Reference: *Biol. Bull.* **214:** 284–302. (June 2008)

cDNA Sequences for Transcription Factors and Signaling Proteins of the Hemichordate *Saccoglossus kowalevskii*: Efficacy of the Expressed Sequence Tag (EST) Approach for Evolutionary and Developmental Studies of a New Organism

R. M. FREEMAN, JR.[1], M. WU[2], M-M. CORDONNIER-PRATT[3], L. H. PRATT[3], C. E. GRUBER[4], M. SMITH[4], E. S. LANDER[1,5,6], N. STANGE-THOMANN[5], C. J. LOWE[7], J. GERHART[2,*], AND M. KIRSCHNER[1]

[1]*Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02115;* [2]*Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3200;* [3]*Department of Plant Biology, University of Georgia, Athens, Georgia 30602;* [4]*Express Genomics, Inc., 5 South Wisner Street, Frederick, Maryland 21701;* [5]*Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142;* [6]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139; and* [7]*Department of Organismal Biology and Anatomy, University of Chicago, Chicago, Illinois 60637*

Abstract. We describe a collection of expressed sequence tags (ESTs) for *Saccoglossus kowalevskii*, a direct-developing hemichordate valuable for evolutionary comparisons with chordates. The 202,175 ESTs represent 163,633 arrayed clones carrying cDNAs prepared from embryonic libraries, and they assemble into 13,677 continuous sequences (contigs), leaving 10,896 singletons (excluding mitochondrial sequences). Of the contigs, 53% had significant matches when BLAST was used to query the NCBI databases ($\leq 10^{-10}$), as did 51% of the singletons. Contigs most frequently matched sequences from amphioxus (29%), chordates (67%), and deuterostomes (87%). From the clone array, we isolated 400 full-length sequences for transcription factors and signaling proteins of use for evolutionary and developmental studies. The set includes sequences for fox, pax, tbx, hox, and other homeobox-containing factors, and for ligands and receptors of the TGFβ, Wnt, Hh, Delta/Notch, and RTK pathways. At least 80% of key sequences have been obtained, when judged against gene lists of model organisms. The median length of these cDNAs is 2.3 kb, including 1.05 kb of 3′ untranslated region (UTR). Only 30% are entirely matched by single contigs assembled from ESTs. We conclude that an EST collection based on 150,000 clones is a rich source of sequences for molecular developmental work, and that the EST approach is an efficient way to initiate comparative studies of a new organism.

Received 29 October 2007; accepted 4 March 2008.

* To whom correspondence should be addressed. E-mail: jgerhart@berkeley.edu

Abbreviations: EST, expressed sequence tag; UTR, untranslated region.

Introduction

We began an EST (expressed sequence tag) project five years ago to obtain gene sequences we could use in the comparison of hemichordates to chordates, with regard to their development and body plans. We now report on the properties (cDNA lengths, EST number, contig formation) and sequence diversity of this collection, and on our success in obtaining sequences of interest for developmental and evolutionary studies. Though little studied, hemichordates are important for comparisons to chordates because of their evolutionary position and body plan. They are now thought to be the sister phylum of the pentaradial echinoderms (Fig. 1A), and both phyla make up the sister group of the chordate phylum, which includes urochordates, cephalochordates, and vertebrates (Castresana *et al*., 1998; Bromham and

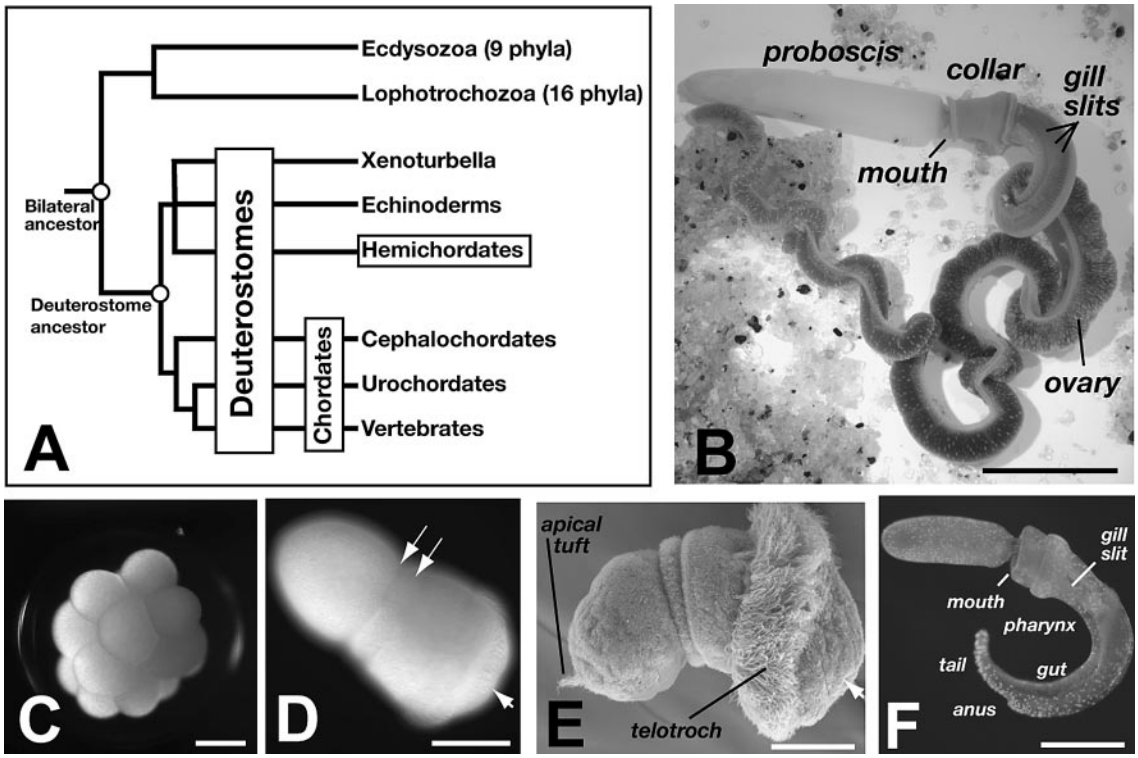

Figure 1. The phylogeny and development of *Saccoglossus kowalevskii*, a direct-developing harrimaniid enteropneust. (A) A simplified phylogeny of deuterostomes based on base sequence comparisons (Cameron *et al*., 2000; Bourlat *et al*., 2006). Note that the deuterostome ancestor is the most recent common ancestor of hemichordates and chordates. (B) Mature female *S. kowalevskii*, gravid (scale bar = 1 cm). (C) A 16-cell embryo (5 h post-fertilization) in its fertilization envelope (scale bar equals 0.1 mm). (D) A neurula-stage embryo, 3 days post-fertilization (scale bar = 0.1 mm), surface view, anterior toward the left top corner. Two parallel arrows mark the two grooves defining the prosome, mesosome, and metasome. A short white arrow marks the closed blastopore. (E) Scanning electron micrograph of a 4.5-day embryo. Anterior is to the left and dorsal to the top. The mouth forms at this stage, and the body elongates and bends ventrally at the posterior end. A white arrow marks the site of the closed blasopore (scale bar = 0.1 mm). (F) Hatched juvenile, 2 weeks post-fertilization, surface view, with body parts indicated (scale bar = 0.1 mm).

Degnan, 1999; Cameron *et al*., 2000; Furlong and Holland, 2002; Smith *et al.,* 2004; Bourlat *et al*., 2006). Recent sequence-based phylogenies place cephalochordates (amphioxus) as the sister group of all other chordates and *Xenoturbella* as a new phylum related to hemichordates and echinoderms (Bourlat *et al*., 2006; Putnam *et al*., 2008). All these groups are within the deuterostome supertaxon and descended from an enigmatic deuterostome ancestor. Of all the deuterostomes, hemichordates emerge as the phylum of bilateral adults that is evolutionarily most closely related to chordates and most suitable for evolutionary developmental comparisons.

Since chordates develop directly, we chose a direct-developing hemichordate, *Saccoglossus kowalevskii* (Agassiz), to facilitate comparisons of adult development and anatomy. Like other enteropneust hemichordates, *S. kowalevskii* is a vermiform animal with three body parts (Fig. 1B, F), a proboscis (prosome), collar (mesosome), and pharynx-gut region (metasome). Like other hemichordates, its probable chordate-like anatomical traits include gill slits and an endostyle in the pharynx (Ogasawara *et al*., 1999; Ruppert, 2005; Rychel *et al*., 2006; Rychel and Swalla, 2007), and a post-anal tail-like extension of the metasome, at least in juveniles of harrimaniid enteropneusts (Burdon-Jones, 1952; Lowe *et al*., 2003). Other chordate traits such as the notochord and dorsal hollow nerve cord, though attributed to hemichordates by Bateson (1885), are doubtful; they instead have a diffuse subepidermal nerve net (Bullock, 1965; Lowe *et al*., 2003) and a stomochord lacking notochord characteristics (Peterson *et al.,* 1999; Gerhart *et al.,* 2005). Suspected but largely unstudied traits include a pituitary-like rudiment, a podocyte kidney, and a heart-like contractile vessel (Goodrich, 1917; Balser and Ruppert, 1990; Gerhart *et al*., 2005); their development resembles that of chordates such as amphioxus (Fig. 1C, D, E; Gerhart *et al*., 2005). Molecular analyses are clearly needed to clarify and extend the classical anatomical comparisons. Like others studying novel organisms, we hoped that such

comparisons would reveal developmental processes and traits shared by both groups, in our case hemichordates and chordates, that could then be attributed to their ancestor, in our case the deuterostome ancestor. From the differences of the two groups, we sought to elucidate what evolutionary innovations, at the molecular level, occurred in the separate lineages.

We needed a large set of gene sequences to analyze by whole-mount *in situ* hybridization the spatiotemporal expression of various genes known to be important in chordate development and to test the functional consequences of altered expression of these genes by siRNA knockdown experiments. Furthermore, we wanted a set of nonredundant full-length cDNAs for overexpression studies, microarrays, and expression screens. In recent years, EST sequencing has become the method of choice for obtaining large collections of identified cDNAs for model organisms (*e.g.*, Strausberg *et al.*, 1999; Zhu *et al.*, 2001; Gilchrist *et al.*, 2004; Sekiguchi *et al.*, 2007), and it has reached a price range accessible to investigators of new organisms (*e.g.*, Gyoja *et al.*, 2007). New sequencing methods promise to reduce the cost still further. The method for obtaining ESTs entails preparing cDNA libraries from a variety of developmental stages, inserting cDNAs into plasmids propagated in bacterial hosts, arraying bacterial clones, and sequencing the ends of the cDNAs of individual clones.

An EST project for hemichordates no longer needs to be a broad survey of coding sequences but can be directed to the acquisition of specific gene sets, as the result of the past two decades of informative studies of the conserved components used in the development of chordates, echinoderms, *Drosophila,* and nematodes (Scott, 1994, 2000; De Robertis and Sasai, 1996; Veraksa *et al.*, 2000; Barolo and Posakony, 2002; Hsia and McGinnis, 2003; Swalla, 2006). Several hundred cDNA sequences (let's say 400) for known transcription factors and signaling ligands would enable an in-depth comparison to other groups. Specifically, one set should include ligands and members of the signaling pathways of Wnts, Hedgehogs, TGFβs (Bmps and Nodal), Notch/Delta, RTK ligands, G-coupled receptors, and nuclear hormone receptors (Gerhart, 1998; Pires-daSilva and Sommer, 2003; Massague, 2004; Logan and Nusse, 2004). Another set should include members of various transcription factor families, such as the HOX, NK, TBX, FOX, PAX, and bHLH families (for references, see Davidson *et al.,* 2006, on the sea urchin gene models). A further set should include conserved neural-patterning genes (for example, *otx, gbx, pax6*; see Wilson and Houart, 2004) and neural-cell-type–specific genes (*e.g.*, *atonal, neuroD, neurogenin*, genes for enzymes of GABA and serotonin synthesis), muscle-specific genes (for example, *myoD, myosin light chain*), and genes associated with chordate-distinctive traits such as the notochord (for example, *bra, xnot, chordin*), the dorsal hollow nerve cord (*nkx2.2, gsx, nkx6.1, msx, emx, otx, vax*), gill slits (*pax1/9, six1, casr*), post-anal tail (*hox10-13*), left-right asymmetry (*nodal, vg1, lefty*), pituitary (*pitx, pomc, pit*), heart (*nkx2.5*), kidney (*tcf21, nephrin*), thyroid (*ttf1, ttf2*), and Spemann's organizer (*chordin, noggin, dkk1/2*; see Weinstein and Hemmati-Brivanlou, 1999; De Robertis, 2006). Hence, ours was a directed search for certain kinds of transcribed gene sequences, not a survey of all coding sequences of the animal. Furthermore, since many of these gene sequences are conserved across a wide range of bilateral phyla, it was reasonable to expect that the *S. kowalevskii* homologs could be identified by standard BLAST analysis of the EST sequences. The question for us, then, was how large an arrayed EST collection was needed to identify and obtain these high priority cDNAs?

