BMC Genomics

Photo courtesy of Michael Scholl - Save Our Seas Foundation

Characterization of the heart transcriptome of the white shark (*Carcharodon carcharias*)

Richards *et al.*

BioMed Central

Richards *et al. BMC Genomics* 2013, **14**:697
http://www.biomedcentral.com/1471-2164/14/697

BMC Genomics

RESEARCH ARTICLE Open Access

Characterization of the heart transcriptome of the white shark (*Carcharodon carcharias*)

Vincent P Richards[1], Haruo Suzuki[1,3], Michael J Stanhope[1*] and Mahmood S Shivji[2*]

Abstract

Background: The white shark (*Carcharodon carcharias*) is a globally distributed, apex predator possessing physical, physiological, and behavioral traits that have garnered it significant public attention. In addition to interest in the genetic basis of its form and function, as a representative of the oldest extant jawed vertebrate lineage, white sharks are also of conservation concern due to their small population size and threat from overfishing. Despite this, surprisingly little is known about the biology of white sharks, and genomic resources are unavailable. To address this deficit, we combined Roche-454 and Illumina sequencing technologies to characterize the first transciptome of any tissue for this species.

Results: From white shark heart cDNA we generated 665,399 Roche 454 reads (median length 387-bp) that were assembled into 141,626 contigs (mean length 503-bp). We also generated 78,566,588 Illumina reads, which we aligned to the 454 contigs producing 105,014 454/Illumina consensus sequences. To these, we added 3,432 non-singleton 454 contigs. By comparing these sequences to the UniProtKB/Swiss-Prot database we were able to annotate 21,019 translated open reading frames (ORFs) of ≥ 20 amino acids. Of these, 19,277 were additionally assigned Gene Ontology (GO) functional annotations. While acknowledging the limitations of our single tissue transcriptome, Fisher tests showed the white shark transcriptome to be significantly enriched for numerous metabolic GO terms compared to the zebra fish and human transcriptomes, with white shark showing more similarity to human than to zebra fish (i.e. fewer terms were significantly different). We also compared the transcriptome to other available elasmobranch sequences, for signatures of positive selection and identified several genes of putative adaptive significance on the white shark lineage. The white shark transcriptome also contained 8,404 microsatellites (dinucleotide, trinucleotide, or tetranucleotide motifs ≥ five perfect repeats). Detailed characterization of these microsatellites showed that ORFs with trinucleotide repeats, were significantly enriched for transcription regulatory roles and that trinucleotide frequency within ORFs was lower than for a wide range of taxonomic groups including other vertebrates.

Conclusion: The white shark heart transcriptome represents a valuable resource for future elasmobranch functional and comparative genomic studies, as well as for population and other biological studies vital for effective conservation of this globally vulnerable species.

Keywords: White shark, *Carcharodon carcharias*, Heart transcriptome, Microsatellites, Positive selection, Enrichment

* Correspondence: mjs297@cornell.edu; mahmood@nova.edu
[1]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA
[2]Save Our Seas Shark Research Center and Guy Harvey Research Institute, Nova Southeastern University, 8000 North Ocean Drive, Dania Beach, FL 33004, USA
Full list of author information is available at the end of the article

Background

Cartilaginous fishes (Class Chondrichthyes: sharks, skates, rays, chimaeras) provide a notable example of successful evolutionary perseverance, with a fossil record extending to at least the Lower Devonian over 400 million years ago [1]. Given their extraordinary evolutionary history and basal phylogenetic origin relative to other jawed vertebrates, chondrichthyians have been proposed as an important comparative model for understanding vertebrate genome evolution in general and various specific evolutionary and mechanistic aspects of vertebrate development, physiology and immune function [2-5].

One group of chondrichthyians, the modern sharks (subclass Elasmobranchii), comprise over 500 extant species displaying an impressive diversity of form and function, including a broad spectrum of sizes (e.g. 20-1200 cm as adults), functional morphologies (e.g. fusiform heads to the novel, widened heads of hammerhead sharks), physiology (e.g. ectothermy to regional endothermy), reproduction (e.g. egg laying to live births) and habitat use (marine to freshwater; shallow waters to abyssal depths). Sharks have also become a major target of human exploitation for their fins [6], resulting in widespread concerns that their rapidly declining populations coupled with unique life history characteristics will not permit recovery if ongoing exploitation rates continue.

Despite representing a major vertebrate lineage of evolutionary uniqueness and ecological and conservation importance, sharks remain the least explored vertebrate group at the genome level. The handful of genome level studies conducted on sharks have already revealed some distinctive features, including the absence of the *HoxC* cluster of developmental pattern genes found in all other non-elasmobranch vertebrate lineages [7], and the presence of a substantial number of expressed sequence tags for which no homologues in other organisms could be identified [8]. These apparent distinctions hint that other genomic novelties are possible in this lineage and await discovery.

The white shark, *Carcharodon carcharias* (Lamnidae), a large apex predator, is one of the highest-profile marine species, capturing extraordinary attention from the public and media. Although it demonstrates a cosmopolitan distribution, the species is believed to have a low abundance throughout its range, leading to international concerns about its conservation (IUCN Red List Category: Vulnerable A2cd+3cd) in the face of known market utilization for its body parts and widespread shark overfishing practices [9-11]. Arguably, the white shark may be a “poster child” for marine, large animal conservation attention. The white shark also possesses some notable physical and physiological characteristics that make it an interesting biological study, including an estimated genome size (C-value = 6.45 pg) nearly twice that of humans, large adult sizes reaching up to ~6 m in length, a thermal regulatory capability uncommon in fishes, a slow reproductive cycle with oophagous embryos, extensive migratory capabilities, and an ability to utilize a wide thermal niche including diving to near 1000 m depths [12-14].

Despite the high public profile of white sharks, their serious conservation needs, and their noteworthy evolutionary and life-history characteristics, this species is still largely uncharacterized at the molecular level, and no genomics resources for it exist. Given the white shark’s rather large genome size, a transcriptome characterization using next-generation sequencing technology provides a tractable entry into providing the first genomic view and genome resource for this remarkable species. However, obtaining white shark tissue is extremely difficult (see Methods), and as a consequence our study was restricted to one tissue type (heart) from one individual. This precluded examination of expression differences among tissue types, and we acknowledge the obvious limitation of a single transcriptome that may not be typical of the species.

Typically, *de-novo* transcriptomes for non-model organisms where no reference genome exists have been obtained using Roche 454 pyrosequencing technology because of the generation of longer sequencing reads e.g. [15-22]. However, recent advances in *de-novo* assembly for shorter Illumina reads are now making this approach a more viable alternative [23]. In addition, some workers have combined both approaches e.g. [15,24], and here we adopt this latter approach for deriving the first transcriptome dataset for the white shark. Specifically, Illumina reads were aligned to 454 contigs to produce a 454/Illumina consensus sequence. By utilizing the strengths of both sequencing technologies, this approach yielded a considerable increase (~20%) in transcriptome annotation when compared to 454 alone. We utilize this sequence dataset to provide a general characterization of the heart transcriptome with regards to gene discovery and annotation, identification and characterization of multiple microsatellite markers, and detection of genes under positive selection.

Results and discussion

Assembly

Roche 454 sequencing of the white shark heart cDNA produced 665,399 reads ranging in size from 100-931 bp (median = 387 bp) for a total of 240,894,914 bp. The *de-novo* assembly produced 141,626 contigs (unigenes) ranging in size from 101–12,997 bp, with a mean of 503 bp. The distribution of the number of reads per contig was as follows: 87,500 contigs (62%) = 1 read (singletons), 37,915 contigs (27%) = 2–5 reads, 6,595 contigs (5%) = 6–10 reads, and 9,616 contigs (7%) >10 reads (max = 568). The Illumina HiSeq run produced 78,566,588 100 bp reads. Aligning these data to the 454 contigs produced 105,014 454/Illumina consensus sequences (36,612 454 contigs lacked a consensus sequence). A total of 86,785 (82.6%) of the consensus sequences contained an ORF of 20 amino acids or longer. Of the 454 contigs lacking a 454/Illumina consensus sequence, 3,432 (9.4%) were non-singletons and 2,750 contained an ORF of 20 amino acids or longer. These ORFs were combined with the 86,785 ORFs obtained from the consensus sequences resulting in a total of 89,535 ORFs that were subsequently annotated. For purposes of quantitative evaluation of our combined 454/Illumina approach (e.g. number and length of contigs and number of annotated ORFs), we also processed the 454 data exclusively. Non-singleton 454 contigs (54,126) contained 52,841 ORFs of 20 amino acids or longer (97.6%). The 454 and Illumina derived short read files were deposited in the Sequence Read Archive at NCBI under the study accession number SRP016555. The 454 contigs, 454/Illumina consensus sequences, and 454/Illumina consensus ORFs (89,535) are included as Additional files 1, 2, and 3 respectively.

