The draft genome of Ciona intestinalis: Insights into Chordate and Vertebrate Origins

Paramvir Dehal, Yutaka Satou, Robert Campbell, Jarrod Chapman, Bernard Degnan, Anthony de Tomaso, Brad Davidson, Anna Di Gregorio, Maarten Gelpke, David Goodstein, et al.

The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins

Paramvir Dehal,[1]* Yutaka Satou,[2]* Robert K. Campbell,[3,4] Jarrod Chapman,[1] Bernard Degnan,[5] Anthony De Tomaso,[6] Brad Davidson,[7] Anna Di Gregorio,[7] Maarten Gelpke,[1] David M. Goodstein,[1] Naoe Harafuji,[7] Kenneth E. M. Hastings,[8] Isaac Ho,[1] Kohji Hotta,[9] Wayne Huang,[1] Takeshi Kawashima,[10] Patrick Lemaire,[11] Diego Martinez,[1] Ian A. Meinertzhagen,[12] Simona Necula,[1] Masaru Nonaka,[13] Nik Putnam,[1] Sam Rash,[1] Hidetoshi Saiga,[14] Masanobu Satake,[15] Astrid Terry,[1] Lixy Yamada,[2] Hong-Gang Wang,[16] Satoko Awazu,[2] Kaoru Azumi,[17] Jeffrey Boore,[1] Margherita Branno,[18] Stephen Chin-bow,[19] Rosaria DeSantis,[18] Sharon Doyle,[1] Pilar Francino,[1] David N. Keys,[1,7] Shinobu Haga,[9] Hiroko Hayashi,[9] Kyosuke Hino,[2] Kaoru S. Imai,[2] Kazuo Inaba,[20] Shungo Kano,[2,18] Kenji Kobayashi,[2] Mari Kobayashi,[2] Byung-In Lee,[1] Kazuhiro W. Makabe,[2] Chitra Manohar,[1] Giorgio Matassi,[18] Monica Medina,[1] Yasuaki Mochizuki,[2] Steve Mount,[21] Tomomi Morishita,[9] Sachiko Miura,[9] Akie Nakayama,[2] Satoko Nishizaka,[9] Hisayo Nomoto,[9] Fumiko Ohta,[9] Kazuko Oishi,[9] Isidore Rigoutsos,[19] Masako Sano,[9] Akane Sasaki,[2] Yasunori Sasakura,[2] Eiichi Shoguchi,[2] Tadasu Shin-i,[9] Antoinetta Spagnuolo,[18] Didier Stainier,[22] Miho M. Suzuki,[23] Olivier Tassy,[11] Naohito Takatori,[2] Miki Tokuoka,[2] Kasumi Yagi,[2] Fumiko Yoshizaki,[13] Shuichi Wada,[2] Cindy Zhang,[1] P. Douglas Hyatt,[24] Frank Larimer,[24] Chris Detter,[1] Norman Doggett,[25] Tijana Glavina,[1] Trevor Hawkins,[1] Paul Richardson,[1] Susan Lucas,[1] Yuji Kohara,[9]† Michael Levine,[7,26]† Nori Satoh,[2]† Daniel S. Rokhsar[1,7,26]†

The first chordates appear in the fossil record at the time of the Cambrian explosion, nearly 550 million years ago. The modern ascidian tadpole represents a plausible approximation to these ancestral chordates. To illuminate the origins of chordate and vertebrates, we generated a draft of the protein-coding portion of the genome of the most studied ascidian, *Ciona intestinalis*. The *Ciona* genome contains □16,000 protein-coding genes, similar to the number in other invertebrates, but only half that found in vertebrates. Vertebrate gene families are typically found in simplified form in *Ciona*, suggesting that ascidians contain the basic ancestral complement of genes involved in cell signaling and devel-opment. The ascidian genome has also acquired a number of lineage-specific innovations, including a group of genes engaged in cellulose metabolism that are related to those in bacteria and fungi.

Introduction

There has been considerable debate regarding the evolutionary origins of the chordates and, within them, of vertebrates, and the relationships between the first chordates and invertebrate deuterostomes (*1–3*). There is a consensus that the three major deuterostome phyla (echinoderms, hemichordates, and chordates) arose from a common ancestor more than 550 million years ago. The chordates subsequently diverged into three subphyla: urochordates (also known as tunicates), cephalochordates, and vertebrates (Fig. 1). It is generally believed that cephalochordates and vertebrates are sister groups, with the urochordate lineage most basal among the chordates.

The most prevalent modern urochordates are the ascidians, known familiarly as sea squirts. The analysis of ascidians offers the simultaneous prospects of providing insights into two compelling problems in evolutionary biology: the origins of the chordates from an ancestral deuterostome, and the origins of the vertebrates from a simple chordate.

Ascidians, or sea squirts, are sessile, hermaphroditic marine invertebrates that live in shallow ocean waters around the world. The adults are simple filter feeders encased in a fibrous tunic supporting incurrent and outcurrent siphons, and were classified by Aristotle as related to mollusks (Fig. 2A). The close kinship of ascidians with vertebrates was discovered over 130 years ago by Kowalevsky (*4*), who recognized that the ascidian larva has the general appearance of a vertebrate tadpole (Fig. 2L). Most notably, the tail of the ascidian tadpole contains a prominent notochord and a dorsal tubular nerve cord. These findings elevated ascidians to the ranks of the

[1]U.S. Department of Energy Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA. [2]Department of Zoology, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan. [3]Marine Biological Laboratories, Woods Hole, MA 02543, USA. [4]Serono Reproductive Biology Institute, One Technology Place, Rockland, MA 02370, USA. [5]Department of Zoology and Entomology, University of Queensland, Brisbane, Qld 4072, Australia. [6]Department of Pathology, Stanford University School of Medicine, Stanford, CA 93950, USA. [7]Department of Molecular and Cellular Biology, Division of Genetics, 401 Barker Hall, University of California, Berkeley, CA 94720, USA. [8]Montreal Neurological Institute, McGill University, Montreal, H3A 2T5, Canada. [9]National Institute of Genetics, Mishima 411-8540, Japan. [10]Bioinformatics Center, Institute for Chemical Research, Kyoto Univeristy, Uji, Kyoto 611-0011, Japan. [11]LGPD, IBDM, Case 907, Campus de Luminy, F-13288 Marseille Cedex 09, France. [12]Life Sciences Centre, Dalhousie University, Halifax, Nova Scotia, B3H 4J1, Canada. [13]Department of Biological Sciences, Graduate School of Science, University of Tokyo, Tokyo 113-0033, Japan. [14]Department of Biological Sciences, Graduate School of Science, Tokyo Metropolitan University, Hachiohji, Tokyo 192-0397, Japan. [15]Department of Molecular Immunology, Institute of Development, Aging and Cancer, Tohoku University, Seiryo-machi, Aoba-ku, Sendai 980-8575, Japan. [16]Drug Discovery Program, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA. [17]Department of Biochemistry, Graduate School of Pharmaceutical Sciences, Hokkaido University, Sapporo 060-0812, Japan. [18]Stazione Zoologica "Anton Dohrn," Villa Comunale, 80121 Napoli, Italy. [19]IBM Watson Research Center, Post Office Box 218, Yorktown Heights, NY 10598, USA. [20]Asamushi Marine Biological Station, Graduate School of Science, Tohoku University, Asamushi, Aomori 039-3501, Japan. [21]University of Maryland, College Park, MD 20742, USA. [22]Department of Biochemistry and Biophysics, Programs in Developmental Biology, Genetics, and Human Genetics, University of California, San Francisco, CA 94143, USA. [23]Wellcome Trust Centre for Cell Biology, Institute of Cell and Molecular Biology, University of Edinburgh, Edinburgh EH9 3JR, UK. [24]Oak Ridge National Laboratory, Post Office Box 2008, Oak Ridge, TN 37831, USA. [25]Los Alamos National Laboratory, Los Alamos, NM 87545, USA. [26]Center for Integrative Genomics, University of California, Berkeley 94720, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: ykohara@lab.nig.ac.jp (Y.K.), mlevine@uclink4.berkeley.edu (M.L.), satoh@ascidian.zool.kyoto-u.ac.jp (N.S.), dsrokhsar@lbl.gov (D.R.)

chordates and convinced Darwin that the ascidian tadpole is a legitimate relative of vertebrates, including humans (*5*).

In addition to their unique evolutionary position as invertebrate chordates, ascidians provide a simple experimental system to investigate the molecular mechanisms underlying cell-fate specification during chordate development (*6*–*10*). First, the ascidian tadpole is composed of only about 2500 cells, and there is extensive information on the cell lineage of most major tissues and organs (*11*, *12*). Second, the blastomeres of early embryos are large and easy to manipulate; they permit the detailed visualization of differential gene expression during development (*6*–*10*). Third, embryogenesis is rapid (~18 hours from fertilization to a free-swimming tadpole at 18°C) (Fig. 2, B to L) and the entire life cycle takes less than 3 months, thereby facilitating genetic analyses (*13*, *14*). Finally, it is possible to introduce transgenic DNAs into developing embryos by using simple electroporation methods (*15*, *16*) that permit the simultaneous transformation of hundreds or even thousands of synchronously developing embryos. The method has been used to characterize cis-regulatory DNAs (*15*, *16*), and produce mutant phenotypes by misexpressing or overexpressing a variety of regulatory genes that encode transcription factors (*17*) or signaling molecules (*18*).

The genome of the ascidian *Ciona intestinalis* is among the smallest of any experimentally accessible chordate; at an estimated 160 million base pairs (*19*), it is nearly 20 times smaller than the human genome. Among the chordates, only larvaceans (another group of urochordates) have a smaller genome (*20*), but they are not easily manipulated using modern molecular methods. It has been proposed that large-scale gene duplications occurred in the vertebrate lineage after its divergence from cephalochordates and urochordates (*21*). The ensuing increase in gene number seems to underlie some of the novel and increasingly complex developmental processes seen in vertebrates. These expansions suggest that the more basal urochordates and cephalochordates contain an approximation to the "ancestral complement" of nonduplicated chordate genes. Here we report the draft sequence of the protein-encoding portion of the *C. intestinalis* genome and compare it with the gene complement of other animals to gain insight into the evolutionary origins of chordates and vertebrates.

