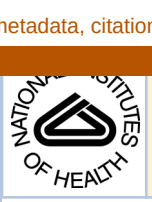


Published in final edited form as:

Science. 2006 November 10; 314(5801): 941–952. doi:10.1126/science.1133609.

The Genome of the Sea Urchin *Strongylocentrotus purpuratus*

Sea Urchin Genome Sequencing Consortium[*,†]

Abstract

We report the sequence and analysis of the 814-megabase genome of the sea urchin *Strongylocentrotus purpuratus*, a model for developmental and systems biology. The sequencing strategy combined whole-genome shotgun and bacterial artificial chromosome (BAC) sequences. This use of BAC clones, aided by a pooling strategy, overcame difficulties associated with high heterozygosity of the genome. The genome encodes about 23,300 genes, including many previously thought to be vertebrate innovations or known only outside the deuterostomes. This echinoderm genome provides an evolutionary outgroup for the chordates and yields insights into the evolution of deuterostomes.

The genome of the sea urchin was sequenced primarily because of the remarkable usefulness of the echinoderm embryo as a research model system for modern molecular, evolutionary, and cell biology. The sea urchin is the first animal with a sequenced genome that (i) is a free-living, motile marine invertebrate; (ii) has a bilaterally organized embryo but a radial adult body plan; (iii) has the endoskeleton and water vascular system found only in echinoderms; and (iv) has a nonadaptive immune system that is unique in the enormous complexity of its receptor repertoire. Sea urchins are remarkably long-lived with life spans of *Strongylocentrotid* species extending to over a century [see supporting online material (SOM)] and highly fecund, producing millions of gametes each year; and *Strongylocentrotus purpuratus* is a pivotal component of subtidal marine ecology and an important fishery catch in several areas of the world, including the United States. Although a research model in developmental biology for a century and a half, for most of that time, few were aware of one of the most important characteristics of sea urchins, a character that directly enhances its significance for genomic analysis: Echinoderms (and their sister phylum, the hemichordates) are the closest known relatives of the chordates (Fig. 1 and SOM). A description of the echinoderm body plan, as well as aspects of the life-style, longevity, polymorphic gene pool, and characteristics that make the sea urchin so valuable as a research organism, are presented in the SOM.

The last common ancestors of the deuterostomal groups at the branch points shown in Fig. 1 are of Precambrian antiquity [>540 million years ago (Ma)], according to protein molecular phylogeny. Stem group echinoderms appear in the Lower Cambrian fossil assemblages dating to 520 Ma. Cambrian echinoderms came in many distinct forms, but from their first appearance, the fossil record illustrates certain distinctive features that are still present: their water vascular system, including rows of tube feet protruding through holes in the ambulacral grooves and their calcite endoskeleton (mainly, a certain form of $CaCO_3$), which displays the specific three-dimensional structure known as “stereom.” The species sequenced, *Strongylocentrotus purpuratus*, commonly known as the “California purple sea urchin” is a representative of the thin-spined “modern” group of regularly developing sea urchins (euechinoids). These evolved to become the dominant echinoid form after the great Permian-Triassic extinction 250 million years ago.

Transcription regulatory factors: Eric H. Davidson (leader), Maria Ina Arnone, Margherita Branno, C. Titus Brown, R. Andrew Cameron,[3] Lili Chen,[3] Rachel F. Gray,[3] Meredith Howard-Ashby,[3] Sorin Istrail,[46] Pei Yun Lee,[3] Annamaria Locascio,[5] Pedro Martinez,[73,74] Stefan C. Materna,[3] Jongmin Nam,[3] Paola Oliveri,[3] Francesca Rizzo,[5] Joel Smith[3]