For a 454 contig, if there were nucleotide sites lacking consensus with Illumina data (possibly due to lack of coverage), the consensus sequence would contain Ns at the relevant positions. This in turn would lead to Xs (unspecified or unknown amino acids) in the subsequent translated ORF. The 86,785 ORFs generated from the 105,014 consensus sequences contained 7,674,130 amino acids (AA) including 783,158 Xs (10.2%). To place this apparent loss of AA data in perspective, the 52,841 ORFs generated from the 454 data alone, contained 5,579,487AA. Therefore, despite the ~10% loss of AA data in the consensus approach, we were still able to generate 1,311,485 more AA data, an increase of approximately one third, using the combined platform approach.

Lengths of ORFs generated using 454/Illumina consensus sequences and 454 data exclusively, showed in general, similar distributions (Figure 1). A noticeable difference, however, was an increase in the number of shorter ORFs (20AA – 169AA) for the consensus data. The number of shorter reads for the consensus data could be even higher as some consensus sequence ORFs contained X homopolymers at their 3′ end that might have masked a stop codon, which may in turn have erroneously increased the ORF length. Therefore, the length comparison was also performed excluding Xs in the consensus sequence ORFs (Figure 1). Mean ORF lengths were as follows: 454 only = 105.6AA, consensus = 88.4AA, consensus (Xs removed) = 79.4AA. The higher mean for the 454 ORFs most likely reflects the exclusion of singletons, which ranged in size from 101 bp to 931 bp (mean = 313 bp).

Annotation and comparative gene ontology

The ORF annotation was performed by searching the Swiss-Prot database using BLAST2GO [25], and a total of 21,019 consensus ORFs (23.5%) had blast hits with the database, with 19,277 (21.5%) of these receiving annotation with Gene Ontology (GO) terms (see Additional file 4). In comparison, 16,996 454-derived ORFs (32.2%) had blast hits with the Swiss-Prot database, with 15,597 (29.5%) annotated with GO terms (see Additional file 5). Consequently, although the mean ORF length for the consensus data was lower, there was a considerable increase in the number of annotated ORFs obtained (approximately one fifth), highlighting the improvement gained when 454 and Illumina data are combined.

The ORFs were also annotated with GO-Slim terms using the generic GO Slim (http://www.geneontology.org/GO_slims/goslim_generic.obo). GO Slim is a reduced version of the full GO that contains a sub-set of more general

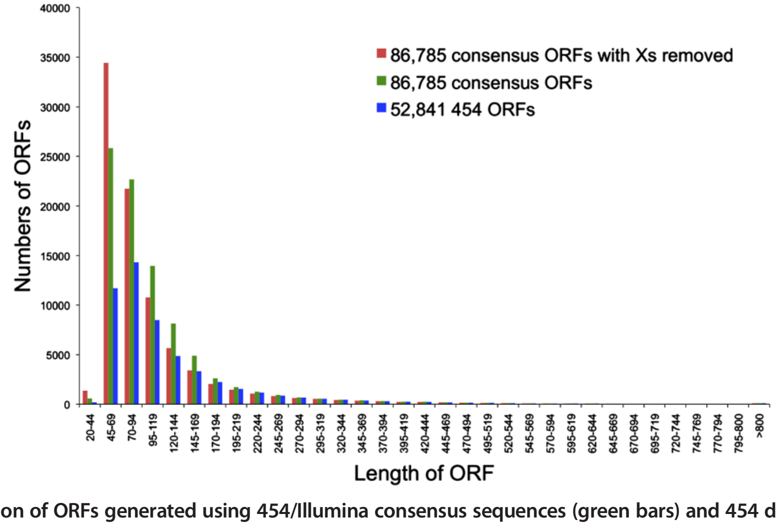

Figure 1 Length distribution of ORFs generated using 454/Illumina consensus sequences (green bars) and 454 data exclusively (blue bars). Red bars show distribution for consensus ORFs with unspecified or unknown amino acids (X) removed (see text).

GO terms and excludes the more fine-grained specific terms. This approach provides a broad overview of the ontology and gene product functions for genomic data. The genome sequence of the zebrafish (*Danio rerio*) is perhaps the most extensively studied of all fishes; consequently, its corresponding transcriptome sequence data should be the most complete for a fish and thus provide an appropriate evolutionary model for comparative characterization of the white shark transcriptome. However, the white shark also possesses certain endothermic capabilities more characteristic of mammals [26,27]; therefore, we compared the proportion of the white shark 19,277 consensus ORFs assigned to each GO Slim term to the proportion of zebrafish and human transcripts assigned to each GO Slim term (Figure 2, 3, 4). Distributions showed generally similar proportions of GO term assignments for each species, suggesting that we obtained a good representation of the heart transcriptome for the white shark. Closer inspection of the GO term proportions in the Biological Process domain (Figure 2) showed that the white shark heart transcriptome had the highest proportion of genes for most of the 18 metabolic process terms, with the zebrafish having the lowest. Fisher tests showed virtually all of these higher white shark proportions to be statistically significant (i.e. 14 of the 18 metabolic terms when compared to the zebrafish and 12 of these terms when compared to human; significance determined using a *FDR* correction of 0.05). Although this comparison is tempered by the fact that the shark transcriptome was derived exclusively from heart tissue and may already be enriched for metabolism relative to the complete transcriptome, it opens the possibility that some aspects of white shark metabolism, at least at the level of gene expression, might be more similar to that of a mammal than to that of an ectothermic teleost. The comparison is tempered because for each term, the Fisher test compares the relative proportion of genes assigned and not assigned the term between a particular species pair (i.e. white shark vs. human and white shark vs. zebrafish). Consequently, the relative proportion of genes assigned a term for the white shark might have been inflated because transcripts derived from other tissue types were absent. Although somewhat speculative, without a complete white shark transcriptome, this apparent higher gene proportion in metabolic process terms compared to zebrafish might be explained partly by the fact that the white shark is not a true poikilotherm. The white shark is among a very small group of fishes that have the ability to physiologically regulate their body temperature and maintain a substantially higher temperature than ambient seawater [26,27], which in turn is associated with elevated metabolic rates and aerobic and anaerobic capacities [28].

For the Molecular Function domain (Figure 3), comparison to the zebrafish showed 20 (32%) GO terms to be significantly enriched (i.e. had a significantly higher proportion of ORFs assigned) in the white shark, whereas comparison to human showed 18 (29%) terms to be enriched. There were 11 terms enriched in the zebrafish comparison that were not enriched in the human comparison. In general, these enriched terms described ion/nucleic acid/RNA binding, and enzyme/peptidase/nuclease/hydrolase/electron carrier activity. In turn, there were nine terms enriched in the human comparison that were not enriched in the comparison to zebrafish. In general, these terms described pyrophosphatase/phosphotransferase/hydrolase/nucleoside/transferase/kinase activity. While many of these enzymatic terms are likely involved in metabolic processes, two terms for the zebrafish comparison are perhaps particularly noteworthy: electron carrier and peptidase activity. Enrichment in these may again reflect the endothermic nature of the white shark. For example, electron carrier term enrichment suggests elevated oxidative metabolism, which is consistent with the increased energetic needs of an endothermic physiology, a continuous swimming lifestyle (required to obtain sufficient ventilation and hydrostatic lift) and the very long distance migratory capability of white sharks [29,30]. Enrichment for the peptidase activity term suggests increased digestive rates in white sharks, consistent with previous hypotheses of this capability based on the elevated temperatures observed in the stomach and other viscera of white sharks [27].

The Cellular Component domain (Figure 4) describes where a gene product is active; and a notably large number of GO terms in this domain were enriched for the white shark: 56% were enriched compared to human and 77% were enriched compared to zebrafish. This unexpectedly large difference in GO term enrichment in the white shark - zebrafish comparison compared to the white shark - human comparison also hints at the possibility that a component of the white shark transcriptome may be more similar to human than zebrafish. Similarities between another chondrichthyian and humans were also apparent in the genome sequence comparisons of Venkatesh et al. [4,31], in which the elephant shark (a chimaera; subclass Euchondrocephali) surprisingly shared a higher degree of gene synteny and more conserved non-coding elements (CNEs) with humans than with either the zebrafish or puffer fish (*Fugu rubripes*).