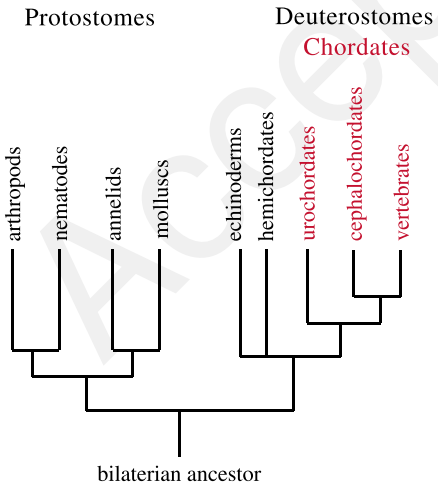

Fig. 1. Phylogeny of bilaterian animals (*1*–*3*). *Ciona intestinalis* is a member of the urochordates, the most primitively branching clade of chordates. Chordates in turn are deuterostomes, one of two great divisions (along with protostomes) of bilaterian animals.

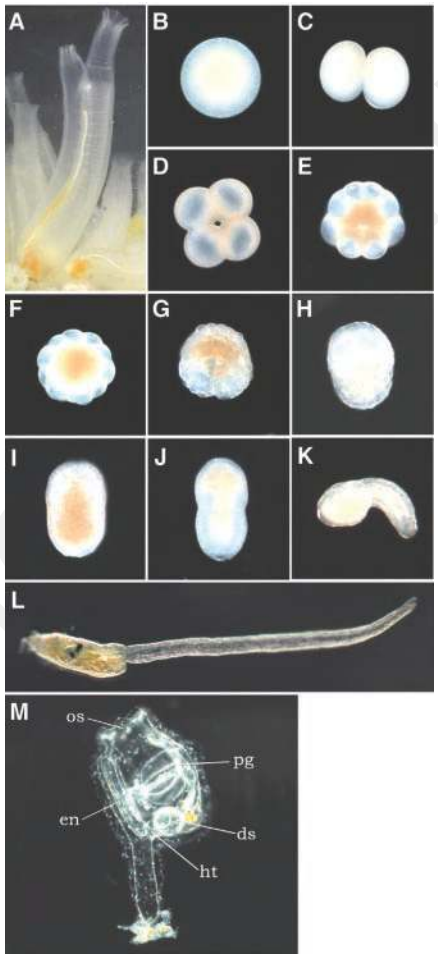

Fig. 2. The sea squirt *Ciona intestinalis*. (**A**) Adults with incurrent and outcurrent siphons. The white duct is the sperm duct, while the orange duct paralleling it is the egg duct. (**B** to **L**) Embryogenesis. (B) Fertilized egg, (C) 2-cell embryo, (D) 4-cell embryo, (E) 16-cell embryo, (F) 32-cell embryo, (G) gastrula (about 150 cells), (H) neurula, (I to K) tailbud embryos, and (L) tadpole larva. Embryos were dechorionated to show their outer morphology clearly. (**M**) A juvenile a few days after metamorphosis, showing the internal structures. ds, digestive system; en, endostyle; ht, heart; os, neuronal complex; and pg, pharyngeal gill.

Sequencing and Global Analysis of the *Ciona intestinalis* Genome

To sequence the *Ciona* genome, we used a whole-genome shotgun approach (*22*, *23*). Most of the sequence data in this study was derived from a single individual in Half Moon Bay, California. Sperm was isolated from this individual, and DNA was extracted and sheared to create 3-kb fragments, which were subsequently cloned into plasmids and end-sequenced to cover the genome eightfold (*24*). More than 2 million sequence fragments were obtained. Bacterial artificial chromosome (BAC) and cosmid libraries were end-sequenced to 0.2× genome coverage to provide longer range linking information. The DNA for these libraries was obtained from a pool of Japanese individuals (BAC) and a different Californian (cosmid).

A high level of allelic polymorphism was found in the sequences from the single individual, with 1.2% of the nucleotides differing between alleles—nearly 15-fold more than in humans (*24*), and three-fold more than in *fugu* (*22*, *23*). Allelic polymorphism is easily distinguished from sequencing error, because at eightfold shotgun coverage, each allele is redundantly sampled by multiple shotgun fragments (Fig. 3). Observed variations include single-nucleotide polymorphisms (SNPs), as well as small insertions and deletions. The variation is not uniform, with local peaks as high as 10 to 15% polymorphic sites within a window of 100 nucleotides. The high degree of allelic variation is in part a consequence of the large effective population size that results from the breeding mode of ascidians, which release sperm and eggs into the ocean for external fertilization.

Genome assembly was carried out using the JAZZ suite of assembly tools developed for large whole-genome shotgun projects (*23*–*25*). The output of JAZZ is a set of sequence "contigs" (segments of the genome reconstructed from overlapping shotgun fragments) linked together into sequence "scaffolds" (reconstructed sequence with internal gaps of measurable size) (*22*). For haploid genomes, or diploid genomes with a low rate of polymorphism (like human, or inbred strains of laboratory animals like flies and mice), overlapping shotgun fragments derived from a given region of the genome differ only by random sequencing error. The extensive polymorphisms found in pufferfish (0.4%) (*23*) and *Ciona* (1.2%), however, produce higher levels of sequence mismatch (Fig. 3). To overcome this problem we tolerated imperfect alignments that were consistent with the order and orientation of the paired end sequences derived from each plasmid insert (*23*–*25*).

The assembled sequence is a mosaic of the maternal and paternal haplotypes of the sequenced individual. Using the consensus as a reference, one can extract long stretch-

es of either haplotype (Fig. 3), as will be described elsewhere (*25*). The distinctiveness of the two haplotypes raises the concern that some highly polymorphic regions of the genome could be represented twice in the assembly. To eliminate any residual redundancy, we aligned the initial assembly with itself using BLASTN and removed small sequence scaffolds that contain more than 95% identity with other scaffolds. These identified artifacts typically involved small scaffolds (shorter than 3 kb). For uniformity we removed all scaffolds shorter than 3 kb from the present assembly release, which is deposited in GenBank under project accession number AABS00000000, version AABS01000000. The removed scaffolds are available at www.jgi.doe.gov/ciona. After removing these identified duplications from the assembly, we estimate that up to 2 to 3% sequence redundancy persists. In some cases, short, ~500–base pair (bp) duplications are present on either side of an intrascaffold gap due to allelic variation; this problem has been corrected in subsequent assemblies, which are continually improving with additional data and algorithmic developments. Updated assemblies are available at www.jgi.doe.gov/ciona.

The *C. intestinalis* assembly presented here spans 116.7 Mbp of nonrepetitive sequence in 2501 scaffolds longer than 3 kbp. Sixty Mbp, or about half the assembly, is reconstructed in only 177 scaffolds longer than 190 kbp. More than 85% of the assembled sequence (total of 104.1 Mbp) is found in 905 scaffolds longer than 20 kb, and therefore longer than the span of a typical gene (5 to 10 kbp). These scaffolds include 4264 contigs that account for 100.9 Mbp. There are 3359 captured gaps that total 3.2 Mbp, which account for less than 3% of the net scaffold sequence. Families of high–copy number tandem repeats identified from initial sample sequencing (ribosomal RNAs, clusters of tRNAs, and so forth) were withheld from assembly (*24*) and account for nearly 11% of the raw data, for an additional 16 to 18 Mbp of genome sequence; these are typically problematic to assemble. Another 15 to 17 Mbp of genome sequence (10% of the raw shotgun data) remains unassembled. This shotgun sequence contains additional low-copy transposable elements and repeats, and a relatively small amount of unique sequence. We estimate (*24*) that the *C. intestinalis* genome is ~153 to 159 Mbp in total length, in accord with earlier estimates (*19*). The genome is notably AT rich (65%) compared with the human genome (*26*, *27*). Available mapping data do not yet allow us to place the assembled sequence on *Ciona*'s 14 chromosomes.

We can confirm the completeness of the euchromatic assembly by comparing it with other samplings of *Ciona* sequence. From the current collection of 5647 full-length cDNA sequences from the Kyoto project, 5359 aligned with the assembled sequence using BLAST (*28*, *29*). These results suggest that 95% of the protein-coding genes are contained in the assembly. Similarly, 146 of 151 (96.7%) previously known *Ciona* genes were recovered (*24*). These figures are common for draft whole-genome shotgun projects, because certain segments of the genome cannot be cloned and/or sequenced using present methods, or remain unassembled. A gene missing from the current assembly is probably absent from the genome itself. Two of the 151 known genes appear to be misassembled in the present draft, suggesting that a few hundred genes might be interrupted in this manner.

Annotation of gene content

The genome sequence is complemented by an extensive collection of over 480,000 5′ and 3′ expressed sequence tags (ESTs) and cDNAs produced from a large-scale EST project in Kyoto (http://ghost.zool.kyoto-u.ac.jpindexr1.html; see also supporting online material). These were essential for gene identification and provide a resource for future functional studies. ESTs are available from cDNA libraries derived from 10 different developmental stages or adult tissues, including fertilized egg, early cleaving embryos, gastrula/neurula-stage embryos, tailbud embryos, mature larvae, and whole juvenile adults, as well as adult heart, neural complex, endostyle, and testis (*28*, *29*). A partially redundant set of 5647 cDNA clones have been sequenced by directed walking.

