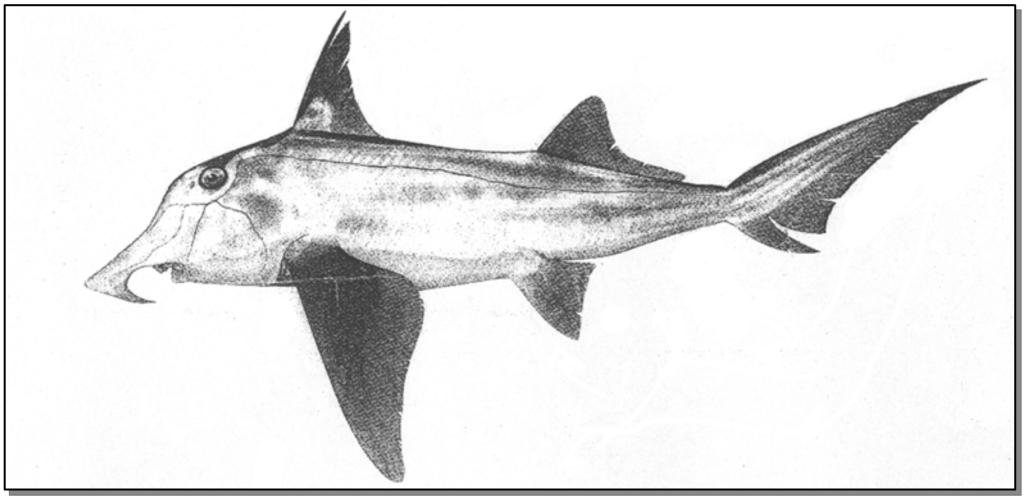
Proposal to generate a draft assembly of the compact elephant shark genome

Ewen Kirkness and Robert Strausberg
J. Craig Venter Institute (JCVI), Rockville, MD 20850, USA

Byrappa Venkatesh and Sydney Brenner
Institute of Molecular and Cell Biology (IMCB), Biopolis, Singapore 138673



Introduction

Cartilaginous fishes (Chondrichthyes) are the most basal extant jawed vertebrates that diverged from the common ancestor of tetrapods and teleost fishes (Osteichthyes) approximately 530 Myr ago (Kumar and Hedges, 1998; Fig.1). Cartilaginous fishes possess a body plan and complex physiological systems that are typical of all jawed vertebrates, but are lacking in the jawless vertebrates (e.g. lampreys and hagfishes). They therefore constitute an important basal vertebrate 'reference genome' that can help us better understand the structure, function and evolution of human and other vertebrate genomes. Cartilaginous fishes are a monophyletic group comprising two sister groups, the elasmobranchs (sharks, rays and skates) and the chimeras (Kikugawa et al., 2004; Fig.1). A major impediment to the characterization of genomes from cartilaginous fish is their large size. The dogfish shark (*Squalus acanthias*), nurse shark (*Ginglystoma cirratum*), horn shark (*Heterdontus francisi*) and little skate (*Raja erinacea*), which are all subjects for biological research, have genome sizes that range from 3.5 Gb to 7 Gb (Hinegardner, 1976; Schwartz and Maddock, 2002). Measurements of genome sizes for cartilaginous fishes have indicated that chimeras have smaller genomes than the elasmobranchs (Hinegardner, 1976; Stingo, 1979). Among the chimeras, we have identified the elephant shark (*Callorhinchus milii*) as having the smallest known genome (Venkatesh et al., 2005). Based on FACScan flow cytometry and survey sequencing of the genome (see below), we estimate the size of the elephant shark genome to be less than 1 Gb, at least 3 times smaller than the human genome. In contrast, the only other cartilaginous fish that is currently being considered for genome sequencing, the little skate (*Raja erinacea*), has an estimated genome size of approximately 3.5 Gb (http://www.genome.gov/10002154; Schwartz and Maddock, 2002). Thus, the elephant shark is an attractive candidate for yielding an economical and high-quality sequence assembly of a cartilaginous fish genome. The value of such an assembly for annotation and interpretation of the human genome is discussed below.