We obtained 202,175 ESTs representing 163,633 arrayed clones from cDNA libraries of a variety of embryonic and juvenile stages, the latter stage possessing adult organization and physiology but lacking gametogenesis. The purposes of this report are first, to characterize the entire EST collection in terms of contig number and size, the EST membership of contigs, and the annotation and putative functional identification of contigs and singletons; and second, to evaluate a subset of the cDNAs of the collection in regard to their length and coding sequence, the composition of the untranslated region (UTR), and their abundance in the collection. This subset was chosen as containing the longest ones available for development studies.

Overall we conclude that a collection of this size indeed affords a majority (>80%) of useful long sequences for developmental and evolutionary comparisons, and that the systematic collection of ESTs provides an efficient way to approach a newly studied organism. As published elsewhere (Lowe *et al.*, 2003, 2006; Aronowicz and Lowe, 2006), these sequences are proving useful in discovering previously unsuspected similarities in the development and body plans of hemichordates and chordates. Such similarities were not discerned in anatomical comparisons published in the previous century (Bateson, 1886; Colwin and Colwin, 1953), or they may have been assigned incorrectly. Sequencing of the *S. kowalevskii* genome is now well advanced, and the ESTs and contigs will further assist genome assembly and the construction of useful gene models.

Materials and Methods

Animals, eggs, and embryos

Adults of *Sacoglossus kowalevskii* were collected intertidally in September from a single salt pond near Woods Hole, Massachusetts. Ovulation and fertilization were achieved in the laboratory by the methods of Colwin and Colwin (1950, 1962) with modifications (Lowe *et al.*, 2004). Embryos were staged by the normal tables of Bateson (1884, 1885, 1886) and Colwin and Colwin (1953).

Library construction

Six plasmid cDNA libraries were prepared for EST sequencing, two from mixed blastula and gastrula stages, two from mixed late gastrula and neurula stages, and two from juveniles with 2–3 gill slits. For each developmental interval, one library was normalized, the other not. Libraries were prepared and normalized at Invitrogen and Express Genomics (Bethesda, MD). For each library, 800–1000 embryos were rapidly frozen with liquid nitrogen, and total RNA was isolated using TRIzol Reagent (Invitrogen). Poly(A)$^+$ RNA was then isolated by two rounds of oligo(dT) selection with oligo(dT)-coated magnetic particles (Seradyn, Inc.).

From the poly(A)$^+$ RNA, a cDNA library was constructed by using an oligo dT primer-adapter containing a Not I site and Moloney Murine Leukemia Virus Reverse Transcriptase (M-MLV RT) to prime and synthesize first-strand cDNA. After the second strand was synthesized, the double-stranded cDNA was size-fractionated (>1.4 kb) and cloned directionally into the Not I-Eco RV sites of the pExpress-1 vector, or into the Not I-Sal I or Not I-Eco RV sites of the pCMVSport6 vector. Primary clones were produced from a bulk ligation (300 ng of vector, cut as described above, and 120 ng of Not I digested cDNA per 120 μl of ligation), which was followed by electroporation into T1 phage-resistant *E. coli*. The cDNA inserts from 24 randomly picked clones were sized by digestion with Not I and Eco RI.

A normalized cDNA library was produced from the primary cDNA library. Biotinylated driver RNA produced from the T7 RNA polymerase promoter and single-stranded (ss) target DNA produced from the F1 *ori* were hybridized to each other at a low value of C_{ot} (concentration of driver multiplied by the time of hybridization). The RNA:DNA hybrids were removed by phenol extraction, and the remaining ss target DNA was converted to double-stranded DNA (dsDNA) with a repair oligo and *Taq* DNA polymerase. After electroporation of the dsDNA into T1 phage-resistant *E. coli*, primary clones were produced. The cDNA inserts from 24 randomly picked clones were sized by digestion with Not I and Eco RI. The normalization process reduces the level of abundant and moderately abundant genes and enriches for rare genes. To determine the reduction of an abundant transcript, we compared the relative representation of the actin sequence in the primary library with the normalized library by hybridization of different dilutions of the primary and normalized dsDNA with an actin probe. Using this method, we determined that the actin sequence was reduced at least 20-fold in the normalized library.

Table 1 gives details of the libraries, including the identification of the EST sequencing runs done on each. The EST sequences, which have been deposited in GenBank, bear these identifiers (G50, G125, G142, *etc*).

Table 1

Saccoglossus kowalevskii libraries used for expressed sequence tag (EST) sequencing

Developmental stage of library (time from fertlization)	Library type	Vector (all in *E. coli* DH10B)	Primary clones	Average insert length, kb	EST run size and direction of sequencing
Blastula and gastrula, 14–20 h	Non-normalized	pCMVSport6	50×10^6	1.9	G125 and G613: 5′ reads only (24,363 ESTs)
Gastrula and neurula, 18–36 h	Non-normalized	pCMVSport6	22×10^6	1.7	G50: 5′ and 30% paired 3′ reads (12,288 ESTs) G178: 5′ reads only (15,770 ESTs)
Juvenile, 2 wk	Non-normalized	pCMVSport6	44×10^6	1.5	G142: 5′ reads only (17,686 ESTs)
Late gastrula, 20–24 h	Normalized (90-fold actin sequence reduction)	pExpress1	40×10^6	1.4	G708: 5′ and 10% paired 3′ reads (17,665 ESTs) G825: 5′, 3′ paired reads (39,399 ESTs)
Dorsal flexure, 4 d	Normalized (80-fold actin sequence reduction)	pExpress1	42×10^6	1.05	G709: 5′ and 13% paired 3′ reads (18,681 ESTs)
Juvenile, 2 wk	Normalized (90-fold actin sequence reduction)	pExpress1	28×10^6	1.0	G710: 5′ and 10% paired 3′ reads 21,137 ESTs) G826: 5′, 3′ paired reads (37,974 ESTs)

In total, 206,715 ESTs were obtained, from which mitochondrial and contaminant sequences were removed as described in the text, leaving 182,607 ESTs for contig assembly and annotation.

EST sequencing

Sequencing was completed at the Whitehead Institute/MIT and Broad Institute Sequencing Centers, by standard procedures (Lander *et al.*, 2001). Clones were picked and arrayed in 96-well plates prior to sequencing. Plate and well numbers were matched with EST reads so that particular clones could be sampled later for preparing *in situ* hybridization probes and for characterizing inserts more fully. Plates were stored at −80°C. Chosen inserts were sequenced completely (UC Berkeley Sequencing Facility) using primers spaced approximately 500 bases apart.

PCR-based isolation of cDNAs with longer inserts

This method served as a nonradioactive alternative to plate lift/autoradiographic methods and was particularly useful for the isolation of rare clones (*e.g.,* <1 in 30,000 bacteria of a library). Bacteria from unamplified libraries were seeded, 250 per tube, into 64 tubes, each containing 1.0 ml of nutrient broth, and all arranged in an 8 × 8 racked array on which rows and columns were numbered. Thus each rack contained approximately 16,000 different insert-bearing bacteria. In total, 10 racks of tubes were prepared (160,000 bacteria sampled). After a day of bacterial growth at 37 °C, 0.1 ml from each tube of a given row was collected, and these aliquots were mixed to make eight pooled row cultures; likewise, tubes of each column were sampled, 0.1 ml each, to make eight pooled column cultures. Minipreps of plasmid DNA were made from these 16 pooled cultures and frozen (−20°C). In addition, aliquots of the row and column DNA samples were combined to make one sample (kept at −20°C) representing all cDNAs of the rack. To each of the 64 bacterial cultures of the rack (each of 0.8 ml volume), an equal volume of 50% sterile glycerol was added and all were frozen at −80°C.

For the PCR assay, we designed exact primers (Operon) based on an existing short EST sequence for which we needed a longer cDNA insert (more 5′ sequence) from the libraries. Rack samples, and then row and column samples, were assayed by PCR to locate the tube containing the particular clone, and from that tube about 400 bacteria were spread on a plate (LB agar plus ampicillin, 100 mg/ml). After overnight colony growth (37 °C), the bottom of this master plate was marked with lines to divide it into eight sectors. Two replica plates were prepared, also marked into eight sectors and oriented relative to the master plate. After the replica plates had grown, the colony-bearing agar medium of the second replica was cut into eight sectors with a scalpel blade (#11), and each sector was placed in a 50-ml tube and washed with 1 ml of LB medium to suspend the bacteria, which were then lysed with 10 ml of 20 mmol l^{-1} NaOH for 5 min to release DNA, and the lysate was neutralized with 20 ml of 20 mmol l^{-1} acetic acid. Lysates were assayed by PCR to find the location of the sector bearing the desired clone. That sector of the master plate was divided into 10–15 patches (3–4 colonies per patch), outlined on the plate bottom, and the colonies and agar of each patch were removed into a 1.5-ml tube and washed with 200 μl of LB. Alkaline lysates were prepared as described above and assayed by PCR to find the patch bearing the clone. The first replica plate was matched with the master plate to locate the 3–4 colonies of the patch, and each colony was picked and streaked to obtain single, well-separated colonies. Portions of single colonies were suspended, lysed by NaOH/acetic acid as described above, and assayed by PCR to locate the desired clone. The procedure takes about 10 days to complete, and several searches can be run in parallel. More than 70 cDNA clones with long inserts have been obtained by this method.

One drawback to this method is that slow-growing bacteria, initially seeded at 1 in 250 in a tube, fall behind others during growth in the tube and are thereafter too rare for isolation on plates (*e.g.,* 1 in 2500). This happened in about 10% of the attempted isolations. Still, the identified tube served as a good source of DNA for PCR amplification of insert DNA.

Bioinformatics

EST processing. Trace files from the EST sequencing were processed by the MAGIC system, ver. 0.92 (Liang *et al.*, 2006). Briefly, MAGIC performed base-calling *via* Phred 0.020425.c (Ewing *et al.*, 1998; Ewing and Green, 1998), trace quality gap elimination, vector and adapter trimming, and finally the designation of contaminating sequences. At the project start, *Balanoglossus carnosus* mitochondrial DNA was used to identify mitochondrial contaminants; thereafter, when it became available, *S. kowalevskii* mitochondrial DNA (GenBank AY336131 by Smith, Beckenbach, and Scouras) was relied upon to identify mitochondrial EST sequences using SSAHA ver. 3.2 (Ning *et al.*, 2001). Furthermore, 4400 eukaryotic entries from the Ribosomal Database Project (Maidak *et al.*, 2001) were used for ribosomal RNA screening. An additional manual step for removing vector and adapter sequences from all the pExpress-1 clones sequenced from the SP6 primer site was performed utilizing regular expression text searches. This step was required to compensate for a known issue with Phred miscalling particular bases in a 20-base window close to the sequencing primer; removal ensured that these sequences would not interfere with clustering and assembly. Only the resulting trimmed, high-quality reads (referred to as *processed ESTs*) were used for further analyses. These reads have been deposited in GenBank, with accession numbers FF418995-534157 and FF602128-677500.

To complement MAGIC, the processed ESTs were imported into a small, cross-platform, custom FileMaker Pro (ver. 8.0) database system, which facilitated the data man-

agement, mining, and presentation activities of the research team. The software is available by request.

Annotation. Using BLASTx, processed ESTs were compared against both GenBank NR and Uniprot (SwissProt+ TrEMBL) sequences, the latter to facilitate the assignment of gene ontology (GO) terms and the functional identification (when available) of sequences from high-quality curations. Up to 20 top BLASTx hits were retained for each. To reduce the amount of data, BLASTx hits were routinely screened against a blacklist of genes deemed uninteresting to the project (*e.g.,* mitochondrial genes, ribosomal sequences, tubulin, actin). Hits were discarded if the *E* value was $>10^{-10}$ or if they were on the blacklist; however, the first hit of a blacklisted set was retained. Original reports showing the BLASTx hits and alignments were retained and made accessible for clarification purposes.

Removal of contaminating sequences. Processed ESTs were compared by BLASTn against *S. kowalevskii* WGS traces in GenBank. ESTs that did not elicit a hit, whether or not they themselves scored as significant ($\leq 1 \times 10^{-10}$) BLASTx hits, were tagged as contaminants and were not included in further analyses.

Clustering and contig assembly. Processed ESTs were clustered in two rounds using TGICL (Pertea *et al*., 2003) and were assembled using CAP3 (Huang and Madan, 1999). All ESTs (5′ + 3′) from the normalized libraries were clustered and assembled first, as these sequences represented broader coverage of the cDNA libraries. In the second round, all output (both contigs and singletons) from the first round was combined with a nonredundant set of sequences from the non-normalized libraries, and were then clustered and assembled. Any large contigs having more than 10,000 members were subsequently reclustered and reassembled. The final contig set represents the second-round contigs and any first-round contigs that were marked as singletons in the second round; true singletons are the non-contig (*i.e.,* EST) sequences from the second-round output.