To model genes in the *Ciona* genome, the assembled draft sequence was translated in all reading frames and compared with that of all proteins in GenBank using ungapped BLASTX (*30*) with default parameters. Best hits at each locus were selected using a parsing scheme that finds collinear sequence-pair alignments with the highest sum of scores. To augment the EST collection, gene models based on these homologies with known proteins were created with GeneWise (*31*). The resulting partial gene models are typically incomplete at the 5′ and 3′ ends due to reduced amino acid conservation near the beginning and end of proteins. The cDNA and EST sequences were combined with these partial homology-based models to obtain a more complete set of gene models. Ultimately, each predicted peptide was analyzed for domain content using InterPro (*32*) and aligned with the human, fly, and worm proteomes along with the remainder of the GenBank proteins.

An annotation workshop (*33*) was held at the U.S. Department of Energy Joint Genome Institute to bring together experts from the worldwide ascidian research community to review gene models, assign putative names and functional annotations, and examine the *Ciona* genome from the perspective of metabolic and regulatory networks and pathways. Automated analyses were refined by inspection of gene families and alignments with known genes in other organisms at the workshop and subse-

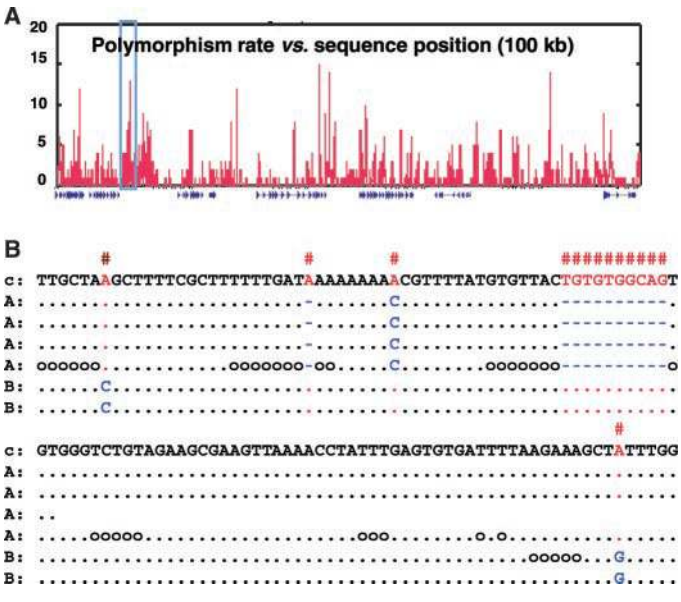

Fig. 3. Allelic polymorphism in the *C. intestinalis* genome. (**A**) The number of variable nucleotides in a running 100-bp window across a 100-kb segment of genomic sequence. Note the wide variation in polymorphism rate on this scale. (**B**) Multiple sequence alignment of six shotgun sequence fragments showing sequence-level variation in the blue box in (A). The top line ("c") represents the assembled consensus sequence, with polymorphic positions highlighted in red and by "#" above. Each subsequent line shows individual shotgun-sequencing fragments aligned with the consensus. Dots indicate match to the consensus at high-quality positions, and "o" indicates matches to the consensus at low-quality positions; ACTG indicates discrepancies with consensus, and dashes indicate nucleotides that are in the consensus but missing from a sequence fragment. Evidently the first four fragments come from one haplotype ("A") and the next two come from the other ("B"), because the discrepancies with the consensus fall into two distinct sets. Note the presence of single-nucleotide variation, and insertions or deletions of various lengths.

quently in a distributed collaboration over the Web. To date, nearly 2000 genes have been examined in this manner. This work is ongoing and can be accessed (and contributed to) at www.jgi.doe.gov/ciona.

Gene complement and global comparisons

A total of 15,852 distinct gene models were obtained. These models have been stringently screened to eliminate any residual redundancies that may remain in the assembly (*24*), which may also have the effect of removing copies of some recently duplicated genes. Genes that span the boundaries of two scaffolds are counted twice in this tally; by examining known genes we estimate that 5% of the genes may be split in this manner (*24*). The gene complement in this simple chordate is thus comparable in size to that of the other completely sequenced invertebrates [14,000 in fruit flies (*22*) and 19,000 in nematodes (*34*)] and somewhat less than the number in vertebrates [31,000 for pufferfish (*23*) and 30,000 to 35,000 for human (*26*, *27*)]. Three-quarters of the predicted *Ciona* genes are supported by EST evidence.

The overall organization of the protein-coding genes in the *Ciona* genome is intermediate between that of protostomes and vertebrates. *Ciona* genes are generally compact and densely packed. The average gene density is somewhat higher than in *Drosophila* (one gene per 7.5 kb versus one per 9 kb in fly), and far greater than the density in human (one gene per 100 kb). *Ciona* genes contain an average of 6.8 exons per gene (versus 5 in *Drosophila* and 8.8 in human). Notable large genes in *Ciona* include the putative ortholog of the vertebrate cardiac ryanodine receptor, which spans 63 kb and encodes a protein of 4978 amino acids in 113 exons that matches its human counterpart at 51% of amino acids over the entire length of both peptides. The intron size distribution (*24*) has a sharp peak near 60 bp, compared with 87 bp in human and 59 bp in fly. *Ciona* exhibits a broad second peak near 300 bp that accounts for the majority of introns, in contrast to the monotonic decrease in the intron size distribution observed in fly and human. The exon size distribution (*24*) peaks at 125 bp, close to the human peak and somewhat less than the *Drosophila* peak at 150 bp.

The draft *Ciona* genome allows us to infer gene origins by comparison with the genomes of flies, nematodes, pufferfish, and mammals. Nearly 60% of predicted *Ciona* gene models (9883) have a detectable protostome homolog, as found by Smith-Waterman alignment of the predicted *Ciona* protein over more than 60% of the length of fly and/or nematode peptides (*24*). These are presumably ancient bilaterian genes, with core physiological and/or developmental roles common to all animals.

A few hundred *Ciona* genes have stronger similarity to fly and/or worm genes than to any genes in the current vertebrate set (*24*). These are candidates for ancient bilaterian genes that were preserved through the ancestral chordate but lost along the vertebrate lineage. An alternative hypothesis is, of course, that these genes have evolved rapidly in vertebrates and no longer have recognizable similarity, or that these genes have been horizontally transferred from protostomes to ascidians. These are mostly genes of unknown function in fly and worm, but include genes encoding chitin synthase, phytochelatin synthase (involved in metal detoxification), and hemocyanin (the primary oxgen carrier in protostomes).

Conversely, nearly one-sixth (2570) of *Ciona* genes lack a clear protostome homolog yet possess a recognizable vertebrate counterpart as measured by a Smith-Waterman alignment over 60% of the length of a fugu or human protein. Given the phylogeny shown in Fig. 1, the most straightforward interpretation is that these genes arose in their modern form on the deuterostome branch some time before the last common chordate ancestor. (An alternative hypothesis is that these genes are of bilaterian origin but have rapidly diverged after the split between protosomes and deuterostomes, or were even lost in flies and nematodes.) Because the known deuterostome genes are essentially all from chordates, we cannot know at this time how many genes in this set are chordate- rather than deuterostome-specific. This ambiguity will be partly removed when the sea urchin genome becomes available (*35*), which will allow more accurate assessment of the origins of these genes along the deuterostome branch.

Finally, nearly one-fifth of the present set of *Ciona* genes (3399) have no clear homolog in fly, worm, pufferfish, or human by the criterion of an alignment over 60% of the target protein (*24*). Some of these are no doubt simply poorly modeled or fragmentary genes. Most (80%) are supported by one or more ESTs from the Kyoto set, however, and 72% have partial alignments with known proteins (BLOSUM62 score above 150, but aligning over less than 60% of the target protein). Although some of these genes will no doubt join the protostome or chordate ranks upon further refinement of the genome, this set contains candidates for tunicate- or ascidian-specific genes (i.e., emerging along the urochordate lineage since its divergence from the common ancestor with vertebrates), and/or rapidly evolving genes whose homologs cannot be reliably detected using simple alignment methods.

The *Ciona* gene set can also be used as a point of comparison to identify genes in vertebrates that appeared between the base of the chordate clade and the radiation of bony vertebrates, as represented by teleost fish and tetrapods. These are genes found in *fugu* and humans, but absent in *Ciona*, and include a wide range of neural genes, as discussed below. Other vertebrate-specific innovations that apparently emerged in the interval between the divergence of the urochordates and the tetrapods include the bulk of the core genes of the adaptive immune system. The tumor necrosis factor (TNF) gene family is widely expanded in vertebrates relative to *Ciona*, as are the genes involved in steroid hormone metabolism, as described below.

We used InterProScan (*36*) to assess the domain content of the *C. intestinalis* proteome, and compared it with that of other animal genomes (*24*). Notable expansions in humans relative to *Ciona* are found in genes containing zinc-finger DNA binding domains, immunoglobulin domains, and the rhodopsin-like heterotrimeric GTP-binding protein (G protein)–coupled receptor family. The marked expansion of immunoglobulins and sensory transmembrane proteins reflects the two most dramatic innovations in the vertebrate lineage: adaptive immunity and the nervous system.

The overall twofold difference in gene number between ascidians and vertebrates is thought to reflect gene duplications in the vertebrate lineage since the common chordate ancestor (*21*). The extent to which modern ascidians represent the ancestral chordate depends in part on the extent to which *Ciona* genes can be unambiguously assigned to individual genes and gene families in vertebrates. Conversely, expansions in the ascidian lineage reflect divergences from the ancestral gene complement of chordates. To examine this question, we grouped *Ciona* and human genes into clusters representing gene families and subfamilies (*24*). This analysis demonstrates that there are many more lineage-specific duplications in vertebrates than in ascidians. Some examples of these are discussed below. These results suggest that the gene complement of *Ciona* is a reasonable approximation to that of the ancestral chordate.

Comparison of select gene families in ascidians and vertebrates

Vertebrate-specific gene duplication events are responsible for the increased gene number relative to protostomes such as *Drosophila* and *Caenorhabditis elegans* (*21*). But protostomes are often too distant from vertebrates to permit direct assignments of orthology on a gene-by-gene basis, which is confounded by protostome-specific divergences and duplications. When comparing *Ciona* to vertebrates, however, we often find that the sea squirt genome appears to represent the ancestral gene content, in the sense that a paralo-

gous family in vertebrates is represented by a single gene in *Ciona* (*24*). However, apparent lineage-specific duplications in ascidians are also found. Thus, *Ciona* combines both ancestral and derived features. We illustrate this for several families of developmental signaling molecules and transcription factors below.