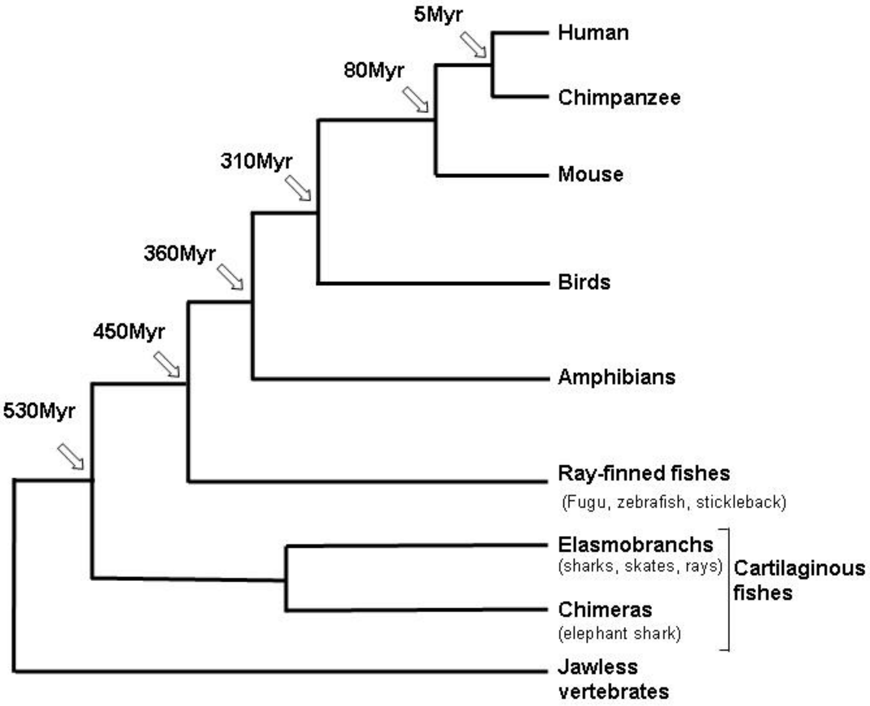


Figure.1 Phylogenetic tree of vertebrates

About the elephant shark

The elephant shark is also known as ‘elephant fish’, ‘ghost shark’, ‘reperepe’ and ‘whitefish’. The natural habitat of elephant sharks lies within the continental shelves of New Zealand and southern Australia at depths of 200 to 500 m. They grow to a maximum length of 120 cm. Mature adults migrate into large estuaries and inshore bays for spawning during spring and summer. Females lay eggs on sandy or muddy substrates. The egg cases are large, about 25 cm long and 10 cm wide. Like shark and skate eggs, elephant shark eggs take 6 to 8 months to hatch (Last and Stevens, 1994). There is no information about the chromosome number in the elephant shark. However, a closely related chimera, the ratfish (*Hydrolagus colliei*), has been shown to contain 29N dot-like chromosomes (Ohno et al., 1969) that resemble microchromosomes in birds.

Elephant sharks are exploited commercially, particularly during spring and summer when they migrate into shallow coastal waters. About 1,000 tonnes of elephant shark are caught in New Zealand alone. The white flesh fillets of elephant shark are very popular with ‘fish-and-chips’ restaurants in New Zealand and Australia.

Specific biological/biomedical rationales for the utility of new sequence data

A compact cartilaginous fish genome

Cartilaginous fishes are the most basal living group of jawed vertebrates. They possess complex physiological systems such as a central nervous system, adaptive immune system, circulatory system and neuroendocrine system that are typical of vertebrates. Like ray-finned fishes, cartilaginous fishes are amenable to physiological experiments under laboratory conditions. Several cartilaginous fishes such as spiny dogfish shark, nurse shark, horn shark, little skate and clearnose skate, are being used as model vertebrates in biomedical research to better understand the physiology of the immune system, neurobiology, pharmacology, and tumor suppression. Considering their importance to biological and biomedical research, two species of cartilaginous fishes, the little skate and the spiny dogfish shark, (both are elasmobranchs), have been proposed as models for whole genome sequencing (http://www.genome.gov/Pages/-Research/Sequencing/SeqProposals/SharkSkateSeq.pdf).

Chimeras are the other main branch of cartilaginous fishes. They possess physiological features very similar to elasmobranchs, but morphological characteristics such as a single gill opening, smooth skin, and the upper jaw fused to the head that distinguish them from elasmobranchs (Nelson, 1994). Chimeras are particularly attractive models for comparative genomic studies because of their compact genome sizes. The species of chimera proposed here, the elephant shark, has the smallest known genome size among cartilaginous fishes, and is thus an ideal genome model. In addition to its intrinsic value for comparison with other vertebrate genomes, the genome sequence of this compact cartilaginous fish can provide a framework for assembling the larger genomes of the little skate and spiny dogfish shark, which may pose a major challenge for genome assembly because of the higher content of repetitive sequences.