Model organisms BLAST. We assembled from Genbank a database of 1,187,343 protein entries representing 13 model organisms (*Nematostella vectensis*, 49,985 entries; *Caenorhabditis elegans*, 56,572; *Aedes aegypti*, 18,398; *Tribolium castaneum*, 10,144; *Drosophila melanogaster*, 79,085; *Strongylocentrotus purpuratus*, 43,929; *Branchiostoma floridae*, 50,817; *Danio rerio*, 64,544; *Xenopus laevis/tropicalis*, 45,110; *Gallus gallus*, 31,544; *Rattus norvegicus*, 113,019; *Mus musculus*, 241,757; and *Homo sapiens*, 411,523). Contigs were compared against this database using BLASTx. Either the top hit or the top five hits were scored. In situations where two hits from different organisms scored similarly, both were kept at the same rank and the next lower rank was skipped. Also, organisms were scored only once in the set of five hits; a second hit from the same organism was skipped and did not affect the ranking.

Results

Characterization of the total EST collection

We have obtained 202,175 ESTs, representing 163,633 clones. Of the cDNAs of these clones, 44,144 have both 5′ and 3′ read pairs, whereas 113,887 (56%) have only 5′ reads. Of the total number of ESTs, 21.8% were 3′ reads, and 65.3% were from normalized libraries. All sequences have been deposited with GenBank. We intentionally sought more 5′ than 3′ reads of inserts because we anticipated that 3′ reads would often fall within 3′ UTR regions and hence would less often provide useful functional annotation *via* BLAST analysis than would 5′ reads. Indeed, as discussed later in more detail, 50.4% of the 3′ reads gave no BLASTx hit, whereas 27.7% of the 5′ reads gave no hit. As the project continued, we increased the percentage of 3′ reads to facilitate the clustering and assembly of ESTs into contigs (Heber *et al*., 2002; Wang *et al*., 2004).

Contig assembly. Prior to assembling contigs from the EST collection, we removed ESTs containing mitochondrial and mitochondrial-like sequences. We used the *Saccoglossus kowalevskii* mitochondrial genome sequence (GenBank AY336131) for identifications and found that 19,568 ESTs, 9.7% of our total, represented mitochondrial and mitochondrial-like sequences. With their removal, we were left with a set of 182,607 ESTs for contig assembly and BLASTx analysis. From these ESTs, we assembled 13,677 contigs (*i.e.,* containing two or more overlapping ESTs), while 10,896 ESTs remained as singletons—that is, 94% of the ESTs were contained by contigs. In a *Xenopus tropicalis* EST collection of comparable size, 93% of the ESTs were contained by contigs, an equivalent frequency of entry (Gilchrist *et al.*, 2004).

When the 13,677 contigs were associated with one another on the basis of their containing one or the other partner of a 5′-3′ read pair, the number of contig clusters was found to be to 10,722. This number may approach the number of different transcripts ("virtual transcripts," or "unique expressed sequences," or "unigenes") represented in our collection, but still may overestimate it because 56% of our ESTs were obtained as unpaired reads.

The length distribution of *S. kowalevskii* contigs is shown in Figure 2. Contig lengths range from 106 to 5795 bases, with a median length of 883 bases and a mean of 1048 bases, shorter than the average insert length of 1.33 kb in the libraries.

Contigs differ widely in the number of ESTs contained in each, as shown in Figure 3. Most contigs have low EST membership. For example, 65% contain five or fewer ESTs,

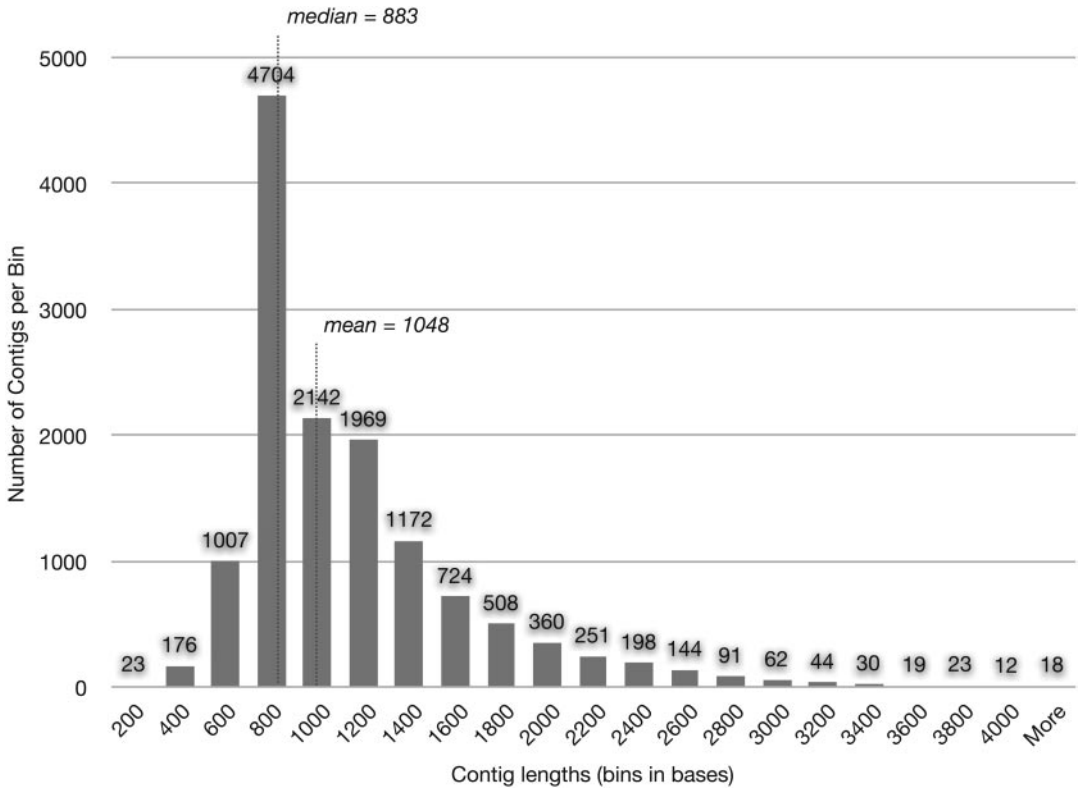

Figure 2. Distribution of the lengths of 13,677 contigs assembled from 182,607 expressed sequence tags (ESTs), from which mitochondrial sequences and contaminants had been removed (see Methods). Lengths are grouped in bins increasing by 200 base intervals (horizontal axis), and the number of contigs falling in that length bin is shown on the vertical axis. The median EST read length after trimming was 611 bases. The median length of contigs is 883 bases, and the mean is 1048 bases. The mean size of inserts in the libraries used for the ESTs is 1.33 kb, after correcting length for the contribution of each library to the total collection of 182,607 ESTs, suggesting that most inserts are incompletely covered by ESTs.

and 32% contain only two; however, the tail of the distribution of membership is long and monotonic, with 1.7% of the contigs containing 100 or more ESTs.

For the top 5 high-membership contigs as well as 10 mitochondrial sequences, we assessed the contribution of ESTs from the normalized *versus* non-normalized libraries in order to evaluate the efficacy of removing high-abundance cDNAs by normalization. As shown in Figure 4, most of the ESTs of these contigs indeed derived from non-normalized libraries; for example, the representation of alpha tubulin, beta tubulin, cytochrome *c* oxidase subunit 1, EF1alpha, and cytoplasmic actin were reduced by 96%, 93%, 96%, 95%, and 88%, respectively. Furthermore, whereas mitochondrial and mitochondrial-like sequences accounted for 22.4% of the total non-normalized sequences, they constituted only 2.77% of the total normalized sequences, an 88% reduction. For comparison, Figure 4 also displays 17 middle- and low-abundance cDNAs that were unaffected or enriched by the normalization procedure. Despite the efficacy of normalization, we chose to derive 34% of our ESTs from non-normalized libraries because the cDNAs are longer (approximately 1.7 kb, *versus* 1.2 kb for normalized libraries; see Table 1), and therefore they favor the recovery of full-length cDNAs.

Annotation. Contigs and singletons were provisionally identified by BLASTx comparison against GenBank's NR and Uniprot (SwissProt+TrEMBL) databases. The distributions of BLASTx scores for the contigs and singletons are shown in Figure 5. Of the 13,677 contigs, 53% gave Expect (*E*) values $\leq 10^{-10}$, indicative of significant similarity; 16% gave BLASTx hits of low significance ($>10^{-10}$), and 31% gave no BLASTx hit. Of the 10,896 singletons, 51% gave significant BLASTx hits ($\leq 10^{-10}$), 14% gave hits of low significance ($>10^{-10}$), and 36% gave no hit. Of the singletons receiving no hit, 75% of them derived from 3′ reads and probably represent 3′ UTRs.

When contig sequences are compared in more detail with the sequences of their top hits, it appears that 26% of all contigs ($n = 13{,}677$) and 47% of contigs with significant BLASTx hits ($\leq 10^{-10}$; $n = 7{,}292$) overlap the start ATG of the top hit's sequence, or reach to within 20 amino acids of the start codon.

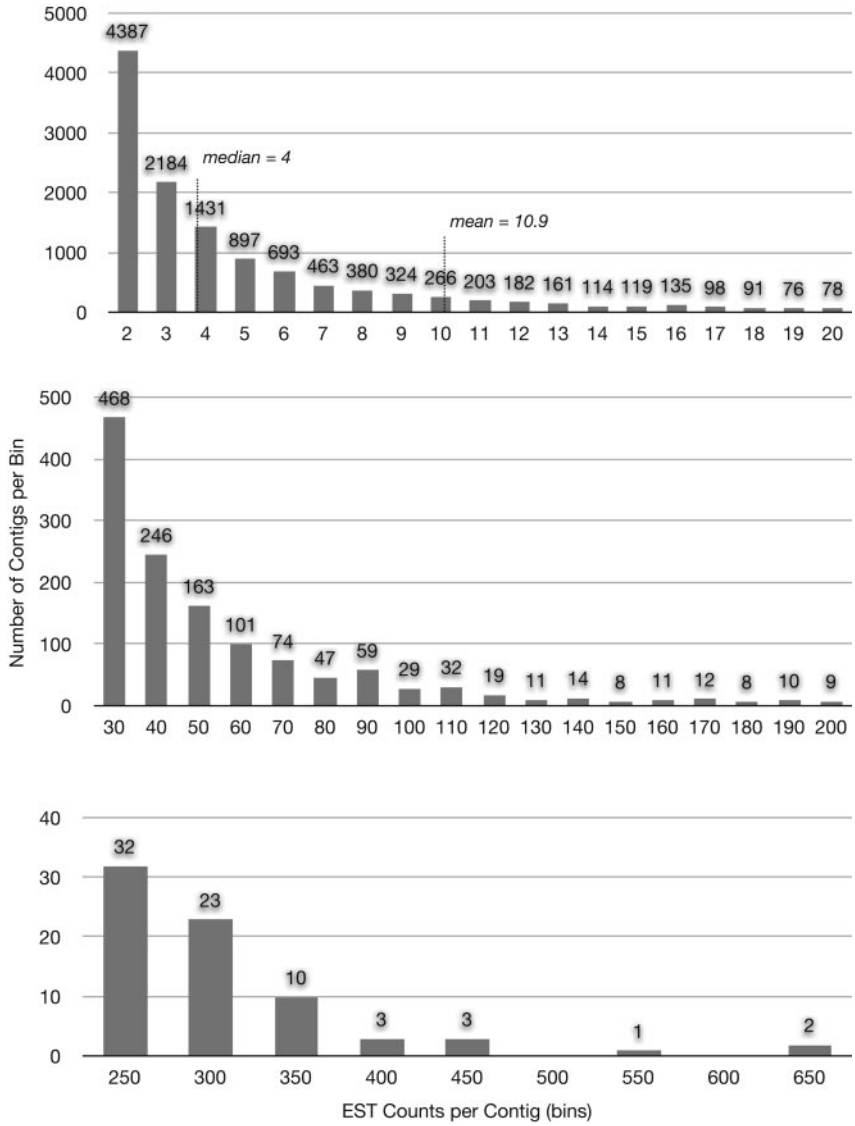

Figure 3. Distribution of the number of expressed sequence tags (ESTs) contained by contigs. Contigs are ordered on the horizontal axis in three groups: upper panel, those containing 2–20 ESTs; middle panel, those containing 21–200 ESTs; and lower panel, those containing 201–650 ESTs. The number of contigs having that number of ESTs is shown on the vertical axes. The number of contigs adds up to 13,677, and the number of non-redundant ESTs is 148,581 (see Methods).