DNA sequencing: Donna Muzny[1,2] (leader), Erica Sodergren[1,2] (leader), Richard A. Gibbs[1,2] (leader), George M. Weinstock[1,2] (leader), Stephanie Bell,[1,2] Joseph Chacko,[1,2] Andrew Cree,[1,2] Stacey Curry,[1,2] Clay Davis,[1,2] Huyen Dinh,[1,2] Shannon Dugan-Rocha,[1,2] Jerry Fowler,[1,2] Rachel Gill,[1,2] Cerrissa Hamilton,[1,2] Judith Hernandez,[1,2] Sandra Hines,[1,2] Jennifer Hume,[1,2] LaRonda Jackson,[1,2] Angela Jolivet,[1,2] Christie Kovar,[1,2] Sandra Lee,[1,2] Lora Lewis,[1,2] George Miner,[1,2] Margaret Morgan,[1,2] Lynne V. Nazareth,[1,2] Geoffrey Okwuonu,[1,2] David Parker,[1,2] Ling-Ling Pu,[1,2] Ruifeng Shen,[1,2] Rachel Thorn,[1,2] Rita Wright[1,2]

We present here a description of the *S. purpuratus* genome and gene products. The genome

[1]Human Genome Sequencing Center, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [2]Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [3]Division of Biology, California Institute of Technology, Pasadena, CA 91125, USA. [4]National Institute of Dental and Craniofacial Research, NIH, Bethesda, MD 20892, USA. [5]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Napoli, Italy. [6]Department of Biology, Boston College, Chestnut Hill, MA 02467, USA. [7]Department of Biology, Department of Biochemistry and Microbiology, University of Victoria, Victoria, BC, Canada, V8W 3N5. [8]Mount Desert Island Biological Laboratory, Salisbury Cove, ME 04672, USA. [9]Human Genetics Section, Laboratory of Genomic Diversity, National Cancer Institute–Frederick, Frederick, MD 21702, USA. [10]School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS, UK. [11]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, 15213, USA. [12]Department Molecular, Cellular and Developmental Biology and the Marine Science Institute, University of California, Santa Barbara, Santa Barbara, CA 93106–9610, USA. [13]Hopkins Marine Station, Stanford University, Pacific Grove, CA 93950, USA. [14]Howard Hughes Medical Institute, Center for Cancer Research, Massachusetts Institute of Technology (MIT), Cambridge, MA 02139, USA. [15]Departments of Biochemistry and Molecular Biology, University of Texas, M. D. Anderson Cancer Center, Houston, TX, 77030, USA. [16]Molecular Biology and Biotechnology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [17]Department of Biology, Duke University, Durham, NC 27708, USA. [18]Department of Biology, Wheaton College, Norton, MA 02766, USA. [19]Stowers Institute for Medical Research, Kansas City, MO 64110, USA. [20]Department of Microbiology, Kansas University Medical Center, Kansas City, KS 66160, USA. [21]Sunnybrook Research Institute and Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada M4N 3M5. [22]Department of Immunology, University of Toronto, Toronto, Ontario, Canada, M4N 3M5. [23]Department of Biological Sciences, George Washington University, Washington, DC 20052, USA. [24]Royal Swedish Academy of Sciences, Kristineberg Marine Research Station, Fiskebackskil, 450 34, Sweden. [25]Marine Biology, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093–0202, USA. [26]Department of Molecular and Cellular Biology and Biochemistry, Brown University Providence, RI 02912, USA. [27]Department of Biology and Institute for Genome Sciences and Policy, Duke University, Durham, NC 27708, USA. 28Department of Animal Science, Texas A&M University, College Station, TX 77843, USA. [29]National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, MD 20894, USA. [30]Department of Ecology, Evolution, and Marine Biology, University of California Santa Barbara, Santa Barbara, CA 93106, USA. [31]National Center for Biotechnology Information, NIH, Bethesda, MD 20892, USA. [32]Penn Genomics Institute, University of Pennsylvania, Philadelphia, PA 19104, USA. [33]Evolution and Development Group, Max-Planck Institut für Molekulare Genetik, 14195 Berlin, Germany. [34]Royal Holloway, University of London, Egham, Surrey TW20 0EX, UK. [35]Center for Cancer Research, MIT, Cambridge, MA 02139, USA. [36]Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA 94720–3200, USA. [37]Department of Biology, University of South Florida, Tampa, FL 33618, USA. [38]Université Pierre et Marie Curie (Paris 6), UMR 7150, Equipe Cycle Cellulaire et Développement, Station Biologique de Roscoff, 29682 Roscoff Cedex, France. 