Microsatellites

Roche 454 sequencing has become an effective alternative to established protocols for the isolation of microsatellite markers [32-34], and this approach is increasingly being used to develop such markers for teleost fishes of economic and conservation interest [35,36]. The use of this technology is in its infancy for sharks, with three reports thus far [37-39], all of which were based on 454 genome shotgun sequencing. Here, we provide the results of the first

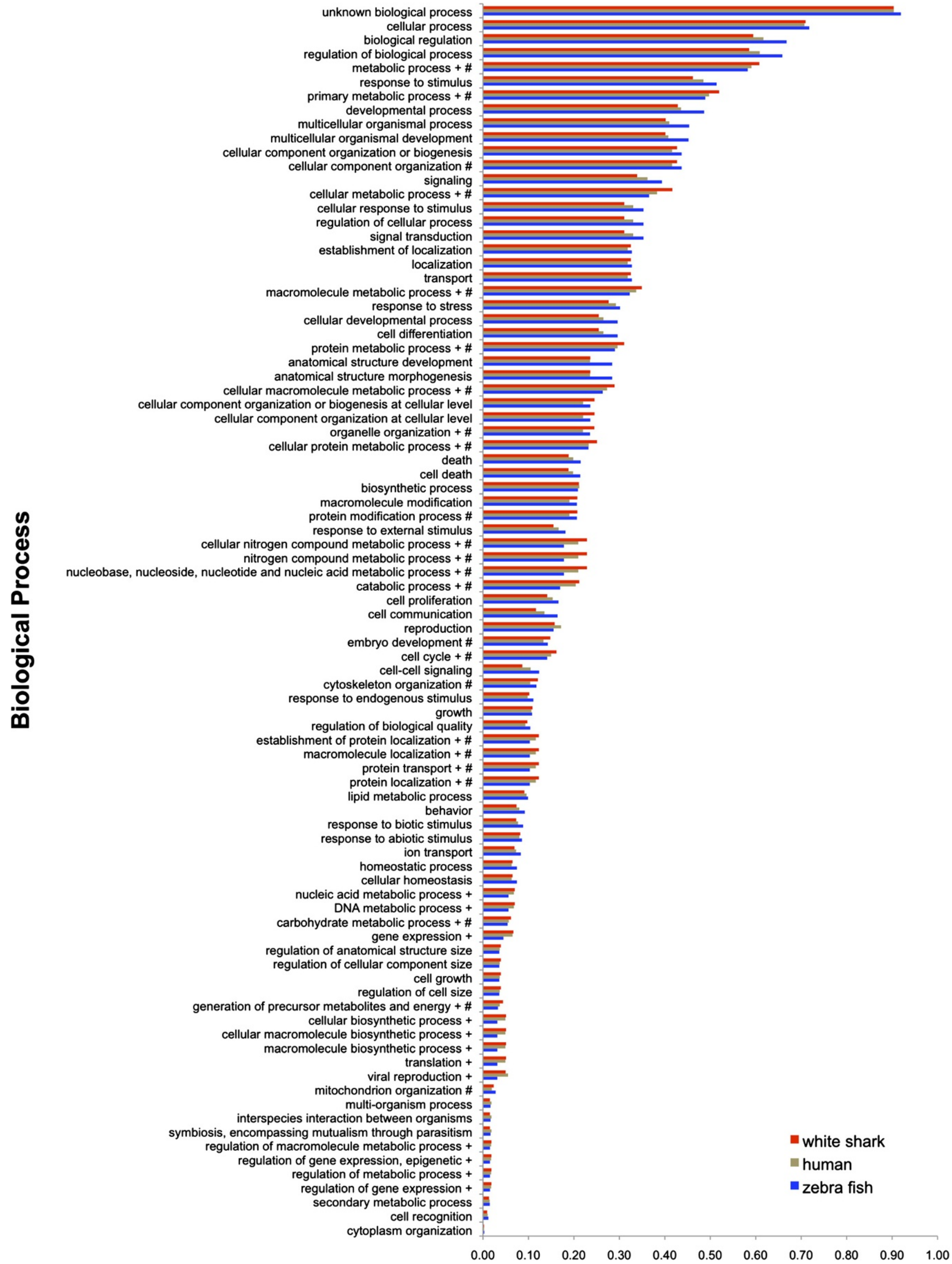

Figure 2 Proportion of open reading frames assigned to each gene ontology (GO) slim Biological Process term for white shark heart, zebrafish, and human transcriptomes. The + symbol shows GO terms significantly enriched for the white shark when compared to the zebrafish. The # symbol shows GO terms significantly enriched for the white shark when compared to human.

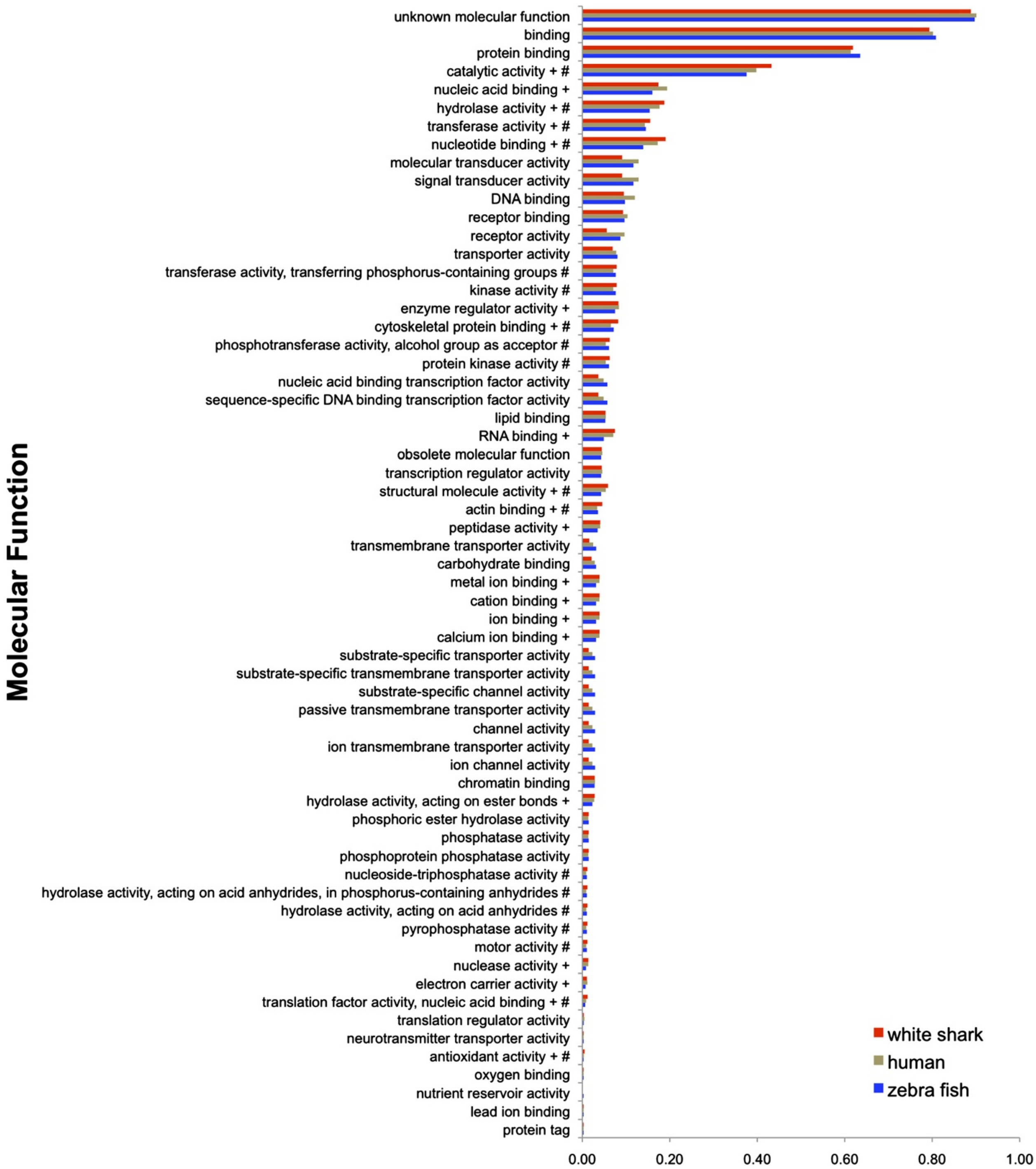

Figure 3 Proportion of open reading frames assigned to each gene ontology (GO) slim Molecular Function term for white shark heart, zebrafish, and human transcriptomes. The + symbol shows GO terms significantly enriched for the white shark when compared to the zebrafish. The # symbol shows GO terms significantly enriched for the white shark when compared to human.

transcriptome based discovery of microsatellite markers and their distributional characteristics in an elasmobranch. Of the 141,626 contigs derived from the 454 white shark data, 6,555 (4.6%) contained one or more dinucleotide, trinucleotide, or tetranucleotide microsatellites of five perfect repeats or more. In total, we detected 8,404 microsatellites with the following motifs: di = 7,467 (88.9%), tri = 864 (10.3%), tetra = 73 (0.9%). The maximum number of repeats for each motif was: di = 63 (average = 13), tri = 31 (average = 6), tetra = 21 (average = 7). See Additional file 6