That is, at least 3400 contigs can be used to identify arrayed clones with full-length or nearly full-length cDNA inserts. The actual number of ESTs overlapping the start ATG of the top hit sequence is, of course, greater than 3400 (a number to be estimated later) because many contigs contain more than one EST at the 5′ end. All such ESTs represent full-length cDNA inserts, and their high frequency reflects the success of the reverse transcription step of cDNA production.

We then asked which organisms have protein sequences most similar to those encoded by the *S. kowalevskii* contigs, limiting our inquiry to contigs with BLASTx matches of an *E* value $\leq 10^{-10}$. This inquiry was not an effort to illuminate phylogeny but merely to determine which of the well-studied animals have sequences most similar to those of this hemichordate. Using BLASTx, we compared the *S. kowalevskii* contigs against the sequences of 13 animals for which extensive genomic information and gene models are now available (*Nematostella, Caenorhabditis elegans, Drosophila, Aedes, Tribolium*, sea urchin [*Strongylocentrotus purpuratus*], amphioxus [*Branchiostoma floridae*], zebrafish, *Xenopus laevis* and *tropicalis*, chick, rat, mouse, and human). As shown in Figure 6, amphioxus is the animal with sequences most often giving the top hit. Of the *S. kowalevskii* contigs, 29% are best matched by amphioxus

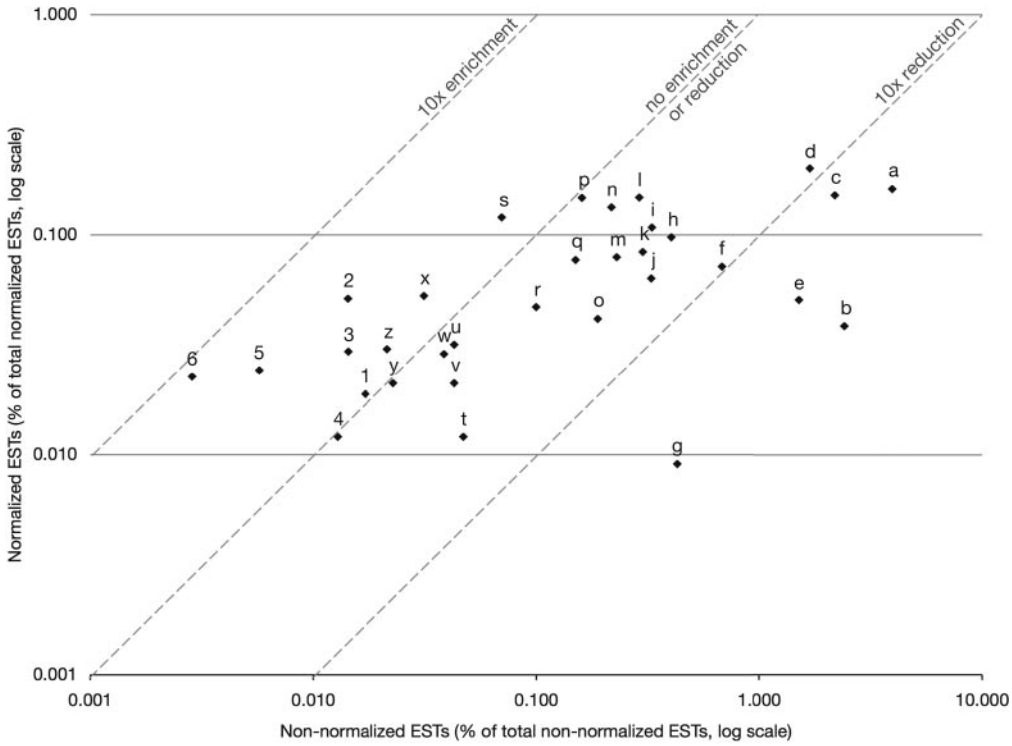

Figure 4. Reduction or enrichment of expressed sequence tags (ESTs) by normalization of the cDNA libraries. Thirty-two full-length cDNAs, or coding sequences in the case of mitochondrial transcripts, were chosen because of their high, medium, or low abundance and were used to call up matching ESTs from the entire collection (202,175 ESTs), that is, including mitochondrial sequences; and then the ESTs were identified as deriving from non-normalized or normalized libraries. The numbers were then divided by the total number of ESTs obtained from normalized (n = 70,107, including mitochondrial sequences) or non-normalized (n = 132,068 including mitochondrial sequences) libraries to obtain percent values, which were then plotted on a log-log scale. The three dashed lines indicate hypothetical cDNAs for which normalization causes 10× reduction, 10× enrichment, or no effect. Letters and numbers identify the assessed sequences as (a) alpha tubulin, (b) 16S mitochondrial rRNA, (c) beta tubulin, (d) cytoplasmic actin, (e) cytochrome *c* oxidase subunit 1, (f) EF1alpha, (g) NADH dehydrogenase subunit 1, (h) cytochrome *c* oxidase subunit 2, (i) cytochome *b*, (j) ATPase F0 subunit 6, (k) NADH dehydrogenase subunit 2, (l) NADH dehydrogenase subunit 4, (m) NADH dehydrogenase subunit 4L, (n) cytochrome *c* oxidase subunit 3, (o) major yolk protein, (p) muscle lim, (q) sox2/3, (r) NADH dehydrogenase subunit 5, (s) musashi, (t) hairyA, (u) foxA2, (v) ornithine decarboxylase, (w) ftz-f1, (x) ets, (y) orthopedia, (z) 18S rRNA, (1) dynein beta chain, (2) six3, (3) iroquoix, (4) dlx, (5) dan, (6) foxC.

sequences, and 69.2% of the contigs have an amphioxus sequence among their top 5 hits. As was shown in Figure 1A, amphioxus represents Cephalochordata, the likely sister group of tunicates plus vertebrates within the chordate phylum (Bourlat *et al*., 2006; Putnam *et al*., 2008). Chordates as a group have sequences giving 66.6% of the top matches to *S. kowalevskii* contigs, and deuterostomes (chordates plus *S. purpuratus*) give 86.6% of the top hits. Despite the evidence that hemichordates share a more recent common ancestor with echinoderms than with chordates (see Sodergren *et al*., 2006), sequences of *S. purpuratus* are the top BLASTx hit for 20% of the *S. kowalevskii* contigs, compared to 29% for amphioxus. This seeming discrepancy may reflect, not phylogenic relationships, but the more rapid evolution (sequence divergence) of sea urchins relative to other groups such as amphioxus, as reported by Putnam *et al.* (2008). Paralog groups in vertebrates may in sum be larger than in amphioxus and *S. purpuratus*, a factor which may contribute to the high frequency of vertebrate matches for *S. kowalevskii*. The animals with the fewest sequences to give top BLASTx matches with *S. kowalevskii* are the ecdysozoan protostomes, *Drosophila* (2.3%), *C. elegans* (0.8%), *Aedes aegypti* (1.2%), and *Tribolium* (2.6%). The cnidarian *Nematostella*, a non-bilaterian diploblast, gives a higher frequency of top hits (6.7%) than do the ecdysozans, consistent with the proposed gene losses within the latter group (Putnam *et al*., 2007).

Recovery of developmental gene sequences from the EST collection

We searched the contigs and singletons for *S. kowalevskii* homologs of gene sequences known to have developmental

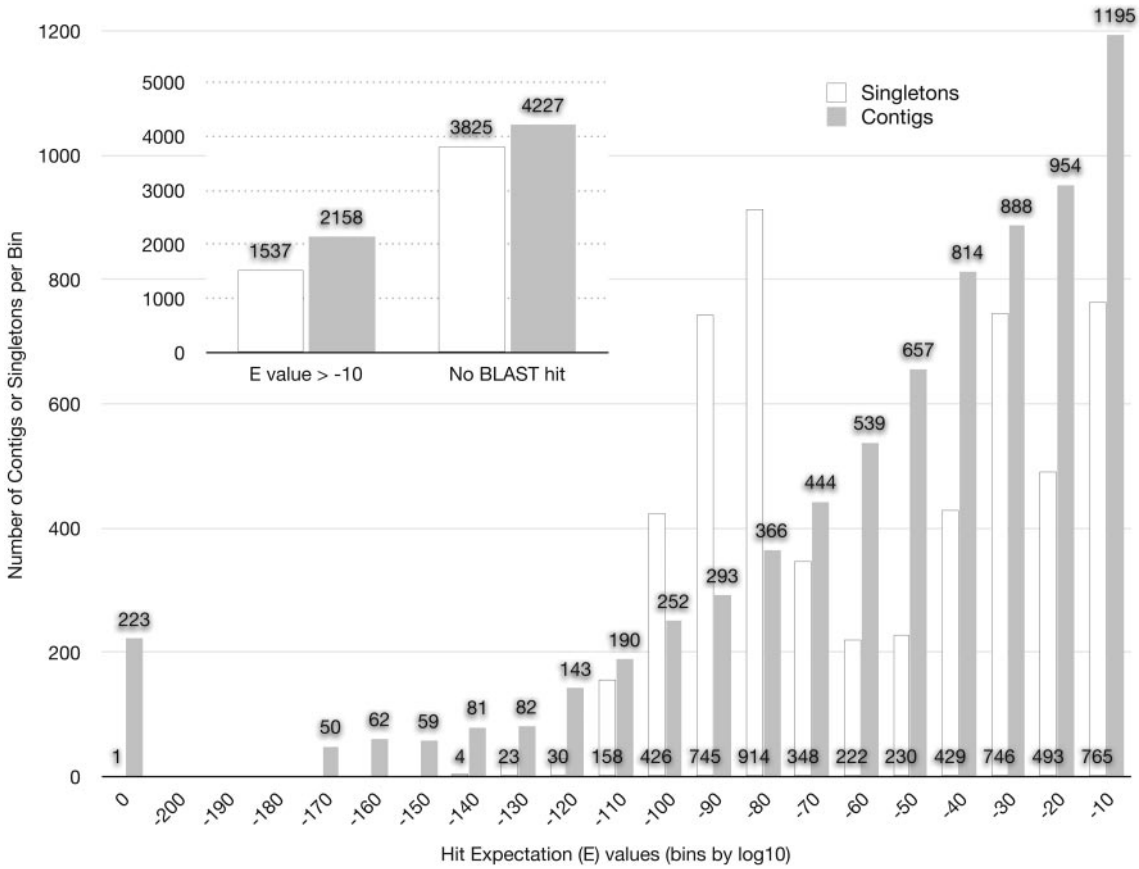

Figure 5. Distribution of BLASTx scores for contigs and singletons from the *Saccoglossus kowalevskii* expressed sequence tag (EST) collection. Mitochondrial and mitochondrial–like sequences were removed prior to the BLASTx analysis. The horizontal axis displays increasing E values binned by orders of magnitude. The zero value on the x-axis indicates $E < 10^{-200}$. The vertical axis indicates the number of contigs (total of 13,677) or singletons (total of 10,896) in each bin of score values. Note that a separate vertical scale is used in the insert (upper left) for contigs and singletons with the scores of lowest significance ($>10^{-10}$) or with no BLASTx hit.

roles in well-studied animals, especially chordates, sea urchin, fruit fly, and nematodes. The success of our recovery can be estimated by comparing the variety of our candidate homologs to the lists of members of transcription factor families and signaling pathways recently tabulated from the annotated genomes of chordates and echinoderms. In Table 2, 200 full-length or nearly full-length sequences from the arrayed clones of the EST collection are listed with their top BLASTx scores. These are compared to sequences of the sea urchin, S. *purpuratus*, since for this closely related organism, even more than for amphioxus at this time, extensive gene lists have been published (*e.g.,* see Croce *et al*., 2006; Howard-Ashby *et al*., 2006a, b; Lapraz *et al*., 2006; Materna *et al*., 2006; and Tu *et al*., 2006).

For example, Bmp ligands, receptors, and transduction intermediates, as well as Bmp antagonists and modulators, are important in dorsoventral patterning in a variety of bilateral model organisms such as *Drosophila* and *Xenopus*, and also for *S. kowalevskii* (Lowe *et al*., 2006). Components identified from the arrayed clones of the EST collection include *bmp2/4, bmp5-8, twg, cv-2, xolloid/tolloid, chordin, noggin, follistatin, gremlin1, gremlin2, neuralin, dan, kielin-like, bmpR1B, bmpR2, bambi, smad1/5, smad4, smad6/7, smurf,* and *tak*. Sequences of possible interest, but missing from the EST collection, are *sizzled* and *cerberus*, which may be chordate-specific proteins. In comparison to the gene model lists from *S. purpuratus* (Lapraz *et al*., 2006), we have recovered >85% of members in our EST collection.

