Evolution of the human proteome: Completing the Chordate Nodes

Prepared by members of the Comparative Genome Evolution Committee
(John Gerhart, Marianne Bronner-Fraser, Scott Edwards and Peter Holland)

August 7, 2006

Summary: We propose the sequencing of **9 chordate species** to complete the coverage of all major nodes of chordate evolution with at least two sequenced species. These include;

1. **Reptile 1**: *Alligator mississipiensis* (2.5 Gb genome), the American alligator, for high quality draft genomic sequence (approx. 6x) plus ESTs.
2. **Reptile 2:** *Chrysemys picta* (2.6 Gb genome), the painted turtle, for high quality draft genomic sequence (approx. 6x) plus ESTs.
3. **Amphibian**: *Ambystoma mexicanum* (>32 Gb genome), the Mexican axolotl, a urodele, for 100,000 ESTs only.
4. **Sarcopterygian fish 1**: *Latimeria chalumnae* (2.75 Gb genome*)*, the African coelacanth, for high quality draft genomic sequence (approx. 6x) plus ESTs.
5. **Sarcopterygian fish 2**: *Neoceratodus forsteri* (>30 Gb genome), the Australian lungfish, for 100,000 ESTs only.
6. **Basal actinopterygian fish**: *Lepisosteus oculatus* (1.4 Gb genome), the spotted gar, for high quality draft genomic sequence (approx. 6x) plus ESTs.
7. **Cartliaginous fish**: *Raja erinacea* (3.3 Gb genome), the little skate, for high quality draft genomic sequence (approx. 6x) plus ESTs.
8. **Jawless fish**: *Eptatretus burgeri* (2.5-3.0 Gb genome), the Japanese Hagfish, for high quality draft genomic sequence (approx. 6x) plus ESTs.
9. **Basal chordate:** *Branchiostoma lanceolatum* (0.6 Gb genome), the European lancelet or amphioxus, for high quality draft genomic sequence (approx. 6x) plus 50,000 ESTs

Introduction

The overarching goal of this proposal is to assemble comparative genomic data to enable stepwise reconstruction of the evolution of the human proteome. This involves tracing every human gene back to its progenitor genes or modules, at each stage of its ancestry, from mammals to the first multicellular animals. These data will then enable resolution of a fundamental question: how has the evolution and deployment of new proteins, as well as expansion of protein families, contributed to each evolutionary step in human ancestry?

The proposal builds upon previous proposals from the CGE to sequence the genomes of various non-mammalian animals at key phylogenetic positions. These previous proposals, several of which are already generating informative data, form the baseline onto which the current proposal is built. Here we propose additional taxa, to fill in key “nodes” that are as yet unsampled and also to widen sampling at each node, thereby facilitating reconstruction of ancestral states. In some cases, whole genome sequencing is proposed; where this may be difficult due to large genome sizes, alternative strategies are proposed.

In Part I below we explain the scientific rationale behind the proposal, outlining why reconstructing the evolution of the human proteome is of fundamental interest. In Part II, we describe the phylogenetic nodes in the most recent 500-600 million years of our evolutionary history (within the chordate phylum), and explain where our proposed species for sequencing fit in relation to the nodes. In Part III, we describe each species in detail, including (where appropriate) discussion of existing data for other taxa at that node. In one case, a white paper (coelacanth) is appended for more information.

Part I: Rationale

One popular line of thought that is currently prominent in the evolutionary developmental biology community, is that cis-regulatory change has been paramount in the evolutionary diversification of animals. While the importance of cis-regulatory change should not be overlooked, concentration on this view detracts attention from some major recent findings in comparative genomics. One finding is that despite widespread enthusiasm surrounding genes such as Hox genes, hedgehog, and others, only a limited subset of genes is widely conserved across the animal kingdom. The proteome is actually a mix of conserved and changing sequences. In the conserved subset are many proteins of metabolism, ribosomes, the cytoskeleton, signal transduction, and the cell cycle. For example, in a comparison of yeast and humans, separated in evolution by perhaps a billion years, alpha actins are 89% identical in amino acid sequence, alpha tubulins are 76% identical, and glyceraldehyde 3-phosphate dehydrogenases are 65% identical. In particular, the core domains of proteins are highly conserved, and these often serve to identify proteins in blast searches. But every genome sequenced to date, including the human genome, also contains species-specific or taxon-specific genes and gene families. These genes have arisen through a combination of gene duplication, accelerated protein sequence divergence, and domain transposition or copying. New combinations of protein domains can be found in each genome, while even the most conserved regulatory proteins can show alterations in accessory binding sites and in regions affecting activity by post-translational modification, localization, or stability, separate from the core conserved domains.

Furthermore, these changes to the proteome architecture of each species are occuring over all time scales. At one extreme, many human proteins have recognisable orthologues right across the eukaryotes, such as the ribosomal proteins. Other proteins are found throughout the multicellular animals, but not beyond, such as specific families of receptor tyrosine kinases and hedgehog signalling proteins. Many well-characterised transcription factors and signalling molecules have unambiguous orthologues across bilaterians (e.g. from human to flies), while others are more restricted in their distribution, emerging gradually through the course of chordate evolution. Finally, at the other extreme, a small proportion of human proteins do not have clear orthologues even within the mouse genome. The key point is that there is no single evolutionary distance of comparison where all insights about proteome evolution can be obtained. Some proteins change rapidly, some slowly, and some episodically. Also, proteins change more rapidly in some lineages than others. A rough approximation of rate for chordates would be 0.1 amino acid substitution per site per 300 million years, averaged over a group of 50 proteins (Blair and Hedges, 2005). What is required is a stepwise approach, where the complete proteome is predicted for every node on the evolutionary lineage leading to humans.

Relatively short evolutionary timescales (that is, the most recent 100 millions years of our history) can be addressed by comparative analysis of mammalian genomes. The AHG working group is dedicated to mammalian evolution as part of their annotation of the human genome. Since mammals are a tight knit group (relative to the entire chordate phylum), this approach promises to deliver long-range synteny, plus also identification of introns, cis-regulatory sequences, and other conserved non-coding sequences. Comparison between mammals is likely to shed less light on protein family evolution, because of the high level of functional and sequence conservation in many proteins, although it will be useful for identifying the most rapidly changing protein

sequences, identification of residues subject to positive or negative selection, and cases of recent gene cluster expansion and contraction.

This proposal is complementary to those of the AHG working group, and takes advantage of more widely diverged animals, but still within our phylum, the Chordata. These evolutionary distances are ideal for gaining insight into the pathways of evolution of coding sequences, and also those transcribed RNAs for which there is some degree of conservation. Our questions, then, focus on the evolution of the proteome and transcriptome across the Chordata. Ultimately, we aim to provide a data set from which researchers can trace the origin of every human gene and gene family. Furthermore, these changes in the protein repertoire can be related to major steps in the evolution of the human lineage.

A few examples are useful for emphasizing these different time scales. The beta/gamma crystallins of the vertebrate eye lens provide a good example of how duplication and fusion of domains contributed to evolution within chordates. The ability to domain pair within or between protein polypeptide chains in this protein family was only acquired within vertebrates, after complex changes to these proteins. Other protein types that have so far been found only in vertebrates include: calcium phosphate-associated bone proteins, several proteins of the adaptive immune system and of myelination, and the Cerberus signaling antagonist. In these cases, and more, investigating the organisation of homologous proteins in vertebrates and basal chordates will be important to elucidate the pathway of proteome evolution. Moving closer to humans, there are also proteins that have so far only been found in mammals. Thus, comparison of the whole genome sequence of chicken with those of mammals highlights many groups of potentially mammal-specific proteins, including a group of alpha-interferons, the high/ultra sulfur-hair keratins, the DUX double homeodomain proteins, the casein milk proteins, lactose producing alpha lactalbumin, salivary-associated proteins (statherin and histatins), brown fat uncoupling proteins, and type 2 taste receptors. In these and other cases, genome data from other amniotes (e.g. non-avian reptiles) and non-amniotes (amphibian, coelacanth, lungfish) will clarify the course of evolution and reveal ancestral domains. Furthermore, the increased sampling would increase confidence in the assertion that particular proteins are indeed mammal specific and not simply lost from birds.

Changes to the protein repertoire can be described in terms of (a) duplications and losses, (b) domain copying, movements and fusions, and (c) adaptive change to protein sequences. What is the most effective strategy to identify all these events in the evolution of the human proteome? Key to the strategy is the careful choice of species. In particular, we argue that it is important to have information on every node in human evolutionary history. This can be obtained by genomic analysis of species that diverged from our lineage at each of these nodes, carefully choosing species that are basal and less “derived” representatives of their respective groups. Discounting nodes within mammalian radiation (for the reasons given above), there are only about 8 of these nodes in the chordate period of our history, and about as many in the non-chordate period. Here we concentrate on the chordate period.

A second factor to be taken into account is that it will be important to have genomic information from at least two deeply diverged species at each node. This is because if only a single species is sampled, insights into ancestral states can be confounded by radical lineage-specific changes and gene losses. These become much easier to overcome if there is information from two derivatives from a node, particularly if they are divergent species (i.e. the two species chosen diverged from each other early after their shared ancestor with humans). A few examples illustrate this point. (1) The nematode *Caenorhabditis elegans* would have been a very poor model for the

'protostome node' if it were the only protostome sampled. Many genes have radically diverged such that they are barely recognizable (e.g. hedgehog gene family), while some gene families have suffered such extensive rearrangement and loss that the ancestral situation is impossible to deduce (e.g. Hox genes). Additional nematode data, and other protostome genomes, have already clarified this picture. (2) The ray-finned fish node is currently represented by teleost fish genomes (e.g. pufferfish). But teleosts have clearly undergone a whole genome duplication, which in turn has facilitated the dissociation of otherwise conserved gene assemblages. Hence, it is markedly difficult to reconstruct the ancestral state for the 'ray-finned fish node' using teleost fish data alone. Sequence information from a non-teleost ray-finned fish would be informative to overcome this problem. (3) A third example is lamprey. This occupies a crucial position as one of the earliest lineages to have diverged from the rest of the vertebrates, so insight into this node is critical. Reconstructing the ancestral state at this 'cyclostome' node from the genome of a single living species (compared to other vertebrates of course) will be greatly aided if a second cyclostome, notably hagfish, is included.

