PROPOSAL TO SEQUENCE THE AMPHIOXUS GENOME

Jeremy J. Gibson-Brown[*] and Linda Z. Holland[†]
In consultation with:
Pieter de Jong[‡], John McPherson[¶] and Robert Waterston[¶]

[*]Dept of Biology, Washington University, St. Louis, MO
[†]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA
[‡]BACPAC Resources, Children's Hospital Oakland Research Institute, Oakland, CA
[¶]Genome Sequencing Center, Washington University, St. Louis, MO

I. Introduction

The cephalochordate amphioxus (*Branchiostoma*) is the closest living invertebrate relative of the vertebrates. Morphologically and developmentally it closely resembles the last common ancestor of all vertebrates, possessing a notochord, a hollow dorsal neural tube, segmented muscle blocks, and a perforated pharyngeal (branchial) region (Fig.1). The organization of several large genomic regions (e.g., Hox, MHC) has been determined, and closely parallels that of the homologous regions in vertebrates. Although a number of cephalochordate lineage-specific gene duplications have been identified, amphioxus lacks the large-scale gene duplications characteristic of vertebrates (e.g. only one Hox cluster). Thus the amphioxus genome, of all living animals, most closely approximates that of the last common ancestor of all vertebrates. The other invertebrate chordates, the urochordates (tunicates), are less closely related to vertebrates, and have a highly derived genomic composition and structure making direct comparisons to vertebrates problematic. As the sister group to vertebrates within the phylum Chordata, amphioxus is therefore the most appropriate outgroup for understanding the origins of vertebrate genome content and structure, and how duplicated genes in vertebrates have evolved new functions.

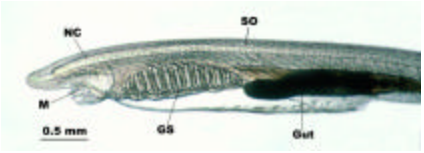

Fig. 1. Anterior half of juvenile amphioxus (3.5 weeks old). Anterior to the left, showing the dorsal hollow nerve cord (NC), pharyngeal gill slits (GS), gut (G), mouth (M), and muscular somites [one delineated by white dashes (SO)]. Sexual maturity is at 6 weeks. Adults add additional gill slits as they grow to a maximum length of 5 cm.

The amphioxus genome has a haploid content of about 500 megabases, around 17% that of the mouse or human. The sequencing strategy we will use will be to assemble 6-fold whole-genome shotgun coverage of the genome from a single animal, and to order and orient the resulting sequence scaffolds by alignment to end-sequenced BACs in a comprehensive contig map. A 15-fold coverage, large-insert (175-200kb) sequence-quality amphioxus BAC library is currently under construction (de Jong). The generation of finished sequence will be readily achievable using the combined BAC map and sequence scaffolds. The genome sequence generated will provide an invaluable resource not only for the amphioxus community but also for other evolutionary, developmental, cellular and molecular biologists as well.

II. Biological Rationales for the Utility of the Amphioxus Genome Sequence

a. Connecting the sequences of non-human organisms and the human sequence

Currently the only basal metazoans for which complete genome sequences are available are *Drosophila melanogaster* and *Caenorhabditis elegans*. Although these species have proven to be outstanding genetic model organisms, the relevance of the experimental data to understanding human genetics has often been complicated by the facts that (1) as protostomes, the divergence of their genomes from that of humans is so ancient that it is often difficult to determine homologous gene relationships, and (2) both of these organisms have evolved highly derived developmental and biochemical mechanisms, even compared to other taxa within the ecdysozoan (molting animal) clade of protostomes. For example, the determinate development of *C. elegans* and the long-germ-band embryonic patterning mechanism of *Drosophila* are both derived features specific to these organisms. Additionally, genetic approaches alone have failed to identify thousands of genes as witnessed by the many computer-predicted genes of no known function revealed by the genome projects. However, amphioxus, at the boundary between invertebrates and vertebrates, is *the* key link between vertebrate and invertebrate genomes required to understand the evolution of these uncharacterized genes and their functions. Moreover, many vertebrate model organisms have highly derived genomes that do not reflect the ancestral condition from which the human genome evolved. For example, *Xenopus laevis* is a rediploidized tetraploid, and teleost fish (e.g. zebrafish, medaka and *Fugu*) all derive from a common ancestor that apparently underwent independent whole-genome duplication after separation of the bony fishes from the lineage leading to humans. Importantly, amphioxus appears to have the same number of genes as other invertebrates such as Drosophila, but in phylogenetic analyses, amphioxus genes typically branch at the base of their multiple vertebrate homologs. Thus, given its phylogenetic position at the root of the vertebrate lineage, and the "primitive", unduplicated condition of its genome, the amphioxus genome provides the ideal comparative data set for understanding the evolutionary origins of human genes and their functions, and for annotating the human genome sequence.

