Sequencing the genome of the hemichordate *Saccoglossus kowalevskii*

John Gerhart[1], Christopher Lowe[1,2], Nicole Stange-Thomann[3], Eric S. Lander[3], and Marc Kirschner[2]

[1]Department of Molecular and Cell Biology, LSA 142, University of California, Berkeley, CA 94720-3200
Tel: 510-642-6382;gerhart@socrates.berkeley.edu; clowe@uclink4.berkeley.edu

[2]Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, MA 02115-5730
Tel: 617-432-2250; marc@hms.harvard.edu

[3]Whitehead Institute/MIT Center for Genome Research. 320 Charles Street, Cambridge, MA 02139-2023
Tel: 617-252-1906; sthomann@genome.wi.mit.edu; lander@genome.wi.mit.edu

Summary

Hemichordates are the phylum of bilateral animals most closely related to chordates, the phylum that includes vertebrates such as human. The hemichordate is a three-part worm-like animal, but possesses many chordate-like traits long thought to hold insights into chordate evolution. These traits include gill slits (1), a post-anal tail-like extension (2), a notochord-like stomocord (3), a neurulated portion of the nervous system, a podocyte-containing kidney (3), possibly a pituitary-like proboscis pore (4), and possibly an endostyle-like pharyngeal region with homology to the thyroid (5). In addition, recent *in situ* hybridization studies have revealed extensive hemichordate–chordate similarities in the expression domains of 22 neural patterning genes (6).

We propose the generation of a 6X assembly of the hemichordate *Saccoglossus kowalevskii* (the acorn worm) genome to advance hemichordate–chordate comparisons and to illuminate chordate origins. The *S. kowalevskii* genome sequence will also accelerate work on the development and organization of members of the hemichordate phylum.

In Section A of this White Paper, we discuss the specific biological advances anticipated from the availability of a hemichordate genome by highlighting its potential contribution to informing human biology, informing the human sequence, providing a better connection between the sequences of non-human organisms and the human sequence, expanding our understanding of basic biological processes such as development and neurobiology, and expanding our understanding of evolutionary processes (biological innovation) in general and human evolution in particular. We address the phylogenetic position of hemichordates and their numerous chordate-like characteristics.

In Section B we consider the strategic issues associated with producing a *Saccoglossus* genome sequence. We comment on the value of such sequence information, its anticipated widespread use, and the rapid expansion of the research community once the genome sequence is available. In addition, we highlight the suitability of *Saccoglossus* for experimentation. We then discuss our proposed sequencing strategy and summarize current genomic resources that will significantly complement the sequencing effort.

Section A: Specific Biological Rationales for the Utility of New Sequence Data

The compelling arguments for sequencing the genome of the hemichordate *Saccoglossus kowalevskii*, in regard to human biology and evolution, are the phylogenetic position of its phylum close to chordates (7, 8) and its numerous chordate-like characteristics despite its classification as a non-chordate. A full genome analysis of a hemichordate promises to shed light on the evolutionary origins of the chordate body plan and its development, the origins of chordate organs, tissues, and cell types, and the origins of chordate gene families and gene regulation. Vertebrates, including humans, evolved from a non-vertebrate chordate ancestor, perhaps as early as the Cambrian, and earlier evolved from a non-chordate deuterostome ancestor. Within the chordates, the evolutionary step to vertebrates is large (e.g., many gene duplications, neural crest, sclerotome, etc.) and will be illuminated by the recently obtained ascidian genomes. The step to chordates, however, is much larger and very poorly understood (9). The origins of the chordate body plan and all of its key morphological innovations can only be determined by examining groups outside the chordates. The only groups relevant to addressing this transition are the echinoderms and hemichordates. To appreciate the magnitude of this step, we must summarize the traits shared by all chordates and compare these with the traits of hemichordates. The first four characteristics listed below are frequently cited key chordate traits; the remaining characteristics are less-noted traits:

1. a dorsal hollow nerve cord with a rich topography of gene expression domains in the anteroposterior and dorsoventral dimension;
2. gill slits;
3. a rod-like notochord;
4. postanal tail;
5. iterated somites of the tail and trunk (ascidians lack these);
6. various left/right asymmetries, such as the direction of the heart and gut coiling;
7. a dorsoventral organization inverted with respect to most non-chordate phyla;
8. an endostyle/thyroid complex in the pharynx;
9. a podocyte-based kidney;
10. a pituitary/Hatcheck's pit neurosecretory organ; and
11. a signaling and morphogenetic cell population of the embryo involved in the embryonic development of most of the above traits, called Spemann's organizer.