For the above reasons, the choice of species, and the double sampling of each node, will be vital. We must then consider what type of genomic data is required. Whole genome sequence will always be the most exhaustive data source for a given species, and this is recommended for most species listed in this proposal. This is partly because of the obvious reason of completeness of sampling, such that weakly expressed genes are not missed in EST screens. In addition, there is a second benefit to whole genome data that will be central in the more problematic cases of proteome evolution. This is the identification of orthologues through synteny. When gene sequences undergo extensive change, for example through positive selection, it can be difficult to identify true orthologues between species. This is particularly troublesome in complex gene families, where several genes could be candidates for being the true orthologue. In several recent cases, syntenic analysis has proven the key to finding these divergent or 'cryptic' orthologues. That is, the neighbors of the problematic gene are first found in the human genome, and then these are more easily located in the genome sequence of the target animal. Analysis of the scaffold around those neighbor orthologues can often reveal the identity and sequence of the 'cryptic' orthologue. For example, the mouse orthologue of the rapidly evolving human *TPRX1* homeobox gene was found in the same way (Booth and Holland, 2006).

Parenthetically it should be said that while this is a proteome proposal, we fully appreciate that there will also be benefits from the whole genome sequencing of these diverse chordates for the ongoing tracking of conserved non-transcribed sequences, regulatory and others, back to the chordate ancestor, and we support this sequencing objective. Furthermore, we appreciate that the information gained on cis-regulatory sequences will have experimental value to researchers using these species, or wanting to test such sequences in other species, and again, we support this objective.

Returning to the proteome proposal as such, we note that in two of the species proposed, enormous genome size (10-30x human) precludes whole genome sequencing with current technologies. These are, however, species that diverged from important nodes on our lineage. In these two cases, we propose extensive sequencing of ESTs. While this approach will never give the same depth of insight as gained by whole genome sequencing, it is still a powerful way to address particular questions. A perceived problem with EST sequencing is that genes will be missed, leading to an inaccurate view of proteome evolution. In reality, the problem is less serious. When considering the evolution of a particular protein family, it is not a simple matter of counting how many members are found in each family (i.e., 47 zinc fingers, 23 Lim, etc), but rather of drawing trees of domains, and making inferences from those trees. With a

sufficient sampling of species (and assuming most of these have whole genome sequence data), then the fact that some genes will be missed in some EST screens does not compromise all of the biological insight. It is still often possible to infer the ancestral condition for a gene family by extrapolation from the tree topology, particularly for genes with well conserved domains (less so for rapidly evolving genes). What cannot be achieved with the use of ESTs alone is the use of synteny to identify orthologues, and the examination of ancestral gene linkages.

References for Part I:

Blair JE and Hedges SB. (2005) Molecular phylogeny and divergence times of deuterostome animals. Mol. Biol. Evol. 22: 2275-2284.

Booth HAF and Holland PWH (2006) Annotation, nomenclature and evolution of four novel homeobox genes expressed in the human germ line. Gene**.** In press. doi:10.1016/j.gene.2006.07.034

Delsuc F, Brinkmann H, Chourrout, D, and Philippe H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439: 965-968.

Part II: Chordate nodes

An accompanying figure presents 8 major nodes in the evolution of chordates, as the lineage from mammals is followed back to the chordate ancestor. Moving back in time, the significant nodes are in turn denoted by the ancestors of (1) amniotes (mammals and reptiles, including birds), (2) tetrapods (amniotes and amphibia), (3) sarcopterygian [lobe-fin] fish (tetrapods plus coelacanth and lungfish), (4) bony fish (sarcopterygians plus ray-finned fish), (5) jawed fish (bony fish plus cartilaginous fish [sharks, skates, rays, chimeras], (6) jawless fish, which would be the node of the vertebrate ancestor (jawed vertebrates plus hagfish and lamprey), (7) an enigmatic urochordate-vertebrate ancestor, and (8) the chordate ancestor (cephalochordates plus urochordates plus vertebrates). This ordering reflects the recent genomic evidence (Delsuc et al, 2006) that cephalochordates (amphioxus) are the basally split chordate group, not urochordates (ascidians, larvaceans), as had been widely thought for the past century.

At present, the nodes are unevenly characterized by the sequencing of chordate species, and our intention in this proposal is to bring all nodes to the level where each is informed by sequencing data from at least two well diverged species. As noted, above, the evolutionary differences of branches from each node makes necessary such two-fold coverage, just to reduce the chance of conclusions based on idiosyncracies of a particular species. The following list is a summary of the status of the current coverage of nodes and our proposed species:

The mammalian node: already well covered by 44 sequenced mammalian species, completed and pending, some at 2x, some at high quality draft, and a few at finishing levels.

The amniote node: in addition to the mammals, two birds have been done (chicken) or in progress (zebra finch), and the green anole lizard is the only non-avian reptile accepted for sequencing, now in progress at the Broad Institute. We propose two more non-avian reptiles:

1. **The American alligator**, *Alligator mississipiensis* (2.5 Gb genome; high quality draft genomic sequence plus ESTs), which is on the reptilian branch with avians,

2. **The painted turtle**, *Chrysemys picta* (2.6 Gb genome; high quality draft genomic sequence plus ESTs). Note that turtles are deeply split from lizards and alligators, but are not basal among reptiles, as was thought a few years ago before sequence comparisons were made.

The tetrapod node: In addition to the above species, only one amphibian amphibian genome has been sequenced for this node, the anuran (frog) *Xenopus tropicalis*. We propose another amphibian,

3. **The Mexican axolotl**, *Ambystoma mexicanum* (>32 Gb genome), a urodele (salamanders, newts), for 100,000 ESTs only, since the genome is prohibitively large.

The sarcopterygian node: To date, nothing has been sequenced for this node. We recommend both a coelacanth and a lungfish, which are the only two possibilities.

4. **The African coelacanth**, *Latimeria chalumnae* (2.75 Gb genome) high quality draft genomic sequence plus ESTs, and
5. **The Australian lungfish**, *Neoceratodus forsteri* (>30 Gb genome); for 100,000 ESTs only.

The bony fish node: At present 4 teleosts have been sequenced and one is in progress (the stickleback). Since teleosts have undergone a group-specific genome duplication and extensive gene loss, we propose a holostean species, which is a basal actinopterygian (ray-finned) fish that has not undergone such changes, namely,

6. **The spotted gar**, *Lepisosteus oculatus* (1.2Gb genome; high quality draft genomic sequence plus ESTs).

The jawed vertebrate node: Beyond the species above, nothing has been done here. We recommend a skate with a middle-sized genome,

7. **The little skate**, *Raja erinacea* (3.3 Gb genome; high quality draft genomic sequence plus ESTs). Although we considered the elephant "shark" (0.9 Gb genome, actually a chimera) as an alternative, we disfavored it because so little experimental use has been made of it.

The jawless vertebrate node, that is, the vertebrate ancestor. The lamprey genome is currently in progress. We recommend a member of the only other jawless fish group, the hagfish, namely

8. **The Japanese hagfish**, *Eptatretus burgeri* (2.5Gb genome; high quality draft genomic sequence plus ESTs).

The urochordate/vertebrate node: The genomes of two ascidian species have been sequenced, namely, of *Ciona savignyi* and *Ciona intestinalis*, and of one larvacean species, *Oikopleura dioica*. We consider this to be an adequate representation of this node for now and make no further recommendation.

The chordate ancestor node: Cephalochordates are now seen as the basally split chordate group, not urochordates. One species of amphioxus has been done recently, *Branciostoma floridae*, and we propose a second species,

9. **The European lancelet**, *Branchiostoma lanceolatum* (0.6Gb genome; high quality draft genomic sequence plus ESTs).

Part III. Descriptions of the proposed species and their existing resources:

Recommendation 1: Reptile 1, *Alligator mississipiensis*, the American alligator (2.49 Gb genome) for high quality draft genomic sequence (approx. 6x) plus ESTs.

.

Phylogenetic position:

Reptilia (non-avian reptiles + birds) are the sister group of mammals, and as such, occupy an important position for illuminating the ancestry of the human proteome. Lizards (including *Anolis*) and snakes comprise the lepidosaurs, which as a group are basally split within reptiles relative to the the "Archosauriomorpha', which now include crocodilians, birds, and turtles. Although paleontology had long held turtles to be the basal branch within the Reptilia, recent phylogenetic studies of nuclear and mitochondrial genes firmly place turtles close to archosaurs (birds + crocodilians), whereas lizards and snakes (lepidosaurs) are the basal split (Iwabe et al. 2005). Thus, within the reptiles, the alligator is phylogenetically intermediate between *Anolis* and birds, and closer to birds than are turtles. Alligators and birds diverged approximately 220Mya, and for a variety of reasons, including slow rates of genomic and chromosmal evolution, good synteny is expected between alligator, chicken and other birds, and *Anolis* (Edwards et al. 2005).

Sequencing of the alligator will increase the information about the amniote node from which reptiles, birds, and mammals descended. This node is the key node at which the first adaptations to fully terrestrial life arose, including the reproductive modifications of a cleiodoic egg (that is, enclosed in shells and membrane), massive yolk stores, and the development of four major extraembryonic tissues, the amnion (as in the term "amniotes") being but one of these. Among non-avian reptiles, only the green anole has been chosen, and its sequencing is now in progress at the Broad Institute. We recommend that more reptiles be sequenced to inform this amniote node.

Research and health relatedness:

Research on the American alligator has provided numerous advances in human health and welfare, such as studies of physiology and blood chemistry (Coulson & Hernandez 1983); antiviral and antimicrobial blood proteins (Merchant et al. 2005); microbial induction of the immune response (Brown et al. 2001); the effects of environmental contaminants as endocrine disruptors (Guillette et al. 2000); temperature dependent sex determination (Western et al. 2000); and the development of the heart (Crossley & Altimiras 2005), scales and skin (Alibardi 2004), and feathers (Sawyer et al. 2005). Alligators harbor West Nile virus, Mycoplasma, and other disease vectors that endanger human health (Brown 2002, Jacobson et al. 2005; Merchant et al. 2005). Large numbers of eggs and animals are available for research (e.g., ~ 500,000 alligators are on farms in Louisiana alone). Its early development has been studied, and gastrulation proceeds via a primitive streak, like birds, and not via a blastopore, as in lizards, snakes, and amphibia.