39CNRS, UMR 7150, Station Biologique de Roscoff, 29682 Roscoff Cedex, France. [40]CNRS, UMR7628, Banyuls-sur-Mer, F-66650, France. [41]Université Pierre et Marie Curie (Paris 6), UMR7628, Banyuls-sur-Mer, F-66650, France. [42]Center for Bioinformatics, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA. [43]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. [44]Tethys Research, LLC, 2115 Union Street, Bangor, Maine 04401, USA. [45]Department of Molecular, Cellular, and Developmental Biology, University of California, Berkeley, Berkeley, CA 94720, USA. [46]Center for Computational Molecular Biology, and Computer Science Department, Brown University, Providence, RI 02912, USA. [47]Genome Research Facility, National Aeronautics and Space Administration, Ames Research Center, Moffet Field, CA 94035, USA. [48]Systemix Institute, Cupertino, CA 95014, USA. [49]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6. [50]Department of Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC, Canada, V5A 1S6. [51]Department of Biology, Center for Cancer Research, MIT, Cambridge, MA 02139, USA. [52]Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089–0740, USA. [53]Department of Biology, University of Central Florida, Orlando, FL 32816–2368, USA. [54]Department of Biological Sciences, Dartmouth College, Hanover, NH 03755, USA. [55]Center for Computational Regulatory Genomics, Beckman Institute, California Institute of Technology, Pasadena, CA 91125, USA. [56]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204, USA. [57]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC, Canada, V5Z 4E6. [58]Department of Biology and the Institute of Systems Research, University of Maryland, College Park, MD 20742, USA. [59]Laboratory of Cellular and Molecular Biology, National Institute on Aging, NIH, Baltimore, MD 21224, USA. [60]Department of Biological Sciences, Macquarie University, Sydney NSW 2109, Australia. [61]Center of Marine Biotechnology, UMBI, Columbus Center, Baltimore, MD 21202, USA. [62]Department of Cell Biology and Anatomy, Louisiana State University Health Sciences Center, New Orleans, LA 70112, USA. [63]Department of Biology, University of Victoria, Victoria, BC, Canada, V8W 2Y2. [64]Department of Neuroscience, Uppsala University, Uppsala, Sweden. [65]Laboratory of Cellular and Molecular Biophysics, National Institute of Child Health and Development, NIH, Bethesda, MD 20895, USA. [66]Developmental Unit, EMBL, 69117 Heidelberg, Germany. [67]Computational Unit, EMBL, 69117 Heidelberg, Germany. [68]Biotechnology Institute, Universidad Nacional Autónoma de Mexico (UNAM), Cuernavaca, Morelos, Mexico 62250. [69]Department of Cellular and Developmental Biology “Alberto Monroy,” University of Palermo, 90146 Palermo, Italy. [70]Laboratoire de Biologie du Développement (UMR 7009), CNRS and Université Pierre et Marie Curie (Paris 6), Observatoire Océanologique, 06230 Villefranche-sur-Mer, France. [71]Department of Biology, University of Patras, Patras, Greece. [72]Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA. [73]Departament de Genetica, Universitat de Barcelona, 08028–Barcelona, Spain. [74]Institució Catalana de Recerca i Estudis Avancats (ICREA), Barcelona, Spain. [75]Institut Jacques Monod, CNR-UMR 7592, 75005 Paris, France. [76]Consiglio Nazionale delle Ricerche, Istituto di Biomedicina e Immunologia Molecolare “Alberto Monroy,” 90146 Palermo, Italy. [77]Razavi-Newman Center for Bioinformatics, Salk Institute for Biological Studies, La Jolla, CA 92186, USA. [78]Department of Zoology, University of Hawaii at Manoa, Honolulu, HI 96822, USA.

*Present address: GlaxoSmithKline, 1250 South Collegeville Road, Collegeville, PA 19426, USA.