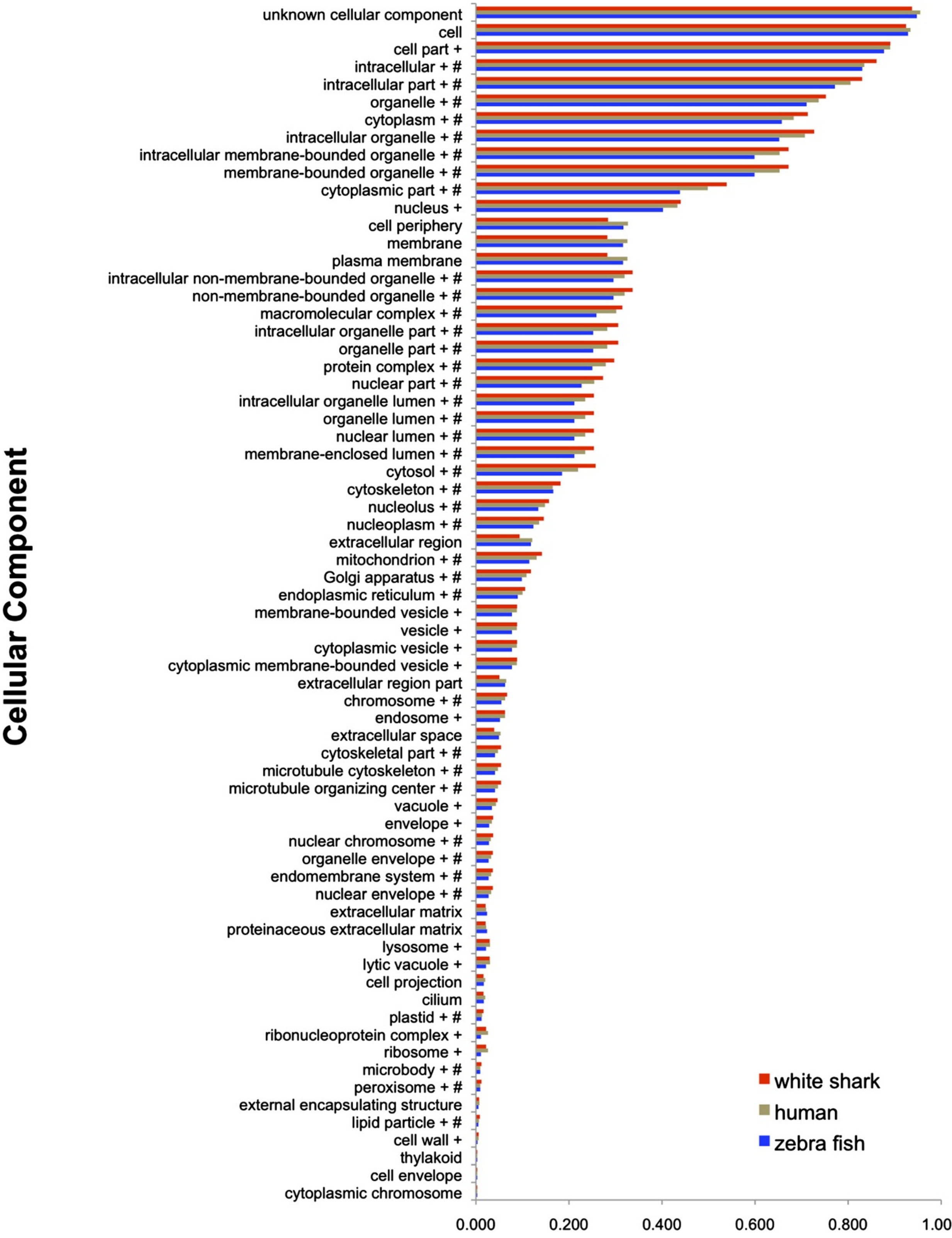

Figure 4 Proportion of open reading frames assigned to each gene ontology (GO) slim Cellular Component term for white shark heart, zebrafish, and human transcriptomes. The + symbol shows GO terms significantly enriched for the white shark when compared to the zebrafish. The # symbol shows GO terms significantly enriched for the white shark when compared to human.

for a description of the microsatellites. Of the 6,555 contigs containing a microsatellite, 854 were singletons with no consensus sequence. Of the remaining 5,701 contigs, 762 (13.4%) lacked an open reading frame and were therefore possibly non-coding transcripts or transcript fragments. The remaining 4,939 contigs (86.6%) contained one or more microstatellites within the ORF, 5′ untranslated region, and 3′ untranslated region. The proportion of microstatellites within the ORFs and untranslated regions (UTRs) was approximately equal (ORF = 31%, 5′UTR = 31%, 3′ UTR = 34%) (Figure 5). A small proportion of microstatellites (4%) straddled the ORF and UTRs.

The proportion of di, tri, and tetranucleotide repeat motifs within the ORFs, 5′ and 3′ UTRs, and putative non-coding transcripts are shown in Figure 6. The vast majority of motifs in all transcript types and regions were dinucleotides (ORF = 84%, 5′ UTR = 90%, 3′ UTR = 91%, and putative non-coding transcript = 92%). The only published microsatellites to date from white sharks are also dinucleotides, and were isolated using total genomic DNA and conventional enrichment protocols [40]. A majority of dinucleotide repeat microsatellite motifs were also found in the three shark species (all ectotherms in the order Carcharhiniformes) subject to whole genome 454 sequence analysis [37-39], hinting that a high frequency of dinucleotides may be a general feature of shark genomes. In the white shark, the frequency of repeat motifs within annotated ORFs showed a similar strong bias for dinucleotides (di = 68.8%, tri = 29.5%, tetra = 1.7%). Our finding that dinucleotides were the most frequent repeat motif in white shark transcripts irrespective of transcript region is typical for a wide range of taxonomic groups including other vertebrates [41]. However, when Toth *et al.* [41] only considered exons, trinucleotide repeats were found to be the most frequent. This finding contrasts sharply with that for the white shark where ORFs were strongly dominated by dinucleotide repeats. Another fish, the teleost *F. rubripes*, is a partial exception to the trend of high frequency of trinucleotides in vertebrate coding regions. For example, when repeat motifs of one through eight within ORFs were examined in *Fugu rubripes*, dinucleotides occurred in almost equal proportions to trinucleotides (di = 33.8%, tri = 31.7%) [42]. While the teleost proportions appear to be the most similar to the white shark (perhaps due to a closer evolutionary relationship relative to other vertebrates in this comparison), the white shark remains the most distinctive due to its high proportion of dinucleotides relative to trinucleotides within ORFs.

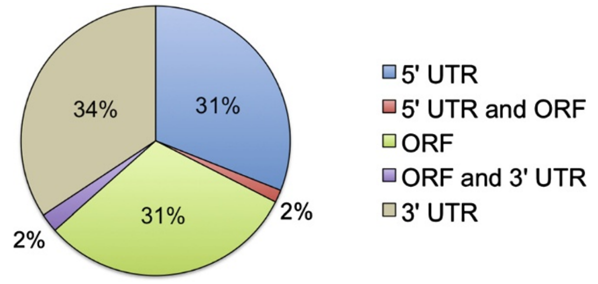

Figure 5 The proportion of white shark microstatellites within open reading frames (ORF) and untranslated regions (UTR). Red and purple sections show the proportions for a small number of microsatellites that straddled (i) the 5′ end of the ORF and 5′ UTR and (ii) the 3′ end of the ORF and 3′ UTR.

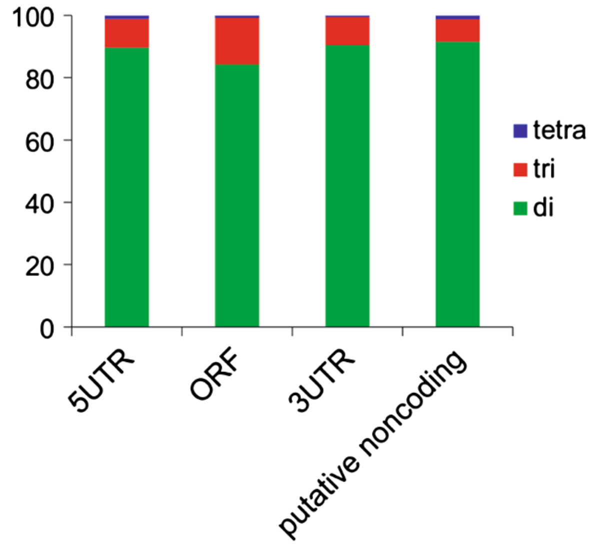

Figure 6 The proportion of white shark di, tri, and tetranucleotide microsatellite repeat motifs within the ORFs, 5′ and 3′ UTRs, and putative non-coding transcripts.

Expansion of trinucleotide repeats within ORFs has been implicated in human neurodegenerative disorders and some cancers [43-47]. Notably, elasmobranchs allegedly have the lowest incidence of malignant neoplasia (tumors) of any vertebrate group [48], although this claim remains controversial due to a lack of sufficient study [49]. If further studies demonstrate that elasmobranchs do indeed have a lower susceptibility to cancer, the relatively lower proportion of trinucleotide microsatellite repeats within ORFs, as seen here for the white shark, may provide a genetic mechanism hypothesis for further exploration.