Fgf genes. The fibroblast growth factor (FGF) gene family comprises secreted proteins that control cell proliferation and differentiation (*37*). There are at least 22 members of the FGF family in mammals. In contrast, only one *Fgf* gene (*branchless*) has been identified in *Drosophila* and just two (*egl-17* and *let-756*) in *C. elegans*. The invertebrate FGF proteins (35 to 84 kD) are considerably larger than vertebrate FGFs (17 to 34 kD), and the relationships between the vertebrate and invertebrate FGF proteins are not fully understood.

The *Ciona* cDNA project has identified six *Fgf* genes in *C. intestinalis* (*38*). Phylogenetic analysis indicates that two of the genes correspond to vertebrate *Fgf8/17/18* and *Fgf11/12/13/14*, respectively. Three of the *Ciona Fgf* genes represent orthologs of vertebrate *Fgf3/7/10/22*, *Fgf4/5/6*, and *Fgf9/16/20*, respectively. The sixth *Ciona Fgf* gene cannot be assigned to any of the vertebrate *Fgf* genes. A survey of the *C. intestinalis* genome identified all six *Fgf* genes, but no others. This analysis of *Fgf* duplication and divergence is typical of the *Ciona* genome. The number of gene family members in *Ciona* falls somewhere between those in invertebrate and vertebrate species, and in most cases each of the *Ciona* genes clearly corresponds to one or more vertebrate genes.

Smad genes. Smad transcription factors are the key mediators of bone morphogenetic factor (BMP) and transforming growth factor–β (TGF-β)/activin signaling in nematodes, flies, and vertebrates (*39*). There are eight different *Smad* genes in mouse and human. These fall into three distinct classes: I-Smads (inhibitory Smads: Smad6 and Smad7), R-Smads (receptor-regulated Smads: Smad1, Smad2, Smad3, Smad5, and Smad8/9), and C-Smad (the common Smad: Smad4). Two of the R-Smads (Smad2 and Smad3) function downstream of TGF-β/activin, while three of the R-Smads (Smad1, Smad5, and Smad8/9) function downstream of BMP. The I-Smads and C-Smad mediate both BMP and TGF-β/activin signaling.

The *C. intestinalis* genome contains five *Smad* genes. Like vertebrates, *Ciona* contains two *Smad2/3* genes that may function downstream of TGF-β/activin, although they were duplicated independently in the sea squirt genome. (That is, the two *Ciona Smad2/3* genes cannot be placed in a one-to-one correspondence with the vertebrate *Smad2/3* genes.) Except for Smad2/3, there is at least one *Ciona* gene for each of the three classes, including one for the I-Smads, one for the remaining R-Smads, and one for C-Smad. As in the case of the *Fgf* genes, *Ciona* contains a single copy of genes that are duplicated in vertebrates. It is notable that both types of Smads, which function downstream of BMP and TGF-β/activin, are seen in the *Ciona* genome.

T-box genes. Members of the T-box family of transcription factors share an evolutionarily conserved DNA-binding domain first identified in the product of the mouse *Brachyury* (T) gene (*40*). Mammals have at least 18 T-box genes, whereas *Drosophila* and *C. elegans* possess 8 and 20 T-box genes, respectively (*41*). Many of the T-box genes in *Drosophila* and *C. elegans* arose from lineage-specific duplication events, and it is difficult to align them with their vertebrate counterparts.

The *C. intestinalis* genome contains 10 T-box genes. Six of these are single-copy representatives of six of the seven major subfamilies found in mammals: *Brachyury*, *Tbx1/10*, *Tbx2/3/4/5*, *Tbx15/18/22*, *Tbx20*, and *Eomes/Tbr/Tbx22*. The remaining subfamily, which is represented by a single gene in mammals, *Tbx6*, has four copies in *Ciona*. (These copies are distinct at the nucleotide level, and are not artifacts of sequence polymorphism.) The vertebrate *Tbx2/3/4/5* subfamily can be divided into two subsets: *Tbx2/3* and *Tbx4/5*. The *Ciona* T-box gene that represents the *Tbx2/3/4/5* subfamily is actually more closely related to *Tbx2/3* than to *Tbx4/5*. These observations suggest that *Tbx4/5* arose after the separation of the ascidian and vertebrate lineages. This makes sense because *Tbx4/5* has been implicated in the development of vertebrate appendages, which are obviously lacking in *Ciona*. There are also no *Tbx4/5* orthologs present in the genomes of *Drosophila* and *C. elegans* (*41*, *42*).

Vertebrate Characters Viewed from the *Ciona intestinalis* Genome

Classic studies suggest that ascidians possess organs homologous to the vertebrate thyroid, pineal, and gill slits (*43*, *44*). Comparative genome analyses are broadly consistent with these homologies, as described in subsequent sections.

Genes for the endocrine system in *Ciona intestinalis*

Metazoans use a number of common mechanisms of cell-fate specification during development. In contrast, there seems to be little conservation of the mechanisms that control physiological processes (*43*, *44*). Although many endocrine organs and hormone-receptor systems are conserved among vertebrates, it is difficult to identify corresponding systems in protostomes. The *C. intestinalis* genome provides a unique opportunity to determine which of the vertebrate endocrine systems may have been present in ancient chordates.

Inspection of the *C. intestinalis* genome identifies a number of genes and gene families implicated in vertebrate endocrine processes. In particular, the *Ciona* genome contains genes encoding all the major endocrine receptors that bind peptide or protein ligands, with the exception of the growth hormone family of cytokine receptors. For example, genes encoding the gonadotropin-releasing pathway, the insulin signaling system, and a glycoprotein hormone are all found in the *Ciona* genome. These basic receptor families are also represented in protostome genomes. The *Ciona* genome, however, also contains genes for the synthesis of thyroid hormones, including a Na/I synporter and thyroid peroxidase, as well as a thyroid hormone receptor (see below); similar genes are not observed outside the chordate clade (*45*). In contrast to the conservation of the thyroid hormone system between ascidians and vertebrates, the *Ciona* genome lacks clear orthologs for some of the P450 enzymes essential for the synthesis of vertebrate steroid hormones such as androgens (CYP17), estrogens (CYP19), and corticosteroids and mineralocorticoids (CYP21) (*46*). It also lacks genes encoding steroid hormone receptors (see below).

Although *Ciona* lacks steroid hormones, it contains multiple genes that encode nuclear receptors (NRs). Figure 4 shows a molecular phylogenetic analysis of NRs (*47*). There is a single NR gene in *Ciona* that is equally related to the vitamin D receptors, ecdysone receptors, RORs/Rev-erbs, retinoid X receptors, COUP, HNF4, TR2, TR4, NR4A orphan receptors, FTZ-F1 orphan receptors, and GCNF1 orphan receptor (Fig. 4, black lines). These NR genes are common to bilaterian animals including flies, ascidians, and humans. In contrast, genes encoding thyroid hormone receptors, retinoic acid receptors, and peroxisome proliferator-activated receptors are evident in both *Ciona* and vertebrate genomes, but not in the fly genome (Fig. 4, green lines). These genes are likely chordate (or possibly deuterostome) innovations. Finally, as mentioned above, the *Ciona* genome does not contain genes encoding steroid receptors such as estrogen receptors, androgen receptors, mineralocorticoid receptor, glucocorticoid receptor, and progesterone receptor; these genes thus appear to represent vertebrate innovations (Fig. 4, red line). However, a member of the estrogen-related receptor family—orphan receptors that do not bind any known, naturally occurring steroid hormone—is present in *Ciona*, as well as in fly and human.

Innovation in ascidian endocrine systems?

The *Ciona* genome contains genes with combinations of two or three different protein domains that have not been reported in other organisms. Glycoprotein hormone receptors (GLHRs) are mosaic proteins that contain an extracellular domain with multiple leucine-rich repeats (LRRs) and a G protein–coupled receptor (GPCR) transmembrane and cytosolic domain. Protostome genomes contain orthologs of these hormone receptors, and there is also an ortholog in *C. intestinalis*. The *Ciona* genome also contains a single ortholog of a vertebrate family of orphan receptors related to GLHR—the leucine-rich repeat G protein–coupled receptors, or LGRs (*48*). In addition, the *Ciona* genome contains a large, previously unidentified family of GLHR-related receptors. A further example predicts a polypeptide composed of 333 amino acid residues. This protein contains a caspase domain in the NH_2-terminal half, whereas the COOH-terminal half closely resembles deltex and rhysin, which are involved in Notch signaling (*49*). This is not due to an artifact of the gene prediction program because the corresponding mRNA was identified in the EST collection.