Compact genomes with less repetitive sequences are relatively easy to sequence and assemble using the whole-genome sequencing strategy. Due to the compact structure of genes, the signal-to-noise ratio in gene predictions is much higher in such genomes. The short intergenic regions and introns uncluttered with dispersed repetitive sequences are ideally suited for identifying regulatory regions through comparative genomics. The power of the compact genome size in comparative genomics has been amply demonstrated with the fugu genome which has a small genome size of 0.4 Gb. A draft genome sequence of the fugu was generated by a purely whole-genome shotgun sequencing strategy with a relatively small budget (Aparicio et al., 2002). About 1,000 novel genes in the human genome that were undetected by other approaches, as well as a large number of conserved regulatory elements in the human genome have been identified by comparing fugu and human genome sequences (Aparicio et al., 2002; Venkatesh and Yap, 2005; Woolfe et al., 2005).

Informing the human sequence and providing an outgroup for tetrapods and ray-finned fishes

With the availability of the genome sequences of several vertebrates, comparative genomics has become a powerful tool for identifying functional elements in the human genome. Although comparative analysis of genomes from human and other mammals has proved to be informative, such comparisons between closely related species do not offer adequate resolution for unambiguously distinguishing conserved functional sequences (particularly the non-coding elements) that are evolving slowly from those that are similar because of the lack of adequate divergence period. Comparison of evolutionary distant vertebrates can provide such a resolution to illuminate functional elements that are evolving slowly due to functional constraints. Basal vertebrates such as fugu have proved valuable in identifying conserved coding and non-coding sequences in the human genome (Venkatesh and Yap, 2005; Woolfe et al., 2005). However, cartilaginous fishes, which diverged from their common ancestor with mammals about 530 Myr ago (Kumar and Hedges, 1998), are the most distant jawed vertebrate to humans and offer the highest stringency to highlight conserved functional elements in the human genome. Interestingly, in spite of the large evolutionary distance between human and cartilaginous fishes, a higher level of non-coding sequences has been found to be conserved between human and cartilaginous fishes, than between human and ray-finned fishes. For example, a significant number of putative regulatory elements that are conserved at the HoxA locus in human and the horn shark were found to be either lost or very divergent in the HoxA locus from ray-finned fishes such as zebrafish and striped bass (Chiu et al., 2002). Even coding sequences appear to be more similar between humans and cartilaginous fishes than between human and ray-finned fishes such as fugu and zebrafish. Comparisons of the elephant shark sequences with the human, fugu and zebrafish sequences have indicated that a significant number of coding sequences in the human genome are more similar to elephant shark sequences than to ray-finned fish sequences (Venkatesh et al., 2005; Venkatesh et al., submitted). Furthermore, some of the genes that are highly conserved between the elephant shark and human were found to be missing in fugu and zebrafish (Venkatesh et al., 2005; Venkatesh et al., submitted). These comparisons suggest that ray-finned fishes have been evolving rapidly after they split from the common ancestor with mammals. A whole-genome duplication proposed to have occurred in the ray-finned fish lineage (Christoffels et al., 2004; Jaillon et al., 2004; Postlethwait et al., 2004) might have contributed to the accelerated evolution and divergence of ray-finned fish genome sequences. The whole-genome duplication might have also led to a higher rate of chromosomal rearrangements resulting in the disruption of synteny blocks that are conserved in other vertebrates. For example, the primitive synteny block of the MHC class I and class II molecule genes are highly conserved in cartilaginous fishes and mammals. However, these two related clusters of gene families are located on different chromosomes in ray-finned fishes (Ohta et al., 2000). Such rearrangements could have been mediated through divergent loss of duplicate chromosome segments or through inter- and intrachromosomal rearrangements by homologous recombination between duplicate segments. As such, the genomes of ray-finned fishes appear to be very divergent from 'mainstream' vertebrates. This underscores the importance of a cartilaginous fish genome as a basal 'mainstream' vertebrate genome for illuminating the human genome.