Genomic and genetic resources:

The chromosome number of most crocodilians has long been known (Cohen & Gans 1970); the *A. mississipiensis* karyotype consists of 12 macrochromosomes and no microchromosomes (in contrast to birds). The complete mtDNA sequences have been obtained (Janke & Arnason 1997). Genetic variation within and among American alligator populations has been investigated using many techniques, and polymorphism is generally low, perhaps due to a recent bottleneck (reviewed by Dessauer et al. 2002).

High molecular weight DNA is available from JGI. Genomic resources recently developed for Alligator include a BAC library from JGI (http://evogen.jgi.doe.gov/second_levels/BACs/Our_libraries.html; see also http://www.benaroyaresearch.org/investigators/amemiya_chris/libraries.htm) with ~ 1900 pairs of clone-end reads (Edwards et al. unpubl.), full sequences of targeted BACs (Green et al. unpubl.; http://www.nisc.nih.gov/open_page.html?/projects/comp_seq.html), thousands of expressed sequenced tag sequences from liver and testis libraries (Guillette et al., unpubl.), and characterization of repetitive elements from BAC end reads (Shedlock et al., submitted). In addition to the Alligator BAC library, there are also BAC libraries from JGI for emu, tuatara, garter snake and gila monster, if they are wanted to add proteome information at this node.

As sequencing methods improve, we suggest for the future that at least two additional species be considered, such as a basal bird, perhaps an Emu (*Dromaius novaehollandiae*), and a snake, such as the garter snake (*Thamnophis*). The proteome complement of the ancestral amniote will improve further with sampling within Reptilia, and in turn, the distinct features of the proteome complement of mammals, when compared with their reptile-bird sister group, can be better distinguished.

Information on Alligator biology was provided by Dr. Travis Glenn glennt@biol.sc.edu and other members of the Reptile Genome Working Group www.reptilegenome.com.

References for the alligator can be found in the Appendix.

2. Reptile 2: *Chrysemys picta* (2.57 Gb), the painted turtle, for high quality draft genomic sequence (approx. 6x) plus ESTs.

Phylogenetic position:

Within the amniotes, and within the Reptilia**,** turtles and tortoises are classified in the order Testudines. From an evolutionary perspective, they represent a unique clade defined by a large number of important morphological novelties. Chief among these are a) the shell (the carapace, which itself is a combination of fused vertebrae, ribs and girdle elements, and several novel bones found in no other vertebrates), b) the placement of the girdles medial to the ribcage (a unique vertebrate feature), c) the absence of teeth (also seen in birds), and d) the ability to retract the head within the ribcage (a feature which has evolved independently in the side-neck and hidden-neck turtles; Pough et al., 2001). The development of these unique morphological features, their tempo and mode of evolution, and the lessons they tell us about amniote development are venerable problems that persist today (Rieppel 2001, Gilbert et al. 2001).

One of the key confusions in turtle evolution has been their phylogenetic relationship to other amniotes (Rieppel and Reisz 1999, Hill 2005). They used to be considered the most basally split reptile group, but now, based on both molecular and morphological evidence, turtles are viewed as a highly derived group of diapsid reptiles, which are either the sister-group to archosauria (birds plus crocodilians; Iwabe et al. 2005) or to crocodilians (Cao et al. 2000), with the bulk of current evidence favoring turtles as the sister-group to archosauria (Meyer and Zardoya 2003). Within turtles, interrelationships of the major lineages are becoming resolved (Sasaki et al. 2004, Krenz et al., 2005), as are the species-level relationships of several of the most speciose families (Engstrom et al. 2004, Spinks et al. 2004). Based on their exquisite fossil record, the living turtles are known to span some 210 million years of evolutionary history, and recent analyses provide a well-supported time frame for the diversification of the major

clades of turtles (Near et al. 2005). During that period of time, certain features of the group have been evolutionarily conservative (e.g. all turtles have shells, and all lay eggs), while others have been extremely labile (e.g. the transition from genetic to temperature-dependent sex determination).

As of 2004, an astonishing 40% of the world's approximately 300+ species of turtles were "Red listed" by the International Union for the Conservation of Nature (http://www.redlist.org/). As a clade, this makes turtles the world's most endangered higher group of vertebrates (amphibians have 32.5% of their species red listed, mammals 23%, birds 12%).

Research and health relatedness:

The advantage of *Chrysemys* over other turtles is that it is the most important North American model species of turtle for developmental biologists, physiologists and geneticists. Below we highlight a sample of important, novel research directions involving this and other turtle species.

Genome and sex chromosome evolution: Turtles are becoming a model system for genome evolution in reptiles, with recent research indicating fundamental differences in isochore structure from mammals (Kuraku et al. 2006). Turtles have both temperature-dependent and genetic sex determination, making them an ideal system in which to study the evolution of sex determination. Recent mechanistic research has mainly looked at the effect of hormones on temperature sex reversal and has recently begun to quantify the levels of transcripts of candidate sex determining genes and the evolution of sex chromosomes in turtle species (Ezaz et al. 2006, Murdock and Wibbels 2006, Matsuda et al. 2005).

Aging and oxidative damage: Turtles are extremely long-lived with very little age-related fitness decreases. Recent work shows that there is interspecific variation in turtle hemoglobin antioxidant properties (Congdon et al. 2001, 2003), and in recent tests turtle hemoglobin was shown to mitigate damage by hydrogen peroxide to the human erythrocyte wall by up to 30% (Torsoni et al. 2000).

Cryogenics: Some turtles from northern climates can undergo super-cooling and tolerate freezing. Recent studies on cyroprotectants, gene expression, and physiological (metabolic and enzymatic) responses to supercooling and freezing have established turtles as a model system for vertebrate cryogenic research (see Packard and Packard, 2004 for a review and Storey, 2006 for gene expression work).

Evo-Devo: Turtle development is a textbook example of how early embryonic cues can lead to radical evolutionary novelties. Work in development has concentrated on shell formation (Alibardi and Toni 2006, Cebra-Thomas et al. 2005, Gilbert et al. 2001, Loredo et al 2001, Vincent et al 2003, Kuraku et al. 2005, Ohya et al. 2005, Nagashima et al. 2005), eye (Francisco-Morcillo et al, 2006) and brain (Kalman et al. 1997, Hemmings and Storey 1999) development.

Metabolism and anoxia: As opposed to the mammalian brain, the turtle brain is protected against lactate or pH damage during anoxia. Investigations of gene expression, protein physiology, and cell biology in a gradient of anoxic-tolerant species have shed light on the fundamental basis of tolerance to the lack of O2 in vertebrates (Reese et al. 2004, Lutz et al. 1984).

Genomic Resources:

From a proteome point of view, the painted turtle is a suitable choice for sequencing since like all turtles, it has a genome size of about 2.57 Gb, and it is fully

representative of the monophylic turtle group, which branched from of reptile line leading to birds and crocodilians, rather than from the lizard-snake branch.

A *Chrysemys* BAC library is available through JGI (http://evogen.jgi.doe.gov/second_levels/BACs/Our_libraries.html), and various cDNA libraries are available. To our knowledge, however, no extensive EST database has been generated for the turtle. Preliminary BAC-end sequencing of ~1900 BACs (~2.5 Mb) reveals a diversity of CR1 LINE elements which are phylogenetically related to those of chickens and alligators but also show turtle-specific evolution (S. Edwards, unpubl.). Given the need to better define the proteome of the ancestral amniote, additional reptiles that will complement the upcoming *Anolis* sequence are needed if we are to understand how the mammalian proteome is unique relative to its reptile-avian sister group.

Turtle consultants for this recommendation: Drs. Brad Shaffer hbshaffer@ucdavis.edu and Shigehiro Kuraku venezia@cdb.riken.jp.

References for the turtle can be found in the Appendix

3. A urodele amphibian: *Ambystoma mexicanum*, the Mexican axolotl.

We propose the sequencing of a large set of ESTs (100,000) from cDNA libraries from embryonic and larval stages of the axolotl. The genome is too large (>30Gb) for sequencing of it at present.

Phylogenetic position: From the amphibian node evolved modern amphibia and amniotes (reptiles, birds, and mammals). The ancestor at this node would be the tetrapod ancestor, the first vertebrate adapted to terrestrial life, though still partial compared to reptiles. The evolutionary transition to land involved numerous physiological, anatomical, and hormonal changes for air breathing, weight bearing without water support, fin to limb changes especially in the wrist and hand regions, and probably a metamorphosis from a water-dwelling juvenile to a land dwelling adult. Living amphibia, which seem but shadows of their great ancestors, are the three orders of lissamphibia, that is, the Urodeles (newts and salamanders, including the axolotl), the Anurans (frogs and toads), and the Caecelians (limbless salamanders). Anurans may have split from the Urodele/Caecelian branch at least 300 Mya (San Mauro et al. 2005; Zhang et al. 2005).

The only amphibian sequenced so far is the anuran (frog) *Xenopus tropicalis,* recently completed by JGI. Further sequence information is desirable to illuminate this node, but the choices are limited due to the large genome size of many species. We have at present decided against recommending a second anuran which is deeply diverged from *Xenopus* (such as a *Rana* or *Bufo* species, with tractable genomes at 2.5-3 Gb), and have chosen instead to represent the anciently-diverged urodele side of the amphibian clade, despite the enormous genomes of all urodeles (many >30GB, some even 100 Gb). Therefore, we recommend obtaining a large set of ESTs from a urodele at his time, namely, the Mexican axolotl.

Research and health relatedness:

Urodeles, including the axolotl, have long been an attractive experimental model for studies of development. Hans Spemann's classic work on the Organizer was all done on newts. They are the favored model for studies of limb and tail regeneration, and this work has recently taken a strong molecular turn with significant results (Da Silva et al, 2002; Mercader et al, 2005) Even parts of the nervous system can regenerate at certain

stages. Transgenic axolotls can now be produced at will, and a variety of mutants exist, as do inbred lines, available form the Ambystoma Genetic Stock Center.

