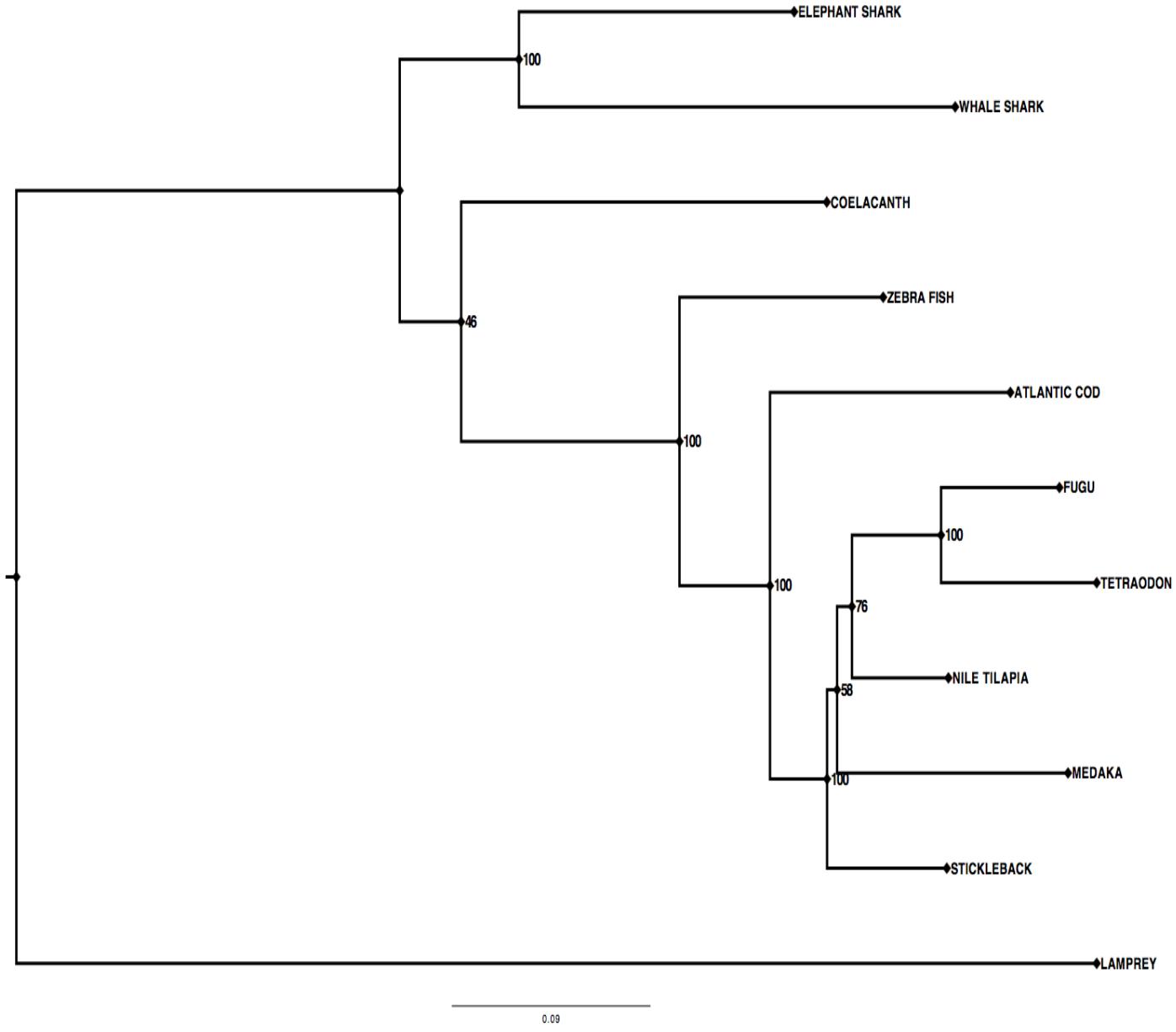


Figure 4. Overview of best matches to the protein database that map to the Chordata taxonomy group



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Figure 5 Phylogeny based on alignment of conserved core proteins.



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015

Figure 6. Top blast alignment to the putative whale shark TLR13/4 protein

PREDICTED: toll-like receptor 13-like [Ictidomys tridecemlineatus]