b. Informing the human genome sequence

One of the most exciting prospects stemming from the widespread availability of genomic sequence from different organisms is the ability to detect functionally important regulatory elements in non-coding DNA using bioinformatic approaches. Traditionally, the detection of these regions has depended on their *in vivo* characterization using reporter constructs, a very time-consuming and expensive process. Recent studies, focusing primarily on the *b-globin* and *Hox* gene loci, have shown that comparing the genomic sequence of different organisms can identify non-coding regions of known biological significance, a technique known as "phylogenetic footprinting". Comparisons of mouse-to-human, and even shark-to-human DNA (G. Wagner, Yale) have revealed numerous short regions, typically 100 to 200bp in length, conserved up to 90% at the nucleotide level between these organisms. These conserved "modules" are presumably under strong selective pressure to maintain their sequence since they contain numerous transcription factor binding sites required for regulating target gene expression. Comparison of vertebrate genomic loci with those of classical invertebrate genetic models such as *Drosophila* and *C. elegans* have failed to detect similarly conserved regulatory motifs. Indeed sequence comparisons between closely related species of *Drosophila* have failed to detect such motifs, suggesting that protostome genomes can rapidly diverge in sequence while conserving important regulatory functions. *In vivo* reporter constructs using amphioxus genomic DNA in transgenic mice have already revealed the presence of conserved regulatory modules within chordates, at least within the *Hox* clusters, indicating that the amphioxus genome is likely to be an invaluable resource for characterizing and annotating the human genome sequence. Another exciting new area of research involves the detection and characterization of non-coding RNAs (ncRNAs). The importance of ncRNAs in such diverse processes as transcription, gene

silencing, replication, RNA processing/modification/stability, translation and protein stability has only recently been fully appreciated (reviewed by Storz, Science 2002, 296:1260-12630). Given the absence of a diagnostic open reading frame, the identification of these genes can be very difficult. As with non-coding regulatory elements, comparative genomics probably offers the best opportunity for identifying such elements. In particular, the amphioxus genome probably offers the best resource for characterizing the origin and evolution of ncRNA function in vertebrates, and their identification in the human genome sequence.

c. Expanding our understanding of evolutionary processes in general, and human evolution in particular

A longstanding question concerning the origin of vertebrate genomes has been whether one or more whole-genome duplications (tetraploidizations) early in vertebrate evolution led to the dramatic increase in gene numbers apparent in modern vertebrates. Whereas invertebrate genomes typically contain around 15,000 genes, the mouse and human genome projects indicate around 30-40,000 genes in mammals. To attempt to resolve this issue, one of us (Horton et al., submitted) recently compared all amphioxus genes for which sequence is available to the complete set of relevant mammalian genes (Figure 2).

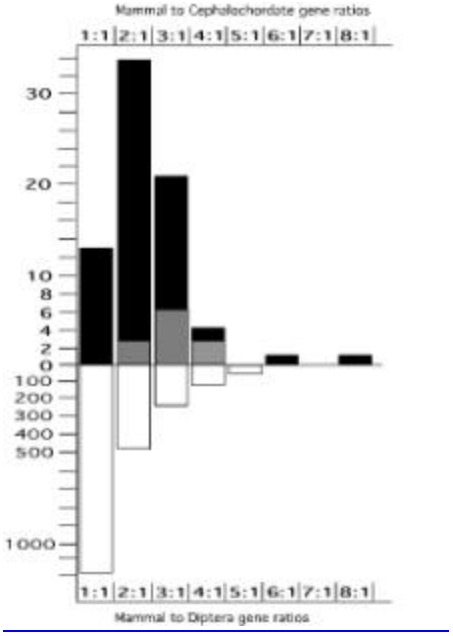

FIGURE 2. -- Summary of gene ratios. Vertical bars represent number of occurrences for each gene ratio. Black, mammalian-to-cephalochordate gene ratios for 75 amphioxus genes. Grey, Hox genes form a subset of the genes included in this study. White, human-to-Drosophila gene ratios. Horton et al., submitted.

We find that whereas mammalian-to-fly gene ratios peak at 1:1, mammalian-to-amphioxus gene ratios peak at 2:1. Although a peak at 2:1 is suggestive of at least one whole-genome duplication followed by some local duplications, the results are equally consistent with two whole-genome duplications followed by a greater level of gene loss. We conclude that "phylogenetic" analyses alone will *never* determine whether whole-genome duplications were responsible for the increase in vertebrate gene numbers. Instead we propose that a "phylogenomic" approach, in which paralogous clusters of genes (i.e., syntenic regions) are compared between the complete amphioxus and human genomes will also be required to resolve this issue, as illustrated in Figure 3.