There were 1,600 ORFs that contained one or more microsatellite (1,888 microsatellites in total). Of these, 1,331 ORFs contained one or more dinucleotide repeat, 255 ORFs contained one or more trinucleotide repeat, and 14 ORFs contained one or more tetranucleotide repeat. A total of 413 (~ 26%) of these microsatellite-containing ORFs were annotated (motif distribution: di = 284, tri = 122, and tetra = 7). For these ORFs, we investigated whether any of the GO terms assigned to them, appeared in significantly higher proportions (i.e. were relatively enriched) compared to the remainder of the transcriptome's non-microsatellite containing ORFs. For ORFs containing dinucleotide or tetranucleotide repeats, a Fisher test showed that no GO term was significantly enriched (*FDR* = 0.05). For ORFs containing trinucleotide repeats, however, terms within the Molecular Function domain (nucleic acid/DNA binding and transcription factor/regulator activity) and Cellular Component domain (nucleoplasm) were significantly enriched.

The Molecular Function domain enriched terms described gene products that (i) interacted selectively and non-covalently with nucleic acids, and (ii) interacted selectively and non-covalently with specific DNA sequences in order to modulate transcription. These results suggest that white shark ORFs containing trinucleotide repeats may have regulatory roles involved in the control of transcription (see Additional file 7 for a list of these ORFs).

Previous studies have shown that certain types of trinucleotide repeat, coding for specific amino acid homopolymers, have specific functions. For example, poly-glutamic acid homopolymers are common in nuclear localization signal proteins [50] and have been implicated in transcription activation/de-activation [41,51-53], whereas proline homopolymers may provide a domain for DNA binding and affect protein-protein interactions [51,52]. In general, the white shark was concordant with these findings, as the most frequent amino acid homopolymers within ORFs for the enriched nucleic acid/DNA binding GO term were poly-glutamic acid (28.3%), poly-aspartic acid (19.6%), and poly-proline (17.4%) (Figure 7). There was a similar pattern for the enriched transcription factor/regulator activity GO term with poly-glutamic acid (21.4%) and poly-proline (21.4%) being the most frequent (Figure 7). The white shark was distinctive however, in that there was a large proportion of poly-aspartic acid homopolymers within ORFs with regulatory roles (i.e. nucleic acid/DNA binding).

Finally, the large pool of microsatellites discovered here provides the potential to greatly expand the limited microsatellite marker resources available for this vulnerable species. To this end and as part of a separate study, we are developing microsatellite PCR primers on a global set of white shark fin tissue samples. To date, we have tested 35 loci (mostly dinucleotide and trinucleotide repeats). Of these, 14 are scorable (an individual can be genotyped), suggesting good prospects for the development of additional loci (A. Bernard, VPR, MJS, MSS; data not shown).

Positive selection

We searched the white shark transcriptome for genes showing signs of positive selection by comparing it to embryo transcriptomes of two additional elasmobranch species: *Scyliorhinus canicular* (cat shark) and *Leucoraja erinacea* (little skate). For each of the three species, we tested each species' lineage for positive selection using the branch-site test as implemented in PAML [54]. Before correction for multiple testing, there were ten, three, and five genes on the white shark, cat shark, and skate lineages respectively that had significant results for positive selection (Additional file 8 shows results for white shark). After correction (*FDR* = 0.05), four white shark genes remained significant: UN031816 (TIP41-like protein), UN034361 (mediator of RNA polymerase II transcription subunit 20), UN050025 (protein MIS12 homolog), and UN034642 (uncharacterized protein C12orf12 homolog). None of the cat shark or skate genes remained significant after correction for multiple testing.

In yeast, TIP41 indirectly regulates cell growth by regulating SIT4 (serine/threonine-protein phosphatase 2A) activity

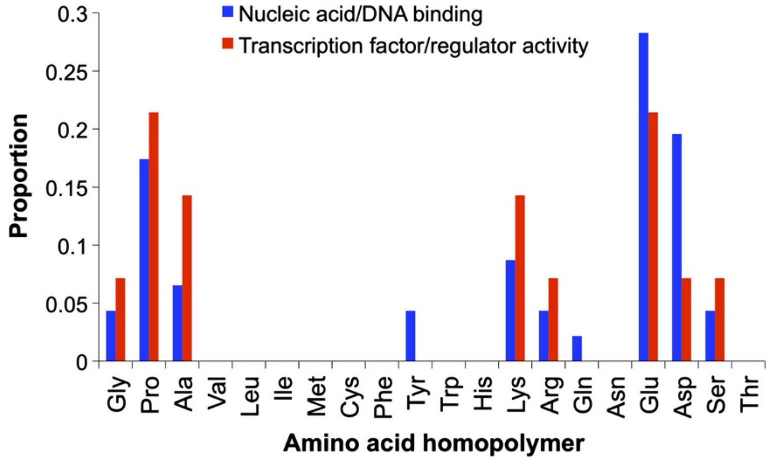

Figure 7 Proportion of amino acid homopolymers within ORFs for (i) the enriched nucleic acid/DNA binding GO term, (ii) the enriched transcription factor/regulator activity GO term.

[55]. More specifically, when nutrients such as nitrogen or carbon are abundant, the rapamycin-sensitive TOR signaling pathway promotes binding of the inhibitor protein TAP42 to SIT4 thereby inhibiting its activity. However, when nutrients are low, TOR does not promote binding of TAP42 and the inhibitor disassociates from SIT4 and TIP41 binds to the inhibitor, which in turn permits SIT4 activity. The regulatory role of protein phosphatases is debated in the literature, and they may function to both up and down regulate cell growth [56]. Furthermore, given the important role of these enzymes in cell growth, they have been actively studied by cancer researchers, with some studies suggesting that they might possess tumor suppressive capabilities [56]. However, other studies have emphasized their requirement for active cell growth and survival [56]. Nevertheless, the finding here of positive selection for a white shark gene involved in their regulation warrants further investigation given (i) an apparent low incidence of malignant tumors reported for elasmobranchs [48], (ii) the high levels of nitrogen (urea) in shark tissue, and (iii) the unique ability of elasmobranchs when compared to higher vertebrates to regenerate kidney tissue [57,58].

Mediator of RNA polymerase II transcription subunit 20 is a component of the Mediator complex. This large multi protein complex, which is conserved among eukaryotes, binds RNA polymerase II and regulates transcription of class II genes [59-61]. In addition to controlling cell growth, the TOR signaling pathway has also been implicated in the regulation of transcription. For example, in yeast, TOR limits transcription when nitrogen levels are low [62]. Perhaps the elevated level of urea in shark tissue is a factor contributing to positive selection for the Mediator subunit 20.

Using Blast2GO, we were able to assign GO terms to three of the genes showing signs of positive selection (UN034361, UN031816, and UN050025) (14 terms in total) (Additional file 8). Montoya-Burgos [63] compared GO terms enriched for genes under positive selection among two teleosts and six eutherian mammals (including humans). Comparison of these results to the GO terms for the white shark genes under positive selection is shown in Table 1. Similarities were regulation of transcription for the Biological Process domain and protein binding for the Molecular Function domain. A notable difference, however, was the complete absence of Biological Process response terms for the white shark (and also the cat shark and skate). In contrast, multiple response terms were shared between the teleosts and mammals (e.g. response to stimulus, stress, and wounding; defense response; and immune response). Furthermore, numerous studies involving a variety of additional teleosts have reported detection of positive selection for genes involved in stress and immune system response [64-68]. Elasmobranchs are the most primitive jawed vertebrate to possess an adaptive immune system based on immunoglobulins, T cell receptors, and major histocompatibility complex molecules (Ig/TCR/MHC). However, this system is genetically distinct from the higher vertebrates [69,70] and has a restricted antibody response when compared to teleosts and mammals [71,72]. The lack of positive selection for genes involved in immune and defense response reported here for three elasmobranch species might in part reflect this elasmobranch immune system distinctiveness.

Table 1 Comparison of GO terms for genes under positive selection for the white shark, two teleosts, and six mammals

Biological process	**Biological process**
(white shark)	**(shared between teleosts and mammals)**
Biological regulation	Response to stimulus
Regulation of transcription from RNA polymerase II promoter	Defense response
Cell communication	Response to stress* or wounding**
Chromosome segregation	Immune system process, Immune response
Kinetochore assembly	Signal transduction
	Regulation of transcription DNA-dependent
	Ion transport
Molecular function	**Molecular function**
(white shark)	**(shared between teleosts and mammals)**
Protein binding	DNA binding, or mismatch DNA binding
RNA polymerase II transcription mediator activity	Protein binding**, chemokine receptor binding**, interleukin binding**, interleukin-1 receptor binding*
DNA-directed RNA polymerase activity	Metal ion binding

GO terms shared between teleosts and mammals are from Montoya-Burgos [63]. A single asterisk shows terms found in teleosts. Two asterisks shows terms found in mammals.