The traits outlined above are classically used to define chordates, however, it is likely that some of these traits are very ancient and were present in the deuterostome ancestor and even perhaps in the ancestor of the bilaterians. Certainly one of the biggest surprises in the last 20 years was the discovery of conserved genes with conserved roles in the axial patterning of both chordates and various invertebrate groups. Such cases have been the *hox* complex, involved in patterning the trunk of both groups, and the *otx* and *emx* genes, involved in patterning the head of both groups. While there have been extremely useful genetic comparisons between arthropods and chordates, morphological comparisons are inappropriate since *Drosophila* shares few if any of the defining morphological traits of the chordates. These questions about chordate origins can be illuminated by further comparison of chordates with extant members of other phyla — ones closer to chordates than arthropods and nematodes. According to current sequence-based phylogenies (Fig. 1), only two phyla are closely related to chordates: the hemichordates and the echinoderms (7, 8). Together, the three phyla constitute the supertaxon of deuterostomes, which is deeply separated from protostomes, the remaining 20–25 bilateral phyla that includes ecdysozoa and lophotrochozoa (7). All of these phyla evolved from the bilaterian ancestor.

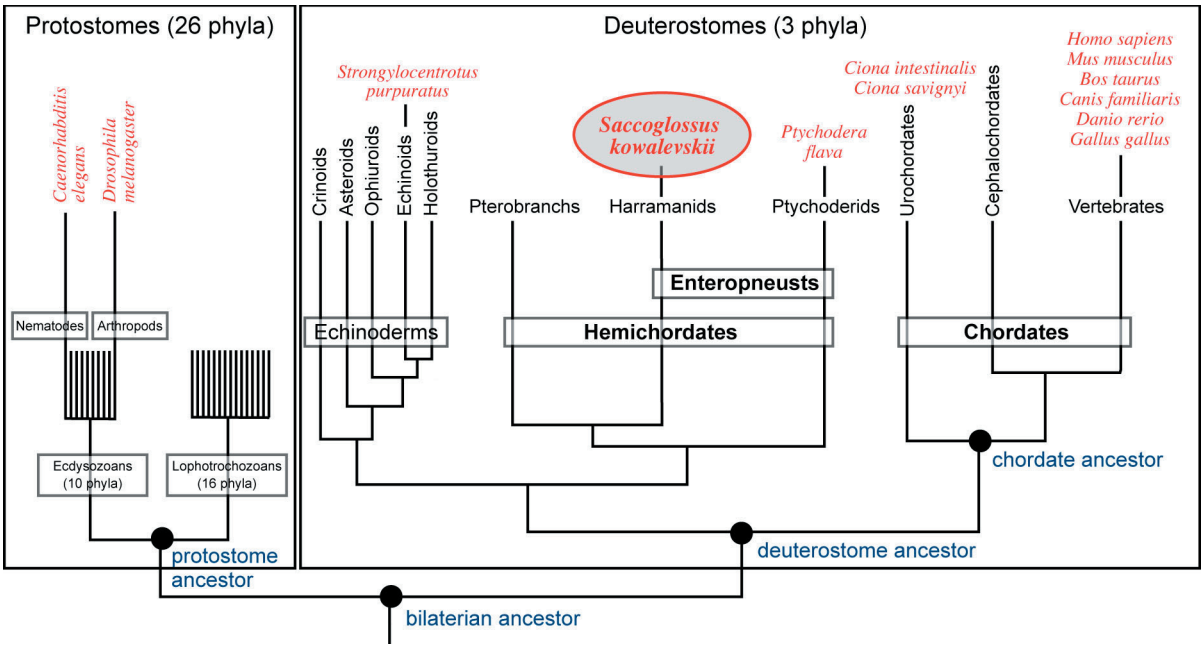

Figure 1. Placement of hemichordates and chordates in the deuterostomes.

Hemichordates — the key group for understanding early chordate evolution

Hemichordates are the appropriate phylum for comparison to chordates because they are a phylum of bilateral adult animals, just as chordates are bilateral, and they have a variety of chordate-like morphological traits. Echinoderms are equally close — they are the sister phylum of hemichordates (Fig. 1) — but their adult body organization has undergone extensive evolutionary modification to a radial body plan with derived organ systems, defying comparison to the bilateral chordates. Thus, despite extensive research on sea urchin development, there have been few insights to enrich our understanding of early deuterostome evolution and the evolution of chordates. Hemichordates, despite more apparent chordate similarities, have been largely overlooked in recent decades after morphological comparative studies reached an impasse. Modern molecular, cellular, and developmental studies have focused on other organisms, and the old comparative questions about hemichordates and chordates have been put aside. The time has come to apply new, incisive approaches. To this end, our group and others have initiated molecular studies of hemichordate organization and development (6), and we now propose the genome sequencing of *Saccoglossus kowalevskii*, a direct-developing enteropneust hemichordate, to enrich and facilitate these comparisons. When the hemichordate genome is complete, the information will complement echinoderm and chordate genome information to reveal the molecular traits of the deuterostome ancestor from which chordates, as well hemichordates and echinoderms, evolved.

Chordate–hemichordate similarities

William Bateson (the biologist who coined the word "genetics") studied *S. kowalevskii* in 1884–1886 and concluded that hemichordates are so like chordates that they should be included in the chordate phylum (10). This classification was accepted for 50–60 years, but by 1940 (11) hemichordates were placed in their own phylum. Their morphological traits, while suggestive, were not definitive enough, and the light microscopy methods of the time were insufficient to resolve the ambiguities. Our group and others started a molecular analysis of the hemichordate body plan, developmental processes, and components, with the expectation that new methods, among which genome sequencing is center-placed, can now yield answers. Some of our new results are listed below in conjunction with older morphological comparisons. Potential chordate similarities to the hemichordate include:

1. Up to 70 pairs of gill slits, with gill bars, similar to chordate gill slits in morphology. *Pax1/9* is expressed in these gill slits as they are in chordate branchial arches (1, 6).