The dramatic difference in gene ratio distributions between mammalian-to-fly and mammalian-to-cephalochordate comparisons (Figure 2) clearly illustrates the problem of using distantly related protostome species such as *Drosophila* when attempting to understand the evolution of human and other vertebrate genomes, and reemphasizes the importance of amphioxus as the most appropriate outgroup for such studies.

d. Facilitating the ability to do experiments in additional organisms

An enormous benefit of the complete amphioxus genome sequence is that comparative sequence analyses will allow the identification of genes, regulatory elements, other non-coding features, and their *locations* in the genomes of *all* vertebrate species. This informatic data will allow other researchers using classical

vertebrate model systems to positionally map, and target specific regions of the genome for functional studies (e.g., gene knockout, enhancer knockout, targeted enhancer-trap, misexpression studies, etc.).

e. Expanding our understanding of basic biological processes relevant to human health

Amphioxus is an ideal system for gaining insights into the genetic bases of birth defects and genetic diseases. It is the most vertebrate-like of all invertebrates with a dorsal hollow nerve cord, pharyngeal gill slits, notochord, and a segmented body musculature. Fertilization is external, facilitating experimental manipulations. The eggs can be easily microinjected (Figs. 4, 5). Antisense morpholino oligonucleotides appear to block translation effectively, and the expression of injected reporter constructs is less mosaic than is typical for transient transgenics in other species. Moreover, the embryos and larvae are transparent and can be raised in large numbers through metamorphosis.

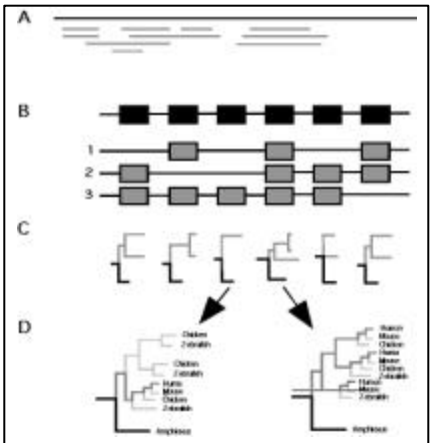

FIGURE 3. – Schematic diagram representing the phylogenomic approach that will be required if the issue of whole-genome duplications within chordates is to be resolved. **A,** An amphioxus-human synteny map reconstructed by aligning **B,** clusters of paralogy groups. **C,** Phylogenetic trees of paralog groups. Black terminal branches represent amphioxus genes; dark gray branches represent human genes. Trees appear to support different duplication histories for different paralog groups. **D,** Phylogenetic trees containing orthologs from additional vertebrate groups not only serve to date duplications, but also allow detection of lineage-specific gene death that may otherwise result in misleading evolutionary inferences. Horton et al., submitted.

Amphioxus embryos have the advantage of being structurally simple compared to those of vertebrates. Nevertheless, the genetic controls of embryogenesis are largely like those in vertebrates. In both amphioxus and vertebrates, signaling by Wnt/ß-catenin, Notch and retinoic acid patterns the anteroposterior axis, while BMP2/4 is involved in dorsoventral patterning, and Sox1/2/3, neurogenin, iroquois and islet are involved in specifying the neural plate and neurogenesis. The effects of teratogens such as retinoic acid (RA) on amphioxus development and gene expression are like those on vertebrate embryos (Escriva et al., 2000). Thus while lack of duplicate gene copies in amphioxus makes understanding the molecular networks mediating development easier to understand than in vertebrates, the results from experiments with amphioxus embryos can typically be extrapolated to vertebrates.

Because of its simplified development compared to vertebrates, amphioxus is also an excellent model system for studying such birth defects as failure of the neural tube to close (*spina bifida*), early pharyngeal jaw development, muscle development, and other early developmental defects, or those caused by maternal intake of retinoic acid (the acne medicine Accutane), cadmium, mercury, dioxins or alcohol. Other invertebrate chordates (i.e., tunicates) have highly-derived ontogenies and have secondarily lost many adult features still present in amphioxus. Amphioxus also promises to be valuable for understanding the function(s) of genes involved in carcinogenesis such as DRAL (down-regulated in rhabdomyosarcoma) (Schubert et al., 1998). Its role in vertebrate embryogenesis has not been studied, but amphioxus DRAL is co-expressed with BMP2/4 during the gastrula stage, pointing to a critical role in embryogenesis for this gene. Thus amphioxus is an excellent simplified model for understanding how signaling pathways pattern the vertebrate embryo, for understanding structure/function relations in proteins, and for understanding the evolutionary question of how gene duplication has led to the diversification of vertebrate gene functions.