†Present address: Massachusetts General Hospital Cancer Center, Charlestown, MA 02129, USA.

provides a wealth of discoveries about the biology of the sea urchin, Echinodermata, and the deuterostomes. Among the key findings are the following:

- The sea urchin is estimated to have 23,300 genes with representatives of nearly all vertebrate gene families, although often the families are not as large as in vertebrates.
- Some genes thought to be vertebrate-specific were found in the sea urchin (deuterostome-specific); others were identified in sea urchin but not the chordate lineage, which suggests loss in the vertebrates.
- Expansion of some gene families occurred apparently independently in the sea urchin and vertebrates.
- The sea urchin has a diverse and sophisticated immune system mediated by an astonishingly large repertoire of innate pathogen recognition proteins.
- An extensive defensome was identified.
- The sea urchin has orthologs of genes associated with vision, hearing, balance, and chemosensation in vertebrates, which suggests hitherto unknown sensory capabilities.
- Distinct genes for biomineralization exist in the sea urchin and vertebrates.
- Orthologs of many human disease–associated genes were found in the sea urchin.

Sequencing and Annotation of the *S. purpuratus* Genome

Sequencing and assembly

Sperm from a single male was used to prepare DNA for all libraries (tables S1 and S2) and whole-genome shotgun (WGS) sequencing. The overall approach was basedonthe "combined strategy" used for the rat genome (1), where WGS sequencing to six times coverage was combined with two times sequence coverage of BAC clones from a minimal tiling path (MTP) (fig. S1). The use of BACs provided a framework for localizing the assembly process, which aided in the assembly of repeated sequences and solved problems associated with the high heterozygosity of the sea urchin genome, without our resorting to extremely high coverage sequencing.

Several different assemblies were produced during the course of the project (see SOM for details). The Sea Urchin Genome Project (SUGP) was the first to produce both intermediate WGS assemblies and a final combined assembly. This was especially useful, not only for the early availability of an assembly for analysis, but also because WGS contigs were used to fill gaps between BACs in the combined assembly. The pure WGS assembly was produced (v 0.5 GenBank accession number range AAGJ01000001 to AAGJ01320773; also referred to as NCBI build 1.1) and released in April 2005. The final combined BAC-WGS assembly was released in July 2006 as version (v) 2.1 and submitted to GenBank (accession number range AAGJ02000001 to AAGJ02220581).

A second innovation in the SUGP was the use of the clone-array pooled shotgun sequencing (CAPSS) strategy (2) for BAC sequencing (fig. S2). The MTP consisted of 8248 BACs, and rather than prepare separate random libraries from each of these, the CAPSS strategy involved BAC shotgun sequencing from pools of clones and then deconvoluting the reads to the individual BACs. This allowed the BAC sequencing to be performed in 1/5th the time and at 1/10th the cost.

The principal new challenge in the SUGP was the high heterozygosity in the outbred animal that was sequenced. It was known that single-copy DNA in the sea urchin varied by as much as 4 to 5% [single nucleotide polymorphism (SNP) plus insertion/deletion (indel)], which is much greater than human (~0.5%) (3). Moreover, alignment of WGS reads to the early v 0.1 WGS assembly revealed at least one SNP per 100 bases, as well as a comparable frequency of indel variants. This average frequency of a mismatch per 50 bases or higher prevented merging by the assembly module in Atlas, the Phrap assembler, and also made it difficult to determine if reads were from duplicated but diverged sections of the genome or heterozygous homologs. This challenge was met by adding components to Atlas to handle local regions of heterozygosity and to take advantage of the BAC data, because each BAC sequence represented a single haplotype (see SOM). High heterozygosity has been seen in the past with the *Ciona* genomes (4, 5) and is likely to be the norm in the future as fewer inbred organisms are sequenced. Moreover, the CAPSS approach makes BAC sequencing more manageable for large genomes. Thus, the sea urchin project may serve as a paradigm for future difficult endeavors.

Combining the BAC-derived sequence with the WGS sequence generated a high-quality draft with 4 to 5% redundancy that covered more than 90% of the genome while sequencing to a level of 8× base coverage (table S2). The assembly size of 814 Mb is in good agreement with the previous estimate of genome size, 800 Mb ± 5% (6). The assembly is a mosaic of the two haplotypes, but it was possible to determine the phase of the BACs on the basis of how many mismatches neighboring BACs had in their overlap regions. This information will be used to create a future version of the genome in which the individual haplotypes are resolved.