Cartilaginous fishes represent the ancestral lineage that gave rise to both tetrapods and ray-finned fishes, the two branches of bony vertebrates. Besides the human genome, genomes of several tetrapods (chimpanzee, mouse, dog, cow, chicken, frog, etc.) and ray-finned fishes (fugu, *Tetraodon*, zebrafish, medaka and stickleback) are the subject of genome sequencing and comparative analysis. Comparative analyses of genomes from the two bony vertebrate lineages have identified several genetic features that are unique to each of the lineage. However, without the genome sequence of a common outgroup, the ancestral states of these elements will remain obscure. Genetic information of the elephant shark will constitute such an outgroup for understanding the evolutionary changes in the genomes of the divergent lineages of tetrapods and ray-finned fishes and help to highlight the lineage-specific changes that have contributed to their unique morphology and physiology.

Expanding understanding of evolutionary processes

A big leap in the complexity of vertebrate genomes appears to have occurred during the evolution of jawed vertebrates from jawless vertebrates within a short period of about 70 Myr. Genetic components that gave rise to several complex morphological and physiological features of jawed vertebrates were invented during this period. Genome sequence of a basal jawed vertebrate, such as the elephant shark, should inform the ancestral state of various genetic components of the vertebrate physiological systems and shed light on the evolutionary origin of vertebrate-specific genes. It has been proposed that the availability of additional genetic material was the major driving force behind the origin of novel genes in jawed vertebrates and that the additional genetic material was generated through whole genome duplication(s) (Ohno, 1970). Based on the number of Hox gene clusters in invertebrate chordates such as Amphioxus, which contains a single Hox cluster, and in mammals, which contain four Hox clusters (designated HoxA, B, C and D), and the presence of a large number of human gene families with up to four members, it has been proposed that two rounds of genome duplication (so called '2R' hypothesis) occurred during the evolution of early vertebrates (Abi-Rached et al., 2002; Holland, 2003; Lundin et al., 2003). However, comparisons of whole genome sequences of human, fugu and invertebrates have not provided unequivocal evidence to support two rounds of genome duplication. Instead, the evidence support scenarios ranging from a single round of genome duplication, to single genome duplication followed by several rounds of segmental duplications and to two rounds of genome duplication in quick succession (Gu et al., 2002; McLysaght et al., 2002; Panopoulou et al., 2003; Vandepoele et al., 2004). Cloning of Hox gene clusters in a cartilaginous fish, the horn shark, has so far identified only two Hox clusters (Kim et al., 2000). Since these clusters have been identified as orthologs of mammalian HoxA and D clusters, it is predicted that cartilaginous fishes may contain more than two clusters. The elephant shark genome sequence should help to identify signatures of gene and genome duplications in this basal vertebrate and to resolve the major question regarding the role of gene and genome duplications in the evolution of vertebrates.

Strategic issues in acquiring new sequence data

1. The demand for the new sequence data. The research community that will be most enthusiastic about a draft elephant shark genome sequence have interests in comparative vertebrate genomics and vertebrate evolution. In order to study the recent evolution of gene families, protein domains, and to identify functional non-coding sequences in vertebrate genomes, a high quality genome sequence from the most basal class of jawed vertebrates will be extremely valuable. Comparison with distantly related genomes can highlight conserved elements that likely play fundamental roles in vertebrate development and physiology. Among the vertebrate taxa that are most distant from human, teleost fishes have been valuable for discovery of novel human genes (Aparicio et al., 2002; Jaillon et al., 2004) and regulatory elements (Shin et al., 2005; Woolfe et al., 2005). However, such comparisons between human and teleost fish are complicated by many genes and chromosomal segments that are duplicated in the teleost genomes. Preliminary studies with survey sequence (see below) suggest this is not a complicating factor for the elephant shark genome.

2. The suitability of the organism for experimentation. Historically, the elephant shark has not been studied in the laboratory, and its suitability for future laboratory studies is unknown. Although elephant shark genes may find uses in gene transfer or gene mutation studies, the principal objective of sequencing the elephant shark genome is not for direct experimental purposes, but to obtain a high-quality genome sequence for a cartilaginous fish. Such a sequence would be extremely valuable for comparative vertebrate genomics, and thereby greatly inform future experimental designs in established model organisms (e.g. functional evaluation of conserved non-coding sequences in transgenic mice).

3. The rationale for the complete sequence of the organism. The principal attractions of the elephant shark genome are its compactness, and apparent absence of complex repeats or significant heterozygosity. Consequently, the elephant shark can yield a very economical route to the first high-quality genome assembly for a cartilaginous fish. Such a high-quality assembly is important for accurate prediction and ordering of exons, for associating conserved non-coding sequences with nearby genes, and for delineating chromosome architecture. Therefore, to fully exploit a basal genome sequence for comparative studies (e.g. evolution of vertebrate coding and conserved non-coding sequences), a high-quality assembly is critical. Furthermore, a high-quality assembly of the elephant shark genome could be a useful framework for aiding assembly of other cartilaginous fish genomes, such as that of the little skate (*Raja erinacea*) which has already been approved for draft-level sequencing by NHGRI. Notably, our proposal to generate an additional 4.5x coverage of the 0.9 Gb elephant shark genome can be achieved for a similar cost as 1.2x coverage of the 3.5 Gb little skate genome.