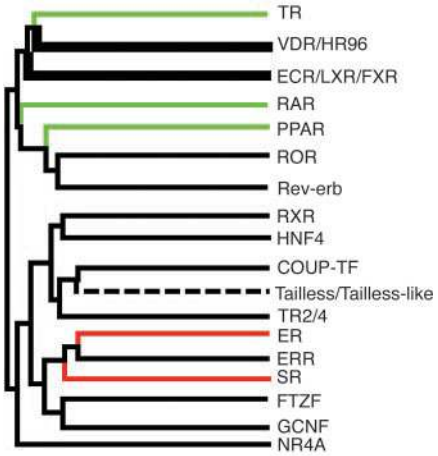

Fig. 4. A molecular phylogenetic tree of nuclear receptor (NR) proteins generated by the neighbor-joining method with both the ligand-binding domain and the DNA-binding domain. The tree was constructed using all of *Ciona* NRs, human NRs, and *Drosophila* NRs. Black lines show the presence of the genes in the fly, ascidian, and human genomes. Green lines show chordate innovations, i.e., genes found in ascidians and vertebrates but not in worms and flies. Red lines show vertebrate innovation, i.e., genes found in vertebrates but not in ascidians. A dotted line shows ascidian lineage-specific innovation, in which the gene was lost only in this lineage. Within a branch indicated by thick lines, two *Ciona* genes were found, suggesting the branches can be further divided into two or more subfamilies. TR, thyroid hormone receptors; VDR, vitamin D receptors; LXR, liver X receptor; FXR, Farnesoid X-activated receptor; RAR, retinoic acid receptors; PPAR, peroxisome proliferator-activated receptors; ROR, RAR-related orphan receptor; Rev-erb, RXR, retinoid X receptors; HNF4, hepatocyte nuclear factor 4; COUP-TF, chicken ovalbumin upstream promoter transcription factor; *Drosophila* Tailless, a fly nuclear receptor; TR2/4, TR2 and TR4 orphan receptors; ER, estrogen receptor; ERR, estrogen-related receptor; SR, steroid hormone receptor; FTZF, fushi tarazu factor; GCNF, germ cell nuclear factor; and NR4A, nuclear receptor 4A. Note that ECR (ecdysone receptor) is the *Drosophila* ortholog of the vertebrate VDRs, and HR96 is the *Drosophila* ortholog of vertebrate VDRs.

The ascidian endostyle and the vertebrate thyroid gland

One of the most prominent classical homologies is the proposed relationship between the vertebrate thyroid gland and the ascidian endostyle, based on their common use of iodine, and continuity with the endostyle of cephalochordates (*43*, *44*). The vertebrate thyroid secretes the potent iodine-containing metabolic hormones triiodothyronine and thyroxine in response to signals from the pituitary and in a feedback loop with the hypothalamus, and these hormones play important roles in vertebrate development, maturation, and homeostasis by regulating gene expression through a nuclear hormone receptor family transcription factor (*45*). The ascidian endostyle is an iodine-sequestering strip that secretes mucus into the pharyngeal feeding apparatus of the adult sea squirt (Fig. 2, A and M). We sought evidence for thyroid-specific proteins in the *Ciona* genome.

Thyroid hormone in vertebrates is produced by the action of thyroid peroxidase (Tpo), which iodinates tyrosine residues on the carrier protein thyroglobulin, using hydrogen peroxide produced by a pair of thyroid oxidases (also known as dual oxidases, Duox1/2) (*50*). Clear homologs of Tpo, and orthologs of Duox1 and Duox2, are encoded in the *Ciona* genome, but no homolog of thyroglobulin was found, suggesting that a different source of tyrosine is used in *Ciona*. Tpo is a member of the myeloperoxidase family that generates hydrogen peroxide for various purposes across bilaterians. Notably, multiple Tpo homologs are encoded in the *Ciona* genome, suggesting an ascidian-specific expansion of this gene that is found in only a single copy in vertebrates. No evidence for the thyroxine-binding globulin and other carrier molecules of vertebrate blood were found.

In vertebrates, the conversion of thyroxine to the active hormone triiodothyronine is catalyzed by a pair of iodothyronine deiodinases expressed in the thyroid (type I) and in body tissues (type II and type I); a third related enzyme (type III) inactivates both forms of thyroid hormone as a means of regulating hormone action (*51*). The *Ciona* genome encodes a pair of closely related enzymes that are slightly more closely related to type I/III than to type II. Phylogenetic analysis shows that the *Ciona* enzymes are more closely related to each other than to the vertebrate enzymes, suggesting that the original deiodinase may have multiplied independently in the ascidian and vertebrate lineages, or that the split between I/III and II occurred close to the last common chordate ancestor. This is a recurring theme in our analysis of the *Ciona* genome: Ascidians and vertebrates have mobilized the same raw materials present in the ancestral chordate genome to their unique purposes and needs.

Apoptosis genes in *Ciona intestinalis*

Apoptosis is triggered by the activation of cysteine proteases called caspases (*52*, *53*). Caspases are initially synthesized as inactive zymogens, but become activated by proteolytic processing during apoptosis. CED-3 is the only nematode caspase involved in apoptosis. In contrast, there are at least 14 distinct caspases in mammals, including four "initiator" (−2, −8, −9, and −10) and three "effector" caspases (−3, −6, and −7). The *Ciona* genome contains 11 caspases; 4 correspond to initiator caspases and 7 to effector caspases.

In *C. elegans*, CED-3 becomes activated by autoprocessing through interaction with the adapter protein CED-4 in cells fated to die during development. In vertebrates, however, there are two major pathways for the activation of caspases, involving mitochondria (intrinsic pathway) or death receptors (extrinsic). The intrinsic pathway is controlled by a CED-4 ortholog called Apaf-1, and by anti-apoptotic proteins Bcl-2 and Bcl-XL, which prevent the release of apoptogenic molecules from mitochondria and thereby block caspase activation (*52*, *53*). Vertebrates also contain two proteins, Bax and Bok, that promote these events. Two Bcl-2 family members, CED-9 and EGL-1, have been identified in *C. elegans*, but there are no counterparts of the vertebrate Bax and Bok genes in worms. The *Ciona* genome contains three genes related to CED-4 and Apaf-1, as well as one copy of the BCL-XL–like anti-apoptotic gene. In contrast to the situation seen in *C. elegans*, *Ciona* also contains two genes that encode proteins with similarities to the pro-apoptotic proteins Bax and Bok.

The extrinsic pathway of caspase activation depends on TNF receptors and is not present in *C. elegans*. Vertebrates contain a variety of TNF receptors, including TNFR1, Fas, DR-3, DR4, and DR5. Four putative members of the TNF family and three possible TNFR-like genes have been detected in the *Ciona* genome. These results suggest that *Ciona* may use both intrinsic and extrinsic pathways for caspase activation.

Immunity-related genes in *Ciona intestinalis*

All Metazoa possess a variety of innate mechanisms to resist infection by pathogens. In contrast, the lymphocyte-based adaptive immune system seems to have suddenly arisen in the jawed vertebrate lineage (*54*, *55*). It is still not clear, however, how this highly sophisticated system involving hundreds of specific genes has evolved. The genome-wide identification of immunity-related genes in nonvertebrate chordates is expected to help elucidate the evolution of both the innate and adaptive immune systems in vertebrates.

A systematic search of the *C. intestinalis* genome failed to identify any of the pivotal genes implicated in adaptive immunity, such as immunoglobulin, T cell receptor, and major histocompatibility complex (MHC) class I and II genes, although we cannot exclude the possibility that *Ciona* has highly divergent orthologs of one or more of these genes. A more convincing "negative" result was obtained by analyzing the genes that encode the 20*S* proteasome, which destroys misfolded proteins (*56*). Eukaryotic 20*S* proteasomes are composed of 14 different gene products; three possess catalytic activity. Mammals contain a second copy of each of the genes that encode these three catalytic subunits. These duplicated genes encode components of an immunoproteasome that is essential for the presentation of antigen to T cells. The *Ciona* genome contains orthologs for each of the 14 vertebrate proteasome genes, but none for the immunoproteasome-specific genes. These observations strongly suggest that *Ciona* lacks the antigen-presenting system for T cells. Putative *Ciona* homologs of the vertebrate MHC-encoded genes do not exhibit an extensive linkage among them, nor syntenic conservation with the vertebrate MHC.

Although there is no evidence for adaptive immunity, a search of the *Ciona* genome reveals a variety of genes that are likely to mediate innate immunity. There are a large number of possible complement genes, including C1q-like and C6-like genes, three Toll-like receptor genes, and a variety of lectin genes. No interleukin or interleukin-receptor genes were identified except for an interleukin-1 (IL-1) receptor and an IL-17 receptor gene. It is possible that *Ciona* has evolved distinctive innate-immunity genes, because a search of the protein domains found in vertebrate innate-immunity genes identifies a number of *Ciona* genes that contain these domains in previously unknown combinations.

Muscle-related genes in *Ciona intestinalis*

Vertebrates produce multiple isoforms of many muscle contractile proteins, whose differential expression contributes to the functional specializations of different muscle cell types such as fast skeletal, slow skeletal, cardiac, and smooth muscles. Ascidians also develop several distinct muscle types, including sarcomeric muscle in the tail of the larva and in the adult heart, and nonsarcomeric, but troponin-regulated, muscle in the adult body-wall (siphon and mantle) (*6*).

A search of the *C. intestinalis* genome suggests that in most cases, the vertebrate gene families encoding multiple isoforms of thick- and thin-filament proteins arose in the vertebrate lineage following the ascidian/vertebrate divergence. For example, vertebrates contain three differentially expressed genes (fast skeletal, slow skeletal, and cardiac) for each of the troponin subunits TnI and TnT. However, the *Ciona* genome contains one TnI gene and one TnT gene, each of which is equally related to all three of the corresponding vertebrate isoforms. Thus, the vertebrate TnT and TnI gene families appear to have arisen within the vertebrate lineage following the ascidian/vertebrate divergence by duplication of a single ancestral TnI or TnT gene, respectively. For some muscle proteins, e.g., tropomyosin, sarcomeric myosin heavy chain, and essential (alkali) myosin light chain, the *Ciona* genome contains a family of two or more related genes. In each of these cases, however, the vertebrate and ascidian gene families appear to have arisen independently; the vertebrate genes form a clade that does not include any of the *Ciona* genes, suggesting duplication from a single ancestral gene following the ascidian/vertebrate divergence. Thus, in terms of molecular genetic specializations and tissue evolution, vertebrate and ascidian muscles have followed largely separate trajectories.