Additionally, the Mexican axolotl has attracted study among urodeles because of its neoteny; it become sexually mature while retaining larval characteristics. If given thyroxin, it metamorphoses into a salamander-like form resembling a related Texas species. It is thought that neoteny is an adaptation to the iodine deficient conditions of the lakes of the Mexican highlands (such as Xochimilco) where the axolotl lives, a condition precluding thyroxin synthesis.

Genomic and genetic resources:

The contact person for cDNA libraries, embryos, and tissues is Dr. Randal Voss who directs the Ambystoma Genetic Stock Center at the University of Kentucky (funded by NSF). An inbred axolotl strain is available, with estimates of its low DNA polymorphism.

The animal has been long-favored for cytogenetics, with 1N = 14 and huge chromosomes containing >30Gb of DNA.

cDNA libraries, EST collections, and BAC libraries: Approximately 30,000 ESTs now exist (Putta et al, 2004) and many are catalogued on the Ambystoma website (http://www.ambystoma.org). Many more ESTs are needed, and researchers of the axolotl plan to pursue various chip analyses that would benefit greatly from ESTs. Although several cDNA libraries are available for further EST sequencing, Dr. Voss judges that new and better libraries should be prepared for additional EST development. He would consider undertaking the cost of new library construction, if this axolotl EST sequencing project is approved.

No BAC or Fosmid library is currently available.

Axolotl researchers to make use of the sequence information: Dr. Voss' laboratory would be a heavy user, as would the following researchers: Dr. Jeremy Brockes (University College, London); Dr. Elly Tanaka (Max-Planck, Dresden), Dr. John Kauer (Tufts), Dr. Panagiotis Tsonis (U. Miami-Ohio), Dr. David Gardiner (UC Irvine), Dr. Susan Bryant (UC Irvine), Dr. Craig Crews (Yale University), Dr. John Postlethwaite (U. Oregon), Dr. Chris Beachy (Minot St U), and Dr. Linda Barlow (UColorado HSC), Dr. Julie Drawbridge (U. Rider), Dr. Vincent Laudet (Lyon, France), Dr. Andrew Storfer (Washington State U), Dr. Juan Carlos Izpisua Belmonte (Salk Institute), Dr. Marianne Bronner-Fraser (CalTech), Dr. Hans Epperlein (University of Dresden), and Dr. David Paricy (University of Washington).

References for the axolotl can be found in the Appendix.

4. Coelacanth: *Latimeria chalumnae*, the African coelacanth.

We recommend the high quality draft sequencing (approx. 6x) of the coelacanth genome (2.75Gb) and the acquisition of a large set of ESTs (100,000).

The CGE committee reviewed the recent white paper by CT Amemiya, ES Lander, and RM Myers (appended to this report), and we fully support the request. That paper should be consulted for detailed information about the animal's history, usage, and readiness for sequencing.

Phylogenetic position and research:

Sequencing the coelacanth will inform the sarcopterygian fish node, from which tetrapods (land vertebrates), lungfish, and coelacanths evolved. Sarcopterygians are the sister group of the actinopterygians, the ray finned fish, and both are bony fish. As the authors note, the coelacanth is the only possibility for full sequencing at this node; this is because lungfish have enormous genomes, >30Gb, precluding them. Separately (see

below) we will propose obtaining a set of ESTs from the Australian lungfish. Although the coelacanth is an animal of very restricted experimental availability, a few individuals are found each year in South African fish markets, making possible the intermittent collection of tissue for further libraries.

One of the most significant events of the water-land transition was the modification of the lobe-fin into the tetrapod limb. Substantial research has been done on tetrapod limb development and on teleost fin development, including identifications of genetic regulatory circuits and signaling pathways, but the lobe-fin has been the missing element of the comparison. Recent paleontological discoveries reveal a variety of stem forms and branches for the water-land transition of 360-380Mya (Shubin et al, 2006).

References for the coelacanth are given in the White paper.

4. Lungfish: *Neoceratodus forsteri*, the Australian lungfish

We propose the sequencing of a large set of ESTs (100,000) from the Australian lungfish.

Lungfish and coelacanths are the closest living relatives of tetrapods, the land-dwelling vertebrates. They all share a sarcopterygian ancestor (a lobe fin fish), a major node on the evolutionary tree to humans. Recent paleontological discoveries reveal a variety of stem forms and branches for the water-land transition of 360-380Mya (Shubin et al, 2006).

The arguments for sequencing the coelacanth apply as well to the lungfish, that is, to gain genomic information related to the transition to land, with all the anatomical and physiological changes of fin/limb anatomy, weight bearing, gas exchange, protection from dessication, sense organs, and kidney function. In various phylogenetic studies, lungfish may be slightly more closely related to tetrapods than are coelacanths (Brinkman et al, 2004), and their fins may be slightly more limb-like. The body of the Australian lungfish is quite similar to Devonian fossil forms (moreso than are the South American and African species). As the name implies, they have lungs as well as gills (as have other primitive fish, too). Moreso than the coelacanth, the Australian lungfish is amenable to experimental study, for example, in studies of lobe-fin development, and the Australian species can be bred in captivity.

As a major impediment to sequencing, however, all six living species of lungfish (Africa, South America, and Australia) have very large genomes (>30Gb), precluding whole genome sequencing at this time. Nonetheless, in light of the importance of the node, we propose the sequencing of a large collection of EST sequences (100,000) to inform the node and to have the reagents for future selective sequencing of BACs related to tetrapod innovations, such as limb development.

The Australian lungfish (*Neoceratodus forsteri*) is the species most suitable for sequencing projects at this time. Dr. Jean Joss of Marquarie University (Sydney) obtains this species from the Burnett and Mary Rivers in Queensland; she and her colleagues are the only researchers to have succeeded in breeding lungfish (in Olympic sized pools) and in obtaining all developmental stages. Although a few cDNA libraries from larvae are available, she recommends that new cDNA libraries be made, and she can provide fresh, staged material for these. The spawning season begins mid September, so from October on, the relevant stages will be available. Near hatching and larval stages would be most useful. A lobe-fin library would be desirable, given the interest in limb evolution.

Other researchers of lungfish include Dr. Cushla Metcalfe (with Didier Casane in Paris) and Dr. Jen Rock in Bangor, N. Wales. Dr. Chris Amemiya has expressed an interest in making a lungfish BAC library.

References for the lungfish can be found in the Appendix

6. Basal Actinopterygian: *Lepisosteus oculatus*, the spotted gar

The proposal is for high quality draft coverage (approximately 6x) of the whole genome of spotted gar (genome of 1.4 Gb), together with 50,000 ESTs to aid construction of gene models.

Phylogenetic position and advantages:

The actinopterygians (ray-finned fish) are the sister group to sarcopterygians (mammals and other tetrapods, coelacanth and lungfish). Comparing these genomes should reveal which proteins and modules were assembled at the origin of 'bony vertebrates' (Osteichthyes). There are already five ray-finned fish genome sequences (zebrafish, two pufferfish, stickleback, medaka) - but there is a problem that limits their utility for investigating evolution of the proteome. All these species are teleosts, and underwent a fish-specific genome duplication (FSGD) in their evolution (reviewed by Meyer and Van de Peer 2005). There are several consequences. First, each genomic region is duplicated giving "double conserved synteny", rather than simple syntenic relations to human chromosomes (Postlethwait et al., 1998; Jaillon et al. 2004). Second, following FSGD many redundant genes were lost - some gene families have a 2:1 ratio to humans, some have reverted to 1:1 and some are now 0:1, because other fish genes have taken their roles (Postlethwait et al., 1998; Mulley et al. 2006). Third, gene loss means that some ancient and conserved linkages between genes, which could be functional in the human genome, have been disassembled by differential gene loss: an example being the absence of the ParaHox gene cluster in zebrafish and pufferfish (Mulley et al. 2006). Additional genomic scrambling has also occurred, for example the MHC is split into several loci in teleosts, complicating comparison to tetrapods (Bingulac-Popovic et al. 1997; Sambrook et al. 2005). Fourth, and perhaps surprisingly, rates of molecular evolution have accelerated in teleost fish. This is already clear in non-coding DNA, where some conserved modules present at the base of bony vertebrates have been lost from teleosts (Chiu et al. 2004).

The solution to these problems is to examine the genome of ray-finned fish that diverged prior to the FSGD. There are a few candidates, notably sturgeons, paddlefish, bichir, bowfin and gar. All these animals diverged before the fish-specific genome duplication shown by teleost fish (Meyer and Van de Peer 2005; Mulley et al. 2006). There are some important consequences of this phylogenetic position. For example, sequencing of the ParaHox gene cluster in bichir (*Polypterus*) and bowfin (*Amia*) has revealed that an ancient linkage is intact in these basal ray-finned fish, with the same gene order and orientations as in human, but quite unlike teleost fish (Mulley et al. 2006). Similarly, the Hox clusters of bichirs share conserved sequences with human that have been lost from teleosts (Chiu et al. 2004). It is clear, therefore, that a genome sequence from one of the basal ray-finned fish species would be extremely useful for comparison to the human genome. There would also be added value in comparison to teleost genomes, as it would allow insight into the genomic consequences of genome duplication.

Choice of the ideal 'basal actinopterygian' for whole genome sequencing is driven by considerations of genome size, DNA availability, and practicality of keeping adults and rearing embryos. The sturgeons and paddlefish are not likely to be good candidates, because although developing stages are readily available there is evidence that these species underwent their own genome duplications, which would complicate assembly and analysis. Furthermore, sturgeon genomes are relatively large, between 3 and 5 Gb. The bichirs (genus *Polypterus*) are interesting candidates, as some very informative studies of gene clusters have already been undertaken (Chiu et al. 2005;

Mulley et al. 2006); however, these fish are difficult to breed and they have relatively large genomes (5 Gb). The two remaining taxa – bowfin and gars – have smaller genomes (1.2 to 1.4 Gb) and represent the more feasible candidate species. Of these, the bowfin (*Amia calva*) has also proved informative in studies undertaken to date, and would be a feasible choice; however, obtaining embryos and developing stages currently relies on locating spawning sites in the wild. In contrast, several gars can be spawned reliably in captivity. This would provide a supply of embryos and developing stages for experimental work, such as gene expression studies and morpholino experiments.