Wnt ligands, Wnt pathway intermediates, and Wnt antagonists are important for anteroposterior patterning in chordates, and recent evidence indicates such a role in *S. kowalevskii* (Lowe *et al*., unpubl. data). In chordates, 13 Wnt ligands have been identified, and in *S. purpuratus*, 11 Wnt ligands. We have recovered sequences for 11 of 13 Wnt ligands (including WntA) in the EST collection, and for three of four receptors (*fz5, 8,* and *10*), the co-receptor *lrp5/6*, and the wnt antagonists *dkk1/2, dkk3, sfrp1/5, wish/sclerostin, wif,* and *wingful*. Wnt pathway components that have been recovered include β*-catenin, axin, dsl, gsk3beta,*

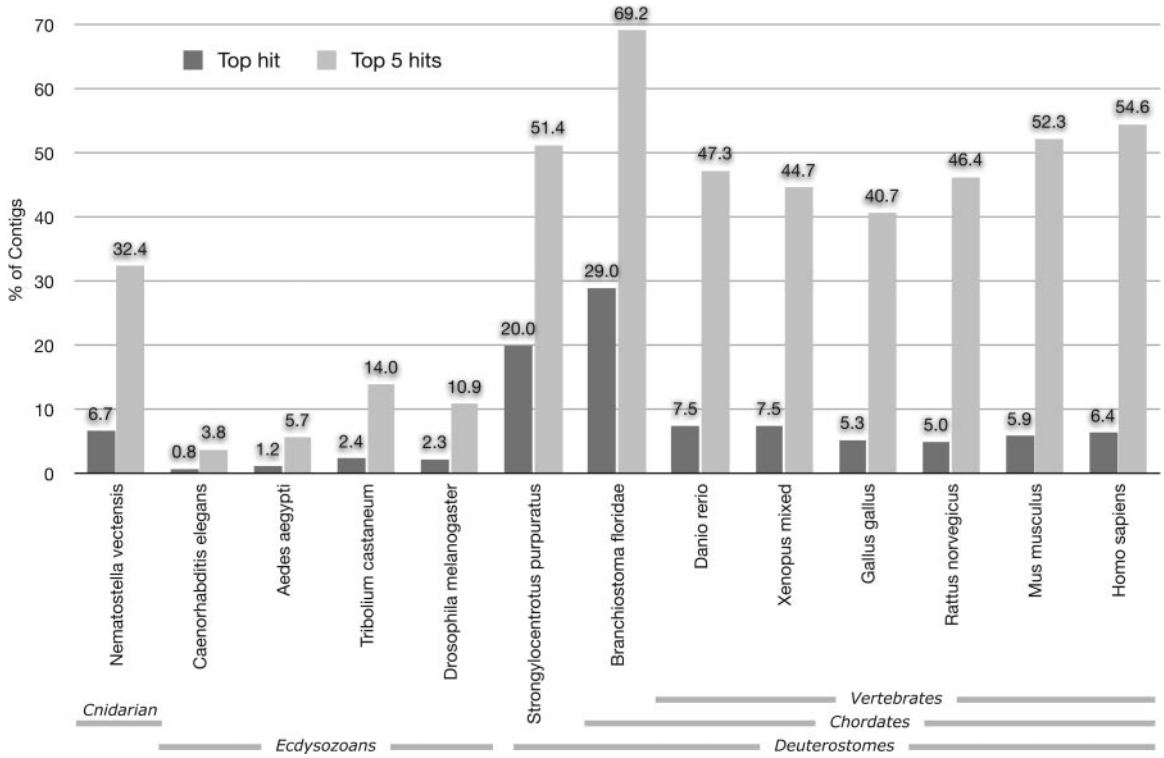

Figure 6. Frequency of matches of *Saccoglossus kowalevskii* contigs to the sequences of 13 animals with sequenced genomes and comprehensive sets of gene models. The animals are ordered along the horizontal axis to reflect phylogenetic relatedness. The cnidarian diploblast, *Nematostella*, is on the left; four ecdysozoans are grouped next, and the deuterostomes are arranged to the right as a group with the chordates, vertebrates, and mammals nested therein (human on the right). *Xenopus* "mixed" means data from *X. laevis* and *X. tropicalis* were combined. The vertical axis is the percentage of the *S. kowalevskii* contigs that have a top BLASTx match to a sequence of the designated animal (dark bars) for 7409 contigs that gave 10,315 significant top hits ($E \leq 10^{-10}$), since some contigs had two top hits of identical score. Also shown (light bars) is the percentage of contigs having that animal's sequence among their top five BLASTx hits. For this analysis, the top five hits of each contig were inspected to eliminate multiple hits by the same animal (*e.g.*, we took only one of three human sequences ranked in succession), and hits were counted until 5 of the 13 animals were represented. For some contigs, fewer than five significant hits were obtained, and so the total number of hits for all contigs was 35,071. The percentages for "top 5 hits" are were obtained by taking the number of contigs having a particular animal's sequences in the top five and dividing by 7409, the number of total contigs.

gbp, and *tcf7/LEF*. Missing members of interest from the EST collection include *wnt6, wnt10, fz4, adenomatous polyposis coli protein, frzb, and sfrp4* (which have been obtained by other means from *S. kowalevskii*). Matched against the *S. purpuratus* gene models, we have recovered >75% of the Wnt family members and their signaling modifiers from our EST collection (Croce *et al*., 2006).

Among members of transcription factor families (Table 2), we have found 11 Hox family members in the EST collection (Aronowicz and Lowe, 2006), perhaps a full set when compared to the 11 of *S. purpuratus* (Howard-Ashby *et al.*, 2006a, b). Among other homeobox-containing gene sequences, we have recovered NK family members 1.1, 2.1, 2.2, 2.3/2.5, 3, 5.1, 6.1, and 7.1, a number equal to those known in *S. purpuratus*; and we found most of the Pax family members, namely, 1/9, 2/5/8, 6, and 2/B (compare to Howard-Ashby *et al*., 2006a). From the Fox family of transcription factors, we retrieved homologs of 14 of the 19 kinds known in *S. purpuratus* (Tu *et al*., 2006) and 12 of the 17 subfamilies of vertebrates (FoxA through Q); and from the Tbx family, we retrieved all six members known in *S. purpuratus* (Howard-Ashby *et al*., 2006b). Homologs absent from the EST collection yet of interest for developmental studies include *pax3/7, evx, lox,* and *gsx* (which have been recently isolated from *S. kowalevskii* cDNA libraries by other means).

Further examples of recovered homologs are given in Table 2 and in previous publications on anteroposterior and dorsoventral development (Lowe *et al*., 2003, 2006; Aronowicz and Lowe, 2006). Of course, many other genes have been recovered and are not indicated in Table 2 because they are not of current interest for our developmental

Table 2

cDNA sequences for evolutionary and developmental studies

Pathway or gene family of interest for comparisons	Number of genes in *S. purpuratus*	Examples of cDNAs looked for, and found, in the *S. kowalevskii* EST collection (*E* values [exponent base10] from BLASTx analysis are in parentheses	Examples of missing sequences
Bmp signaling pathway, antagonists, and modifiers	(a) 11 ligands, 8 receptors, 15 modulators and antagonists	bmp2/4(-159), bmp3(-38), bmp5-8(-85), bmpR1b(-158), bmpRII(-33), smad1/5(<-200), admp(-96), smad6/7(-96), smurf(-42), bambi(-25), dullard(-91), tak(-130), chordin(<-200), noggin(-31), follistatin(-73), gremlin(-40), neuralin(-66), kielin(<-200), dan(-29), twg(-44), cv-2(-160), xolloid(-68)	cerberus?
Nodal signaling pathway, and other TBFβs		nodal1(-33), nodal2(-36), univin(-56), alk4 receptor(-62; partial), smad2/3(-118), cripto(-10), dally(-91), nomo(-126), activin(-26), myostatin(-31), tgfβ2(-65)	lefty
Wnt signaling pathway, antagonists, and modifiers	(b) 11 ligands, 4 receptors, and >5 antagonists	wnt1(-112), wnt2/13(-141), wnt3(-125), wnt4(-118), wnt5(-140), wnt7(-144), wnt8(-104), wnt9(-97), wnt11(-100), wnt16(-85), wntA(-75), fz2/7(<-200), fz5/8(<-200), fz10(<-200), lrp5/6(-70), ryk(-142), axin(-137), β-catenin(<-200), casein kinase1(-165), dsl(<-200) gskbeta(<-200), tcf7/lef(-106), prickle(-171), van gogh(-122), β-trcp(<-200), dkk1/2(-20), dkk3(-19), sfrp1/5(-72), wish/sclerostin(-30), wif(-45), wingful(-131)	wnt6, wnt10, adenomatous polyposis coli protein, frzb, fz4, sfrp4, cerberus?
Notch/Delta signaling pathway	(c) 2 ligands, 1 receptor, 20 pathway genes	delta(-162), serrate(-92), notch receptor(<-200), fringe(-96), su(H)(<-200), mindbomb(<-200), neuralized(<-200), numb(-101), nicastrin(-110), nkd(-7), noonanptp(-111), nrarp(-29), presenilin(-116), adam10(<-200).	
Hox and parahox families	(d) 11 hox, 3 parahox	hox1(-98), hox2(-33), hox3(-23), hox4(-39), hox5(-71), hox6(-63), hox7(-56), hox9/10(-102), hox11/13a(-80), hox11/13b(-47), hox11/13c(-90), cdx(-57)	lox, gsx, hox8
Other homeobox transcription factors	(e) approx. 70, excluding hox and parahox	alx(-34), barH(-42), brn124(e-96), brn3(-71), bsh(-34), chx10(-19), dachshund(e-50), dbx(-64), dlx(-62), dmbx(-52), drgx(-31), en(-31), emx(-37), gbx(-34), glass, gsc(-32), hex(-42), iroquois(-78), islet(-128), lbx(-48), lhx1/5(-51), lhx3/4(-113), lhx2/9(-106), lmx1(e-21), meis(e-71), mox/gax(-38), mnx(e-40), msx(-49), numb (e-101), olig2/3(-34), otp(-42), otx(-100), pbx(-143), pnox(e-46), pitx(-46), prox1(-80), rx(e-55), six1(e-105), six3(-101), tale(-33), tlx(-39), unc42(-60), vax(-31), vent/vox(-19), xanf(-16)	xnot, evx
Pax family		pax1/9(-146), pax2/5/8(-76), pax6(-122), paxB/2(-55), poxNeuro(-59)	pax3/7
NK family		1(-36), 2.1(-104), 2.2(-66), 2.3-2.5(-34), 3/bagpipe(-45), 5.1/h6(-27), 6/gtx(-42), 7(-15)	
Smad family	4(a,e)	smad1/5(-200), smad2/3(-118), smad4, smad6/7(-96)	
Tbx family	6 (e)	tbx1(-66), tbr(-169), bra(-88), tbx2/3(-123), tbx6/18(-30), tbx20(-126)	
Fox family	19-23 (f)	a2/hnf3(-102), b(-74), c(-87), d(-50), e/ttf2(-64), f(-65), g/bf1(-77), j, n/nude(-60), o(-89), p(-106), q2(e-34), q-like1(-49)	
Zinc-finger proteins	≫30 (g)	buttonhead(-64), churchill(-39), fez/foreheadin(-100), gli(-134), klf2/nepture(-48), klf5(-59), nocA(-69), snail(-75), zic/opal(-167), krox/egr(-61), klf/sp1, ovo(e-46), spalt(e-115), scratch(-81).	
Nuclear hormone receptors	33 (e)	amnr2-like(-83), coup-tf(<-200), thyroid hormone receptor(-127), rxr(e-64), rarβ(-49), ftz-f1(-165), tll(-69), pparγ(-37), photoreceptorTF(-142)	
Sox family	11 (e)	sox2/3(-77), sox7/17(-41), sox11(-56), sox14/21(-66), bbx(-27), tcf7(-106)	

These cDNAs were recovered from an EST collection of 163,633 arrayed clones from *Saccoglossus kowalevskii*, as described in the text. References: (a) Lapraz *et al*., 2006; (b) Croce *et al*., 2006; (c) Walton *et al*., 2006; (d) Arnone *et al*., 2006, Howard-Ashby *et al*., 2006b; (e) Howard-Ashby *et al* 2006a; (f) Tu *et al*., 2006; (g) Materna *et al*., 2006.

studies. Some of these sequences have been used in phylogenetic studies to resolve the relationships of hemichordates, echinoderms, amphioxus, and other chordates (Bourlat *et al*., 2006; Peterson *et al.,* 2006). In summary, from our collection of 202,175 ESTs, representing 163,633 clones, we estimate that our recovery of cDNA sequences encoding members of particular families and pathways has exceeded 80% (1 of 5 missing) in most cases, and in some cases has

reached 90% (1 of 10 missing). The recovered set of gene sequences is certainly sufficient to undertake comprehensive comparative studies.