2. A short stomacord homologized by Bateson to the notochord, though it may have more similarities to the chordate prechordal plate. The stomacord contains notochord-like vacuolated cells (3) and develops from the dorsal archenteron roof, as do both the prechordal plate and notochord of chordates. It doesn't express *bra*, *admp,* or *hh* genes like the notochord, but it does express *otx, dkk,* and *hex* genes like the prechordal plate (C. Lowe et al., unpublished).
3. A dorsal axon tract homologized by Bateson, and others more recently, to the chordate dorsal hollow nerve cord. It is formed in part by neurulation (discovered by T.H. Morgan during his graduate research), similar to the chordate cord. However, the tract is not a neurogenic, information-processing cord. It is a thick bundle of axons (12), and the animal has a second tract ventrally (not formed by neurulation). The hemichordate nervous system is in fact an epidermal nerve net. Cell bodies are evenly distributed throughout the epidermis, and axons extend to the tracts (13). Nonetheless the nerve net has most of the neural patterning gene expression domains of chordates (6, and see below).
4. A post-anal extension in which *hox11/13* is expressed, perhaps related evolutionarily to the post-anal chordate tail, which expresses posterior *Hox* genes (6).
5. Possibly an endostyle (5), like the endostyle/thyroid of the chordate pharynx. *Ttf1* is expressed in both the hemichordate pharynx and the vertebrate thyroid (5, 14).
6. An anterior heart–kidney complex (3). The kidney contains podocytes.
7. A proboscis pore homologized to Hatcheck's pit and the pituitary (4). *Pitx* is expressed in the pore region and in the chordate pituitary.
8. The anterior coelom empties from a duct on the left side only, a left–right asymmetry.

Some of these chordate-like traits will hold up to molecular scrutiny, in our opinion, and new ones will be discovered. Not just morphological traits are likely to be elucidated, but also molecular components, developmental processes, and regulatory circuits. Shared traits can then be attributed to the ancestor of hemichordates and chordates, and the modifications and innovations in the chordate line (and the hemichordate line) can be better adduced.

An enduring hypothesis of chordate origins holds that a distant ancestor — perhaps of all bilateral animals — had a ventrally placed central nervous system, which has been retained by extant protostomes. On the deuterostome side, however, an ancestor underwent an inversion of the body in the dorsoventral dimension such that the old ventral cord became the new dorsal chord of chordates (15). This inversion hypothesis is far from demonstrated. In fact, an ancestor of chordates may have simply centralized a diffuse nervous system in the dorsal direction without inversion. Hemichordates, which in some respects look inverted and in other respects uninverted, occupy a critical place in resolving such hypotheses (16).

We initiated such a comparison 3 years ago with *S. kowalevskii*, literally bringing the organism out of the mud. At the time, it was the hemichordate species on which the most, though still very little, laboratory work had been done. We have now devised procedures to obtain eggs and embryos easily (17) and have procured three good cDNA libraries. From these, 66,512 EST sequences have been completed, providing us with many orthologs of chordate genes whose products are known to have central roles in chordate development. Over 80 of these orthologs have been examined by *in situ* hybridization to determine expression domains and can now be compared with the domain locations in chordates. We know that the worm-like hemichordate body, which has three consecutive parts (a prosome [or proboscis], a mesosome [or collar], and a metasome [or pharynx and gut]) corresponds closely to the anteroposterior dimension of the chordate body in terms of the map of expression domains of 22 neural patterning genes (Fig. 2):

1. those genes expressed in the chordate nervous system within the forebrain are expressed in thehemichordate prosome (*six3, rx, bf1, vax, dlx, nkx2.1, otp, pitx*);
2. those genes expressed in the chordate nervous system as far back as the midbrain are expressed in the hemichordate mesosome and anterior metasome (*emx, dbx, otx, pax6, iro, lim1/5, barH, tll, en*), stopping before the second gill slit (in chordates, expression stops before the second branchial arch);
3. those genes expressed in the chordate hindbrain and spinal cord are expressed in the hemichordate posterior metasome (*gbx, hox1, hox3, hox4, hox8*);

4. those genes expressed in the chordate tail are expressed in the hemichordate post-anal extension (*hox11/13*).

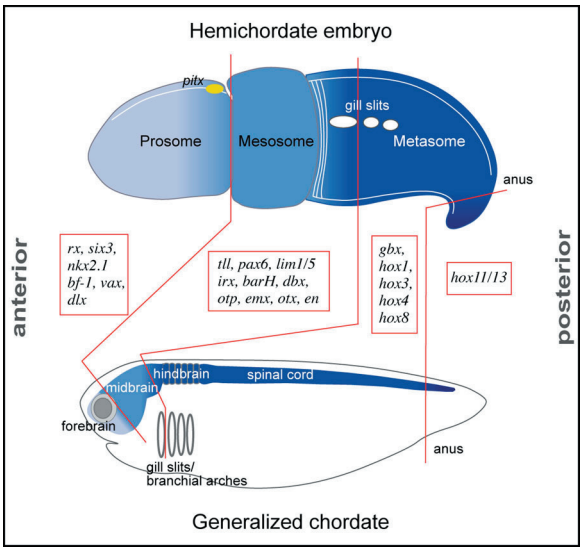

Figure 2. Comparison of expression domains for neural patterning genes in chordates and hemichordates (6).