Sequence ID: gi|532103931|ref|XP_005337562.1| Length: 950 Number of Matches: 1

Range 1: 19 to 929 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
479 bits(1233)	6e-151	Compositional matrix adjust.	330/929(36%)	500/929(53%)	39/929(4%)

```
Query  18   IALAPLVPGYSFRNCIQNANHNGSFQCVHRFLTQVSSAVFDLPPHTIHLNISHNLLTRLP  77
            + L  LV  Y FR C Q   +     C+ + +T ++ A+ D+P +T HLN++ N +  LP
Sbjct  19   LCLLSLVESYGFRKCTQYELNIRHVFCIRKEITNLTDAIGDIPRYTTHLNLTQNQIQVLP  78

Query  78   VGSFSGIPSLLSLRLDYNQIKHVKGGAFSNLSQLLVLNLSYNKIPQLTPSDFLGLENLCQ  137
              SF+ + +L+ LRL++N I  +  GAF  L  L  LNL  NKI  +  S F GL NL
Sbjct  79   PHSFANLSALVDLRLEWNLIWKIDKGAFKELGNLTFLNLVENKIESVNNS-FEGLSNLET  137

Query  138  LLVDHNRLATVQPDAFTTPRSLQTLDMSFNELRNFSGVVSSIAALRKLTFLDLSTNHLES  197
            LL+ HN +  +   AF     L+ L +S N + NFS ++ ++  L  L +LDL+ N++ S
Sbjct  138  LLLSHNFITYIHKTAFVPLVKLKHLSLSRNHITNFSNILEAVQQLPCLEYLDLTNNNIIS  197

Query  198  LQHSAPLPPSLQFLYLCNNLLKALDCQRDFLNRNFTLDLSDNKISQFSKVDLRKVAWL--  255
            L H+     SL  L L  N LK L+     L    TLD+S N+      VDL  +  L
Sbjct  198  LDHNPRSLVSLTHLSLQGNKLKELNFSALSLPNLTTLDVSQNRHQVIQNVDLETLPQLRS  257

Query  256  -----ILKNNPLDISEFLR-FRNVDPQRVEYSGLRLHSQFANLCQHLGN-GTMERLLLQN  308
                 L    + ++++L+  R +D  +++  G   H     +C  L N  T+E L+ Q
Sbjct  258  LNLSGTLVKLEMLLAKYLQNLREMDITKLDLRGG--HLNLNTVCCLLKNLPTLEALVFQK  315

Query  309  NQISLHNLKSMSKCPAIKAWDLSHN---YLIRDDCLDFVPRKEQVRSFILEHNRIQMLSS  365
            N +   ++K ++ C  + + DLS N     + D+    +P  +      L  N+ Q LS
Sbjct  316  NAMDAGDIKHLANCTRLLSLDLSQNSDLVYLNDNEFVAMPSLQS-----LHLNKCQ-LSF  369

Query  366  CSKGGNGVIQFPKLTYLSYRYNRILAIASHAFSHTAYVQTLLLNINNIAVINKTAFANLS  425
             S      +Q   LT L   +N+  +    AFS    +Q+L L+ N I  +N  AF  L
Sbjct  370  VSNKTWSSLQ--NLTALYLSHNKFKSFPDFAFSPLKGLQSLFLSKNPITELNHMAFHGLD  427

Query  426  ELLTLRLDNNLITDIYQETFRELTSLRTLNLRNNRVSIIFRNVFVNLSSLDILDLGGNKI  485
             L  L L    I  I   +F  LT+L +L+LR N +  + +  F  L  L +L  N++
Sbjct  428  LLKELNLAECWIVMIDSSSFVHLTNLESLDLRLNNIHTLKQRTFQFLKKLKVLILSRNRL  487

Query  486  RSLTNLSFSGLHNLTKLYLDRNCITHISREIFGDLVSLKVLDLAKNWIRYNSGLT-EKSP  544
              +   +F GL +L  L L  N ++    ++F  L +L+VL+L+ N I Y +  T +  P
Sbjct  488  EVIEENAFFGLTHLYYLDLAYNSLSGFHLKLFLGLENLEVLNLSYNRITYETTRTLQFPP  547

Query  545  FINLTKLNILKLQAQQPYGINIIPPKFFKGLISLQALYLGENKMS-LSKYEFEDLINLKT  603
            F NL  L  L L+ Q  +GI ++P  FF+GL  LQ L+LG+N M  L   +F+ L+NL
Sbjct  548  FKNLKSLKQLNLEGQN-HGIQVVPTNFFQGLNCLQELFLGKNSMIFLDHLQFDPLVNLTK  606

Query  604  LSMPDTCNGIHSL--NPGVFQKLQHLQRLNLENVGLQFMSVDIFGSLSNLRTLVLGKNAI  661
            L +  T  G  SL  N  +FQKL+ L+ L LEN  ++ ++  +F  L +L+  L  N +
Sbjct  607  LDISGTKAGDRSLYLNTSLFQKLKRLKMLRLENNNMESLTPGMFSGLESLQVFSLRFNNL  666

Query  662  QTVNSTVLEHLSSLSYLDLRKNPFPCICSNSWFQNWCLSNPRVQVVYFYNQTCANQQTE-  720
            + +N + LE+L SL Y DL  N   C C N WF+NW ++   V + Y  +  C    T+
Sbjct  667  KVINQSHLENLKSLMYFDLYGNKLQCNCDNMWFKNWSINTAVVHIPYLGSYPCHQPDTQS  726

Query  721  YLYRFDARVCYLDIGHLMF----EIILPFLLLFTFAPIIYGKGYWHIKYGLYIFRSWLND  776
             L  FD  +C  D+G + F      ++L  ++   F+  I     W   YGLYIFR+W
Sbjct  727  LLIDFDDAMCNFDLGKIYFFSSFSLVLTTMVFSWFSAKIIS-SLW---YGLYIFRAWYLA  782

Query  777  YRGREESQQCYRYDAFISYNSNDERWVLQELVPNLETEGSQCFKLCLHHRDFELGKYIID  836
               R E +  + YDAF+S+ + DE+WV +ELVP LE  G   FKLCLHHRDFE G  I +
Sbjct  783  KWHRTEKE--FIYDAFVSFTATDEQWVYEELVPALEDGGQPKFKLCLHHRDFEPGMDIFE  840

Query  837  NIVDSIYQSRKTICVMSRNYLESEWCSMELELASYRLFHELKDVVILIFLEKIPEAELST  896
            NI ++I  SRKT+CV+S +YL SEWC +E++LAS ++F+E +DV+ILIFLE+IP  +LS+
Sbjct  841  NIQNAIDTSRKTLCVVSNHYLHSEWCRLEVQLASIKMFYEHEDVIILIFLEEIPNYKLSS  900

Query  897  YHKMRKVTKKKTYIQWPTDVEAQKLFWIK  925
            YH++RK+  ++T+I WP +V  + LFW +
Sbjct  901  YHRLRKLVNRQTFITWPDNVHERPLFWAR  929
```


PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.837v1 | CC-BY 4.0 Open Access | rec: 14 Feb 2015, publ: 14 Feb 2015