We suggest the most suitable choice would be the spotted gar *Lepisosteus oculatus*. This animal is widespread in the Southern USA, is readily collected and has an estimated genome size of 1.4 Gb (estimated by flow cytometry and Feulgen densitometry; Hardie and Hebert, 2004; Ojima and Yamamoto, 1990). Juvenile cDNA libraries from several tissues are under construction in the lab of John Postlethwait, University of Oregon; additional tissue samples for genomic libraries can be obtained from Dr Allyse Ferrara (Nicholls State University, Louisiana; also the source of tissue for the cDNA libraries). Dr Ferrara has found that spotted gar can be spawned in the laboratory and are easily raised; she is preparing a manuscript on the laboratory culture of spotted gar. Furthermore, an embryological series for gar is available (Long and Ballard 2001). The only caveat is there is currently sparse molecular data for this animal. It is recommended that preliminary EST and low-coverage genome sequencing is undertaken first, to check for complications. In the unlikely event of unforeseen issues, e.g lineage-specific polyploidy, the bowfin *Amia calva* would be the alternative choice. For *Amia*, BAC libraries are available, embryos can be obtained from the wild and the genome size is 1.2 Gb.

Weighing up all factors, we conclude that the spotted gar *Lepisosteus oculatus* is the ideal basal actinopterygian for whole genome sequencing.

This proposal made after consultation with: John Postlethwait (jpostle@uoneuro.uoregon.edu), Chi-hua Chiu (Chiu@Biology.Rutgers.Edu), Wilbur Long (wlong@mcdaniel.edu), Allyse Ferrara (Allyse.Ferrara@nicholls.edu), Angel Amores (amores@uoregon.edu), Victoria Prince (vprince@midway.uchicago.edu), Axel Meyer (axel.meyer@uni-konstanz.de), Chris Amemiya (camemiya@benaroyaresearch.org).

References for the spotted gar can be found in the Appendix

7. The little skate:

We propose draft sequencing (6x) of the *Raja erinacea*, the little skate (3.3Gb) and of a large set of ESTs (100,000).

We first proposed sequencing the little skate genome for the May 2004 round of Council. The Coordinating Committee endorsed the proposal, but Council decided to defer it until sequencing becomes cheaper. Therefore, it was never entered into the sequencing center pipeline. We will not repeat the arguments here, except to reaffirm the importance of the node represented by this species and our support for the use of an experimental species, the little skate. We will report our deliberations of an alternative species, the elephant shark, *Callorhinchus milii* (a chimera, 0.9 Gb genome), presented to us recently in a white paper by E. Kirkness, B. Strausberg, B. Venkatesh, and S. Brenner (included in the appended materials).

Phylogenetic position: From the cartilaginous fish node evolved the sharks, chimeras, skates, and ratfish, as well as all bony fish and species mentioned previously. This is a very important node, as the ancestor represents the first jawed vertebrate,

presumed to be an active and predatory animal, whereas before its emergence, the chordates, it is thought, were sluggish filter feeders. Cartilaginous fish possess the full suite of major vertebrate traits such as an adaptive immune system, sclerotome cells, paired fins, and a full set of paired sense organs. As discussed in our previous report (2004), a number of cartilaginous fish species, mostly sharks and skates, are used for experimental studies, but most have rather large genomes (1.5-3x human). The little skate provides a good compromise by having a relatively small genome for a skate (3.3 Gb), while still having a record of experimental usage.

As described in the appended whitepaper, the elephant shark, or elephant fish, has a significantly smaller genome (0.9 Gb), and this is an attraction for sequencing. As the authors point out, "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." They point out a greater similarity of human and cartilaginous fish coding sequences, compared to teleosts, and this would be expected to hold as well for little skate as for elephant fish (and most likely also for spotted gar, above). The main problem with elephant fish is that it is has almost no record of experimental work, whereas the little skate has.

Thus, even though the genome of the little skate is larger, we recommend it for sequencing because the community use of the sequence information will be greater.

References to the elephant shark can be found in the appended white paper.

8. Hagfish: *Eptatretus burgeri*, the Japanese hagfish

We propose high quality draft sequencing (approx. 6x) of the Japanese hagfish and sequencing of a large set of ESTs (100,000).

Jawless fish (lamprey and hagfish) occupy an important phylogenetic position due to several characteristics: 1) they are the most basally split vertebrates, and the ancestor at this node would be the ancestor of all vertebrates; 2) their body plan differs significantly from jawed vertebrates, including the lack of jaw structures and other skeletal elements; 3) they appear to lack an adaptive immune system, and 4) whereas jawed vertebrates have undergone two genome-wide duplications, the number of duplications in agnathans may be fewer. Thus, obtaining genome sequence from a hagfish will provide important information at this critical node. The time of branching of hagfish and lamprey is estimated at approximately 500Mya (Cambrian lamprey fossils are known). Sequencing of one jawless fish, the sea lamprey (*Petromyzon marinus*), is currently in progress. We propose additionally the sequencing of a hagfish.

Three species of hagfish are most frequently used for research purposes, *Eptatretus burgeri, Eptatretus stouti*, and *Myxine glutinosa*. Their estimated genome sizes are roughly equivalent (2.5-3Gb) and BAC libraries exist for all. Dr. Shigeru Kuratani at the Riken Institute in Kobe, Japan, is making progress in breeding the Japanese species and obtaining embryos. Therefore, the best choice of species for sequencing is the Japanese species, *Eptatretus burgeri*. In light of the chromosome elimination and perhaps chromatin diminution in this species, it is preferable to use sperm as the DNA source.

Regarding the suitability of hagfish for sequencing, it is known that hagfish genes have 50% GC whereas lamprey genes have 80% GC (Kuraku and Kuratani, 2006). If applied to the whole genome, this difference should make assembly easier for hagfish than lamprey, an important consideration. Sequencing of both is necessary as they deeply diverged (500 MYa), but it is likely that hagfish will not suffer from the same sequencing problems.

Below is a description of the available resources for the two leading species of hagfish, on which is based our choice of the Japanese species.

	Eptatretus burgeri	*Eptatretus stoutii*
Availability	easily obtained in Japan; commercial fisheries in Japan and Korea	easily obtained in Oregon through commercial collectors
Size of animal	> 30 cm when fully grown	> 30 cm when fully grown
Genome size	2.5-3 x 10^9 bp	2.5-3 x 10^9 bp
Culturability	relatively easy to keep in captivity for extended periods	relatively easy to keep in captivity for extended periods
BAC library	Yes, 100-120 kb inserts (Suzuki, T. et al., 2004a)	Yes, 90-100 kb inserts (Pancer, Z. et al., 2005)
cDNA libraries	A buffy coat plasmid library exists (Kasahara) (Suzuki, T. et al., 2005;Suzuki, T. et al., 2004b). No ESTs deposited.	A buffy coat library exists (lambda) in Gary Litman's lab (unpubl).
embryos	Shigeru Kuratani is working on this	Never observed, to our knowledge
chromatin and chromosomal reduction	Hagfishes undergo a loss of chromosomes and genomic content during development (Kohno, S. et al., 1986;Kubota, S. et al., 2001;Kubota, S. et al., 1997;Nabeyama, M. et al., 2000;Nakai, Y. et al., 1995;Nakai, Y. et al., 1991). That is, its germline genome is altered during development so that its somatic genome lacks some chromosomes and certain repetitive elements.	
immune system	The genes of the immune system are probably the best studied of the hagfish genes. The reason is because people had been trying for decades to identify immunoglobulin type molecules despite the fact that hagfishes have no spleen, thymus or bone marrow. No bona fide immunoglobulins have ever been found, although a complement molecule thought to be immunoglobulin was identified in multiple hagfish species (Kobayashi, K. et al., 1985;Raison, R. L. et al., 1978a;Raison, R. L. et al., 1978b;Varner, J. et al., 1991). While lacking immunoglobulins, hagfishes do have variable lymphocyte receptors, the same as those in lampreys (these have leucine rich repeat modules) (Pancer, Z. et al., 2005). The VLR loci of the two Eptatretus are very complicated and may encode over 10^{14} different receptors. Additionally, Masanori Kasahara has done a leucocyte EST project and identified numerous genes that could be part of the immune system by virtue of their structures, including a few that have immunoglobulin type domains (Haruta, C. et al., 2006; Nagata, T. et al., 2002; Suzuki, T. et al., 2005;Suzuki, T. et al., 2004b).	

References for the hagfish can be found in the Appendix.

9. Cephalochordates: *Branchiostoma lanceolatum*, Amphioxus

The proposal is for high quality draft coverage (approximately 6x) of the whole genome of *B. lanceolatum* (0.6 Gb genome) together with 50,000 ESTs to aid construction of gene models.

The cephalochordates (amphioxus) are the basal lineage of chordates and the ideal outgroup for genomic comparisons to all vertebrate genomes. They have three particular characters that make them essential for understanding the evolution of the human proteome. First, the cephalochordates have been recently shown to represent the most ancestral branch of the chordate phylum, rather than the tunicates which traditionally were placed in that position (Delsuc et al. 2006). This implies that characters shared by amphioxus and humans, but differing in tunicates, are ancestral characters for the whole of our phylum. Second, the cephalochordates diverged before the two genome duplications at the base of vertebrates, such that ancestral pre-duplication states can be inferred by comparison to amphioxus. Third, they have undergone much less secondary change to the genome than have tunicates, the other pre-duplication chordate group (which includes *Ciona* and *Oikopleura*). For example, amphioxus protein sequences have undergone less lineage-specific change (e.g., approximately two-thirds the amino acid substitutions found in ascidians), fewer genes have been secondarily lost, and ancient linkage arrangements have been retained (e.g. intact Hox cluster, intact ParaHox cluster).