Gene predictions

The v 0.5 WGS assembly displayed sufficient sequence continuity (a contig N50 of 9.1 kb) and higher-order organization (a scaffold N50 of 65.6 kb) to allow gene predictions to be produced and the annotation process to begin even while the BAC component was being sequenced. We generated an official gene set (OGS), consisting of ~28,900 gene models, by merging four different sets of gene predictions with the GLEAN program (7) (see SOM for details). One of these gene sets, produced from the Ensembl gene prediction software, was created for both v 0.5 and v 2.0 assemblies.

To estimate the number of genes in the *S. purpuratus* genome, we began with the 28,900 gene models in the OGS and reduced this by the 5% redundancy found by mapping to the v 2.0 assembly, then increased it by a few percent for the new genes observed in the Ensembl set from the v 2.0 assembly compared with v 0.5. From manual analysis of well-characterized gene sets (e.g., ciliary, cell cycle control, and RNA metabolism genes), we estimated that, in addition to redundancy, another 25% of the genes in the OGS were fragments, pseudogenes, or otherwise not valid. Finally, whole-genome tiling microarray analysis (see below) showed 10% of the transcriptionally active regions (long open reading frames, not small RNAs) were not represented by genes in the OGS. Taken together, this analysis gave an estimate of about 23,300 genes for *S. purpuratus*. Information on all annotated genes can be found at (8).

The overall trends in gene structure were similar to those seen in the human genome. The statistics of the Ensembl predictions from the WGS assembly revealed an average of 8.3 exons and 7.3 introns per transcript (see SOM). The average gene length was 7.7 kb with an average primary transcript length of 8.9 kb. A broad distribution of all exon lengths peaked at around 100 to 115 nucleotides, whereas that for introns at around 750 nucleotides. The smaller average intron size relative to humans' was consistent with the trend that intron size is correlated with genome size.

Annotation process

Manual annotation and analysis of the OGS was performed by a group of over 200 international volunteers, primarily from the sea urchin research community. To facilitate and to centralize the annotation efforts, an annotation database and a shared Web browser, Genboree (9), were established at the BCM-HGSC. These tools enabled integrated and collaborative analysis of both precomputed and experimental information (see SOM). A variety of precomputed information for each predicted gene model was made available to the annotators in the browser, including expressed sequence tag (EST) data, the four unmerged gene prediction sets, and transcription data from whole-genome tiling microarray with embryonic RNA (see below) (10). Additional resources available to the community are listed in table S4.

Over 9000 gene models were manually curated by the consortium with 159 novel models (gene models not represented in the OGS) added to the official set. If we assume no bias in the curated gene models, the number of novel models added may imply that the official set contains >98% of the protein-coding genes.

Genome features

A window on the genetic landscape is scaffold-centric in *S. purpuratus*, because linkage and cytogenetic maps are not available. The 36.9% GC content of the genome is uniformly low because assessment of the average GC content by domains is consistent (36.8%), and the distribution is tight (see SOM). Genes from the OGS show no tendency to occupy regions of higher- or lower-than-average GC content. In fact, nearly all genes lie in regions of 35 to 39% GC.

The Echinoderm Genome in the Context of Metazoan Evolution

The sea urchin genetic tool kit lends evolutionary perspective to the gene catalogs that characterize the superclades of the bilaterian animals. The distribution of highly conserved protein domains and sequence motifs provides a view of the expansion and contraction of gene families, as well as an insight into changes in protein function. Examples are enumerated in Table 1, which presents a global overview of gene variety obtained by comparing sequences identified in Interpro, and Table 2, which shows the distribution of specific Pfam database domains associated with selected aspects of cell physiology, including sequences identified in the cnidarian *Nematostella vectensis* (11). The Interpro data suggest that about one-third of the 50 most prevalent domains in the sea urchin gene models are not in the 50 most abundant families in the other representative genomes (mouse, tunicate, fruit fly, and nematode), and thus, they constitute expansions that are specific at least to sea urchins, if not to the complex of echinoderms and hemichordates. Two of the most abundant domains make up 3% of the total and mark genes that are involved in the innate immune response. Others define proteins associated with apoptosis and cell death regulation, as well as proteins that serve as downstream effectors in the Toll–interleukin 1 (IL-1) receptor (TIR) cascade. The quinoprotein amine dehydrogenase domain seen in the sea urchin set is 10 times as abundant as in other representative genomes and may be used in the systems of quinone-containing pigments known to occur in these marine animals. The large number of nucleosomal histone domains found agrees with the long-established sea urchin–specific expansion of histone genes. In summary, the distribution of proteins among these conserved families shows the trend of expansion and shrinkage of the preexisting protein families, rather than frequent gene innovation or loss. Gene family sizes in the sea urchin are more closely correlated with what is seen in deuterostomes than what is seen in the protostomes.