Conclusions

Utilizing an approach that combined Roche 454 and Illumina sequencing technologies, we assembled and characterized the first white shark transcriptome. This combined approach yielded a considerable improvement over Roche 454 technology alone, generating 21,019 annotated transcripts. The white shark transcriptome is a valuable resource that adds to the currently nascent field of cartilaginous fish genomics and provides a reference for characterization of genomic datasets from other elasmobranchs, which we anticipate will emerge with increasing frequency. This resource also provides the first large-scale view of the gene content of a major marine apex predator that displays a collection of remarkable physical, physiological, and behavioral properties. Of particular interest is the observation that the proportion of annotated transcripts involved in metabolic processes was more similar between the white shark and humans than between the white shark and a teleost, a finding consistent with those of Venkatesh *et al.* [4,31] who found genomic non-coding elements and the relative position of genes to be more similar between another cartilaginous fish (the elephant shark) and humans than between the elephant shark and a teleost. We also compared the white shark transcriptome to other available elasmobranch sequences, for signatures of positive selection and identified several genes of putative adaptive significance on the white shark lineage. The transcriptome resource also provides a large set of new microsatellites that will be immediately useful as markers in studies of population structure, dispersal dynamics, genetic diversity, and mating system biology to further the conservation and management of this vulnerable species.

Methods

Tissue collection

The white shark is protected by many countries, including the US, and is also a CITES Appendix II listed species [73]. Consequently, obtaining white shark tissue is extremely difficult. However, we were able to obtain tissue from a juvenile white shark illegally landed by an independent fisher off the Delaware, USA coast in 2007. The shark was confiscated from the fisher by the US National Oceanic and Atmospheric Administration Office for Law Enforcement. The heart was collected during a subsequent necropsy of the shark conducted by the National Oceanic and Atmospheric Administration for scientific data collection, and provided to us by this agency for further analysis. The heart was kept frozen at -80°C until sub-sampled for RNA isolation.

cDNA library construction and Roche 454/Illumina sequencing

Total RNA was isolated by homogenization of heart tissue in TRIzol (Invitogen, Carlsbad, USA) followed by phenol chloroform extraction. Full-length cDNA was synthesized using two sets of oligo dT primers in a two step procedure and single-stranded cDNA was used for hybridization instead of double-stranded [74]. After hybridization, reassociated ds-cDNA was separated from ss-cDNA (normalized cDNA) by hydroxyapatite chromatography. Normalized cDNA was re-amplified using an oligo dT specific primer (L4N). cDNA was sequenced using a single run on the Roche 454 GS FLX platform and a single lane of Illumina HiSeq 2000 (100 bp reads, single end).

Sequence assembly and annotation

Roche 454 adaptor sequences were removed using LUCY [75] and the script SeqClean (http://compbio.dfci.harvard.edu/tgi/software). SeqClean was also used to remove reads containing low complexity sequence, reads shorter than 100bp, and to clip low quality read ends (ends rich in undetermined bases). 454 reads were assembled into contigs *de-novo* using iAssembler v1.3 [76]. Contigs were searched for di, tri, and tetra microsatellites of five repeats or more using Phobos v3.3.12 as implemented in Geneious v5.5.3 [77]. Illumina HiSeq reads were aligned to 454 contigs using the program Burrows-Wheeler Aligner (BWA) [78] and consensus sequences built using the pileup format as implemented in SAMtools [79].

Roche 454/Illumina consensus sequences were searched for open reading frames (ORFs) of 20 amino acids or longer (including the start codon for methionine, but omitting unspecified or unknown amino acids [coded as X]) using the script longorf.pl (available at http://search.cpan.org/~cjfields/BioPerl-1.6.901/examples/longorf.pl). Non-singleton 454 contigs lacking Illumina read coverage were also searched for ORFs using the longorf.pl procedure. Annotation for ORFs was obtained using Blast2GO v.2.5.0 [25]. Amino acids were searched against the UniProtKB/Swiss-Prot database using an *E* value cut-off = 1e-6 (retaining best 20 hits), with a minimum amino acid alignment length cut-off (high-scoring segment pair length) of 33. Blast2GO was also used to assign GO terms.

For purposes of quantitative evaluation of our combined 454/Illumina sequencing platform approach, we also processed the 454 data without combining it with Illumina data. After singletons were removed, the contigs were searched for ORFs and annotated using the same procedure as for the combined data.

Blast2GO was also used to annotate ORFs and assign GO terms for the zebrafish and human transcriptomes (same procedure as above) to provide comparison to the white shark. Transcriptomes for zebrafish and human were obtained from Ensembl (Danio_rerio.Zv9.66.cdna.all.fa, Homo_sapiens.GRCh37.67.cdna.all.fa). Note: the Ensemble cdna.all files contain “the super set of all transcripts resulting from Ensemble known, novel and pseudo gene predictions" (see the associated readme file for a complete description). Relative enrichment of GO terms for white shark when compared to zebrafish and human (separate comparisons) was assessed using a Fisher exact test. The test was performed using the Gossip statistical package [80] implemented within Blast2GO. The false discovery rate (FDR) procedure of Benjamini and Hochberg [81] was used to correct for multiple hypothesis testing (*FDR* = 0.05). We did not test for underrepresentation (lower proportion of terms) as the white shark transcriptome was obtained from a single tissue type and may therefore not represent complete genomic expression.

Branch-site test of positive selection

In order to detect genes under positive selection using the branch-site test [82], we obtained embryo transcriptome data for two additional elasmobranch species: *Scyliorhinus canicula* (smallspotted cat shark) and *Leucoraja erinacea* (little skate). The data were downloaded from the Gene Expression Omnibus database at NCBI (accession number GSE26235). Transcripts were searched for open reading frames of 20 amino acids or longer using the same procedure as for the white shark. For each of the three elasmobranch species (*C. carcharias*, *S. canicular* and *L. erinacea*), and their putative homologous loci (procedure described below), we tested genes in each species' lineage for positive selection using the branch-site test as implemented in codeml in PAML (Phylogenetic Analysis by Maximum Likelihood) version 4.4 [54]. The test was performed on homologous core genes (those genes shared among all three species). Homologous genes were delineated using the MCL algorithm [83] as implemented in the MCLBLASTLINE pipeline (available at http://micans.org/mcl). The pipeline uses Markov clustering (MCL) to assign genes to homologous clusters based on a BLASTp search between all species pairs of protein sequences using an *E* value cut-off of 1e-5. The MCL algorithm was implemented using an inflation parameter of 1.2. Only single copy core genes were used (i.e. clusters containing paralogs were excluded). The nucleotide sequences corresponding to each set of homologous core genes were aligned using Probalign [84]. Alignment columns with a posterior probability <0.6 were removed, and alignments with >50% of the sites removed were discarded from the analysis. Using each of the alignments and the three elasmobranch species tree topology, positive selection was assessed for each lineage by performing likelihood ratio tests. We compared two branch-site models: (i) a null model that does not allow positive selection (model M1a) and (ii) an alternative model that allows positive selection (model A). *P* values were calculated under the assumption that the likelihood ratio follows a chi-square distribution with one degree of freedom [82]. Multiple testing adjustment was performed using a false discovery rate approach [85] (significance level = 0.05).

Availability of supporting data

The 454 and Illumina derived short read files are available at the NCBI Sequence Read Archive (SRA) under the study accession number SRP016555.

Additional files

Additional file 1: (454 contigs).

Additional file 2: (454/Illumina consensus nucleotide sequences).

Additional file 3: (454/Illumina consensus ORFs [amino acid sequences]).

Additional file 4: (annotation of 454/Illumina consensus ORFs).

Additional file 5: (annotation of 454 ORFs).

Additional file 6: (microsatellite characteristics).

Additional file 7: (annotated ORFs containing trinucleotide microsatellite repeats associated with enriched GO terms).

Additional file 8: (ORFs significant for positive selection).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MSS and MJS conceived, developed, and supervised the project. VPR and HS conducted the bioinformatics analyses. VPR drafted the manuscript with contributions from MSS and HS. All authors contributed to editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are very grateful to Dr. L. Natanson (NOAA NEFSC) for providing the white shark heart and the NOAA Office for Law Enforcement for making the confiscated shark carcass available for scientific study. We thank the Cornell Life Sciences Core Laboratory for performing the 454 and Illumina sequencing, and we are grateful to Qi Sun and Jaroslaw Pillardy for their assistance with the bioinformatics analyses. This study was supported by grants from the Save Our Seas Foundation, Guy Harvey Ocean Foundation, and Hai Stiftung/Shark Foundation to MSS.