There is an extensive literature supporting the value of amphioxus in comparative genomic studies. This is not recounted here, as their utility in research is very widely accepted. It is important to note, however, that despite the large evolutionary distance between amphioxus and humans (over 545 million years divergence) syntenic relationships are still present, both at local scale and over chromosomal-scale distances. For example, the neighbouring genes around the ParaHox cluster are the same in amphioxus and human, *CHIC* and *PRHOXNB* (Ferrier et al. 2005), many genes mapping around the human MHC complex map to a single chromosome in amphioxus (Castro and Holland 2003) with several examples of local synteny (Abi-Rached et al 2002), while NK, En, Gbx, Hox, Dlx, Msx (and other) homeobox genes map to the precise chromosomes in amphioxus (Castro and Holland 2004) that were predicted from paralogy analyses of the human genome (Pollard and Holland 2000). These findings imply that synteny can be used as an additional guide to deciphering gene homologies between amphioxus and human, an approach that is rarely feasible in tunicates such as *Ciona* due to extensive genome rearrangement.

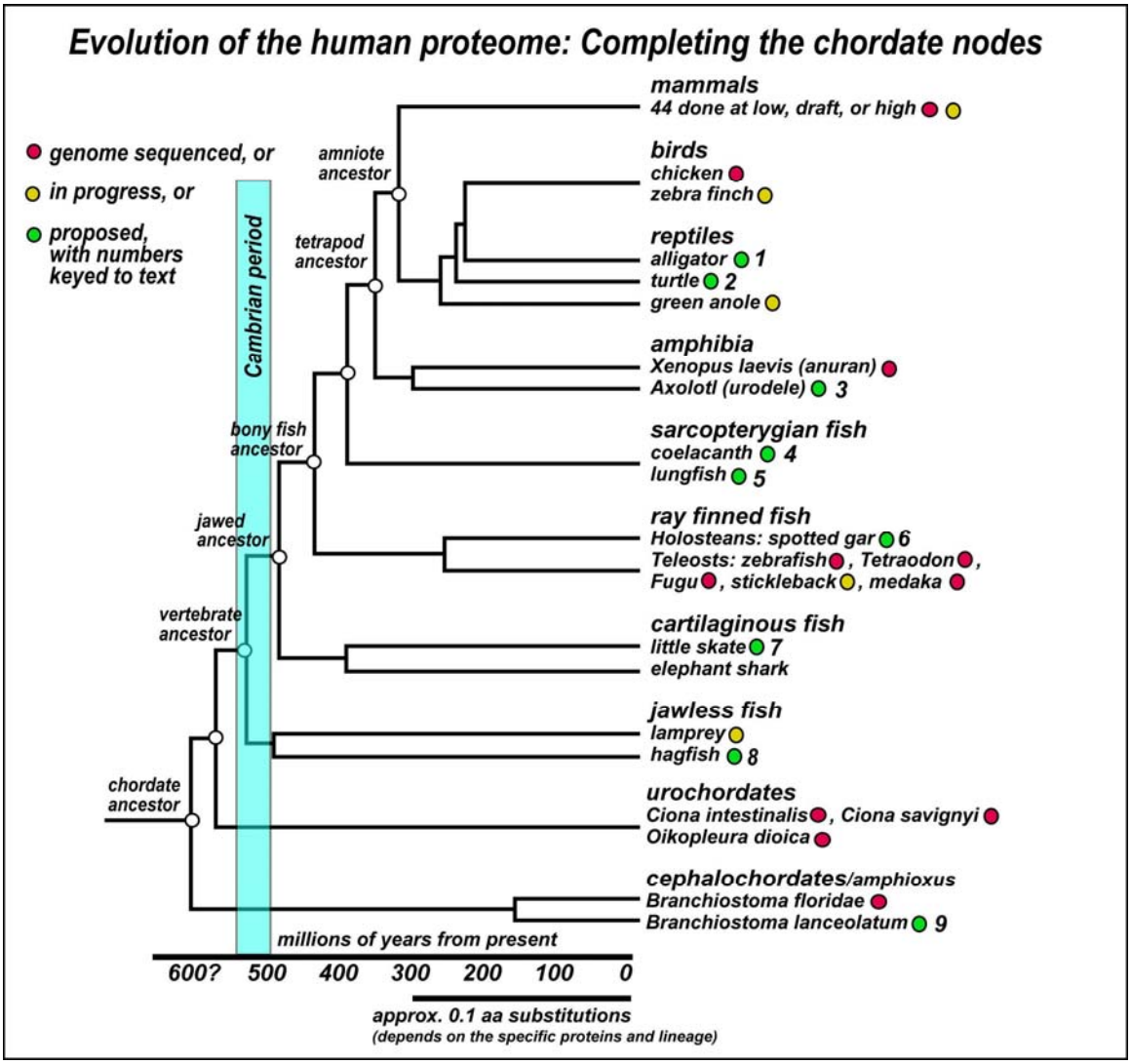
The 500 Mb draft genome sequence of the Florida amphioxus (*Branchiostoma floridae*) has been generated by the Joint Genome Institute, and in May 2006 an initial assembly was released to the community for annotation. An international consortium is currently analysing these data, and uncovering a wealth of findings of significance to the evolution of immune systems, endocrine systems, developmental regulatory genes, signalling molecules and overall genome organisation. As argued earlier, having two genome sequences from one lineage is useful in reconstructing the basal genome condition for that linage, by revealing species-specific gene losses, duplications or protein sequence changes. Furthermore, comparing two species can aid gene model prediction in each genome, and will reveal where chromosomal rearrangements have occurred on each evolutionary lineage. This in turn allows deduction of the ancestral arrangement of genes and will allow investigation of the relation between genome fluidity and protein sequence evolution.

Of the ~24 species of cephalochordate, only three have been used in non-taxonomic biological research: *B. floridae* (genome sequence completed), *B. belcheri* (found in China and Japan) and *B. lanceolatum* (found in Europe). Of these, *B. lanceolatum* is the

only species that can be spawned in a laboratory culture, producing embryos and larvae under controlled conditions (Fuentes et al. 2004). This would be the best candidate to select for a second cephalochordate genome sequence. It has a genome size comparable to *B. floridae* (flow cytometry estimate, 500-600 Mb, Olivier Catrice, Paris). The divergence between the two species dates to the origins of the Atlantic Ocean, over 150 million years ago (Canestro et al 2002), and the animals show several differences. *B. lanceolatum* is much larger than *B. floridae* (~6cm vs. 3 cm), they live in different sediments (gravel vs. sand), at different depths (deep vs. shallow), and they have different reproductive patterns (*B. lanceolatum* is a trickle spawner making it suitable for regular collections of eggs; *B floridae* is a mass spawner making large collections of synchromous stages possible). A small number of developmentally expressed transcription factors and other proteins have been cloned from *B. lanceolatum* (184 protein sequences on GenBank) and these show few differences to date from *B. floridae* (calmodulin is an exception, which is additionally duplicated in *B. lancoelatum*; Karabinos and Bhattacharya, 2000). However, transcription factors and calmodulin are generally highly conserved, and other proteins are expected to be evolving much faster in structure, and therefore differ more between the species. Furthermore, it is expected that gene order will be different at the genomic level, allowing the analyses described above.

In the light of the pivotal position that cephalochordates occupy in the evolution of our own genome, and in view of their relatively conservative genome evolution, there is a strong argument for obtaining the genome sequences of a second cephalochordate species. We therefore propose high quality draft coverage (approximately 6x) of the whole genome of *B. lanceolatum* together with 50,000 ESTs to aid construction of gene models. Genomic DNA, and RNA from embryos and adults, can be provided by Jordi Garcia-Fernandez, Barcelona and Hector Escriva, Banyuls, France.

References to amphioxus can be found in the Appendix.


Evolution of the human proteome: Completing the chordate nodes
genome sequenced, or
in progress, or
proposed, with numbers keyed to text
Cambrian period
amniote ancestor
tetrapod ancestor
bony fish ancestor
jawed ancestor
vertebrate ancestor
chordate ancestor
mammals
44 done at low, draft, or high
birds
chicken
zebra finch
reptiles
alligator 1
turtle 2
green anole
amphibia
Xenopus laevis (anuran)
Axolotl (urodele) 3
sarcopterygian fish
coelacanth 4
lungfish 5
ray finned fish
Holosteans: spotted gar 6
Teleosts: zebrafish, Tetraodon, Fugu, stickleback, medaka
cartilaginous fish
little skate 7
elephant shark
jawless fish
lamprey
hagfish 8
urochordates
Ciona intestinalis, Ciona savignyi
Oikopleura dioica
cephalochordates/amphioxus
Branchiostoma floridae
Branchiostoma lanceolatum 9
millions of years from present
600? 500 400 300 200 100 0
approx. 0.1 aa substitutions
(depends on the specific proteins and lineage)

Appendix

1. References on the alligator

Alibardi, L. 2004. Dermo-epidermal interactions in reptilian scales: Speculations on the evolution of scales, feathers, and hairs J. Exp. Zool. (Mol. Evol. Dev.) 302B:365-383.

Brown, D.R., M.F. Nogueira, T.R. Schoeb, K.A. Vliet, R.A. Bennett, G.W. Pye, and E.R. Jacobson. 2001. Pathology of experimental mycoplasmosis in American alligators. J. Wildl. Diseases 37:671-679.

Cohen, M.M., and C. Gans. 1970. The chromosomes of the order Crocodilia. Cytogenetics 9:81-105.

Crossley, D.A. amd J. Altimiras. 2005. Cardiovascular development in embryos of the American alligator Alligator mississippiensis: effects of chronic and acute hypoxia. J. Exp. Biol. 208:31-39.

Dessauer H.C., TC. Glenn, and LD. Densmore. 2002. Studies on the molecular evolution of the crocodilia: footprints in the sands of time. J. Exp. Zool. (Mol. Dev. Evol.) 294:302-311.

Edwards, S. V., W. B. Jennings, and A. M. Shedlock. 2005. Phylogenetics of modern birds in the era of genomics. Proc. R. Soc. Lond. B 272:979-992.

Guillette, L.J. and T. Iguchi. 2003. Contaminant-induced endocrine and reproductive alterations in reptiles. Pure and Applied Chem. 75:2275-2286.