Of equal interest are the sorts of proteins not found in sea urchins. The sea urchin gene set shares with other bilaterian gene models about 4000 domains, whereas 1375 domains from other bilaterian genomes are not found in the sea urchin set. In agreement with the lack of morphological evidence of gap junctions in sea urchins, there are no gap junction proteins (connexins, pannexins, and innexins). Also missing are several protein domains unique to insects, such as insect cuticle protein, chitin-binding protein, and several pheromone- or odorant-binding proteins, as well as a vertebrate invention—the Krüppel-associated box or KRAB domain, a repressor domain in zinc finger transcription factors (12). Finally, searches for specific subfamilies of G protein–coupled receptors (GPCRs) that are known as chemosensory and/or odorant receptors in distinct bilaterian phyla failed to detect clear representatives in the sea urchin genome. However, this failure more likely reflects the independent evolution of these receptors, rather than a lack of chemoreceptive molecules, because the sea urchin genome encodes close to 900 GPCRs of the same superfamily (rhodopsin-type GPCRs), several of which are expressed in sensory structures (13). A conservative way to compare gene sets is to count the strict orthologs that give reciprocal BLAST matches. Genes that are genuine orthologs are likely to yield each other as a best hit. Comparison of sea urchin, fruit fly, nematode, ascidian, mouse, and human gene sets (Fig. 2) indicates that the greatest number of reciprocal best matches is observed between mouse and human, which reflects their close relation. The numbers of presumed orthologous genes between the ascidian and the two mammals are about equal, but are less than the number counted between these species and the sea urchin. The difference is consistent with the lower gene number and reduced genome size in the urochordates (4).

The number of reciprocal pairs for sea urchin and mouse is about 1.5 times the matches between proteins in sea urchin and fruit fly. The number of nematode proteins matching either sea urchin or fruit fly is even lower. This is likely the result of the more rapid sequence changes in the nematode compared with the other species used in this analysis. More than 75% of the genes that are shared by sea urchin and fruit fly are also shared between sea urchin and mouse. Thus, these genes constitute a set of genes common to the bilaterians, whereas the additional sea urchin–mouse pairs are unique to the deuterostomes.

The sea urchin genome consequently provides evidence for the now extremely robust concept of the deuterostome superclade. A 1908 concept that originated in the form of embryos of dissimilar species (14) is demonstrated by genomic comparisons.

Developmental Genomics

In the 1980s, the sea urchin embryo became the focus of cis-regulatory analyses of embryonic gene expression, and there was a great expansion of molecular explorations of the developmental cell biology, signaling interactions, and regulatory control systems of the embryo. Analysis of the entire genome facilitated the first large-scale correlation of the gene regulatory network for development, which represents the genomic control circuitry for specification of the endoderm and mesoderm of this embryo (15–17)with the encoded potential of the sea urchin.

The embryo transcriptome and regulome

Because of indirect development in the sea urchin, embryogenesis is cleanly separated from adult body plan formation, in developmental process and in time, and therefore, it is possible to estimate the genetic repertoire specifically required for formation of a simple embryo (10). Pooled mRNA preparations from four stages of development, up to the mid-late gastrula stage (48 hours), were hybridized with a whole-genome tiling array. Expression of about 12,000 to 13,000 genes, as conservatively assessed, was seen during this early period, indicating that ~52% of the entire protein-coding capacity of the sea urchin genome is

expressed during development to the mid-late gastrula stage. An additional set of microarray experiments extended the interrogation of embryonic expression to the 3-day pluteus larva stage (see SOM) (18).

The DNA binding domains of transcription factor families are conserved across the Bilateria, and these protein domain motifs were used to extract the sea urchin homologs (see SOM). For each identified gene, if data were not already available, probes were built from the genome sequence and used to measure transcript concentration by quantitative polymerase chain reaction with a time series of embryo mRNAs, as well as to determine spatial expression by whole-mount in situ hybridization.