Author details

[1]Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA. [2]Save Our Seas Shark Research Center and Guy Harvey Research Institute, Nova Southeastern University, 8000 North Ocean Drive, Dania Beach, FL 33004, USA. [3]Current address: Graduate School of Science and Engineering, Yamaguchi University, Yoshida 1677-1, Yamaguchi 753-8512, Japan.

Received: 25 February 2013 Accepted: 20 September 2013
Published: 11 October 2013

References

1. Grogan ED, Lund R: **The origin and relationships of Early Chondrichthyes.** In *Biology of sharks and their relatives [CRC Marine Biology Series]*. Edited by Carrier JC, Musick JA, Heithaus MR, Boca R. London etc: CRC Press; 2004:3–31.
2. Schneider I, Aneas I, Gehrke AR, Dahn RD, Nobrega MA, Shubin NH: **Appendage expression driven by the Hoxd Global Control Region is an ancient gnathostome feature.** *Proc Natl Acad Sci USA* 2011, **108**(31):12782–12786.
3. Tan YY, Kodzius R, Tay BH, Tay A, Brenner S, Venkatesh B: **Sequencing and analysis of full-length cDNAs, 5'-ESTs and 3'-ESTs from a cartilaginous fish, the elephant shark (*Callorhinchus milii*).** *PLoS One* 2012, **7**(10):e47174.
4. Venkatesh B, Kirkness EF, Loh Y-H, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, *et al*: **Survey sequencing and**

comparative analysis of the elephant shark (*Callorhinchus milii*) genome. *PLOS Biology* 2007, **5**(4):0932–0944.

5. Wang Q, Arighi CN, King BL, Polson SW, Vincent J, Chen C, Huang H, Kingham BF, Page ST, Rendino MF, *et al*: **Community annotation and bioinformatics workforce development in concert--Little Skate Genome Annotation Workshops and Jamborees.** *Database (Oxford)* 2012, **2012:**bar064.
6. Clarke SC, McAllister MK, Milner-Gulland EJ, Kirkwood GP, Michielsens CGJ, Agnew DJ, Pikitch EK, Nakano H, Shivji MS: **Global estimates of shark catches using trade records from commercial markets.** *Ecol Lett* 2006, **9**(10):1115–1126.
7. King BL, Gillis JA, Carlisle HR, Dahn RD: **A natural deletion of the HoxC Cluster in Elasmobranch fishes.** *Science (Washington D C)* 2011, **334**(6062):1517.
8. Parton A, Bayne CJ, Barnes DW: **Analysis and functional annotation of expressed sequence tags from in vitro cell lines of elasmobranchs: Spiny dogfish shark (*Squalus acanthias*) and little skate (*Leucoraja erinacea*).** *Comp Biochem Physiol Genom Proteonomics* 2010, **5**(3):199–206.
9. Chapple TK, Jorgensen SJ, Anderson SD, Kanive PE, Klimley AP, Botsford LW, Block BA: **A first estimate of white shark, *Carcharodon carcharias*, abundance off Central California.** *Biol Lett* 2011, **7**(4):581–583.
10. Fergusson I, Compagno L, Marks M: *IUCN 2012. IUCN Red List of Threatened Species. Version 2012.2.* 2009. www.iucnredlist.org.
11. Shivji MS, Chapman DD, Pikitch EK, Raymond PW: **Genetic profiling reveals illegal international trade in fins of the great white shark, *Carcharodon carcharias*.** *Conserv Genet* 2005, **6**(6):1035–1039.
12. Castro JI: *The sharks of North America.* Oxford, New York etc: Oxford University Press; 2011.
13. Gregory TR: *Animal Genome Size Database.* 2005. http://www.genomesize.com.
14. Nasby-Lucas N, Dewar H, Lam CH, Goldman KJ, Domeier ML: **White shark offshore habitat: a behavioral and environmental characterization of the eastern Pacific offshore foraging area.** *PLoS One* 2009, 4(12):e8163. 8161–8114.
15. Cahais V, Gayral P, Tsagkogeorga G, Melo-Ferreira J, Ballenghien M, Weinert L, Chiari Y, Belkhir K, Ranwez V, Galtier N: **Reference-free transcriptome assembly in non-model animals from next-generation sequencing data.** *Mol Ecol Resour* 2012, **12:**834–845.
16. Miller HC, Biggs PJ, Voelckel C, Nelson NJ: **De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*).** *BMC Genomics* 2012, **13:**439.
17. Fraser BA, Weadick CJ, Janowitz I, Rodd FH, Hughes KA: **Sequencing and characterization of the guppy (*Poecilia reticulata*) transcriptome.** *BMC Genomics* 2011, **12:**202.
18. Parchman TL, Geist KS, Grahnen JA, Benkman CW, Buerkle CA: **Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery.** *BMC Genomics* 2010, **11:**180.
19. Meyer E, Aglyamova GV, Wang S, Buchanan-Carter J, Abrego D, Colbourne JK, Willis BL, Matz MV: **Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx.** *BMC Genomics* 2009, **10:**219.
20. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I, Marden JH: **Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing.** *Mol Ecol* 2008, **17**(7):1636–1647.
21. Der JP, Barker MS, Wickett NJ, de Pamphilis CW, Wolf PG: **De novo characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*.** *BMC Genomics* 2011, **12:**99.
22. Reading BJ, Chapman RW, Schaff JE, Scholl EH, Opperman CH, Sullivan CV: **An ovary transcriptome for all maturational stages of the striped bass (*Morone saxatilis*), a highly advanced perciform fish.** *BMC Res Notes* 2012, **5:**111.
23. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome.** *Nat Biotechnol* 2011, **29**(7):644–652.
24. Angeloni F, Wagemaker CA, Jetten MS, Op den Camp HJ, Janssen-Megens EM, Francoijs KJ, Stunnenberg HG, Ouborg NJ: **De novo transcriptome characterization and development of genomic tools for *Scabiosa columbaria* L. using next-generation sequencing techniques.** *Mol Ecol Resour* 2011, **11**(4):662.
25. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36**(10):3420–3435.
26. Katz SL: **Design of heterothermic muscle in fish.** *J Exp Biol* 2002, **205** (Pt 15):2251–2266.
27. Goldman KJ: **Regulation of body temperature in the white shark, *Carcharodon carcharias*.** *J Comp Physiol B Biochem Syst Environ Physiol* 1997, **167**(6):423–429.
28. Dickson KA, Graham JB: **Evolution and consequences of endothermy in fishes.** *Physiol Biochem Zool* 2004, **77**(6):998–1018.
29. Bonfil R, Meyer M, Scholl MC, Johnson R, O'Brien S, Oosthuizen H, Swanson S, Kotze D, Paterson M: **Transoceanic migration, spatial dynamics, and population linkages of white sharks.** *Science (Washington D C)* 2005, **310**(5745):100–103.
30. Carlson JK, Goldman KJ, Lowe CG: **Metabolism, energetic demand, and endothermy.** In *Biology of sharks and their relatives [CRC Marine Biology Series].* Edited by Carrier JC, Musick JA, Heithaus MR. Boca Raton, London etc: CRC Press; 2004:203–224.
31. Venkatesh B, Kirkness EF, Loh YH, Halpern AL, Lee AP, Johnson J, Dandona N, Viswanathan LD, Tay A, Venter JC, *et al*: **Ancient noncoding elements conserved in the human genome.** *Science* 2006, **314**(5807):1892.
32. Hoffman JI, Nichols HJ: **A novel approach for mining polymorphic microsatellite markers in silico.** *PLoS One* 2011, **6**(8):e23283.
33. Babik W, Stuglik M, Qi W, Kuenzli M, Kuduk K, Koteja P, Radwan J: **Heart transcriptome of the bank vole (*Myodes glareolus*): towards understanding the evolutionary variation in metabolic rate.** *BMC Genomics* 2010, **11:**390.
34. Mikheyev AS, Vo T, Wee B, Singer MC, Parmesan C: **Rapid microsatellite isolation from a butterfly by de novo transcriptome sequencing: performance and a comparison with AFLP-derived distances.** *PLoS One* 2010, **5**(6):e11212.
35. Teacher AGF, Kahkonen K, Merila J: **Development of 61 new transcriptome-derived microsatellites for the Atlantic herring (*Clupea harengus*).** *Conserv Genet Resour* 2012, **4**(1):71–74.
36. Saarinen EV, Austin JD: **When technology meets conservation: increased microsatellite marker production using 454 genome sequencing on the endangered Okaloosa Darter (*Etheostoma okaloosae*).** *J Hered* 2010, **101**(6):784–788.
37. Boomer JJ, Stow AJ: **Rapid isolation of the first set of polymorphic microsatellite loci from the Australian gummy shark, *Mustelus antarcticus* and their utility across divergent shark taxa.** *Conserv Genet Resour* 2010, **2:**393–395.
38. Chabot CL: **Characterization of 11 microsatellite loci for the brown smooth-hound shark, *Mustelus henlei* (*Triakidae*), discovered with next-generation sequencing.** *Conserv Genet Resour* 2012, **4**(1):23–25.
39. Chabot CL, Nigenda S: **Characterization of 13 microsatellite loci for the tope shark, *Galeorhinus galeus*, discovered with next-generation sequencing and their utility for eastern Pacific smooth-hound sharks (*Mustelus*).** *Conserv Genet Resour* 2011, **3**(3):553–555.
40. Pardini AT, Jones CS, Scholl MC, Noble LR: **Isolation and characterization of dinucleotide microsatellite loci in the great white shark. *Carcharodon carcharias*.** *Mol Ecol* 2000, **9**(8):1176–1178.
41. Toth G, Gaspari Z, Jurka J: **Microsatellites in different eukaryotic genomes: survey and analysis.** *Genome Res* 2000, **10**(7):967–981.
42. Edwards YJ, Elgar G, Clark MS, Bishop MJ: **The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses.** *J Mol Biol* 1998, **278**(4):843–854.
43. Arzimanoglou II, Gilbert F, Barber HR: **Microsatellite instability in human solid tumors.** *Cancer* 1998, **82**(10):1808–1820.
44. Bates G, Lehrach H: **Trinucleotide repeat expansions and human genetic disease.** *Bioessays* 1994, **16**(4):277–284.
45. Reddy PS, Housman DE: **The complex pathology of trinucleotide repeats.** *Curr Opin Cell Biol* 1997, **9**(3):364–372.
46. Warren ST, Nelson DL: **Trinucleotide repeat expansions in neurological disease.** *Curr Opin Neurobiol* 1993, **3**(5):752–759.
47. Wooster R, Cleton-Jansen AM, Collins N, Mangion J, Cornelis RS, Cooper CS, Gusterson BA, Ponder BA, von Deimling A, Wiestler OD, *et al*: **Instability of short tandem repeats (microsatellites) in human cancers.** *Nat Genet* 1994, **6**(2):152–156.