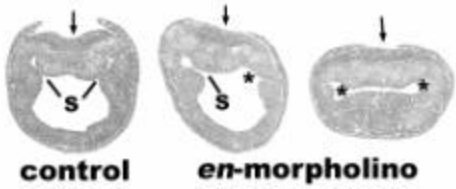

Figure 4. Injection of an *engrailed* antisense-morpholino oligonucleotide prevents segmentation of the paraxial mesoderm in amphioxus embryos. *E* Neurula-stage embryos are seen in cross-section. Arrow shows neural plate. *engrailed* is normally expressed in the posterior half of each nascent somite (S), suggesting a role in segmentation. In injected embryos, somites fail to form on either one or both sides of the embryo (asterisks), supporting a role for engrailed in segmentation of the paraxial mesoderm.

f. Amphioxus as a model for understanding human biology

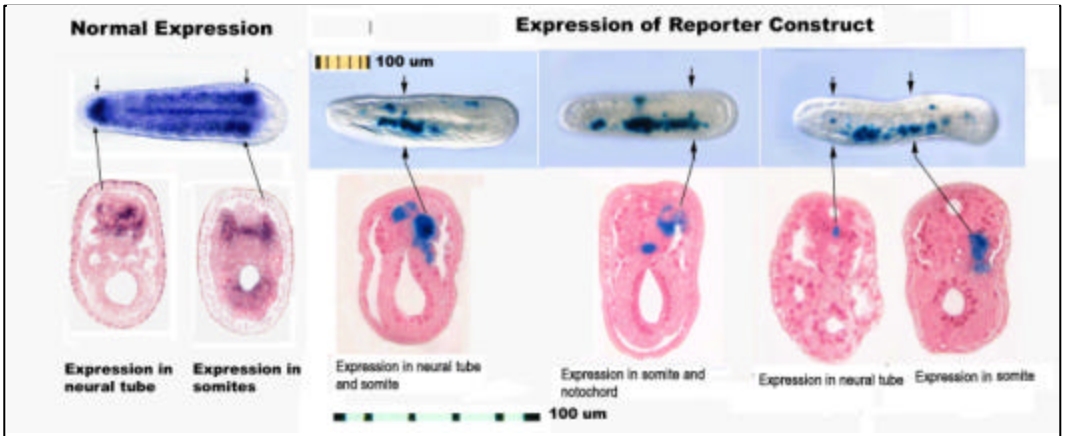

Figure 5. Injection of a FoxD3 lacZ reporter construct into amphioxus eggs recapitulates endogenous expression in nerve cord, somites and notochord. Top, whole mounts. dorsal views of 20 hr neurulae, anterior at left. Bottom, cross sections of whole mounts at the levels indicated by arrows. As is typical in transient transgenics, expression is mosaic and is preferentially on either the right or left side of the embryo, probably because the injected DNA had not completely diffused throughout the egg before first cleavage.

Amphioxus is a prototypical vertebrate, with many homologs of vertebrate organs. These amphioxus structures include the pituitary, pineal organ, striated axial muscles, kidneys, liver, thyroid gland, nerve cord, and pancreatic islet cells. Though much less complex than their vertebrate counterparts, these amphioxus organs share many developmental, cellular and physiological parameters with their vertebrate counterparts. For example, the pituitary homolog of amphioxus labels with antibodies to gonadotropins (lutenizing hormone, chorionic gonadotropin) as well as to thyrotropin-releasing hormone, cholecystokinin and metenkephalin, and expresses homologs of genes like *Pax6,* which are expressed during development of the vertebrate pituitary. Similarly, thyroid-like hormones and an iodoprotein similar to mammalian thyroglobulin have been isolated from the endostyle, which also expresses several of the same genes as the vertebrate thyroid during development (e.g. *NKX2.2*, *NKX2.1*, *Pax2/5/8*). Although amphioxus has no pancreas, the gut has scattered cells that express the gene for insulin-like peptide (ILP), which has features of both insulin and the insulin-like growth factors (IGFs), and others that are immunoreactive for serotonin. Moreover, the amphioxus receptor for ILP has the basic structural determinants necessary for binding and activation by mammalian insulin and IGF1. The nephridia structurally resemble the pronephric (primitive) kidneys in developing vertebrate embryos and express similar genes during development (e.g. *Pax2/5/8*). The striated somitic muscles of amphioxus express a cascade of myogenic factors, beginning with *Pax3/7*, characteristic of developing vertebrate muscle. In addition, the gonads have saturatable receptor activity for mammalian lutenizing hormone (LH) and human chorionic gonadotropin (hCG), while injection of follicle-stimulating hormone (FSH) near the gonads affects oocyte maturation and spawning. Consequently, many vertebrate biologists and biochemists are very interested in amphioxus and several are performing comparative work with amphioxus to gain insights into the vertebrate systems they study. Comments from vertebrate biologists underscore the value of amphioxus for understanding human biology: **Dr. Gerhard Schlosser,** "...amphioxus is able to provide a wealth of phylogenetic information necessary to understand the evolutionary origin of vertebrate novelties. One important example, on which my own research interests are focused, is the neural crest and placodes. Development of these novel tissues makes use of phylogenetically old genes and gene networks, which have just been redeployed in novel contexts during vertebrate evolution, presumably by changes in *cis*-regulatory regions. The proposed amphioxus sequencing project will be an absolutely essential contribution to overcome this present impasse in making not only the coding but also noncoding sequences of the amphioxus genome readily available to the scientific community. Needless to say this will have important medical implications as well." **Dr. Bill**