An example of lineage-specific specialization concerns vertebrate smooth muscle. The *Ciona* genome does not appear to contain smooth muscle–specific genes such as those encoding smooth-muscle myosin heavy chain or actin. There are two major clades of class II myosin heavy chain genes in the metazoa—a sarcomeric clade and a nonmuscle/smooth muscle clade that in vertebrates includes two nonmuscle myosins and the myosin that is specifically expressed in smooth muscle (*57*). The *Ciona* genome has six class II myosin genes; five fall into the metazoan sarcomeric clade and one falls into the nonmuscle/smooth muscle clade. The latter gene falls outside of a clade containing the vertebrate nonmuscle and smooth muscle genes. These relationships suggest that the ascidian/vertebrate common ancestor had a single class II nonmuscle myosin gene, and that the evolutionary invention of smooth muscle–specific myosin, through duplication/divergence of this gene, occurred in the vertebrate lineage after the ascidian/vertebrate divergence. A similar situation exists for the smooth muscle–specific actins. Vertebrates contain four closely related genes encoding muscle-type actins: two for smooth muscle (vascular and enteric) and two for sarcomeric muscle (cardiac and skeletal) (*58*). The smooth muscle actins differ at two diagnostic amino acid residues from the sarcomeric actins (*59*). The *Ciona* genome encodes six muscle-type actins, all of which contain the residues diagnostic of sarcomeric muscle actins. The absence of smooth muscle–specific actin and myosin genes in the *Ciona* genome supports the concept that the tissue called smooth muscle in the vertebrates may be a unique characteristic that arose within that group (*60*).

Origins of the vertebrate heart

The ascidian heart, although rudimentary in structure, exhibits an embryonic origin similar to that of vertebrate hearts (*6*). In most ascidians, a beating heart is first detected after metamorphosis; however, the fusion of bilateral heart primordia occurs during embryogenesis. The *Ciona* genome contains orthologs of many of the critical regulatory genes implicated in vertebrate heart development. These include myocardin, the GATA transcription factors Nkx2.5 and its paralogs, the basic helix-loop-helix (bHLH) factors d and eHand, and the Mef2A, B, C, and D genes (*61*). The presence of these multiple paralogs, many of which seem to exhibit overlapping functions, has complicated efforts to decipher vertebrate heart developmental genetics (*61*). As seen for a number of vertebrate gene families and subfamilies, *Ciona* only contains single copies of these genes. There is only one Nkx2 ortholog, one bHLH Hand ortholog, and one MEF-2 ortholog. Additionally, there are only two putative orthologs of the vertebrate GATA factors, one of which shows relatively poor conservation within the DNA binding domain. These observations raise the possibility that research into *Ciona* heart genetics could be used to assess the functions of vertebrate heart genes without the complications arising from redundancies in gene function.

Genes of the Spemann Organizer

The Spemann organizer coordinates the formation of axial and paraxial meso-endoderm and the induction of neural tissues in a variety of vertebrate embryos. The organizer is a source of secreted antagonists that inhibit two major families of signaling molecules: BMPs and Wnts (*62*). Several BMP and Wnt genes are present in vertebrate and invertebrate genomes, including *Ciona*. However, *Drosophila* and *C. elegans* lack obvious Wnt antagonists and contain only one BMP antagonist, which is related to the vertebrate Chordin gene. In contrast, the *Ciona* genome contains a spectrum of Wnt and BMP antagonists,

similar to those seen in vertebrates. These include Wnt antagonists related to the vertebrate sFRP and Dickkopf gene families, and BMP antagonists related to chordin, noggin, and the DAN/gremlin families. Ascidians also possess a Nodal gene; these encode TGF-β signaling molecules that are an essential component of the vertebrate organizer, but absent in worms and flies. The vertebrate organizer is also a source of Wnt-11, which regulates morphogenesis through convergent extension movements (*63*). Wnt-11 is present in *Ciona* but not in the *Drosophila* and *C. elegans* genomes. Hence, the *Ciona* genome contains the full complement of signaling molecules produced by the Organizer. In addition, a number of the transcription factors that regulate the expression of the signaling molecules are also conserved in *Ciona*, including goosecoid, blimp1, hex, and otx.

The presence of most Organizer genes in *Ciona* does not necessarily imply that it contains a functional Organizer. First, none of the conserved factors are strictly specific to the vertebrate organizer. All have additional functions later in the vertebrate life cycle. Second, we do not know whether ascidian orthologs are expressed in the dorsal mesoderm at the onset of gastrulation. Third, experimental evidence does not support a role for the dorsal mesoderm in the specification of paraxial tissues such as the tail muscles. These tissues can develop in an autonomous manner when separated from the dorsal mesoderm (*64*). Finally, BMPs may be required for the development of the notochord in ascidians, but block notochord development in vertebrates. It is possible that the last common chordate ancestor had a functional Organizer that degenerated in ascidians owing to the limited need for long-range signaling in embryos with relatively small cell numbers. Alternatively, ancestral chordates may have used Organizer molecules for other processes and subsequently recruited them into the Organizer in the cephalochordate/vertebrate lineages.

Neural genes

The central nervous system of the ascidian tadpole is a chordate nervous system in miniature, with a 300-cell cerebral vesicle containing several sensory systems, and a hollow dorsal neural tube with a four-cell circumference above a 40-cell notochord; 10 motor neurons innervate the paired tail muscles (*65*). Several neural genes known in vertebrates have clear homologs in *Ciona* and should now be viewed more broadly as chordate characters. Ciliated ependymal cells and Reissner's fiber in the neural tube appear to bind biogenic amines as in vertebrates, as supported by the presence of a SCO-spondin gene in *Ciona*. Claudins involved in vertebrate tight junctions are also present, as are muscle-type acetylcholine receptors, the first example of such a receptor outside vertebrates (*66*). Noelin, a secreted factor that endows neural crest competence in vertebrates, is also present, although no evidence for neural crest cells is known in ascidians; an ortholog of the tyrosinase gene involved in melanin production in neural crest cells is also found, although such genes are also known in invertebrates (*67*).

Phototransduction differs dramatically in invertebrates and vertebrates. In vertebrates, a dark current of sodium ions is shut off by a reduction in cyclic guanosine monophosphate triggered by the photoisomerization of rhodopsin (*68*). By contrast, no dark current is present in invertebrates, and light absorption triggers release of intracellular calcium by way of a phospholipase C pathway to open cation channels (*69*). Three rhodopsin photoreceptors found in the *Ciona* genome are closely related to the deep brain/pineal opsin of vertebrates, supporting the ancient common origin of the vertebrate eye as derived from the pineal system. Orthologs of many of the components of the vertebrate phototransduction cascade are present in the *Ciona* genome, suggesting that the dark current–based vertebrate scheme is found in this invertebrate. Thus, the dark-current approach to phototransduction is likely a chordate (or possibly deuterostome) rather than vertebrate character. It is easy to speculate that the ancestral bilaterian had a primitive light-sensing capability that was harnessed for amplification to distinct intracellular messenger systems in protostomes and deuterostomes.

The *Ciona* genome apparently lacks many genes involved in the transmission of long-range axonal signals and/or long-range guidance cues, presumably related to the compactness of the few-millimeter-long tadpole. In particular, genes involved in the formation and maintenance of myelin are absent, as are neurotrophins and their receptor, plus neurexophilins and other neuronal glycoproteins involved in axon guidance. The machinery for epinephrine synthesis (expressed in neural crest–derived cells of vertebrates) is missing, as are the enzymes for synthesis of melatonin and histamine. Only weak matches to olfactory receptors were identified, although *Ciona* presumably has chemosensory systems involved in larval attachment. Finally, although *Ciona* exhibits circadian rhythms, no clear ortholog of the *period* gene was identified.

Lineage-Specific Innovations in Ascidians

It is unclear to what extent ascidians represent an ancestral, basal chordate or have diverged from ancient chordates through the acquisition of lineage-specific innovations (*1–3*). For example, the metamorphosis of tadpoles into sessile adults is a hallmark of the ascidians (*6*); this process is not seen in other chordate lineages, not even other Urochordates (*70*). Here we consider three aspects of ascidian evolution in light of the *Ciona* genome: genes that are conserved in other animals, but appear to be missing in *Ciona*; genes that are found in multiple copies in *Ciona* but present in single copies in vertebrates; and genes that are unique to ascidians and not seen in other animals.

Genes missing in the *Ciona intestinalis* genome

Hox genes are characterized by their clustered organization and by collinearity between gene order within the cluster and sequential patterns of expression during development (*71*). They are classified into 13 paralogy groups on the basis of sequence similarity (*72*). Vertebrates have multiple Hox gene clusters, each containing about 10 different genes. By contrast, all invertebrates that have been examined contain a single Hox cluster (with the exception of *Drosophila*, in which the ancestral complex is split into the ANT-C and BX-C). In protostomes, the cluster usually contains 10 or fewer genes because there are only one or two genes from paralogy groups 9 to 13. The sea urchin, an invertebrate deuterostome, possesses a single Hox cluster that contains every class of Hox gene, but with only a single Hox-4/5 gene, and three genes related to the posterior Hox genes 9 to 13 (*73*). The cephalochordate *Amphioxus* has a single cluster that contains each of the 13 Hox genes, as well as an additional 14th gene (*74*, *75*).

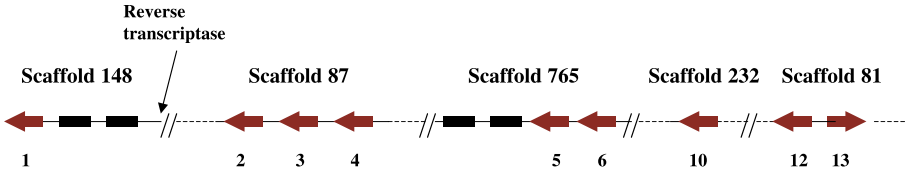

Fig. 5. The diagram represents the different scaffolds containing the *Ciona* Hox genes. Red arrows indicate Hox genes and the orientation of transcription. The black rectangles correspond to nonHox genes. The dashed lines indicate intervals extending to the ends of the respective scaffolds that lack Hox genes but contain additional non-Hox genes. A predicted reverse transcriptase gene is detected at the end of scaffold 1.