All bilaterian transcription factor families were represented in the sea urchin with a few rare exceptions (see below), so the sea urchin data strongly substantiate the concept of a panbilaterian regulatory tool kit (19) or "regulome." We found that 80% of the whole sea urchin regulome (except the zinc finger genes) was expressed by 48 hours of embryogenesis (20), an even greater genetic investment than the 52% total gene use in the same embryo.

Signal transduction pathways

More than 1200 genes involved in signal transduction were identified. Comparative analysis highlights include the protein kinases that mediate the majority of signaling and coordination of complex pathways in eukaryotes. The *S. purpuratus* genome has 353 protein kinases, intermediate between the core vertebrate set of 510 and the fruit fly and nematode conserved sets of ~230. Fine-scale classification and comparison with annotated kinomes (21, 22) reveals a remarkable parsimony. Indeed, with only 68% of the total number of human kinases, the sea urchin has members of 97% of the human kinase subfamilies, lacking just four of those subfamilies (Axl, FastK, H11, and NKF3),whereas *Drosophila* lacks 20 and nematodes 32 (Fig. 3) (23). Most sea urchin kinase subfamilies have just a single member, although many are expanded in vertebrates; thus, the sea urchin kinome is largely nonredundant. The sea urchin therefore possesses a kinase diversity surprisingly comparable to that of vertebrates without the complexity. A small number of kinases were more similar to insect than to vertebrate homologs (including the Titin homolog Projection, the Syk-like tyrosine kinase Shark, and several guanylate cyclases), which indicated for the first time the loss of kinase classes in vertebrates (23). Expression profiling showed that 87% of the signaling kinases and 80% of the 91 phosphatases were expressed in the embryo (23, 24), which emphasized the importance of signaling pathways in embryonic development.

The small guanosine triphosphatases (GTPases) function as molecular switches in signal transduction, nuclear import and export, lipid metabolism, and vesicle docking. Vertebrate GTPase families were expanded after their divergence from echinoderms, in part by whole-genome duplications (25–27). The sea urchin genome did not undergo a whole-genome duplication, yet phylogenies for four Ras GTPase families (Ras, Rho, Rab, and Arf) revealed that local gene duplications occurred (Fig. 4), which ultimately resulted in a comparable number of monomeric GTPases in the human and sea urchin genomes (28). Thus, expansion of each family in vertebrates and echinoderms was achieved by distinct mechanisms (gene-specific versus whole genome duplication). More than 90% of the small GTPases are expressed during sea urchin embryogenesis, which suggests that the complexity of signaling through GTPases is comparable between sea urchins and vertebrates.

The Wnt family of secreted signaling molecules plays a central role in specification and patterning during embryonic development. Phylogenetic analyses from cnidarian to human indicate that of the 13 known Wnt subfamilies, *S. purpuratus* has 11, missing Wnt2 and Wnt11 homologs (Fig. 5). *S. purpuratus* has WntA, previously reported as being absent

from deuterostomes (29). Of 126 genes described as components of the Wnt signal transduction machinery, homologs of ~90% were present in the sea urchin genome, which indicates a high level of conservation of all three Wnt pathways (30). However, of 94 Wnt transcriptional target genes reported in the literature, mostly from vertebrates (31), only 53% were found with high confidence in the sea urchin genome (Fig. 6). The absent Wnt targets include vertebrate adhesion molecules, which were frequently missing from the sea urchin genome (32), as well as signaling receptors, which are more divergent and thus more difficult to identify. In contrast, most transcription factor targets of the Wnt pathway are present in the genome, which reflects a higher degree of conservation of transcription factor families (20). Taken together, the genomic analysis of signal transduction components indicates that sea urchins have signaling machinery strikingly comparable to that of vertebrates, often without the complexity that arises from genetic redundancy.

Sea Urchin Biology

Analysis of the genome allows understanding of parts of the organism that have not been well studied. Several examples of this follow with further details in the SOM. Additional areas such as intermediary metabolism, metalloproteases, ciliary structure, fertilization, and germline specification are presented in the SOM.