Glenn, T.C., JL. Staton, A. Vu, LM. Davis, JR. Alvardo Bremer, WH. Rhodes, IL. Brisbin Jr., and RH. Sawyer. 2002. Low mitochondrial DNA variation among American alligators and a novel non-coding region in crocodilians. J. Exp. Zool. (Mol. Dev. Evol.) 394:312-324.

Iwabe N, Hara Y, Kumazawa Y, Shibamoto K, Saito Y, et al. (2005) Sister group relationship of turtles to the bird-crocodilian clade revealed by nuclear DNA-coded proteins. Molecular Biology and Evolution 22: 810-813.

Jacobson, E.R., A.J. Johnson, J.A. Hernandez, S. J. Tucker, A. P. Dupuis, II, R. Stevens, D. Carbonneau, and L. Stark. 2005. West Nile virus infection in farmed American alligators (Alligator mississippiensis) in Florida. J. Wildl. Diseases 41:96-106.

Janke, A., and U. Arnason. 1997. The complete mitochondrial genome of Alligator mississippiensis and the separation between recent archosauria (birds and crocodiles). Mol. Biol. Evol. 14:1266-1272.

Merchant, M.E., M. Pallansch, R.L. Paulman, J.B. Wells, A. Nalca, and R. Ptak. 2005. Antiviral activity of serum from the American alligator (Alligator mississippiensis). Antiviral Research 66:35-38.

Sawyer, R.H., L Rogers, L. Washington, T.C. Glenn and L.W. Knapp. 2005. The evolutionary origin of the feather epidermis. Developmental Dynamics 232:256-267.

Western, P. S., J. L. Harry, J. A. Marshall Graves, and A. H. Sinclair. 2000. Temperature-dependent sex determination in the American alligator: expression of SF1, WT1 and DAX1 during gonadogenesis. Gene 241:223-232.

2.References on the Turtle:

Alibardi L, Toni M. 2006. Immunolocalization and characterization of beta-keratins in growing epidermis of chelonians. Tissue Cell. 38(1):53-63.

Cao, Y, Sorenson, MD, Kumazawa, Y, Mindell, DP, Hasegawa, M. 2000. Phylogenetic position of turtles among amniotes: Evidence from mitochondrial and nuclear genes. Gene (Amsterdam) 259:139-148.

Cebra-Thomas J, Tan F, Sistla S, Estes E, Bender G, Kim C, Riccio P, Gilbert SF. 2005. How the turtle forms its shell: a paracrine hypothesis of carapace formation. J Exp Zoolog B Mol Dev Evol. 304(6):558-69.

Congdon JD, Nagle RD, Kinney OM, van Loben Sels RC, Quinter T, Tinkle DW. 2003. Testing hypotheses of aging in long-lived painted turtles (*Chrysemys picta*). Exp Gerontol. 38(7):765-72.

Engstrom, TN, Shaffer HB, McCord WP. 2004. Multiple data sets, high homoplasy, and the phylogeny of softshell turtles (Testudines: Trionychidae). Systematic Biology 53:693-710.

Ezaz T, Valenzuela N, Grutzner F, Miura I, Georges A, Burke RL, Graves JA. 2006. An XX/XY sex microchromosome system in a freshwater turtle, *Chelodina longicollis* (Testudines: Chelidae) with genetic sex determination. Chromosome Res. 14:139-50.

Francisco-Morcillo J, Hidalgo-Sanchez M, Martin-Partido G. 2006. Spatial and temporal patterns of proliferation and differentiation in the developing turtle eye. Brain Res. 1103:32-48.

Franz-Odendaal TA. 2006. Intramembranous ossification of scleral ossicles in *Chelydra serpentina*. Zoology (Jena). 109:75-81.

Gilbert SF, Loredo GA, Brukman A, Burke AC. 2001. Morphogenesis of the turtle shell: the development of a novel structure in tetrapod evolution. Evol Dev. (2):47-58.

Hemmings SJ, Storey KB. 1999. Brain gamma-glutamyltranspeptidase: characteristics, development and thyroid hormone dependency of the enzyme in isolated microvessels and neuronal/glial cell plasma membranes. Mol Cell Biochem. 202(1-2):119-30.

Hill RV. 2005. Integration of morphological data sets for phylogenetic analysis of amniota: The importance of integumentary characters and increased taxonomic sampling. Systematic Biology 54:530-547.

Iwabe N, Hara Y, Kumazawa Y, Shibamoto K, Saito Y, Miyata T, Katoh K. 2005. Sister group relationship of turtles to the bird-crocodilian clade revealed by nuclear DNA-coded proteins. Molecular Biology and Evolution 22:810-813.

Kalman M, Martin-Partido G, Hidalgo-Sanchez M, Majorossy K. 1997. Distribution of glial fibrillary acidic protein-immunopositive structures in the developing brain of the turtle Mauremys leprosa. Anat Embryol (Berl). 196:47-65.

Krenz JG, Naylor GJP, Shaffer HB, Janzen FJ. 2005. Molecular phylogenetics and evolution of turtles. Molecular Phylogenetics and Evolution 37:178-191.

Kuraku S, Usuda R, Kuratani S. 2005. Comprehensive survey of carapacial ridge-specific genes in turtle implies co-option of some regulatory genes in carapace evolution Evol. Dev. 7:3-17.

Kuraku, S., J. Ishijima, C. Nishida-Umehara, K. Agata, S. Kuratani, and Y. Matsuda. 2006. cDNA-based gene mapping and GC3 profiling in the soft-shelled turtle suggest a chromosomal size-dependent GC bias shared by sauropsids. Chromosome Res 14:187-202.

Loredo GA, Brukman A, Harris MP, Kagle D, Leclair EE, Gutman R, Denney E, Henkelman E, Murray BP, Fallon JF, Tuan RS, Gilbert SF. 2001. Development of an evolutionarily novel structure: fibroblast growth factor expression in the carapacial ridge of turtle embryos. J Exp Zool. 291:274-81.

Lutz PL, McMahon P, Rosenthal M, Sick TJ. 1984. Relationships between aerobic and anaerobic energy production in turtle brain in situ. Am J Physiol. 1984 247(4 Pt 2):R740-4.

Matsuda Y, Nishida-Umehara C, Tarui H, Kuroiwa A, Yamada K, Isobe T, Ando J, Fujiwara A, Hirao Y, Nishimura O, Ishijima J, Hayashi A, Saito T, Murakami T,

Murakami Y, Kuratani S, Agata K. 2005. Highly conserved linkage homology between birds and turtles: bird and turtle chromosomes are precise counterparts of each other. Chromosome Res. 13:601-15.
Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. Annual Reviews of Ecology, Evolution and Systematics 34:311-338.
Murdock C, Wibbels T. 2006. Dmrt1 expression in response to estrogen treatment in a reptile with temperature-dependent sex determination. J Exp Zoolog B Mol Dev Evol. 306:134-9.
Nagashima, H., K. Uchida, K. Yamamoto, S. Kuraku, R. Usuda, and S. Kuratani. 2005. Turtle-chicken chimera: an experimental approach to understanding evolutionary innovation in the turtle. Dev Dyn 232:149-161.
Near TJ, Meylan PA, Shaffer HB. 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. American Naturalist 165:137-146.
Ohya, Y. K., S. Kuraku, and S. Kuratani. 2005. Hox code in embryos of Chinese soft-shelled turtle Pelodiscus sinensis correlates with the evolutionary innovation in the turtle. J Exp Zoolog B Mol Dev Evol 304:107-118.
Packard GC and MJ Packard. 2004. To freeze or not to freeze: adaptations for overwintering by hatchlings of the North American painted turtle. J Exp Biol. 207(Pt 17):2897-906.
Pough FH, Andrews RM, Cadle JE, Crump ML, Savitzky AH, Wells KD. 2001. *Herpetology*. Prentice Hall, Upper Saddle River, New Jersey.
Reese SA, Ultsch GR, Jackson DC. 2004. Lactate accumulation, glycogen depletion, and shell composition of hatchling turtles during simulated aquatic hibernation Journal of Experimental Biology 207: 2889-2895.
Rest JS, Ast JC, Austin CC, Waddell PJ, Tibbetts EA, Hay JM, Mindell DP. 2003. Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. Molecular Phylogenetics and Evolution 29:289-297.
Rieppel O . 2001. Turtles as hopeful monsters. Bioessays. 23:987-9.
Rieppel O, Reisz RR. 1999. The origin and early evolution of turtles. Annual Reviews of Ecology and Systematics 30:1-22.
Sakai H, Saeki K, Ichihashi H, Kamezaki N, Tanabe S, Tatsukawa R. 2000. Growth-related changes in heavy metal accumulation in green turtle (*Chelonia mydas*) from Yaeyama Islands, Okinawa, Japan. Arch Environ Contam Toxicol. 39:378-85.
Sasaki T, Takahashi K, Nikaido M, Miura S, Yasukawa Y, Okada N. 2004. First application of the SINE (Short interspersed repetitive element) method to infer phylogenetic relationships in reptiles: An example from the turtle superfamily testudinoidea. Molecular Biology and Evolution 21:705-715.
Spinks PQ, Shaffer HB, Iverson JB, Mccord WP. 2004. Phylogenetic hypotheses for the turtle family Geoemydidae. Molecular Phylogenetics and Evolution 32:164-182.
Storey KB. 2006. Reptile freeze tolerance: metabolism and gene expression.
Cryobiology. 52(1):1-16.
Torsoni MA, Ogo SH. 2000. Hemoglobin-sulfhydryls from tortoise (*Geochelone carbonaria*) can reduce oxidative damage induced by organic hydroperoxide in erythrocyte membrane. Comp Biochem Physiol B Biochem Mol Biol. 126:571-7.
Tsai PS, Licht P. 1993. Differential distribution of chicken-I and chicken-II GnRH in the turtle brain. Peptides. 14(2):221-6.
Vincent C, Bontoux M, Le Douarin NM, Pieau C, Monsoro-Burq AH. 2003. Msx genes are expressed in the carapacial ridge of turtle shell: a study of the European pond turtle, *Emys orbicularis*. Dev Genes Evol. 213:464-9.

3. References on the Axolotl

Da Silva SM, Gates PB, and Brockes JP. (2002) The newt ortholog of CD59 Is implicated in proximodistal identity during amphibian limb regeneration. Developmental Cell 3:547–555.