There are nine Hox genes in the *C. intestinalis* genome (*76*, *77*). The *Ciona* Hox genes are located on five different scaffolds containing Hox 1, Hox 2 to 4, Hox 5 and 6, Hox 10, and Hox 12 and 13. Hox 7, 8, 9, and 11 are apparently absent, while Hox 12 and 13 are divergently transcribed (Fig. 5). The sum of the five scaffolds is ~980 kb; if located in a single complex, this would be considerably larger than any of the vertebrate Hox complexes (each of the human complexes is ~125 kb), and even larger than the one present in sea urchins (>500 kb) (*73*). The loss of the genes in paralogy groups 7, 8, and 9 may be specific to the entire urochordate lineage, because these genes appear to be absent in the larvacean genome as well (*78*).

In addition to the apparent loss of Hox 7, 8, 9, and 11, a number of other genes common to *Drosophila* and vertebrates are apparently missing from the *Ciona* genome. These include the gene encoding histidine decarboxylase, which is required for histamine synthesis in vertebrates and *Drosophila*. Moreover, genes encoding orthologs of the *Drosophila* nuclear receptor tailless (Fig. 4, dotted line) and the circadian rhythm protein clock were not found in the *Ciona* genome. LIM class homeobox genes comprise six major groups; Lim1, Lim3, Islet, apterous, LIMX, and Lhx6/7 (*79*). Whereas flies, mice, and humans contain one or more genes in each class, *Ciona* contains members of the first five classes but appears to lack Lhx6/7 genes. Of course, these apparent missing genes are subject to the caveat that the genome sequence is not complete.

Gene duplications found only in the *Ciona intestinalis* genome

A number of gene duplication events appear to be unique to the ascidian lineage (Fig. 6). Several are observed for genes that encode transcription factors, including FoxA, Pax2/5/8, and Prox. Prox is a member of the atypical class of homeobox genes (*80*). Nematodes (*ceh-26*), flies (*prospero*), mice (*PROX1*), and humans (*Prox1*) contain a single copy of the gene, which is involved in eye development. In contrast, the *Ciona Prox* gene was duplicated into two genes that are aligned in tandem within the same scaffold within 20 kb. As mentioned earlier, the single *Tbx6* of vertebrates appears to be duplicated into four genes in *Ciona*. Similarly, the single ancestral gene encoding gonadotropin-releasing hormone receptor has also duplicated into four genes.

Genes that suggest functional innovations

Urochordates are also called tunicates because the adult body is enclosed by a fibrous tunic, which in some species is thick and tough (*6*). The matrix of the tunic contains fibers composed largely of a cellulose-like carbohydrate called tunicin (*81*). Because cellulose is typically produced only by plants and bacteria, its presence in ascidians is a curious lineage-specific evolutionary innovation.

Cellulose synthesis and degradation are controlled by a variety of enzymes, including cellulose synthases and endoglucanases, respectively (*82*). The *Ciona* genome contains at least one potential cellulose synthase and several endoglucanases (Fig. 7). Most of the endoglucanases are related to the Korrigan genes of *Arabidopsis*, which are essential for the biosynthesis of the plant wall (*83*). However, the closest matches are seen for the endoglucanases present in termites and wood-eating cockroaches. Most of these latter endoglucanases are encoded in the genomes of bacterial and fungal symbionts and are essential for the use of wood as a food source. However, there is also evidence that some of the endoglucanases present in termites may be endogenous and presumably arose from a horizontal gene-transfer event between symbiont and host (*84*).

There is little doubt that the endoglucanases present in *Ciona* are endogenous genes and not due to contamination by potential symbionts. For example, one of the *Ciona* endoglucanase genes is present on scaffold 11, which is ~680 kb in length (Fig. 7). This gene is expressed because there is a corresponding cDNA. There are conserved predicted genes located both 5′ and 3′ of the endoglucanase. The 5′ gene encodes a putative glutamate/aspartate

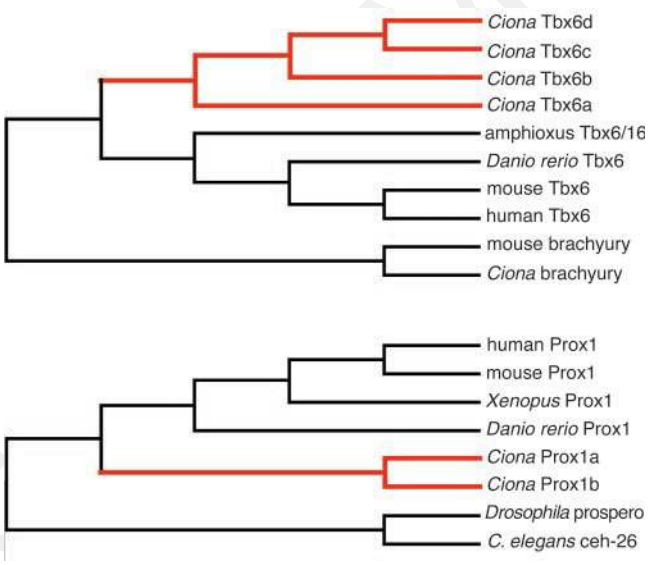

Fig. 6. Lineage-specific gene duplications in *Ciona*. amphioxus, zebrafish (*Danio*), mouse, and humans each appear to contain single copies of the Tbx6 gene (top). However, the *Ciona* genome contains four Tbx6 genes (red lines); these are contained as duplicated genes in a single scaffold. Similarly, there are two copies of the Prox homeobox gene in *Ciona* (red lines), but only single copies of this gene in mouse, human, zebrafish, worms, and flies (bottom).

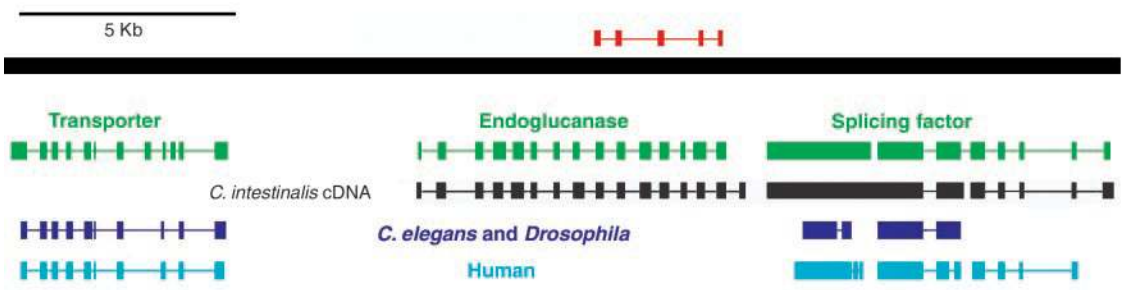

Fig. 7. One of the *Ciona* endoglucanase genes is flanked by conserved "animal" genes, an aspartate/glutamate transporter on the left and a splicing factor on the right. The Korrigan-1 gene from *Arabidopsis* was used to BLAST the *Ciona* genome (red rectangles, top). A predicted gene was identified on scaffold 11, which is 680 kb in length; a portion of the scaffold is shown in this figure. A predicted gene model contains extensive homology with the Korrigan gene used for the BLAST analysis (green rectangles represent predicted exons). It is flanked by at least two additional predicted genes. There are identified cDNAs for the putative endoglucanase and the splicing factor on the right (gray rectangles below the predicted exons in green). Moreover, both the genes on the left and right are conserved in flies, worms (blue), and humans (teal). The rectangles and gaps correspond to exons and introns, respectively.

transporter, whereas the 3′ gene encodes a putative RNA splicing factor. Thus, while the endoglucanase gene is not conserved in worms, flies, or humans, it is flanked by authentic "animal" genes.

Although it is possible that the genomes of wood-eating animals have acquired endogenous endoglucanase genes, no animal reported to date has been shown to contain a bona fide cellulose synthase gene. There appears to be at least one such gene in the *Ciona* genome that is similar to the cellulose synthases seen in a number of nitrogen-fixing bacteria, such as Anabaena and Sinorhizobium (*85*). Future studies will determine whether the encoded protein functions in cellulose biosynthesis. If so, this would represent a dramatic example of horizontal gene transfer.

Hemoglobins are oxygen-transporting proteins found in almost all representatives of every animal phylum (*86*). It is notable, therefore, that genes encoding hemoglobins are not found in the *Ciona* genome. Instead *Ciona* contains two candidate genes encoding an alternate oxygen transport protein, hemocyanin, also found in arthropods and mollusks. Hemocyanins have oxygen-binding centers that are structurally similar to those seen in hemoglobins and tyrosinases (*87*). The *Ciona* hemocyanins are type III copper proteins and have conserved amino acid residues in the upper-binding center, although they are divergent from those of arthropods and mollusks. Genes encoding myoglobin and hemerythrin are, by contrast, not found in the *Ciona* genome.

Conclusions

Our initial reading of the *Ciona* genome provides new insights into the evolutionary origins of the vertebrates. The last shared ancestor of modern chordates probably possessed single-copy genes for the present-day complement of gene families engaged in a variety of signaling and regulatory processes seen in vertebrate development, including FGFs, Smads, and T-box genes. There were also rudiments of key vertebrate organ systems including the heart and the pineal and thyroid glands. Though useful as an approximation of the ancestral chordates, ascidians have experienced notable lineage-specific evolution, including the remarkable acquisition of genes that control cellulose metabolism.

There are two recurring themes in our analysis of the ascidian genome and its relationship to the vertebrate genome. First, we find repeatedly that a family or subfamily of vertebrate genes has only a single representative in *Ciona*. The implication is that the *Ciona* gene content in these families corresponds to the complement of the ancestral chordate. Second, there were notable cases of gene families with multiple members in both *Ciona* and vertebrates that could not be placed in easy correspondence with one another. In these cases, the gene content of the ancestral chordate was apparently mobilized and diversified independently in the two lineages, providing examples of macroevolutionary change within the different branches of the chordate phylum.