Analysis of 120 cDNAs from the EST collection

We have isolated and fully sequenced the cDNAs of about 400 of the 163,633 clones that were arrayed in the course of EST sequencing. Each cDNA is the longest insert available for that particular transcript sequence, as identified by BLASTx. All cDNAs encode transcription factors and signaling components of interest for developmental studies, and their full length or nearly full length qualifies them for overexpression studies, siRNA knockdown, microarrays, and further genomic studies. In about 10% of these cases, the collection did not contain a full-length cDNA, and so a still longer insert clone was isolated from the cDNA libraries by a PCR-based assay (see Methods). For the purpose of this presentation, we have analyzed 120 of the 400 sequenced clones in more detail with regard to the abundance and diversity of ESTs and contigs matching to them, their number of 5′ ESTs, and the lengths of their open reading frames and 3′ UTRs. This subset gives us more information about the availability and quality of sequences of developmental interest in the arrayed collection.

As shown in Figure 7, the number of ESTs matching each full-length sequence varied widely, with a median number of 9.5 ESTs distributed over the sequence and an average of 5.4 ESTs coinciding with the 5′ end. Most (63%) of the ESTs at the 5′ end come from non-normalized libraries, even though only 34.7% of the total ESTs derive from non-normalized libraries, a bias consistent with the greater length of inserts in the non-normalized libraries. For 8 of the 120 cDNAs, only a single EST was available, and in 10 cases only 2 were available, whereas for 55 cDNAs, 10 to 100 matching ESTs were present in the collection, and in 3 cases more than 100 ESTs matched the sequence *(sox2/3, muscle LIM, musashi/nrp)*. The difference in abundance is presumably attributable not only to differences of mRNA abundance in the initially extracted embryos, but also to differences in the suitability of mRNAs for reverse transcription to cDNA and of cDNA sequences for clone growth. Of these 120 cases, 40 (33%) of them were each matched along the entire length by a single contig assembled from ESTs, whereas the remaining 80 cases were matched intermittently along their lengths by multiple con-

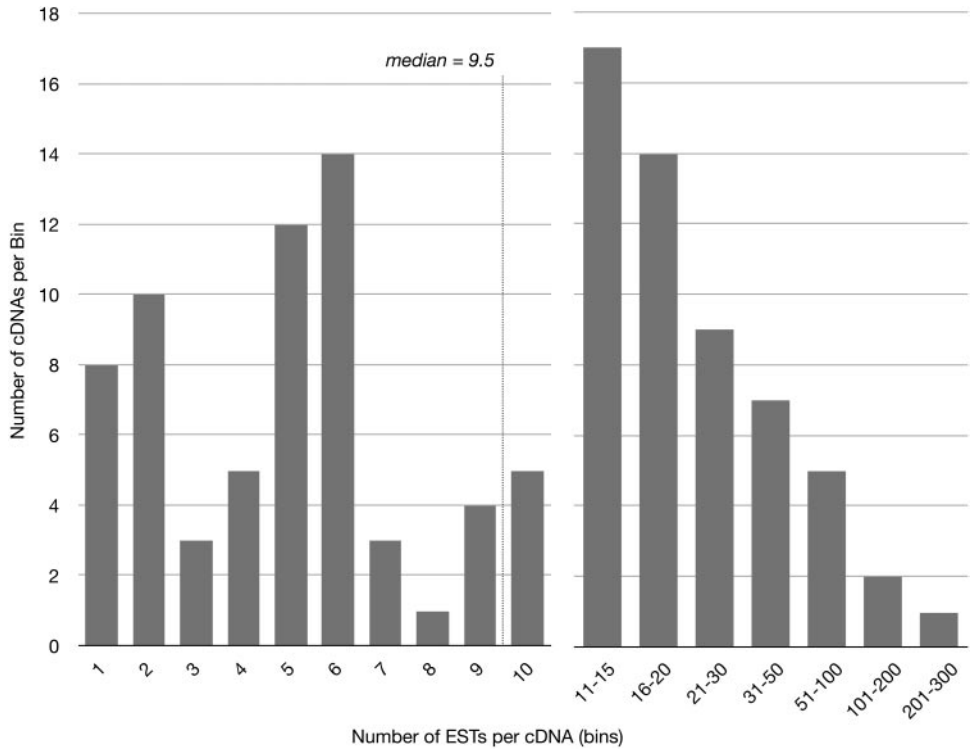

Figure 7. Distribution of the number of expressed sequence tags (ESTs) that match to each of 120 cloned long cDNA inserts of interest for evolutionary and developmental studies of *Saccoglossus kowalevskii*. The ESTs are derived from the *S. kowalevskii* collection totaling 202,175 members. The horizontal axis shows the number of ESTs that match each full-length cDNA, and the vertical axis shows the number of cDNAs (total of 120) matched by that number of ESTs. The median number of ESTs for a sequence is between 9 and 10.

tigs and/or singletons from the EST collection. Overall, the 120 chosen sequences called up 160 contigs, perhaps implying that our entire collection of 13,677 contigs represents only 10,258 unique transcript (mRNA) sequences, roughly equal to the 10,722 contig clusters formed from the total EST collection by associating contigs containing 5′-3′ read pairs. The 120 cDNAs also called up 33 singletons, perhaps indicating that as many as 3000 of our 10,896 singletons of the EST collection do not signify unique transcripts but are non-overlapping members of otherwise identified contig clusters. As expected for ESTs obtained as paired (5′ + 3′) reads of long inserts (*e.g.,* 1.9 kb for our non-normalized blastula/gastrula library), the mid-regions of the cDNAs are infrequently represented by ESTs.

We estimate that 104 of the 120 sequences (87%) include the translation start site, compared to 47% of the entire set of contigs with significant BLASTx *E* values ($\leq 10^{-10}$). The difference simply reflects the fact that we deliberately picked these clones as ones containing the most 5′ ESTs. These full-length coding sequences derive more often (66%; 69 of 104) from the non-normalized libraries, consistent with the fact mentioned above that 63% of ESTs covering translation start sites are derived from non-normalized libraries, even though only 34.7% of the ESTs derived from cDNA inserts from non-normalized libraries. The bias may reflect the median longer length of inserts from non-normalized libraries. Although full-length inserts are, expectedly, more easily found for mRNA species represented by multiple ESTs, even for the 18 cDNAs matched by only one or two ESTs, 12 contained full-length open reading frames, reflecting the high quality of the libraries.

The length distribution of the 120 cDNAs is shown in Figure 8. They range in size from 1.0 to 5.4 kb, with a median length of 2.3 kb, similar to the estimated length of processed transcripts (*i.e.,* introns removed) in other organisms (Samanta *et al.*, 2006; Sodergren *et al.*, 2006). Within this length, the coding and 5′ UTR sequences occupy on average about 1.25 kb, as shown in Figure 9. As shown in Figure 10, the 3′ UTRs of these 120 cDNAs have a median length of about 1.05 kb, and more than 80% of them exceed 611 bases in length, which is the median length of an EST sequence. Thus, in most cases the 3′ UTRs are sufficiently long that sequencing from the 3′ end will not reach a region of coding sequence that can be identified *via* BLASTx, and 5′ and 3′ paired reads will rarely overlap to form a contig. Reciprocally, 5′ UTRs are short enough that 5′ end sequencing will usually reach the region of coding sequence. Anticipating this dilemma, we requested more 5′ than 3′

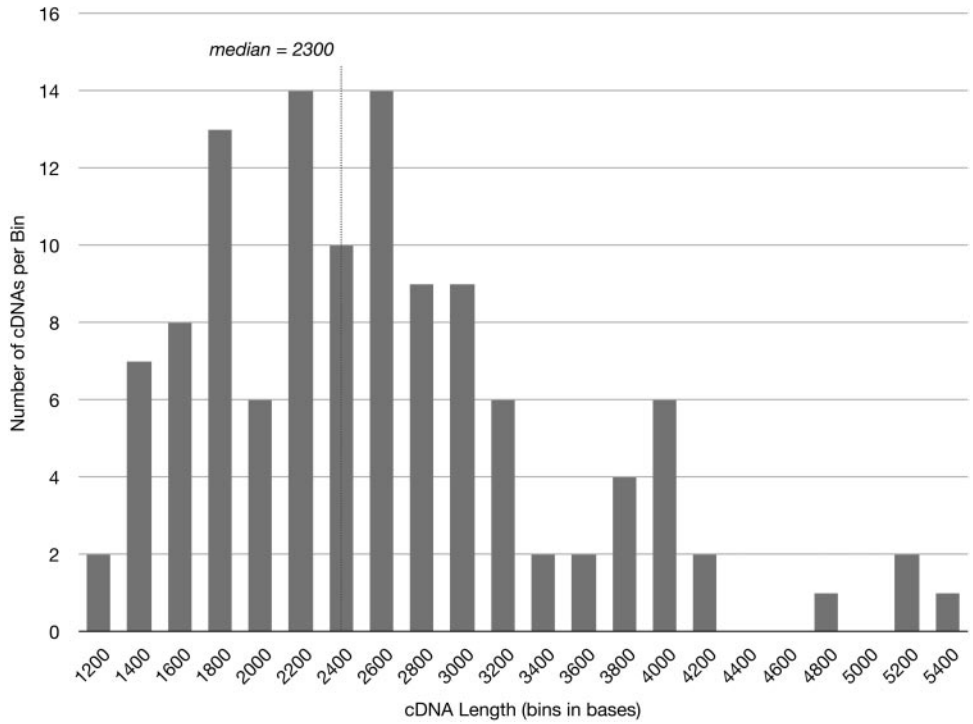

Figure 8. Distribution of lengths (bases) of 120 cloned full-length cDNA inserts chosen for their interest for evolutionary and developmental studies, all derived from the *Saccoglossus kowalevskii* expressed sequence tag (EST) collection, as in Figure 7. Increasing lengths are binned on the horizontal axis in 200-base intervals. The leftmost interval, marked 1200, includes two cDNA with lengths between 1001 and 1200 bases. The vertical axis represents the number of cDNAs (totaling 120) falling within each length interval. The median length of cDNAs is approximately 2300 bases.

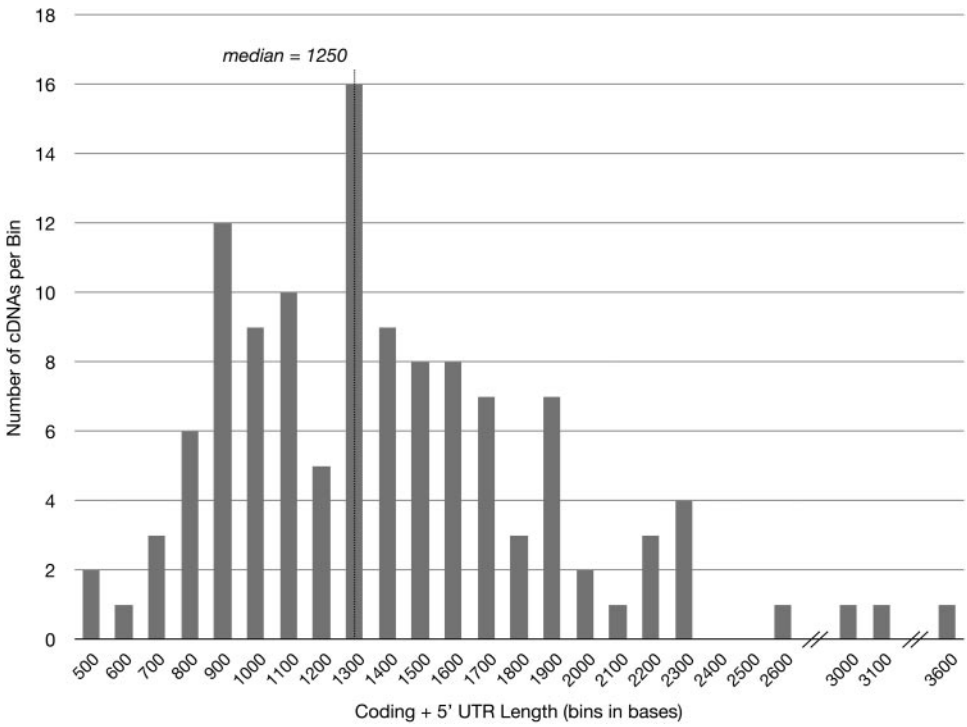

Figure 9. Distribution of lengths of coding sequences plus 5′ untranslated regions (UTRs) of 120 cloned full-length cDNA inserts of interest for evolutionary developmental studies. The cDNAs were derived from the *Saccoglossus kowalevskii* expressed sequence tag (EST) collection, as in Figures 7 and 8. Lengths in bases are binned on the horizontal axis in intervals of 100 bases. For example, two cDNAs have 5′ UTR plus coding sequence of 401–500 bases in length, as shown in the leftmost interval marked 500. The vertical axis shows the number of cDNAs having a length of 5′ UTR plus coding sequence falling in each interval bin.

sequencing early in the project in order to identify more insert sequences; the bias was decreased later. The length of poly-A tracts in the 120 cDNAs varies from 0 to 113 A residues (data not shown), with the majority (76/120) having 15–17 residues in the sequenced poly-A tails. Ten cases (8%) had more than 50 A residues in the poly-A tract, a feature perhaps of importance for the synthesis of antisense strand probes (see Lowe *et al.*, 2003, 2006) since these would contain equally long runs of modified U residues.