Of these 22 genes at least 13 are expressed in the same arrangement in *Drosophila*, indicating ancestry back to the bilaterian ancestor, but at least 5 (and no more than 9) are unique to deuterostome.

Although the anteroposterior dimension of chordates and hemichordates is similar, the dorsoventral dimension shows systematic differences, though in a background of similarities. We suspect they differ because the nervous system was never centralized in the hemichordate line (and it resembles an ancestral nervous system) and that centralization occurred within the chordate line. But it is logically possible that the hemichordate line decentralized the nervous system of a centralized ancestor while the chordates retained centralization. The dorsoventral similarities and differences include:

1. Those neural differentiation genes that are expressed only in the chordate dorsal neural plate and nervous system are expressed in bands around the entire ectoderm of hemichordates. The ectoderm is pervasively neurogenic (*elav/huC, musashi/nrp*, *sox2/3, neurogenin*). Hemichordates do not possess a ventral non-neurogenic (epidermogenic) ectoderm, while chordates do.
2. Those genes expressed in the roofplate of the chordate neural tube and in the chordate neural crest are expressed in the hemichordate dorsal midline ectoderm (*bmp2/4, bmp7, tbx2/3, dlx*). Furthermore, some genes expressed in or near the floorplate of the chordate neural tube are expressed in the ventral midline ectoderm (*netrin*, *mnx*, *admp*) of hemichordates. For these genes, the pattern of the entire hemichordate tube-like body resembles the chordate neural tube pattern.
3. Some genes expressed in the ventral neural tube of chordates are expressed in hemichordates in completely different locations. Namely *hh* (*shh*) is expressed in the chordate floorplate, and the diffusing protein signal activates ventral neural expression of *nkx2.2* and *nkx6.1*. None of these is expressed ventrally in *S. kowalevskii*. The gene *hh* seems to have no dorsoventral neural patterning role in hemichordates.

Even though the nervous system of *S. kowalevskii* is not centralized as it is in chordates, it nonetheless has a well-patterned anteroposterior and dorsoventral topography of gene expression domains (except the *hh*-dependent domains), presumably reflecting patterns of neuronal cell types and axon connections.

Expanding our understanding of development, neurobiology, and evolutionary processes by sequencing

The *S. kowalevskii* genome sequence, when available, will greatly facilitate the answering of questions about the origins of the chordate body plan, organs, tissues, cell types, and their development. It will facilitate the extension of comparisons to the molecular level: to gene families, the interactions of components, and the analysis of key developmental processes and cell biological functions. Furthermore, in conjunction with the echinoderm genome, it will facilitate answering questions about the origins of all the deuterostome phyla, allowing detailed deductions about the ancestor of all deuterostomes, and deductions about what was modified and what originated in each line of evolution to chordates, hemichordates, and echinoderms. The long-term medical impact will come from the deepened understanding of chordates (hence humans) afforded by knowing the ancestry of their components and processes. With the hemichordate sequence, the deuterostome supertaxon, with its immense diversity of forms and lifestyles, will be sampled once over. In more detail, the hemichordate genome will make possible:

1. Expanded comparisons of genes involved in the chordate-like traits of hemichordates for extensive comparison with chordates; e.g., gill slits, stomacord, post-anal tail-like extension, nervous system, various organs such as the pharyngeal endostyle/thyroid, the proboscis pore region being possibly like the pituitary, the podocyte kidney, and the epidermis (which is skin-like, not chitinous or test-like, as implied by the presence of loricrin).
2. A deepened analysis of the development of hemichordates in comparison with chordates; e.g., left–right asymmetry, gill slits, kidney and heart. The development of the nervous system is a key line of inquiry. Since the entire ectoderm is neurogenic, does the embryo lack the vertebrate-type default neuralization pathway based on *bmps* and antagonists such as *chordin*? How does the intricate topography of neural gene expression domains develop in hemichordates — is there a signaling center like the Spemann's organizer of chordates? The chordate organizer forms in the dorsal midline of the gut (as precursors of the prechordal plate and notochord). The dorsal midline of the *S. kowalevskii* gut expresses several genes that are expressed in the organizer. Is this midline in hemichordates also a signaling center, the evolutionary antecedent of the chordate organizer? Also, lateral mesoderm development is worth attention — this mesoderm in hemichordates seems much less diversified in its signaling functions and differentiations than in chordates. What is present and absent?
3. Genome organization questions such as synteny vis-à-vis chordates. Mammalian-like gene contiguity has been detected even in ascidians — does it predate chordates? Gene families can be compared with those of chordates — in hemichordates these are expected to be less duplicated. Signaling pathway members deserve close comparison, such as the *TGF*β family. Are subgroups such as nodal members present, which have key roles in vertebrate development? Are the *TGF*β and *wnt* antagonists as numerous, such as *noggin*, *chordin*, *lefty*, *cerberus*, *crescent*, *dickkopf*, *xnr3*, and *kielen*?
4. All the comparisons contribute to evolutionary inquiries about what is conserved, what is modified, and what is new in deuterostome evolution. Deductions can be made about a) the ancestor of hemichordates and echinoderms, b) the ancestor of chordates, which is also the ancestor of all deuterostomes, and c) in comparison with the *Drosophila* and nematode genomes, the ancestor of bilaterians.
5. Various innovations: Do hemichordates have intermediate filament proteins and desmosomes of the chordate type? Or are these chordate innovations? How many chordate endocrine and neurosecretory components are present in hemichordates, and therefore already present in the chordate ancestor?
6. Evolution of *cis*-regulation of genes — to fit in with such efforts on echinoderms and chordates. It is relevant to note that the hemichordate domains for neural patterning genes are much simpler than those in chordates — usually one large contiguous domain for each gene, without multiple secondary domains. This may indicate that the hemichordate *cis*-regulatory sequences may be informative.