The streamlined nature of the *Ciona* genome should have an enormous impact on unraveling complex developmental processes in vertebrates. For example, information about the function of the major vertebrate FGF subfamilies can be obtained in *Ciona* through the use of simple gene-disruption methods (e.g., morpholinos), unhampered by the complications associated with functional redundancies often encountered in large gene families.

Further analysis of the *Ciona* genome will also provide considerable information about the regulation of vertebrate genes. The *Ciona* cDNA projects have characterized transcripts for more than three-quarters of the genes (*28*). Systematic in situ hybridization assays are being performed to determine the spatial and temporal expression of each gene during embryonic and larval development. Thus far, there is information for nearly one-third of the genes in the *Ciona* genome (http://ghost.zool.kyoto-u.ac.jp/indexr1.html) (*88*, *89*). Finally, all of this information—the gene-disruption data and gene-expression profiles—will be compiled along with large-scale screens for cis-regulatory DNAs to determine complete gene-regulation networks underlying the development of basic chordate features such as the neural tube and notochord (*90*, *91*).

References and Notes

1. H. Gee, *Before the Backbone. Views on the Origin of the Vertebrates* (Chapman & Hall, London, 1996).
2. B. K. Hall, *Evolutionary Developmental Biology* (Chapman & Hall, London, ed. 2, 1998).
3. A. Adoutte *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 4453 (2000).
4. A. Kowalevsky, *Mem. Acad. St. Petersbourg Ser. 7* **10**, 1 (1866).
5. C. Darwin, *The Descent of Man* (Murray, 1871).
6. N. Satoh, *Developmental Biology of Ascidians* (Cambridge Univ. Press, New York, ed., 1994).
7. N. Satoh, *Differentiation* **68**, 1 (2001).
8. J. C. Corbo, A. Di Gregorio, M. Levine, *Cell* **106**, 535 (2001).
9. H. Nishida, *Int. Rev. Cytol.* **176**, 24 (1997).
10. W. R. Jeffery, *Int. Rev. Cytol.* **203**, 3 (2001).
11. E. G. Conklin, *J. Acad. Nat. Sci.* (*Philadelphia*) **13**, 1 (1905).
12. H. Nishida, *Dev. Biol.* **121**, 526 (1987).
13. Y. Nakatani, R. Moody, W. C. Smith, *Development* **126**, 3293 (1999).
14. P. Sordino *et al.*, *Sarsia* **85**, 173 (2000).
15. J. C. Corbo, M. Levine, R. W. Zeller, *Development* **124**, 589 (1999).
16. A. Di Gregorio, M. Levine, *Differentiation* **70**, 132 (2002).
17. H. Takahashi *et al.*, *Genes Dev.* **13**, 1519 (1999).
18. K. S. Imai, N. Satoh, Y. Satou, *Development* **129**, 3441 (2002).
19. M. W. Simmen, S. Leitgeb, V. H. Clark, S. J. M. Jones, A. Bird, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 4437 (1998).
20. H.-C. Seo *et al.*, *Science* **294**, 2506 (2001).
21. P. W. H. Holland, J. Garcia-Fernandez, N. A. Williams, A. Sidow, *Development* **1994** (Suppl.), 125 (1994).
22. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
23. S. Aparicio *et al. Science*, **297**, 1301 (2002).
24. Supplemental methods and data are available on *Science* Online.
25. N. Putnam, J. Chapman, I. Ho, D. Rokhsar, unpublished results.
26. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
27. J.C. Venter *et al.*, *Science* **291**, 1304 (2001).
28. Y. Satou *et al.*, *Genesis* **33**, 153 (2002).
29. W. J. Kent, *Genome Res.* **12**, 656 (2002).
30. S. F. Altschul *et al. J. Mol. Biol.* **215**, 403 (1990).
31. E. Birney, R. Durbin, *Genome Res.* **10**, 547 (2000).
32. N. J. Mulder *et al.*, *Brief Bioinform.* **3**, 225 (2002).
33. E. Pennisi, *Science* **287**, 2182 (2000).
34. The *C. elegans* genome consortium, *Science* **282**, 2012 (1998).
35. R. A. Cameron *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 9514 (2000).
36. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (2001).
37. D. M. Ornitz, N. Itoh, *Genome Biol.* **2**, REVIEWS 3005.1 (2001).
38. Y. Satou, K. S. Imai, N. Satoh, *Dev. Genes Evol.* **212**, 432 (2002).
39. C. S. Hill, *Int. J. Biochem. Cell Biol.* **31**, 1249 (1999).
40. B. G. Herrmann, S. Labeit, A. Poustka, T. R. King, H. Lehrach, *Nature* **343**, 617 (1990).
41. V. E. Papaioannou, *Int. Rev. Cytol.* **207**, 1 (2001).
42. I. Ruvinsky, J. J. Gibson-Brown, *Development* **127**, 5233 (2000).
43. A. S. Romer, *The Vertebrate Body* (Saunders, Philadelphia, ed. 4, 1970).
44. C. K. Weichert, *Elements of Chordate Anatomy* (McGraw-Hill, New York, 1953).
45. D. Kirsten, *Neonatal Netw.* **19**, 11 (2000).
46. A. M. Capponi. *Trends Endocrinol. Metab.* **13**, 118 (2002).
47. V. Laudet, *J. Mol. Endocrinol.* **19**, 207 (1997).
48. P. Joost, A. Methner, *Genome Biol.*, in press.
49. K. Brennan, P. Gardner, *Bioessays* **24**, 405 (2002).
50. L. Lacroix *et al.*, *Thyroid* **11**, 1017 (2001).
51. J. T. Dunn, A. D. Dunn. *Thyroid* **11**, 407 (2001).
52. V. Cryns, J. Yuan, *Genes Dev.* **12**, 1551 (1998).
53. G. S. Salvesen, V. M. Dixit, *Cell* **91**, 443 (1997).
54. J. A. Hoffmann, F. C. Kafatos, C. A. Janeway, R. A. Ezekowitz, *Science* **284**, 1313 (1999).
55. M. F. Flajnik, M. Kasahara, *Immunity* **15**, 351 (2001).
56. H. D. Ulrich, *Curr. Top. Microbiol. Immunol.* **268**, 137 (2002).
57. J. R. Sellers, *Biochim. Biophys. Acta* **1496**, 3 (2000).
58. J. Vandekerckove, K. Weber, *J. Mol. Biol.* **179**, 391 (1984).
59. S. Kovilur, J. W. Jacobson, R. L. Beach, W. R. Jeffery, C. R. Tomlinson, *J. Mol. Evol.* **36**, 361 (1993).
60. S. Oota, N. Saitou, *Mol. Biol. Evol.* **16**, 856 (1999).
61. R. M. Cripps, E. N. Olson, *Dev. Biol.* **246**, 14 (2002).
62. R. Harland, J. Gerhart, *Annu. Rev. Cell Dev. Biol.* **13**, 611 (1997).
63. C. P. Heisenberg *et al.*, *Nature* **405**,76 (2000).
64. T. H. Meedel, J. R. Whittaker, *Dev. Biol.* **105**, 479 (1984).
65. I. A. Meinertzhagen, Y. Okamura, *Trends Neurosci.* **24**, 401 (2001).
66. S. Tsukita, M. Furuse, M. Itoh. *Nature Rev. Mol. Cell. Biol.* **2**, 285 (2001).
67. J. H. Christiansen, E. G. Coles, D. G. Wilkinson, *Curr. Opin. Cell Biol.* **12**, 719 (2000).
68. L. Stryer, *Annu. Rev. Neurosci.* **9**, 87 (1986).
69. C. S. Zuker, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 571 (1996).
70. A. Nishino, N. Satoh, *Genesis* **29**, 36 (2001).
71. D. Duboule, *Development* (Suppl.), 143 (1994).
72. T. R. Burglin, in *Guidebook to the Homeobox Genes*, D. Duboule, Ed. (Oxford Univ. Press, New York, 1994), pp. 27–64.
73. P. Martinez, J. P. Rast, C. Arenas-Mena, E. H. Davidson, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 1469 (1999).

74. J. Garcia-Fernàndez, P. W. H. Holland, *Nature* **370**, 563 (1994).
75. D. E. Ferrier, C. Minguillon, P. W. H. Holland, J. Garcia-Fernandez, *Evol. Dev.* **2**, 284 (2000).
76. A. Di Gregorio *et al.*, *Gene* **156**, 253 (1995).
77. M. Gionti *et al.*, *Dev. Genes. Evol.* **207**, 515 (1998).
78. D. Chourrout, R. Di Lauro, personal communication.
79. O. Hobert, H. Westphal, *Trends Genet.* **16**, 75 (2000).
80. S. I. Tomarev, *Int. J. Dev. Biol.* **41**, 835 (1997).
81. G. Krishnan, *Indian J. Exp. Biol.* **13**, 172 (1975).
82. S. M. Read, T. Bacic, *Science* **295**, 59 (2002).
83. J. Zuo *et al.*, *Plant Cell* **12**, 1137 (2000).
84. N. Lo *et al.*, *Curr. Biol.* **10**, 801 (2000).
85. D. R. Nobles, D. K. Romanovicz, R. M. Brown Jr., *Plant Physiol.* **127**, 529 (2001).
86. R. C. Hardison, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 5675 (1996).
87. K. E. van Holde, K. I. Miller, H. Decker, *J. Biol. Chem.* **276**, 15563 (2001).
88. Y. Satou *et al.*, *Development* **128**, 2893 (2001).
89. T. Kusakabe *et al.*, *Dev. Biol.* **242**, 188 (2002).
90. N. Harafuji, D. N. Keys, M. Levine, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6802 (2002).
91. D. N. Keys *et al.*, in preparation.
92. This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research Program; by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract no. DE-AC03-76SF00098, and Los Alamos National Laboratory under contract no. W-7405-ENG-36; and by MEXT, Japan (grants 12201001 to Y.K., 12202001 to N.S.), Japan Society for the Promotion of Science (to Y.S.), Human Frontier Science Program (to N.S. and M.L.), and NIH (HD-37105 and NSF IBN-9817258 to M.L.)