Finally, we estimated the prevalence of sequence polymorphism in *S. kowalevskii* by examining the 3′ UTR regions of 10 cDNA sequences having enough 3′ ESTs to cover the region to at least 10-fold depth. Although only preliminary, the estimate was of interest because polymorphism of several percent of bases can be a source of difficulties not only for the design of antisense reagents but also for the assembly of whole-genome shotgun reads. In several broadcast-spawning marine animals, polymorphism reaches such levels: ascidians, >5.9% (Small *et al*., 2007); amphioxus, 6.8% (Putnam *et al*., 2008); and sea urchin, 4%–5% (Sodergren *et al*., 2006). Our ESTs are candidates for revealing polymorphisms because they are derived from different libraries of embryos from different *S. kowalevskii* adults collected at different times, though all adults came from the same salt pond near Woods Hole, Massachusetts. With the 10-fold coverage of 3′ UTRs with ESTs, we defined polymorphisms, as opposed to sequencing errors, by their identity in at least 2 ESTs of 10. In all, we surveyed 11,331 bases of 3′ UTR regions in these 10 cDNAs and found a polymorphism frequency of 0.42% (48 polymorphic sites, each with at least 2 ESTs of 10 having an alternative sequence)—that is, about 1 base in 238 (Table 3). Individual 3′ UTRs of these cDNAs varied from 0% polymorphism (for *dlx* and *dan*) to 2.1% (for *muscle lim*), and the frequency of the minority sequence, when present, varied from 17% (2 of 12 ESTs for *foxa2*) to 50% (for *ets* and *otx*). The *muscle lim* 3′ UTR, which contained 16 differences in 776 bases, included both single nucleotide substitutions and small indels (G-A-XXXX-T-C-A-X-T-T-T-T-TC *versus* T-G-AAAC-C-T-G-A-C-C-C-C-CT, where X signifies a missing base). These combined results indicate that polymorphism, though present in *S. kowalevskii*, may be comparatively low in the local population of this direct-developing estuarine organism, more in the range of *Nematostella* (0.65%, Putnam *et al*., 2007) or pufferfish (0.4%, Aparicio *et al*., 2002). From this sampling, we expect fewer

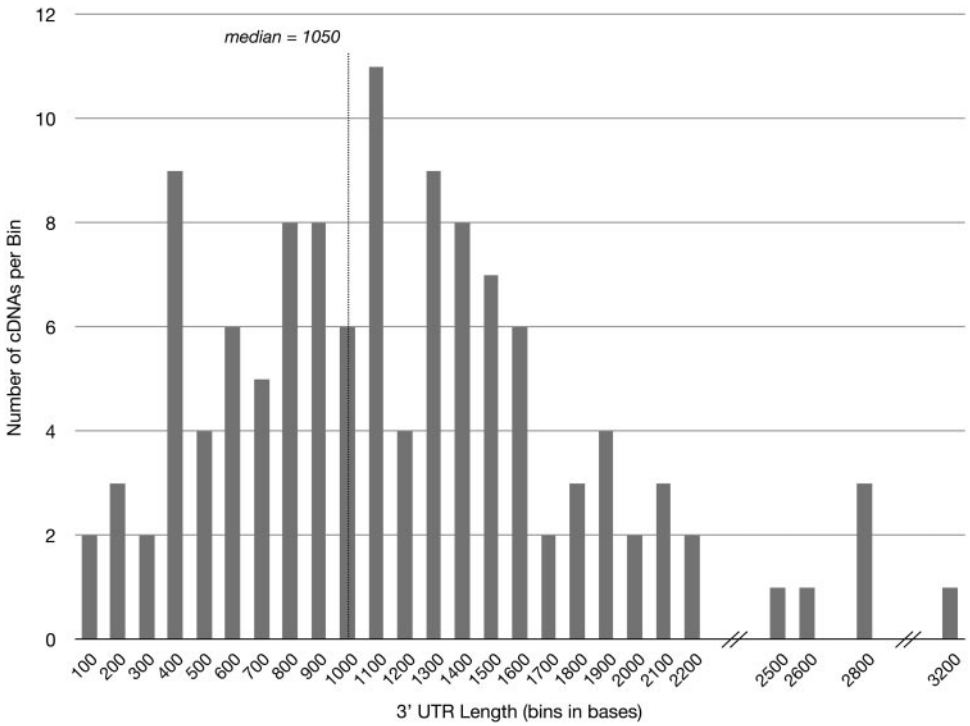

Figure 10. Distribution of the lengths of 3′ untranslated regions (UTRs) of 120 cloned cDNA inserts of interest for evolutionary developmental studies. The cDNAs were derived from the *Saccoglossus kowalevskii* expressed sequence tag collection (EST), as in Figures 7, 8, and 9. Lengths in bases are binned on the horizontal axis in intervals of 100 bases. For example, two cDNAs have 3′ UTRs of 1–100 bases in length, as shown in the leftmost interval marked 100. The vertical axis shows the number of cDNAs having a length of 3′ UTR sequence falling in each interval bin. About 80% of the cDNAs have 3′ UTRs longer than 611 bases, the median length of the sequence reads.

issues with antisense reagent design and genome assembly than were encountered for the broadcast-spawning marine animals cited above. As a caveat, though, although the results are the most incisive we can obtain with ESTs, they may not reflect the status of most parts of the S. *kowalevskii* genome because 3′ UTR sequences do contain conserved

Table 3

Polymorphism levels in 3′ untranslated regions (UTRs) of 10 Saccoglossus kowalevskii *cDNAs*

cDNA	3′ UTR length, bases	ESTs covering 3′ UTR	Number of bases different	Frequency of minority sequence among ESTs	% polymorphism
dan	726	20	0	0/10–0/17	0
dlx	1530	27	0	0/10–0/14	0
ets	1057	25	13	7/14–7/16	1.23
foxa2	939	32	1	2/12	0.11
foxc	1390	22	1	5/10	0.07
ftz-f1	1524	30	8	3/12–5/14	0.52
muscle lim	776	20	16	3/10–3/19	2.06
otp	778	24	2	2/12–2/22	0.13
otx	1428	33	6	5/12–10/20	0.42
sox2/3	1183	26	1	2/10–4/16	0.08
Total	11331		48		0.42

Ten *S. kowalevskii* cDNAs were chosen because they had 10 or more expressed sequence tags (ESTs) covering all regions of the 3′ UTR. The ESTs were aligned and inspected for base sequence alternatives. A polymorphism was defined as an alternate base at a site in two or more ESTs, and the frequency of the minority sequence is given in the table as a fraction of the total ESTs at that site. Single nucleotide differences as well as all positions of indels were counted. The percent polymorphism is defined as the percent of base positions of the 3′ UTR having alternative bases or no base.

subregions; polymporphism in other genome regions may be higher. Furthermore, we have not compared the sequences of this Woods Hole *S. kowalevskii* population with sequences of other populations known to occur on the Atlantic coast from Massachusetts to Florida (Cameron, 2005).

Discussion and Conclusions

Within this collection of 202,175 ESTs, we have found most (>80%) of the gene sequences that we sought for developmental studies, and for most of these sequences (>85%) we have found full-length inserts directly among the 163,633 arrayed clones. In cases where inserts were too short for experimental use, we were able to use exact primers deduced from their ESTs to isolate longer clones. As a small group of researchers, this approach was an efficient way for us to initiate comparative studies of a new animal, whereas the isolation and identification of genes one at a time by degenerate primers and RACE would have been prohibitively laborious.

We can estimate roughly how complete our recovery would have been if we had obtained a smaller EST collection of, say, only 100,000 or 50,000 members, and hence proportionately fewer arrayed clones. The cases about which we know most regarding the frequency and distribution of ESTs are from the set of 120 full- or nearly full-length cDNA sequences described in the Results. These 120 were picked at random from the 400 cDNAs we have thus far chosen from the BLASTx identifications of contigs and singletons on the basis of our interest in them for comparative developmental studies. We have no evidence that they differ in recoverability from any other kinds of cDNA sequences in the collection. Of these 120, 18 (15%) are represented by only two ESTs or one EST. The chance is significant that any of these would be absent from a collection of 100,000 ESTs (*e.g.,* 37% and 61% for the Poisson *P*(0) "empty" class, respectively, for cDNAs with only two or one EST among 200,000), and the chance of their absence is still greater from a collection of 50,000 ESTs (*e.g.,* 0.61 and 0.78 for the Poisson *P*(0) empty class, repectively). Furthermore, 31 cDNAs (26%) of the 120 are represented by five or fewer ESTs; the chance is also significant that some of these would be missing from a collection of 50,000 ESTs (29% for the Poisson *P*(0) class for any one of the cDNAs with five ESTs in the collection of 200,000). On the other hand, 67 (58%) of the 120 cDNA sequences are matched by 10 or more ESTs, and most would likely be recovered even in a collection of 50,000 ESTs. Thus, we estimate that one-half to two-thirds of the sequences would have been obtained even if we had collected only 50,000 ESTs. As an argument in favor of collecting more than 200,000 ESTs, we were still obtaining new EST sequences as our collection approached this number, and some desired genes are still missing—namely, *wnt6, wnt10, fz4, robo, evx, pax3/7, lox, gsx,* and *lefty*—although these have subsequently been isolated by other means. In other cases, only small 3′ fragments of desired genes were found among the ESTs (for example, *notch* and *delta*). Although these last-mentioned sequences are present in the libraries as long clone inserts, the clones themselves are slow growing. As a general conclusion, though, our collection of more than 200,000 ESTs, representing more than 150,000 arrayed clones, has provided an extensive repertoire of developmental gene sequences for comparative studies.

In addition, we can evaluate the benefits of using normalized *versus* non-normalized libraries by surveying our set of 120 full-length or nearly full-length cDNAs. For reasons we do not understand, about a third of the sequences are matched by ESTs primarily, or only, from non-normalized libraries, despite the fact that 1.9-fold more ESTs were obtained from normalized libraries, and despite the expected enrichment of rare sequences in normalized libraries. Furthermore, the 5′-most ESTs of these 120 cDNAs are more frequently (1.7 times) derived from non-normalized libraries. At the same time, about one-third of the sequences show a disproportionate EST representation from the normalized libraries (*e.g.*, 29:1 for *tcf21*, 18:1 for *nato3B*, 14:1 for *dan*, and 11:1 for *dkk1/2*), and 10% of the cDNA sequences have ESTs exclusively from the normalized libraries. These last-mentioned cases presumably exemplify the benefits of normalization for some kinds of sequences. In light of the varied recoveries of different sequences, we conclude that it is beneficial to represent both normalized and non-normalized libraries in the generation of ESTs. In addition, to increase the success of annotation *via* BLAST, it is useful to emphasize 5′ reads when the collection is small. Later in the project, as the number of ESTs increases, 5′-3′ paired reads can be added to facilitate the assembly of putative mRNAs and to generate 3′ reads for genomic usage. The absence of a gene sequence from the EST collection presumably reflects not just the limited size of the collection and the low abundance of the mRNA in embryos but also sequence-specific factors such as poor cDNA synthesis and/or the poor growth of bacteria carrying a given insert.

Acknowledgments

This research was supported by NASA grant FDNAG2-1605 to JG and MK and by USPHS grants HD37277 to MK and HD42724 to JG. We thank Dr. Steven Klein for his support of the project. C. Lowe was supported in part by a Miller Institute fellowship and a Colwin Fellowship (Woods Hole). We thank the staff of the Marine Biology Laboratory, Woods Hole, Massachusetts, for their assistance during our annual September collection of embryos, and Nik Putnam

and Daniel Rokhsar for sharing data on amphioxus in advance of publication.

Literature Cited

Aparicio, S., J. Chapman, E. Stupka, N. Putnam, J.-M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, *et al.* 2002. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* **297:** 1301–1310.

Arnone, M. I., F. Rizzo, R. Annunciata, R. A. Cameron, K. J. Peterson, and P. Martínez. 2006. Genetic organization and embryonic expression of the ParaHox genes in the sea urchin *S. purpuratus*: insights into the relationship between clustering and colinearity. *Dev. Biol.* **300:** 63–73.

Aronowicz, J., and C. J. Lowe. 2006. Hox gene expression in the hemichordate *S. kowalevskii* and the evolution of deuterostome nervous systems. *Integr. Comp. Biol.* **46:** 890–901.

Balser, E. J., and E. E. Ruppert. 1990. Structure, ultrastructure, and function of the preoral heart-kidney of *Saccoglossus kowalevskii* (Hemichordata, Enteropneusta) including new data on the stomochord. *Acta Zool.* **71:** 235–249.

Barolo, S., and J. W. Posakony. 2002. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev.* **16:** 1167–1181.