McGinnis, "I strongly support your effort to sequence the amphioxus genome. Cephalochordates occupy a unique and important position on the evolutionary tree. It is likely that the amphioxus sequence will provide more information about the basic genes needed for programming the development of chordates like us than any other fish or simple chordate genome. I think the sequence would be an enormous boon to a variety of researchers in a wide range of biological disciplines." See also letters in **Appendix I** from C. LaBonne, RG Northcutt, M. Bronner-Fraser, M. Reedy, L. Pezzementi, and R. MacDonald.

g. Amphioxus as a model for understanding human health and disease

All studies show that the major gene/genome duplications within the vertebrate lineage occurred after the divergence of amphioxus, and it is therefore an excellent model for understanding vertebrate signaling pathways. Amphioxus is particularly relevant for understanding genetic diseases and the genetic bases of cancer due to mutations in developmental signaling pathways as well as the mechanisms by which teratogens affect embryogenesis. Bilaterian organisms are largely patterned by the same seven major signaling pathways: Notch, Wnt, TGF-β, Hedgehog (Hh), nuclear receptor (e.g. retinoic acid receptor, estrogen receptor), Jak/STAT, and receptor tyrosine kinase (RTK). *Drosophila* has been a very useful model for elucidating most of these pathways, with some exceptions, such as the retinoic acid signaling pathway, which appears to be chordate-specific. However, there are large differences in body plans between *Drosophila* and vertebrates, suggesting that while the core of these pathways may be largely conserved between them, the terminal events clearly are not. In amphioxus the functions of genes such as *Hh*, *Notch*, *Wnt*s and *RAR* during embryogenesis are directly comparable to those in vertebrates, and experimental data demonstrate that patterning by at least the retinoic acid and Wnt pathways is essentially the same in amphioxus and vertebrates (Holland and Holland, 1996; Escriva et al. 2002). Thus amphioxus is an excellent simplified model for elucidating the downstream targets of signaling pathways and how these pathways pattern the vertebrate embryo. The amphioxus genome sequence will provide the basis for elucidating these pathways in detail. For example, it will make possible the construction of microarrays for determining the downstream targets of signaling pathways, and with many fewer genes in amphioxus than vertebrates, these pathways will be easier to understand. Such an understanding is fundamental to an understanding of the genetic bases of cancer and the mechanisms by which teratogens affect embryogenesis.

III. Strategic Issues in Acquiring New Sequence Data

a. The rationale for the complete genomic sequence

It has become increasingly clear that the operation of a genome is *as* reliant on non-coding sequence components as on protein-coding regions. Information on the former is only available from genomic sequence data. Although it has been suggested that the *Drosophila* genome sequence may be the Rosetta stone for understanding the vertebrate genome (Kornberg and Krasnow, 2000, Science 287:2218-2220), the real Rosetta stone in this regard is the amphioxus genome, since amphioxus is the closest living invertebrate relative of the vertebrates. All of the work on genome organization in amphioxus shows that it has representatives of all of the vertebrate genes except a few that are vertebrate-specific. However it has undergone neither the large-scale gene duplications that characterize the vertebrate genome nor appreciable gene-loss, such as has occurred in the other group of invertebrate chordates, the tunicates. Thus amphioxus has a single *Hox* cluster with 14 *Hox* genes (*Hox14* probably represents an independent tandem duplication in amphioxus) compared to four clusters in humans with a total of 39 *Hox* genes (a loss of 13 genes after cluster duplication). This indicates that a single *Hox* cluster with 13 genes was present in the last common ancestor of amphioxus and vertebrates, and duplicated two or three times in

the lineage leading to mammals. In contrast, tunicates have a degenerated *Hox* cluster, which appears to correlate with a secondary simplification of the body plan, a reduction in the number of cells in the embryo, and a switch from indeterminate to determinate cleavage. Importantly, the fact that the amphioxus genome contains around half the number of genes as that of mammals, yet is only around one sixth the size of that in mouse or humans, indicates that the sequence generated will be at least *three times as gene-rich*, and will probably lack the extensive repeat elements, such as LINEs, SINEs and Alu-repeats, typical of vertebrate genomes. The complete amphioxus genome sequence will finally free investigators from cloning genes, and instead allow them to concentrate their research efforts on structural and functional analyses.