Section B: Strategic Issues in Acquiring New Sequence Data

The demand for the new sequence data

We are providing the NHGRI White Paper Committee nine letters of enthusiasm from researchers who know the promise of the hemichordate studies based on the phylogenetic position of hemichordates vis-à-vis chordates. They uniformly assert the value of such sequence information to their own work and the widespread use they expect it to be put by the research community. Although the current *Saccoglossus* community is small, we expect it will grow rapidly when the hemichordate sequence information becomes available and when it is realized how suitable the organism is for experimental work on the body plan and development. *Saccoglossus* research information and resources are currently far ahead of other hemichordate examples, e.g., regarding the numerous neural patterning genes discussed above. A separate small group of researchers is studying an indirect-developing hemichordate species, *Ptychodera flava*, (no ESTs at present, no BAC library, only a few *in situ* patterns), and a few other researchers worldwide are studying the phylogeny, physiology, and ecology of hemichordates generally, without focus on a specific species.

A much larger community will make use of the sequence information for chordate comparisons and general comparative genomics, looking for ancestral sequences, processes, and regulatory circuits as ways to clarify chordate expression patterns and functions. The new sequence data will positively stimulate expansion of the research community, especially at the hemichordate–chordate interface**.**

The suitability of the organism for experimentation

S. kowalevskii is currently the hemichordate species most suited for laboratory experimentation. The methods were developed by William Bateson (1884–1886), Arthur and Laura Colwin (1951–1961) (18), and our group (in the past 3 years). The eggs, embryos, juveniles, and adults are now suitable for most methods of modern analysis, as outlined below. However, *S. kowalevskii* is not suitable for mutant screens and genetic crosses. The advantages are the following:

1. *S. kowalevskii* is readily found in the intertidal zone along the Atlantic coast from Maine to South Carolina. It is a detritus-feeder, up to 8 inches long, that lives in U-shaped burrows. Cilia of the proboscis and pharynx generate water currents to move small food particles into its mouth. In the Woods Hole area (Marine Biological Laboratory, Cape Cod, MA.), the animal is gravid in April–May and in late August–early October (minimally 2 months of the year). At Woods Hole, we can collect over 50,000 eggs and embryos in a 3-week period. A large female yields up to 1000 eggs after temperature-induced ovulation. A single male contains $>10^8$ sperm and can be used for repeated fertilizations. The spawning period can be extended by at least 2 months by cooling animals (thus, 4–5 months of the year total). Juveniles can be grown in the laboratory for at least 1 month; possibly longer.
2. The species develops directly from the fertilized egg to hatched juvenile in 7 days at 21°C, at which time the adult body organization and functions are fully present. There is no detectable larval period, a condition which facilitates work with the adult traits. All embryonic stages are easily observed and collected. The period from fertilization to sexual maturity is probably ~4 months (May–September), although this has not been determined directly. The animals may live for several years in the wild.
3. Eggs and embryos can be readily injected with fluorescent lineage tracers, siRNAs, morpholinos, and test RNAs. The egg is large, 0.4 mm in diameter, and robust. The fertilization envelope can be removed with DTT, thereby eliminating all surface resistance to the injection needle. Under these conditions, at least 10 eggs can be injected per minute. In test series with lineage tracers, embryos develop normally.
4. Embryos and juveniles can be fixed in formaldehyde and preserved in cold ethanol for over a year for *in situ* hybridization staining analysis. The method resembles that used for *Xenopus*. Double *in situ* hybridization has been done successfully. The embryo is unpigmented. It can be cleared in benzyl alcohol-benzyl benzoate, so the yolk of the egg does not interfere with visualization.
5. Other genomic resources and technologies:
    a. **A high-quality *Saccoglossus* BAC library** has been constructed and arrayed under the direction of Dr. Jeff Tomkins (Clemson University Genomics Institute, BAC/EST