Mercader N, Tanaka EM, Torres M. (2005) Proximodistal identity during vertebrate limb regeneration is regulated by Meis homeodomain proteins.Development 132:4131-42.

Putta S, Smith JJ, Walker JA, Rondet M, Weisrock DW, Monaghan J, Samuels AK, Kump K, King DC, Maness NJ, Habermann B, Tanaka E, Bryant SV, Gardiner DM, Parichy DM,7 and Voss SR. (2004) From biomedicine to natural history research: EST resources for ambystomatid salamanders. BMC Genomics 5: 54-66.

San Mauro D, Vences M, Alcobendas M, Zardoya M, and Meyer A. (2005)Initial Diversification of Living Amphibians Predated the Breakup of Pangaea. *The American Naturalist* 165:590–599.

Zhang P, Zhou H, Chen Y-Q, Liu Y-F, Qu L-H. (2005) Mitogenomic Perspectives on the Origin and Phylogeny of Living Amphibians. Systematic Biology 54: 391–400.

4. Reference for the Coelacanth—see the white paper on coelacanth.

5. References on the lungfish

Brinkmann H, Venkatesh B, Brenner S, and Meyer A. (2004) Nuclear protein-coding genes support lungfish and not the coelacanth as the closest living relatives of land vertebrates. PNAS 101: 4900-4905

Daeschler EB, Shubin NH, Jenkins FA Jr. (2006) A Devonian tetrapod-like fish and the evolution of the tetrapod body plan. Nature 440: 757-63. An example of recent paleontological discoveries about the sarcopterygian-tetrapod transition.

Dores RM, Sollars C, Lecaude S, Lee J, Danielson P, Alrubaian J, Lihrman I, Joss JM, Vaudry H. (2004) Cloning of prodynorphin cDNAs from the brain of Australian and African lungfish: implications for the evolution of the prodynorphin gene. Neuroendocrinology. 79:185-96. (Example of the use of existing cDNA libraries)

Joss JM. (2005) Lungfish evolution and development. Gen Comp Endocrinol. Dec 6th.

Shubin NH, Daeschler EB, Jenkins FA Jr. (2006) The pectoral fin of Tiktaalik roseae and the origin of the tetrapod limb. Nature 440:764-71.

Sirijovski N, Woolnough C, Rock J, Joss JM. (2005) NfCR1, the first non-LTR retrotransposon characterized in the Australian lungfish genome, Neoceratodus forsteri, shows similarities to CR1-like elements. J Exp Zoology Mol Dev Evol. 304B:40-9. (On the possible relationship of neoteny and large genome size in lungfish and urodeles).

6. References on the spotted gar

Bingulac-Popovic J, Figueroa F, Sato A, Talbot WS, Johnson SL, Gates M, Postlethwait JH, Klein J. 1997. Mapping of Mhc class I and class II regions to different linkage groups in the zebrafish, Danio rerio. Immunogenetics 46:129-134.

Chiu, C.-H. et al. (2004) Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. Genome Res. 14, 11-17.

Hardie, D.C. and Hebert, P.D.N. (2004) Genome-size evolution in fishes. Canadian Journal of Fisheries and Aquatic Sciences **61**: 1636-1646.

Jaillon, O. et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature 431, 946-957.

Long, W.L. and Ballard, W.W. (2001) Normal embryonic stages of the Longnose Gar, *Lepisosteus osseus*. BMC Developmental Biology 1:6

Meyer, A. and Van de Peer, Y. (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27, 937-945.
Mulley, J.F., Chiu, C.-H. and Holland, P.W.H. (2006) Break-up of a homeobox gene cluster after genome duplication in teleosts. PNAS 103, 10369-10372
Ojima, Y. and Yamamoto, K. (1990) Cellular DNA contents of fishes determined by flow cytometry. La Kromosomo II 57, 1871-1888
Postlethwait JH et al. (1998) Vertebrate genome evolution and the zebrafish gene map. Nat Genetics 18, 345-349.
Sambrook, J.G., Figueroa, F., Beck, S. (2005) A genome-wide survey of Major Histocompatibility Complex (MHC) genes and their paralogues in zebrafish. BMC Genomics 6:152

8. References on the hagfish:

Escriva, H., Manzon, L., Youson, J., and Laudet, V. (2002). Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Mol Biol Evol 19: 1440-50.
Haruta, C., Suzuki, T., and Kasahara, M. (2006). Variable domains in hagfish: NICIR is a polymorphic multigene family expressed preferentially in leukocytes and is related to lamprey TCR-like. Immunogenetics 58: 216-225.
Kobayashi, K., Tomonaga, S., and Hagiwara, K. (1985). Isolation and characterization of immunoglobulin of hagfish, Eptatretus burgeri, a primitive vertebrate. Mol.Immunol. 22: 1091-1097.
Kohno, S., Nakai, Y., Satoh, S., Yoshida, M., and Kobayashi, H. (1986). Chromosome elimination in the Japanese hagfish, Eptatretus burgeri (Agnatha, Cyclostomata). Cytogenet.Cell Genet. 41: 209-214.
Kubota, S., Ishibashi, T., and Kohno, S. (1997). A germline restricted, highly repetitive DNA sequence in Paramyxine atami: an interspecifically conserved, but somatically eliminated, element. Mol.Gen.Genet. 256: 252-256.
Kubota, S., Takano, J., Tsuneishi, R., Kobayakawa, S., Fujikawa, N., Nabeyama, M., and Kohno, S. (2001). Highly repetitive DNA families restricted to germ cells in a Japanese hagfish (Eptatretus burgeri): a hierarchical and mosaic structure in eliminated chromosomes. Genetica 111: 319-328.
Kuraku S and Kuratani S. (2006) Timescale for cyclostome evolution inferred with a phylogenetic diagnosis of of hagfish and lamprey cDNA sequences. Zoological Science.
Nabeyama, M., Kubota, S., and Kohno, S. (2000). Concerted evolution of a highly repetitive DNA family in eptatretidae (Cyclostomata, agnatha) implies specifically differential homogenization and amplification events in their germ cells. J.Mol.Evol. 50: 154-169.
Nagata, T., Suzuki, T., Ohta, Y., Flajnik, M. F., and Kasahara, M. (2002). The leukocyte common antigen (CD45) of the Pacific hagfish, Eptatretus stoutii: implications for the primordial function of CD45. Immunogenetics 54: 286-291.
Nakai, Y., Kubota, S., Goto, Y., Ishibashi, T., Davison, W., and Kohno, S. (1995). Chromosome elimination in three Baltic, south Pacific and north-east Pacific hagfish species. Chromosome.Res. 3: 321-330.
Nakai, Y., Kubota, S., and Kohno, S. (1991). Chromatin diminution and chromosome elimination in four Japanese hagfish species. Cytogenet.Cell Genet. 56: 196-198.
Pancer, Z., Saha, N. R., Kasamatsu, J., Suzuki, T., Amemiya, C. T., Kasahara, M., and Cooper, M. D. (6-28-2005). Variable lymphocyte receptors in hagfish. Proc.Natl.Acad.Sci.U.S.A 102: 9224-9229.

Raison, R. L., Hull, C. J., and Hildemann, W. H. (1978a). Characterization of immunoglobulin from the Pacific hagfish, a primitive vertebrate. Proc.Natl.Acad.Sci.USA 75: 5679-5682.
Raison, R. L., Hull, C. J., and Hildemann, W. H. (1978b). Production and specificity of antibodies to streptococci in the pacific hagfish, Eptatretus stoutii. Dev.Comp.Immunol. 2: 253-262.
Stadler, P. F., Fried, C., Prohaska, S. J., Bailey, W. J., Misof, B. Y., Ruddle, F. H., and Wagner, G. P. (2004). Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of Eptatretus stoutii. Mol.Phylogenet.Evol. 32: 686-694.
Suzuki, T., Ota, T., Fujiyama, A., and Kasahara, M. (2004a). Construction of a bacterial artificial chromosome library from the inshore hagfish, Eptatretus burgeri: A resource for the analysis of the agnathan genome. Genes Genet.Syst. 79: 251-253.
Suzuki, T., Shin, I., Fujiyama, A., Kohara, Y., and Kasahara, M. (3-1-2005). Hagfish leukocytes express a paired receptor family with a variable domain resembling those of antigen receptors. J. Immunol. 174: 2885-2891.
Suzuki, T., Shin, I., Kohara, Y., and Kasahara, M. (2004b). Transcriptome analysis of hagfish leukocytes: a framework for understanding the immune system of jawless fishes. Dev.Comp Immunol. 28: 993-1003.
Varner, J., Neame, P., and Litman, G. W. (1991). A serum heterodimer from hagfish (Eptatretus stoutii) exhibits structural similarity and partial sequence identity with immunoglobulin. Proc.Natl.Acad.Sci.USA 88: 1746-1750.

9. References on amphioxus

Abi-Rached L, Gilles A, Shiina T, Pontarotti P, Inoko H (2002) Evidence of en bloc duplication in vertebrate genomes. Nat Genet 31, 100–105
Canestro, C. et al. (2002) Ascidian and amphioxus Adh genes correlate functional and molecular features of the ADH family expansion during vertebrate evolution. J. Mol. Evol. 54, 81-89.
Castro, L.F.C. and Holland, P.W.H. (2003) Chromosomal mapping of ANTP class homeobox genes in amphioxus: piecing together ancestral genomes. Evolution and Development 5, 459-465.
Castro, L.F.C., Furlong, R.F. and Holland, P.W.H. (2004) An antecedent of the MHC-linked genomic region in amphioxus. Immunogenetics 55, 782-784
Delsuc, F., Brinkmann, H., Chourrout, D. and Philippe, H. (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439, 965-968.
Fuentes, M. et al. (2004) Preliminary observations on the spawning conditions of the European amphioxus (Branchiostoma lanceolatum) in captivity. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 302 (4), pp. 384-391
Karabinos, A., Bhattacharya, D., 2000. Molecular evolution of calmodulin and calmodulin-like genes in the cephalochordate Branchiostoma. J. Mol. Evol. 51, 141-148