b. The community

The first paper on the developmental genetics of amphioxus was published in 1992 (Holland et al. Development 116:653-61). This paper, one of only ten on amphioxus published that year, was the impetus for a resurgence of interest in amphioxus. Today, there are over thirty laboratories concentrating their efforts on amphioxus developmental and cell biology, and the number of papers per year on amphioxus developmental genetics and biochemistry has increased exponentially. Research on genes and development has focused on the Florida amphioxus, *Branchiostoma floridae*, in the US and Europe, and on *B. belcheri* in Japan and China. Interest within, and *beyond*, the amphioxus community for the sequence of the genome is very high. In **Appendix I** we have excerpted some of the 39 letters of support we have received. In the US and Europe there are at least a dozen laboratories with a major effort devoted to amphioxus as well as many with on-going research projects on amphioxus as part of their overall strategy. (e.g. those of C. Kimmel, Univ. Oregon; M. Bronner-Fraser, Caltech; R. MacDonald, Univ. Texas Southwestern Medical Center; D. Steiner and Shu Jin Chan, Univ. Chicago; A. Spicer, Texas A&M; L. Pezzementi, Birmingham Southern; J. Langeland, Kalamazoo College). An amphioxus genome sequence will greatly facilitate their research and encourage them to devote more of their effort to amphioxus and induce additional laboratories to begin research on amphioxus. The high level of interest in amphioxus genes is reflected in the resources available to the community. J. Gibson-Brown and A. Horton have created AmphiBase (http://www.tbx.wustl.edu), a curated repository for amphioxus/vertebrate phylogenetic trees and sequence alignments. Gridded cosmid and cDNA libraries from adult and several embryonic stages of *B. floridae* have been made by H. Lehrach (Germany) in collaboration with PWH. Holland (UK) and LZ. Holland. Genomic and cDNA libraries in lambda phage have been made by J. Garcia-Fernàndez (Spain), LZ. Holland and J. Langeland (USA). A manuscript describing an EST analysis identifying 14,000 non-redundant clones from gastrula and late neurula cDNA libraries has been submitted for publication (G. Panopoulou and H. Lehrach, Germany). In addition, an EST analysis of notochord-specific genes of *B. belcheri* is underway (N. Satoh, Japan). A small-scale genome-sequencing project of *B. floridae* has been started by P. Pontarotti (France) with T. Shiina and H. Inoko (Japan), who have sequenced 12 cosmids, or approximately 0.1% of the amphioxus genome. This work has shown that the gene order of the MHC region of the amphioxus and human genomes is identical. Recently, NSF approved funding for an amphioxus BAC library to be constructed by P. de Jong in collaboration with LZ. Holland. This library will be freely available to all. Finally, the laboratory of PWH. Holland (Reading, UK) has developed a reliable method for testing for physical linkage between amphioxus cosmids (or larger contigs) using FISH to chromosomes. This could be an invaluable tool in aiding assembly of different regions of the amphioxus genome sequence, or ordering contigs where gaps persist.

The availability of these resources has prompted great interest in sequencing the amphioxus genome in Europe as well as the United States. The Sanger Center and Genoscope (the French genome sequencing

center) in a recent meeting with representatives of several European laboratories expressed strong interest in participating in a joint American/European project to sequence the genome, ***and offered to provide up to half the funds necessary if additional funding could be obtained***. However, a proposal by nine European, Japanese and American laboratories led by Pierre Pontarotti (France) for funds to sequence the *B. floridae* genome which was submitted to the 6th EU Framework Programme has not been approved. At best, it is expected that only about 50% of the estimated $15M cost could be generated from European sources. Given the already high level of interest in the amphioxus genome, there is no question that the amphioxus genome sequence would stimulate an additional increase in the number of labs focusing on amphioxus, and the number of other labs using amphioxus data in their own studies.

c. Beyond the community

The sequence of the amphioxus genome will be highly useful to a very broad spectrum of developmental and evolutionary biologists, geneticists, cancer researchers and biochemists.