48. Ballantyne JS: **Jaws: the inside story. The metabolism of elasmobranch fishes.** *Comp Biochem Physiol B Biochem Mol Biol* 1997, **118B**(4):703–742.
49. Ostrander GK, Cheng KC, Wolf JC, Wolfe MJ: **Shark cartilage, cancer and the growing threat of pseudoscience.** *Cancer Res* 2004, **64**(23):8485–8491.
50. Pearson CE, Nichol Edamura K, Cleary JD: **Repeat instability: mechanisms of dynamic mutations.** *Nat Rev Genet* 2005, **6**(10):729–742.
51. Berger M, Vogt Sionov R, Levine AJ, Haupt Y: **A role for the polyproline domain of p53 in its regulation by Mdm2.** *J Biol Chem* 2001, **276**(6):3785–3790.
52. Gerber HP, Seipel K, Georgiev O, Hofferer M, Hug M, Rusconi S, Schaffner W: **Transcriptional activation modulated by homopolymeric glutamine and proline stretches.** *Science* 1994, **263**(5148):808–811.
53. Perutz MF, Johnson T, Suzuki M, Finch JT: **Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases.** *Proc Natl Acad Sci USA* 1994, **91**(12):5355–5358.
54. Yang Z: **PAML 4: phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24**(8):1586–1591.
55. Jacinto E, Guo B, Arndt KT, Schmelzle T, Hall MN: **TIP41 interacts with TAP42 and negatively regulates the TOR signaling pathway.** *Mol Cell* 2001, **8**(5):1017–1026.
56. Schonthal AH: **Role of serine/threonine protein phosphatase 2A in cancer.** *Cancer Lett* 2001, **170**(1):1–13.
57. Little MH: **Regrow or repair: potential regenerative therapies for the kidney.** *J Am Soc Nephrol* 2006, **17**(9):2390–2401.
58. Elger M, Hentschel H, Litteral J, Wellner M, Kirsch T, Luft FC, Haller H: **Nephrogenesis is induced by partial nephrectomy in the elasmobranch *Leucoraja erinacea*.** *J Am Soc Nephrol* 2003, **14**(6):1506–1518.
59. Blazek E, Mittler G, Meisterernst M: **The mediator of RNA polymerase II.** *Chromosoma* 2005, **113**(8):399–408.
60. Malik S, Roeder RG: **Dynamic regulation of pol II transcription by the mammalian Mediator complex.** *Trends Biochem Sci* 2005, **30**(5):256–263.
61. Asturias FJ, Jiang YW, Myers LC, Gustafsson CM, Kornberg RD: **Conserved structures of mediator and RNA polymerase II holoenzyme.** *Science* 1999, **283**(5404):985–987.
62. Beck T, Hall MN: **The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors.** *Nature* 1999, **402**(6762):689–692.
63. Montoya-Burgos JI: **Patterns of positive selection and neutral evolution in the protein-coding genes of *Tetraodon* and *Takifugu*.** *PLoS One* 2011, **6**(9):e24800.
64. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of positive selection in six Mammalian genomes.** *PLoS Genet* 2008, **4**(8):e1000144.
65. Cohen S: **Strong positive selection and habitat-specific amino acid substitution patterns in MHC from an estuarine fish under intense pollution stress.** *Mol Biol Evol* 2002, **19**(11):1870–1880.
66. Xu T, Sun Y, Shi G, Wang R: **Miiuy croaker hepcidin gene and comparative analyses reveal evidence for positive selection.** *PLoS One* 2012, **7**(4):e35449.
67. Sundaram AY, Consuegra S, Kiron V, Fernandes JM: **Positive selection pressure within teleost toll-like receptors tlr21 and tlr22 subfamilies and their response to temperature stress and microbial components in zebrafish.** *Mol Biol Rep* 2012, **39**(9):8965–8975.
68. van der Aa LM, Levraud JP, Yahmi M, Lauret E, Briolat V, Herbomel P, Benmansour A, Boudinot P: **A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish.** *BMC Biol* 2009, **7**:7.
69. Hinds KR, Litman GW: **Major reorganization of immunoglobulin VH segmental elements during vertebrate evolution.** *Nature* 1986, **320**(6062):546–549.
70. Zhu C, Feng W, Weedon J, Hua P, Stefanov D, Ohta Y, Flajnik MF, Hsu E: **The multiple shark Ig H chain genes rearrange and hypermutate autonomously.** *J Immunol* 2011, **187**(5):2492–2501.
71. Litman GW, Stolen JS, Sarvas HO, Makela O: **The range and fine specificity of the anti-hapten immune response: phylogenetic studies.** *J Immunogenet* 1982, **9**(6):465–474.
72. Makela O, Litman GW: **Lack of heterogeneity in antihapten antibodies of a phylogenetically primitive shark.** *Nature* 1980, **287**(5783):639–640.
73. IUCN: **Review of Migratory Chondrichthyan Fishes. Secretariat of the Convention on the Conservation of Migratory Species of Wild Animals (CMS).** In *CMS Technical Report Series.* 2007:68.
74. Patanjali SR, Parimoo S, Weissman SM: **Construction of a uniform-abundance (normalized) cDNA library.** *Proc Natl Acad Sci USA* 1991, **88**(5):1943–1947.
75. Chou HH, Holmes MH: **DNA sequence quality trimming and vector removal.** *Bioinformatics* 2001, **17**(12):1093–1104.
76. Zheng Y, Zhao L, Gao J, Fei Z: **iAssembler: a package for de novo assembly of Roche-454/Sanger transcriptome sequences.** *BMC Bioinforma* 2011, **12**:453.
77. Drummond A, Ashton B, Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S: *Geneious v5.1.* 2010. Available from http://www.geneious.com.
78. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**(14):1754–1760.
79. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The sequence alignment/map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078–2079.
80. Blüthgen N, Brand K, Cajavec B, Swat M, Herzel H, Beule D: **Biological profiling of gene groups utilizing gene ontology.** *Genome Inform* 2005, **16**(1):106–115.
81. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J R Stat Soc Ser B* 1995, **57**:289–300.
82. Zhang J, Nielsen R, Yang Z: **Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.** *Mol Biol Evol* 2005, **22**(12):2472–2479.
83. van Dongen S: *Graph clustering by flow simulation PhD thesis.* University of Utrecht; 2000.
84. Roshan U, Livesay DR: **Probalign: multiple sequence alignment using partition function posterior probabilities.** *Bioinformatics* 2006, **22**(22):2715–2721.
85. Benjamini Y, Yekutieli D: **The control of the false discovery rate in multiple testing under dependency.** *Ann Statist* 2001, **29**(4):1165–1188.

doi:10.1186/1471-2164-14-697
Cite this article as: Richards *et al.*: Characterization of the heart transcriptome of the white shark (*Carcharodon carcharias*). *BMC Genomics* 2013 **14**:697.