4. The cost of sequencing the genome and the state of readiness of the organism's DNA for sequencing. Based on an analysis of the 1.5x assembly statistics (see below), we estimate the size of the euchromatic elephant shark genome to be ~1 Gb. Previously, the use of survey sequence data to estimate euchromatic genome size has proven to be very accurate (Kirkness et al., 2003; Lindblad-Toh et al., 2005). The survey sequence data indicates that there are few long repetitive elements that cause collapse of sequence reads into deep contigs. We propose that an additional 4.5x sequence coverage (~4 Gb) be generated from the ends of small and medium insert

plasmids (3-12 kb), and combined with the existing fosmid reads (1.5x) to generate a 6x draft WGS assembly. There exists an abundant supply of DNA from the same individual elephant shark that was used to create the fosmid libraries.

5. Are there other (partial) sources of funding available or being sought for this sequencing project? The pilot project that assembled 1.5x sequence coverage from approximately 1 million fosmid clones was funded by Singapore's Agency for Science, Technology and Research. All of the sequence data from that project will be made available, via the NCBI Trace Archive, for incorporation into the proposed 6x assembly.

Preliminary analysis of the elephant shark genome sequence

Scientists at the Institute of Molecular and Cell Biology, Singapore, and the Venter Institute have collaborated to survey the elephant shark genome after limited sequencing. Recently, it has been possible to reduce the cost of fosmid end sequencing to that of conventional plasmids. Consequently, the survey involved the generation and assembly of end reads from approximately 1 million fosmid clones, yielding ~1.5x sequence coverage and ~35x clone coverage. The use of fosmids has provided valuable information on conserved synteny among the genomes of elephant shark and reference species, even after only 1.5x coverage. In addition, the fosmids will provide valuable linking information for future assemblies that are conducted after higher levels of sequence coverage.

Analysis of the assembled elephant shark genome sequence has supported the predicted compact genome size and indicated an absence of complex repetitive elements. Approximately 30% of the genome is repetitive, consisting mainly of four families of retrotransposon-like elements that have not been described previously. There are few long repetitive elements that cause collapse of sequence reads into deep contigs. The 1.5x assembly contains approximately 61,000 'genic regions' (exons or exon clusters) that can be translated and aligned to the predicted protein products of 10,400 unique human genes. This value represents a minimum count and does not include many paralogous elephant shark genes that display their best alignments to the same human protein. Interestingly, many of the predicted elephant shark genes that are homologous to human genes, lack orthologues in any of the sequenced teleost fish. With respect to conserved non protein-coding elements, a significantly higher proportion of non-coding sequence is conserved between the genomes of elephant shark and human compared to that conserved between the genomes of teleost fish and human. Many of the highly conserved non-coding elements are likely to function as tissue-specific enhancers, influencing the regulatory networks that underlie the unique developmental biology and physiology of vertebrates.

In summary, survey-sequencing of elephant shark has demonstrated a relatively compact genome, with no evidence of complex repetitive sequences or significant heterozygosity. Comparison with reference vertebrate genomes has revealed proteins, protein domains and non protein-coding genomic elements that have been highly conserved throughout vertebrate evolution. It also permits us to identify many genes and non-genic elements that likely existed in the common ancestor of all jawed vertebrates, but have been lost selectively for one or more vertebrate lineages over time. In common with previous work on fugu and *Tetraodon*, the elephant shark offers an economical opportunity to make very rapid progress in the field of comparative vertebrate genomics.

References

Abi-Rached, L., Giles, A., Shina, T., Pontarotti, P., and Inoko, H. (2002) Evidence of en bloc duplication in vertebrate genomes. Nat Genet 31, 100-105.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia J., 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.

Chiu, C., Amemiya, C., Dewar, K., Kim, C.B., Ruddle, F.H., and Wagner, G.P. (2002). Molecular evolution of the HoxA cluster in the three major gnathostome lineages. Proc. Natl. Acad. Sci.USA. 99, 5492-5497.