The health sciences community. Because amphioxus is vertebrate-like, but simpler, it is an essential model organism for understanding vertebrate development and disease. **Prof. Ray MacDonald**, Southwestern Medical School writes, "For my own research, a high-quality sequence of the entire *B. floridae* genome would aid immensely my interest in understanding the developmental program for the formation of the pancreas. The complete sequence of the amphioxus genome would help us to identify developmental regulatory genes orthologous to those known to be required for mammalian pancreatic development, and in particular, islet formation." **Dr. Shu Jin Chan**, University of Chicago, Dept. of Biochemistry and Molecular Biology "…the amphioxus genome sequence can provide important insights into the evolutionary origin and ontogeny of a medically important tissue, namely, the endocrine pancreas, which produces insulin."

Evolutionary biologists. Because the amphioxus genome is the best available proxy for the ancestral vertebrate genome it is of key interest to evolutionary biologists. **Prof. Ulrich Welsch**, Dept. Anatomy, Univ. of Munich Medical School, writes, "Amphioxus is a survivor of the first steps of vertebrate evolution. Knowledge of its genome is of utmost importance to understand the key characters of the vertebrates and also specific characters of "higher" vertebrates, including *Homo sapiens*." **Dr. Ricard Albalat**, Dept. de Genètica, Univ. Barcelona. "We have studied repeat elements in the amphioxus genome and shown that amphioxus would be a suitable model to monitor repeat dynamics and analyse repeat instability from an evolutionary perspective. The full genome sequence will be an invaluable tool to the scientific community interested in this type of analysis." **Prof. Thurston Lacalli**, Dept. Biology, Univ. Saskatchewan, Canada. "I think it is now clear that amphioxus is *the* model system for understanding the nature of the chordate genome prior to the duplication events that characterize the vertebrate lineage. As such, our understanding of how various families of developmentally important genes diversified in vertebrates depends on a thorough knowledge of their amphioxus counterparts. From an evolutionary standpoint, amphioxus is the one *crucial* organism, among the lower forms, that we need to know more about." **Dr. Michael T. Ghiselin**, MacArthur Fellow, Dept. Invert. Zool. Geol., Calif. Acad. Sciences, San Francisco. "Not only does *Branchiostoma* occupy a key position within the metazoan tree, classical work on its anatomy and physiology provide a rich source of background material for interpreting the results of new approaches. If there is *any* organism that will tell us something important about our own ancestry, it is *Branchiostoma.*"

Developmental and cancer biology. Amphioxus is ideal for understanding gene networks and how they pattern the vertebrate embryo and become dysregulated in cancer. Thus results from amphioxus are often directly applicable to vertebrates. Collaborations between vertebrate and amphioxus biologists are common. One such is between the laboratory of Prof. V. Laudet (Lyon, France), who studies the

biochemistry, evolution and function of nuclear receptors and that of Dr. LZ. Holland, who studies mechanisms of amphioxus development. This collaboration, involving *in vitro* assays of the function of amphioxus nuclear receptors in mammalian tissue culture systems, and the use of amphioxus as a simple mode to elucidate signaling pathways, has recently identified around 20 previously unknown target genes of retinoic acid signaling involved in anterior-posterior patterning of the nerve cord and pharynx. As observed by **Dr. Hector Escriva,** CNRS, Lyon, one of the collaborators, "Complex gene networks, implicated in metabolism, development or endocrine systems, that have to be looked at as a whole, and are extremely complex in vertebrates, will be much more easily understood once the amphioxus genome is decoded." **Dr. Carole LaBonne,** Northwestern University, comments, "Certainly our own work on the neural crest, a vertebrate-specific cell type with great evolutionary and clinical significance, would be greatly aided by having this information available. Importantly, I would anticipate that knowing this sequence would provide significant insight into the evolution of the human genome and the genetic bases of disease."