Bateson, W. 1884. The early stages in the development of *Balanoglossus* (sp. Incert.). *Q. J. Microsc. Sci.* **24:** 208–236.

Bateson, W. 1885. The later stages in the development of *Balanoglossus kowalevskii*, with a suggestion as to the affinities of the enteropneusta. *Q. J. Microsc. Sci.* **25:** 81–128.

Bateson, W. 1886. Continued account of the later stages in the development of *Balanoglossus kowalevskii*, and of the morphology of the enteropneusta. *Q. J. Microsc. Sci.* **26:** 511–534.

Bourlat, S. J., T. Juliusdottir, C. J. Lowe, R. Freeman, J. Aronowicz, M. Kirschner, E. S. Lander, M. Thorndyke, H. Nakano, A. B. Kohn, *et al.* 2006. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444:** 85–88.

Bromham, L. D., and B. M. Degnan. 1999. Hemichordates and deuterostome evolution: robust molecular phylogenetic support for a hemichordate + echinoderm clade. *Evol. Dev.* **1:** 166–171.

Bullock T. H. 1965. The nervous system of hemichordates. Pp.1567–1577 in *Structure and Function in the Nervous Systems of Invertebrates*. W. H. Freeman, San Francisco.

Burdon-Jones, C. 1952. Development and biology of the larva of *Saccoglossus horsti* (entropneusta). *Proc. R. Soc. Lond. B Biol. Sci.* **236:** 553–589.

Cameron, C. B. 2005. A phylogeny of the hemichordates based on morphological characters. *Can. J. Zool.* **83:** 196–215.

Cameron, C. B., J. R. Garey, and B. J. Swalla. 2000. Evolution of the chordate body plan: new insights from phylogenetic analyses of deuterostome phyla. *Proc. Natl. Acad. Sci. USA* **97:** 4469–4474.

Castresana, J., G. Feldmaier-Fuchs, S.-I. Yokobori, N. Satoh, and S. Pääbo. 1998. The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. *Genetics* **150:** 1115–1123.

Colwin A. L., and L. H. Colwin. 1950. The developmental capacities of separated early blastomeres of an enteropneust, *Saccoglossus kowalevskii*. *J. Exp. Zool.* **115:** 263–296.

Colwin A. L., and L. H. Colwin. 1953. The normal embryology of *Saccoglossus kowalevskii*. *J. Morphol.* **92:** 401–453.

Colwin L. H.., and A. L. Colwin. 1962. Induction of spawning in *Saccoglossus kowalevskii* (Enteropneusta) at Woods Hole. *Biol. Bull.* **123:** 493.

Croce, J. C., S.-Y. Wu, C. Byrum, R. Xu, L. Duloquin, A. H. Wikramanayake, C. Gache, and D. R. McClay. 2006. A genome-wide survey of the evolutionarily conserved Wnt pathways in the sea urchin *Strongylocentrotus purpuratus*. *Dev. Biol.* **300:** 121–131.

Davidson, E., R. Hynes, and D. McClay, eds. 2006. Special issue: Sea urchin genome: implications and insights. *Dev. Biol.* **300(1):** 1–496.

De Robertis, E. M. 2006. Spemann's organizer and self-regulation in amphibian embryos. *Nat. Rev. Mol. Cell Biol.* **7:** 296–302

De Robertis, E. M., and Y. Sasai. 1996. A common plan for dorsoventral patterning in bilateria. *Nature* **380:** 37–40.

Ewing B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8:** 186–194.

Ewing B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8:** 175–185.

Furlong, R. F., and P. W. Holland. 2002. Bayesian phylogenetic analysis supports monophyly of ambulacraria and of cyclostomes. *Zool. Sci.* **19:** 593–599.

Gerhart, J. 1998. Signaling pathways in development. *Teratology* **60:** 226–239.

Gerhart, J., C. Lowe, and M. Kirschner. 2005. Hemichordates and the origin of chordates. *Curr. Opin. Genet. Dev.* **15:** 461–467.

Gilchrist, M. J., A. M. Zorn, J. Voigt, J. C. Smith, N. Papalopulu, and E. Amaya. 2004. Defining a large set of full-length clones from a *Xenopus tropicalis* EST project. *Dev. Biol.* **271:** 498–516.

Goodrich, E. S. 1917. 'Proboscis pores' in craniate vertebrates, a suggestion concerning the premandibular somites and hypophysis. *Q. J. Microsc. Sci.* **62:** 539–553.

Gyoja, F., Y. Satou, T. Shin-i, Y. Kohara, B. J. Swalla, and N. Satoh. 2007. Analysis of large scale expression sequenced tags (ESTs) from the anural ascidian, *Molgula tectiformis*. *Dev. Biol.* **307:** 460–482.

Heber, S., M. Alekseyev, S. H. Sze, H. Tang, and P. A. Pevzner. 2002. Splicing graphs and EST assembly problem. *Bioinformatics* **18 Suppl. 1:** S181-188.

Howard-Ashby, M., S. C. Materna, C. T. Brown, L. Chen, R. A. Cameron, and E. H. Davidson. 2006a. Identification and characterization of homeobox transcription factor genes in *S. purpuratus*, and their expression in embryonic development. *Dev. Biol.* **300:** 74–89.

Howard-Ashby, M., S. C. Materna, C. T. Brown, L. Chen, R. A. Cameron, and E. H. Davidson. 2006b. Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. *Dev. Biol.* **300:** 90–107.

Hsia, C., and W. McGinnis. 2003. Evolution of transcription factor function. *Curr. Opin. Genet. Dev.* **13:** 199–206.

Huang, X., and A. Madan. 1999. CAP3: a DNA sequence assembly program. *Genome Res.* **9:** 868–877.

Lander, E. S., L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.* 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Lapraz, F. E., V. Röttinger, R. Duboc, L. Range, K. Duloquin, S.-Y. Walton, C. Wub, M. Bradham, A. Loza, T. Hibino, K. Wilson, *et al.* 2006. RTK and TGF-beta signaling pathways genes in the sea urchin genome. *Dev. Biol.* **300:** 132–152.

Lemons, D., and W. McGinnis. 2006. Genomic evolution of Hox gene clusters. *Science* **313:** 1918–1922.

Liang C., F. Sun, H. Wang, J. Qu, R. M. Freeman, Jr., L. H. Pratt, and M. M. Cordonnier-Pratt. 2006. MAGIC-SPP: a database-driven DNA sequence processing package with associated management tools. *BMC Bioinformatics* **7:** 115.

Logan, C. Y., and R. Nusse. 2004. The Wnt signaling pathway in development and disease. *Annu. Rev. Cell Dev. Biol.* **20:** 781–810.

Lowe, C. J., M. Wu, A. Salic, L. Evans, E. Lander, N. Stange-Thomann, C. E. Gruber, J. Gerhart, and M. Kirschner. 2003. Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* **113:** 853–865.

Lowe, C. J., K. Tagawa, T. Humpreys, M. Kirschner, and J. Gerhart. 2004. Hemichordate embryos: procurement, culture, and basic methods. *Methods Cell Biol.* **74:** 171–194.

Lowe C. J., M. Terasaki, M. Wu, R. M. Freeman, Jr., L. Runft, K. Kwan, S. Haigo, J. Aronowicz, E. Lander, C. Gruber, *et al.* 2006. Dorsoventral patterning in hemichordates: insights into early chordate evolution. *PLoS Biol.* **4:** e291

Maidak, B. L., J. R. Cole, T. G. Lilburn, C. T. Parker, Jr., P. R. Saxman, R. J. Farris, G. M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. 2001. The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.* **29:** 173–174.

Massagué, J. 2000. How cells read TGF signals. *Nat. Rev. Mol. Cell Biol.* **1:** 169–178.

Materna, S. C., M. Howard-Ashby, R. F. Gray, and E. H. Davidson. 2006. The C_2H_2 zinc finger genes of *Strongylocentrotus purpuratus* and their expression in embryonic development. *Dev. Biol.* **300:** 108–120.

Ning, Z., A. J. Cox, and J. C. Mullikin. 2001. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11:** 1725–1729.

Ogasawara, M., H. Wada, H. Peters, and N. Satoh. 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.

Pertea, G., X. Huang, F. Liang, V. Antonescu, R. Sultana, S. Karamycheva, Y. Lee, J. White, F. Cheung, B. Parvizi, *et al.* 2003. TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics* **22:** 651–652.

Peterson, K. J., J. B. Lyons, K. S. Nowak, C. M. Takacs, M. T. Wargo, and M. A. McPeek. 2004. Estimating metazoan divergence times from a molecular clock. *Proc. Natl. Acad. Sci. USA*: **101:** 6536–6541.

Peterson, K. J., R. A. Cameron, K. Tagawa, N. Satoh, and E. H. Davidson. 1999. A comparative molecular approach to mesodermal patterning in basal deuterostomes: the expression pattern of *Brachyury* in the enteropneust *Ptychodera flava*. *Development* **126:** 85–95.

Pires-daSilva, A., and R. J. Sommer. 2003. The evolution of signaling pathways in animal development. *Nat. Rev. Genet.* **4:** 39–49.

Putnam, N. H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, *et al.* 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317:** 86–94.

Putnam, N. H., U. Hellsten, J. K. Sky-Yu, L. Pennachio, M. Blow, *et al.* 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* (In press).

Ruppert, E. E. 2005. Key characters uniting hemichordates and chordates: homologies or homoplasies. *Can. J. Zool.* **83:** 8–23.

Rychel, A. L. and B. J. Swalla. 2007. Development and evolution of chordate cartilage. *J. Exp. Zool. B Mol. Dev. Evol.* **308B:** 325–335.

Rychel, A. L., S. E. Smith, H. T. Shimamoto, and B. J. Swalla. 2006. Evolution and development of the chordates: collagen and pharyngeal cartilage. *Mol. Biol. Evol.* **23:** 541–549.

Samanta, M. P., W. Tongprasit, S. Istrail, R. A. Cameron, Q. Tu, E. H. Davidson, and V. Stolc. 2006. The transcriptome of the sea urchin embryo. *Science* **314:** 960–965.

Scott M. P. 1994. Intimations of a creature. *Cell* **79:** 1121–1124.

Scott M. P. 2000. Development: the natural history of genes. *Cell* **100:** 27–40.

Sekiguchi, T., T. Kawashima, Y. Satou, and N. Satoh. 2007. Further EST analysis of endocrine genes that are preferentially expressed in the neural complex of *Ciona intestinalis*: receptor and enzyme genes associated with endocrine system in the neural complex. *Gen. Comp. Endocrinol.* **150:** 233–245.

Small, K. S., M. Brudno, M. M. Hill, and A. Sidow. 2007. Extreme genomic variation in a natural population. *Proc. Natl. Acad. Sci. USA* **104:** 5698–5703.

Smith, A. B., K. J. Peterson, G. Wray, and D. T. J. Littlewood. 2004. From bilateral symmetry to pentaradiality: the phylogeny of hemichordates and echinoderms. Pp. 365–383 in *Assembling the Tree of Life,* J. Cracraft and M. J. Donoghue, eds. Oxford University Press, New York.

Sodergren, E., G. M. Weinstock, E. H. Davidson, R. A. Cameron, R. A. Gibbs, *et al.* 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314:** 941–952.

Strausberg, R. L., E. A. Feingold, R. D. Klausner, and F. S. Collins. 1999. The mammalian gene collection. *Science* **286:** 455–457.

Swalla, B. J. 2006. Building divergent body plans with similar genetic pathways. *Heredity* **97:** 235–243.

Tu, Q., C. T. Brown, E. H. Davidson, and P. Oliveri. 2006. Sea urchin forkhead gene family: phylogeny and embryonic expression. *Dev. Biol.* **300:** 49–62.

Veraksa, A., M. Del Campo, and W. McGinnis. 2000. Developmental patterning genes and their conserved functions: from model organism to humans. *Mol. Genet. Metab.* **69:** 85–100.

Walton, K. D., J. C. Croce, T. D. Glenn, S.-Y. Wu, and D. R. McClay. 2006. Genomics and expression profiles of the Hedgehog and Notch signaling pathways in sea urchin development. *Dev. Biol.* **300:** 153–164.

Wang J. P., B. G. Lindsay, J. Leebens-Mack, L. Cui, K. Wall, W. C. Miller, and C. W. dePamphilis. 2004. EST clustering error evaluation and correction. *Bioinformatics* **20:** 2973–2984.

Weinstein, D. C., and A. Hemmati-Brivanlou. 1999. Neural induction. *Annu. Rev. Cell Dev. Biol.* **15:** 411–433.

Wilson, S. W., and C. Houart. 2004. Early steps in the development of the forebrain. *Dev. Cell* **6:** 167–181.

Zhu, X., G. Mahairas, M. Illies, R. A. Cameron, E. H. Davidson, and C. A. Ettensohn. 2001. A large-scale analysis of mRNAs expressed by primary mesenchyme cells of the sea urchin embryo. *Development* **128:** 2615–2627.