Resource Center), and copies of the library will be distributed on request. The library contains ~111,000 BAC clones with an average insert size of 130 kb, providing ~13-fold physical coverage of the *Saccoglossus* genome.

b. **Three large, high-quality cDNA libraries** have been constructed by Dr. Chris Gruber (Invitrogen/Life Technologies). Library 1 was constructed using mRNA preparations from mixed blastula–early gastrula stages and contains 22×10^6 cDNA clones. Library 2 was generated using mRNA from mixed (late) gastrula–neurula stages and contains 50×10^6 clones. Library 3 was generated from mRNA preparations that originated from 2-week old juveniles (two gill slits) and contains 3.7×10^6 cDNA clones. Over 66,000 clones have been sequenced (most of these clones are arrayed), and analysis confirms that each of these libraries contains a diverse set of high-quality cDNAs (in pSPORT6, ~2 kb average size, high diversity).

c. **EST collection and full-length sequences.** Over 66,000 clones from three cDNA libraries have been picked at random for EST sequencing. The sequences have been blasted (BLASTx) and organized in a database for inspection and internal sequence comparison. Information will be transferred to GenBank within the next year as we publish papers. At this time, over 100 clones have been fully sequenced and 80 have been used for *in situ* hybridization analysis. These clones, found in hemichordates, have strong similarity to chordate orthologs. In Table 1, examples are grouped according to their chordate roles to show the richness of the collection:

Table 1. Summary of key patterning and structural genes from *S. kowalevskii* EST collection.

Conserved gene function in chordates	Hemichordate gene present in EST collection
Anterior neural patterning genes	*six3, bf1, vax, rx, dlx, nkx2.1, otp*
Mid-level neural patterning genes	*emx, otx, pax6, tll, barH, iro, lim1/5, dbx, pax5, en*
Posterior neural patterning genes	*gbx, pax1/9, hox1, hox2, hox3, hox4, hox6, hox8, hox9/10, hox11/13, cad*
Dorsoventral neural patterning genes	*bmp2/4, bmp7, dbx, tbx2/3, dlx, msx, islet, mnx, nkx2.2, nkx6.1, olig2-like, netrin, hh, hnf3b, sim, semaphorins, slit*
Neuron markers, neuronal cell types	*elav, musashi, sox2/3, poxneuro, numb, brn3, brn1/2/4, neurofascin, tag1, opioid receptor, neuropeptide Y receptor, serotonin transporter, GABA transporter, GAD (GABA synthesis), GABA receptor, vasotocin, tyrosine hydroxylase*
Other signaling components	*twisted gastrulation, tolloid/xolloid, robo, DCC, fgf4, fgf9, wnt1/14, wnt2/11, wnt3a, wnt4, wnt7a/16, WIF, TGF*β*-binding protein, notch, delta, serrate*
Other transcription factors of regional patterning	*bra, lhx3/4, lhx9, dachshund, buttonhead, aristaless, snail slug, hnf3b, foxD, sim, gata6, vent, zic*
Thyroid/endostyle-related genes	*ttf1 (nkx2.1), ttf2, thyroglobulin, iodide symporter,* various thyroxin forming and destroying enzymes
Prechordal plate and notochord related genes	*dkk, hex, noggin, kielen, hh, bra, fgf4, hnf3b, admp, collagen type II*
Kidney-related gene	*pod1* (podocyte-specific)
Heart/blood-related genes	*nkx2.5, gata2, hex, dHAND,* various heart muscle proteins and *ATPase*
Pituitary-related genes	*pitx, gata2, bmp2/4, lhx3/4, prop1, thyrotropin releasing hormone receptor, corticotropin-releasing hormone receptor*
Gill slits	*pax1/9*
Left–right genes	*pitx, inversin, left-right dynein*
Skeletal muscle specific genes	*myoD, myostatin-like skeletal muscle myosin, MEF1,* various skeletal muscle specific proteins

The rationale for the complete sequence of the organism

Most of the comparative and evolutionary questions we are asking are very broad, and better answered within the context of the full genome sequence. Our current genomic resources consist of 66,000 EST sequences and a recently available BAC library. We could theoretically pull out one genomic region at a time, making slow but inevitable progress with questions of chordate origins. But the answers will be definitive and more rapidly obtained only if we have the whole genome to work from. The reasons for requesting the entire genome as opposed to analyzing ESTs and BACs include:

1. The comparisons of hemichordates with chordates and echinoderms should be as broad as possible — the more genes and regulatory regions the better.
2. Residual synteny vis-à-vis chordates and gene clustering (not just *Hox* but also *NKL*, *Parahox*, and *EGHbox* clusters) will result.
3. Single-copy hemichordate genes can be compared with chordate duplications and paralog diversifications. Chordate researchers looking for a hemichordate ortholog of their chordate gene, in the interest of discerning "ancestral function", can do so with certainty of its presence or absence in hemichordates.
4. Characterization of *cis*-regulatory sequence vis-à-vis chordates and echinoderms and availablility of such sequences for directed expression experiments. Interestingly, the 22 neural patterning genes we investigated by *in situ* hybridization show simpler expression domains than in chordates — usually a single contiguous region, similar to just one of the chordate domains. Possibly *cis*-regulation for spatial expression is simpler and would reveal a "core" or "basal" regulation mode preserved but added to in the chordate line. Some neural domains of hemichordates have boundaries like those in chordates (*pax6/six3, en/bf1, otx/gbx*), perhaps reflecting ancestral regulatory interactions. Thus, hemichordates may contribute to the understanding of the evolution of gene regulation in chordates.
5. Intron–exon patterns giving high-assurance phylogenetic assignment (as recommended by Sydney Brenner) relative to other deuterostomes.
6. The opportunity to complement both the echinoderm and chordate sequences to give a detailed profile of the deuterostome supertaxon.