d. The suitability of amphioxus for experimentation

Amphioxus (*B. floridae*) is a marine invertebrate chordate, commonly found in shallow water along the southeastern coasts of the United States. It is the dominant benthic organism in Old Tampa Bay, Florida, occurring at population densities of up to $1200/m^2$. The life span has been estimated to be 3 years. Although laboratory breeding colonies of amphioxus have not yet been established, it holds great promise as a laboratory model system. Sexes are separate, and each individual breeds at about 12-day intervals throughout the summer. In nature, oocytes undergo the meiotic divisions and arrest at metaphase II during the early afternoon. Spawning is induced by a drop in light level at sundown and can be induced in animals brought into the laboratory by a mild electric shock (Holland and Holland, 1993). Gametes are broadcast into the seawater, with a single female producing 1000-5000 eggs at each spawning. Development is direct, and embryos can be easily raised in the laboratory to adults in about 4 weeks on a diet of mixed algae. Embryos are amenable to experimental manipulation and transient transgenics have been created by microinjection. Sexual maturity occurs at about 6 weeks of age. Gonad growth in the laboratory has been obtained, and adults maintained at 31^0 C (field temperature) on a long-day/short-night cycle have spawned on days on which they would not normally spawn in nature (EE. Ruppert, Clemson Univ.; D. Meulemans, Caltech). It is evident that populations could easily be maintained in the laboratory on this regime and would provide ripe gametes throughout the year. Many amphioxus workers have kept adults in the laboratory for periods of up to 2 years on a diet of mixed algae or commercially available fish food. However no one has yet made a determined effort to establish a laboratory breeding culture, chiefly because none of the amphioxus labs have had the resources required set up a culture facility. Such a facility would require automation of cleaning and feeding. Although the concentration of particulate food is not critical, it must be added to the water at frequent intervals. If food is added in large quantity at infrequent intervals, the adults clear it from the water in a short time and pass it rapidly through the gut, packaging it into fecal pellets without digesting it. Moreover, at present there is no reliable method for controlling the days on which animals will spawn. An effort to clone GNRH from amphioxus is underway (SA. Sower, Univ. New Hampshire). Once the sequence is known, GNRH, which is a relatively small peptide, can be chemically synthesized, and when injected into ripe adults should induce meiotic maturation on demand.

e. The cost of sequencing the genome

Given that the amphioxus haploid genome contains around 500 megabase pairs, spread across 36 chromosomes, the cost of sequencing 6-fold coverage by whole-genome shotgun techniques is estimated

to be around $15M. A 15-fold coverage BAC library, average insert size 180 kb, is currently under construction supported by funds from NSF (Holland and de Jong). For shotgun sequencing we plan to create a whole-genome, short-insert library of about 6-fold coverage from the same animal from which the BAC library is created, as well as a fosmid library with 10-fold coverage. Mapping the BAC library, and the generation of around 75,000 BAC end-sequences for the assembly phase, will add around $500,000 to the total cost. Fosmid end-sequences will also provide linkers for better continuity of the sequence, as well as a substrate for sequencing, and will cost an additional $500,000 for 10-fold coverage. An initial analysis by the Washington University Genome Sequencing Center of a 5-fold mouse whole-genome shotgun sequence indicates that about 98% coverage of the genome has been achieved in reads, and about 90% is covered in assembled scaffolds (McPherson and Waterston, http://genome.wustl.edu). Given the quality and coverage achieved for the mouse sequence, the strategy and level of whole-genome shotgun coverage proposed should be quite sufficient to provide >98% coverage of the amphioxus genome.

One technical issue that has been carefully taken into consideration in developing our library construction strategy is the high level of genetic polymorphism detected within amphioxus genomes, by most estimates around 1.7% at the nucleotide level across the entire genome. To minimize the assembly problem posed by this polymorphism typical of marine invertebrates, we will generate all our sequence-quality libraries from the DNA extracted from a single animal. Preliminary studies (de Jong) have shown it is possible to extract sufficient DNA for a 15-fold coverage BAC library from a single medium-sized (4cm) animal, but not enough to also generate shotgun small-insert and fosmid libraries from the same individual. We therefore intend to extract DNA from a single large (6.5cm) gravid male that should yield at least 4 times as much DNA as those tested in our preliminary studies, and be ample for the simultaneous production of all the libraries while also including genomic information from the heterogametic sex. The amphioxus spawning season begins in the middle of June and continues through early September, so the collection of suitable animals will be achieved by the end of June 2002.

f. Other partial sources of funding

As noted above (Section III b) although there is interest in Europe in sharing the cost of sequencing the amphioxus genome, and the Sanger Center and Genoscope are willing to pay half the cost, there are as yet insufficient funds available from European sources for such a project. Similarly, although a group of Chinese researchers led by Prof. Anlong Xu has also expressed interest in obtaining the amphioxus genome sequence, funds are entirely lacking at present.

g. Informatics and sequence release

Washington University will provide public access to the BAC contig map via its web site (http://genome.wustl.edu) as it does for the human, mouse and *Arabidopsis* genomes. All sequence traces will be deposited in the GenBank Trace Repository. The whole-genome shotgun sequence will be assembled using recently developed, public-domain software packages (ARACHNE, MIT; PHUSION, Sanger Center; JAZZ, Joint Genome Institute). Interactive displays of the assembled sequence and integrated BAC map will be available from the Washington University web site. Views of these data will also be integrated into the Ensemble human and mouse web displays using the DAS protocol. As with all sequence generated by the Washington University Genome Sequencing Center, all data are freely available on a daily basis for the unrestricted use of researchers worldwide.