Christoffels, A., Koh, E.G.L., Chia, J., Brenner, S., Aparicio, S., and Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol. Biol. Evol. 21, 1146-1151.

Gu, X., Wang, Y., and Gu, J. (2002). Age distribution of human gene families shows significant roles of both large- and small-scale duplications in vertebrate evolution. Nat Genet 31, 205-209.

Holland, P.W. (2003) More genes in vertebrates? J Struct. Funct. Genomics 3, 75-84.

Hinegardner, R. (1976). The cellular DNA content of sharks, rays and some other fishes. Comp.Biochem. Physiol. 55B, 367-370.

Jaillon et al., (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 43, 946-957.

Kikugawa, K., Katoh, K., Kuraku, S., Sakurai, H., Ishida, O., Iwabe, N., and Miyata, T. (2004). Basal jawed vertebrate phylogeny inferred from multiple nuclear DNA-coded genes. BMC Biol. 2:3, 1-11.

Kim, C.B., Amemiya, C., Bailey, W., Kawasaki, K., Mezey, J., Miller, W., Minoshima, S., Shimizu, S., Wagner, G. and Ruddle, F. (2000) Hox cluster genomics in the horn shark, Heterodontus francisci. Proc. Natl. Acad. Sci.USA 97, 1655-1660.

Kirkness, E.F. et al. (2003). The dog genome: survey sequencing and comparative analysis. Science 301, 1898-903.

Kumar, S., and Hedges, S.B. (1998). A molecular timescale for vertebrate evolution. Nature 392, 917-920.

Last, P.R., and Stevens, J.D. (1994). Sharks and Rays of Australia. p465-466. CSIRO Australia.

Lindblad-Toh, K. et al. (2005) Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438, 803-19.

Lundin, L.G., Larhammar, D., Hallbook, F. (2003) Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. J. Struct. Funct. Genomics 3, 53-63.

McLysaght, A., Hokamp, K., and Wolfe, K.H. (2002). Extensive genomic duplication during early chordate evolution. Nat Genet 31, 200-204.

Nelson JS (1994) : Fishes of the world. edn. 3. Wiley & Sons, New York; pp 600.

Ohno, S. (1970) Evolution by gene duplication. New York, Springer Verlag.

Ohno, S., Muramoto, J., Stenius, C., Christian, L., Kittrel, W.A. and Atkin, N.B. (1969) Microchromosomes in holocephalian, chondrostean and holostean fishes. Chromosoma (Berl.) 26, 35-40.

Ohta, Y., Okamura, K., McKinney, E.C., Bartl, S., Hashimoto, K., and Flajnik, M.F. (2000). Primitive synteny of vertebrate major histocompatibility complex class I and class II genes. Proc. Natl. Acad. Sci.USA. 97, 4712-4717.

Panopoulou, G., Henning, S., Groth, D. Krause, A. et al., (2003) New evidence for genome-wide duplications at the origin of vertebrates using an *Amphioxus* gene set and completed animal genomes. Genome Res. 13, 1056-1066.

Postlethwait, J., Amores, A., Cresko, W., Singer, A. and Yan, Yi-Lin. (2004). Subfunction partitioning, the teleost radiation and the annotation of the human genome. Trends Genet. 20, 481-490.

Shin, J. T. et al. (2005) Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. Nucleic Acids Res 33, 5437-45 .

Stingo, V. (1979) New developments in vertebrate cytotaxonomy II. The chromosomes of the cartilaginous fishes. Genetica 50, 227-239.

Schwartz, F.J. and Maddock, M.B. (2002) Cytogenetics of the elasmobranchs: genome evolution and phylogenetic implications. Mar. Freshwater Res. 53, 491-502.

Vandepoele, K., De vos, W., Taylor, J.S., Meyer, A. and Van de Peer, A. (2004) Major events in the genome evolution of vertebrates: Paranome age and size differs considerably between ray-finned fishes and land vertebrates. Proc. Natl. Acad. Sci. USA. 101, 1638-161643.

Venkatesh, B., Tay, A., Dandona, N., Patil, J.G. and Brenner, S. (2005) A compact cartilaginous fish model genome. Curr. Biol. 15, R82-R83.

Venkatesh, B. and Yap, W.H. (2005) Comparative genomics using fugu: a tool for the identification of conserved vertebrate *cis*-regulatory elements. BioEssays 27, 100-107.

Woolfe, A., et al., (2005). Highly conserved non-coding sequences are associated with vertebrate development. PLOS Biol. 3, 1-15.