The *Saccoglossus* genome and current status of its genomic resources

The haploid *Saccoglossus* genome is estimated to be about about 1.1 Gb, or 35% that of mammalian genomes (~3 Gb). Karyotyping suggests that the genomic DNA is organized into 20 chromosome pairs (1N = 20). Up to one milligram of high-quality genomic DNA is readily obtained from sperm and gonad tissue of a single individual using standard purification methods.

As described above, several genomic resources are already available that will complement and significantly aid in the assembly and/or in the validation and annotation of the *Saccoglossus* genome sequence: 1) a BAC library providing ~13-fold physical coverage of the *Saccoglossus* genome, 2) three large, high-quality cDNA libraries, and 3) over 66,000 EST sequences and about 100 full-length sequences derived from these three cDNA libraries. Paired-end sequences from BAC clones will be extremely valuable in the whole-genome shotgun assembly by providing the long links necessary to achieve large sequence scaffolds. In addition, the BAC library will provide direct access to cloned DNA for any researcher interested in a specific region of the genome. The EST and protein sequences can be placed on the genome assembly and used to train and validate *Saccoglossus* gene-finding programs.

Whole-genome shotgun sequencing strategy

Our goal is to produce a long-range, high-quality (contig N50 > 10 kb, supercontig N50 > 1 Mb) whole-genome assembly. Specifically, we propose a whole-genome shotgun sequencing strategy to generate an assembly of the *Saccoglossus* genome representing ~6-fold sequence coverage in bases with a Phred quality score of ≥20. The sequence will be generated as paired-end reads from 4-kb and 10-kb plasmids and from Fosmid and BAC clones (Table 2). The ratio of sequences derived from plasmid clones versus large-insert clones will be 9:1. By using different insert sizes and vector types we will minimize cloning bias and allow a hierarchical linking approach in the assembly process to produce the largest possible sequence scaffolds.

Table 2. Sequencing strategy.

Insert size (kb)	Vector type	Attempted reads (M)	Average read length (Phred20)	Phred20 sequence coverage[a] (X)	Physical coverage[b] (X)
4	Plasmid	7.4	790	4.2	10.6
10	Plasmid	2.2	750	1.2	8.0
40	Fosmid	1.0	680	0.5	14.7
130	BAC	0.2	680	0.1	9.6
Total		**10.8**		**6.0**	**42.9**

[a] Sequence coverage is calculated assuming a sequencing pass rate of 80% and a genome size of 1.1×10^9 bases.
[b] Physical coverage of genome represented by shotgun clones used in the assembly.

Our sequencing approach will require a total of 10.8 M attempted reads, corresponding to ~2.5 months of the current sequencing capacity at the Whitehead Institute/MIT Center for Genome Research (WICGR). The combined physical coverage of the sequenced clone inserts will be roughly 43X.

All of the generated paired-end reads derived from the different library types will be assembled using the whole-genome shotgun assembly program ARACHNE (19,20). This software package was developed at WICGR and has been adapted recently to allow for assembly of large (mammalian-size) genomes, such as the mouse genome (21).

Polymorphism rate, repeat content, and segmental duplications are all critical parameters for a successful whole-genome assembly. We estimated the allelic variation of the *Saccoglossus* genome by resequencing randomly selected genomic fragments from multiple organisms. Specifically, we generated and resequenced PCR products based on BAC end sampling from eight individuals (16 chromosomes). We analyzed ~100 loci in the genome with a total length of ~27,000 bp. Our results show that the average heterozygosity rate is ~1/2500. Interestingly, the polymorphism rate of the *Saccoglossus* genome is lower than in human (1/1300), and considerably lower than the rate observed in the genome of *Ciona savignyi* (1–2%), another marine organism. The low polymorphism rate may reflect the facts that *S. kowalevskii*'s habitat is mostly in isolated, protected bays rather than contiguous stretches of coastline. In addition, the adults form burrows in the sand, and the adult female lines the inside of the burrow with oocytes rather than releasing them by broadcast spawning. *S. kowalevskii* does not have a free-swimming larval stage, and embryos develop encapsulated within a vitelline envelope until they hatch, competent to burrow, minimizing embryonic dispersal and gene flow between populations. Thus the effective population size is expected to be extremely small. These results suggest that the assembly of the *Saccoglossus* genome should not be problematic. In addition, we will take care to achieve as low a polymorphism rate as possible by constructing all new libraries from the genomic DNA of a single organism.

We believe that producing a long-range, high-quality shotgun assembly should be the highest priority for this project. If complete sequence in certain regions of interest is desired, a subset of Fosmids or BACs covering these genomic regions could provide templates for finishing. The extent to which finishing should be carried out can be prioritized later, on the basis of evolving assessment of cost and capacity.

Access to the *Saccoglossus* genome data

Genome data will be released in accordance with NHGRI rules. All traces will be submitted to the NCBI trace archive. The whole-genome assembly will be made freely available to the scientific community by submission to GenBank. In addition, WICGR may work with the community to annotate the genome assembly and to display the data via an extensive web-based database.

Are there other (partial) sources of funding available or being sought for sequencing this project?

No other sources of funding are being sought for this project.

Selected References

1. Ogasawara, M., Wada, H., Peters, H., and Satoh, N. (1999) Developmental expression of Pax1/9 genes in urochordate and hemichordate gills: Insight into function and evolution of the pharyngeal epithelium. *Development* **126:** 2539–2550.
2. Burdon-Jones, C. (1952) Development and biology of the larva of *Saccoglossus horsti* (enteropneusta). *Proc. Roy. Soc. London B* **236:** 553–589.
3. Balser, E.J. and Ruppert, E.E. (1990) Ultrastructure and function of the preoral heart-kidney in *Saccoglossus kowalevskii* (Hemichordate; Enteropneusta) including new data on the stomochord. *Acta Zool.* **71:** 235–249.
4. Goodrich, E.S. (1917) "Proboscis pores" in craniate vertebrates, a suggestion concerning the premandibular somites and hypophysis. *Quart. J. Microscop. Sci.* **62:** 539–553.
5. Ruppert, E.E., Cameron, C.B., and Frick, J.E. (1999) Endostyle-like features of the dorsal epibranchial ridge of an enteropneust and the hypothesis of dorsal-ventral axis inversion in chordates. *Invertebr. Biol.* **118:** 202–212
6. Lowe, C., Wu, M., Salic, A., Evans, L., Lander, E.S., Stange-Thomann, N., Gruber, C., Gerhart, J., and Kirschner, M. (2003) Anteriorposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* (in press).
7. Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme. B., and de Rosa, R. (2000) The new animal phylogeny: Reliability and implications. *Proc. Nat. Acad. Sci. USA* 97: 4453–4456.
8. Cameron, C.B., Garey, J.R., and Swalla, B.J. (2000) Evolution of the chordate body plan: New insights from phylogenetic analyses of deuterostome phyla. *Proc. Nat. Acad. Sci. USA* **97:** 4469–4474.
9. Gee, H. (1996) *Before the Backbone: Views on the Origin of the Vertebrates.* (London: Chapman and Hall).
10. Bateson, W. (1886) The ancestry of the chordata. Quart. J. Microscop. Sci. **26:** 535-571.
11. Hyman, L.H. (1959) The enterocoelous coelomates--Phylum Hemichordata. In "The Invertebrates" Vol. 5, McGraw-Hill Book Co., New York, pp. 72-207.
12. Cameron, C.B. and Mackie, G.O. (1996) Conduction pathways in the nervous system of Saccoglossus sp. (enteropneusta). *Canadian Journal of Zoology* **74:** 15–19.
13. Bullock, T.H. (1965) The nervous system of hemichordates. In: *Structure and Function in the Nervous Systems of Invertebrates*. T.H. Bullock and G.A. Horridge (eds). San Francisco: WH Freeman and Co.
14. Takacs, C.M., Moy, V.N., and Peterson, K.J. (2002) Testing putative hemichordate homologues of the chordate dorsal nervous system and endostyle: Expression of NK2.1 (TTF-1) in the acorn worm *Ptychodera flava* (Hemichordata, Ptychoderidae). *Evol. Dev.* **4:** 405–417.
15. De Robertis, E.M., and Sasai, Y. (1996) A common plan for dorsoventral patterning in bilateria. *Nature* **380:** 37–40.
16. Nübler-Jung, K., and Arendt, D. (1999) Dorsoventral axis inversion: Enteropneust anatomy links invertebrates to chordates turned upside down. *J. Zoolog. Syst. Evol. Res.* **37:** 93–100.
17. Lowe, C.J., Tagawa, K., Humphreys, T. Kirschner, M., and Gerhart, J. (2003) Hemichordate embryos; procurement, culture, and basic methods. In: *Methods in Cell Biology*. G.A. Wray, C. Ettensohn, and G. Wessel (eds). San Diego: Elsevier Science.
18. Colwin, L.H. and Colwin, A.L. (1962) Induction of spawning in *Saccoglossus kowalevskii* (Enteropneusta) at Woods Hole. *Biol. Bull.* **123:** 493.
19. Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002) ARACHNE: A whole-genome shotgun assembler. *Genome Res.* **12:** 177–189.
20. Jaffe, D.B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. (2003) Whole-genome sequence assembly for mammalian genomes: ARACHNE 2. *Genome Res.* **13:** 91–96.
21. Mouse Genome Sequencing Consortium (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.