Defense Systems

The need to deal with physical, chemical, and biological challenges in the environment underlies the evolution of an array of defense gene families and pathways. One set of protective mechanisms involves the immune system, which responds to biotic stressors such as pathogens. A second group of genes comprises a chemical “defensome,” anetworkof stress-sensing transcription factors and defense proteins that transform and eliminate many potentially toxic chemicals.

The sea urchin immune system

The sea urchin has a greatly expanded innate immunity repertoire compared with any other animal studied to date (table S5). Three classes of innate receptor proteins are particularly increased (Fig. 7). These make up a vast family of Toll-like receptors (TLRs), a similarly large family of genes that encode NACHT and leucine-rich repeat (LRR)–containing proteins (NLRs), and a set of genes encoding multiple scavenger receptor cysteine-rich (SRCR) domain proteins of a class highly expressed in the sea urchin immune cells or coelomocytes (33, 34). Receptors from each of these families participate in immunity by recognizing nonself molecules that are conserved in pathogens or by responding to self molecules that indicate the presence of infection (35). In contrast, homologs of signal transduction proteins and nuclear factor kappa B (NFκB)/Rel domain transcription factors that are known to function further downstream of these genes were present in numbers similar to those in other invertebrate species. One of the more unexpected findings from our analysis of sea urchin immune genes was the identification of a Rag1/2-like gene cluster (36). The presence of this cluster, along with other recent findings (37), suggested the possibility that these genes had been part of animal genomes for longer than previously considered. Further analysis of the genomic insights into the innate immune system and the underpinnings of vertebrate adaptive immunity can be found in a review in this issue (38).

The complement system

The complement system of vertebrates is a complex array of soluble serum proteins and cellular receptors arranged into three activation pathways (classical, lectin, and alternative) that converge and activate the terminal or lytic pathway. This system opsonizes pathogenic cells for phagocytosis and sometimes activates the terminal pathway, which leads to

pathogen destruction. An invertebrate complement system was first identified in the sea urchin [for reviews, see (39, 40)], and the analysis of the genome sequence presented a more complete picture of this important immune effector system. In chordates, collectins initiate the lectin cascade through members of the mannose-binding protein (MBP)–associated protease (MASP)/C1r/C1s family. Several genes encoding collectins, C1q and MBP, have been predicted (39) and were present in the genome; however, members of the MASP/C1r/C1s family were not identified. There was no evidence for the classical pathway, which links the complement cascade with immunoglobulin recognition in jawed vertebrates. The alternative pathway is initiated by members of the thioester protein family, which, in the sea urchin, was somewhat expanded with four genes. Two of the thioester proteins, SpC3 and SpC3-2, are known to be expressed, respectively, in coelomocytes and in embryos and larvae. Furthermore, there were three homologs of factor B, the second member of the alternative pathway (41).

The terminal complement pathway in vertebrates acts to destroy pathogens or pathogen-infected cells with large pores called membrane attack complexes (MACs). Twenty-eight gene models were identified that encode MAC-perforin domains, but none of these had the additional domains expected for terminal complement factors (C6 through C9). Instead, these are members of a novel and very interesting gene family with perforin-like structure. In vertebrates, perforins carry out cell-killing functions by cytotoxic lymphocytes through the formation of small pores in the cell membranes. If the complement system in the sea urchin functions through multiple lectin and alternative pathways in the absence of the lytic functions of the terminal pathway, the major activity of this system is expected to be opsonization.

Homologs of immune regulatory proteins

Cytokines are key regulators of intercellular communication involving immune cells, acting to coordinate vertebrate immune systems. Genes encoding cytokines and their receptors often evolve at a rapid pace, and most families are known only from vertebrate systems. Although members of many cytokine, chemokine, and receptor families were not identified in the sea urchin genome, a number of important immune signaling homologs were present. These included members of the tumor necrosis factor (TNF) ligand and receptor superfamilies, an IL-1 receptor and accessory proteins, two IL-17 receptor–like genes and 30 IL-17 family ligands, and nine macrophage inhibitory factor (MIF)–like genes. Receptor tyrosine kinases (RTKs) included those that bind important growth factors that regulate cell proliferation in vertebrate hematopoietic systems. Of particular note, from the sea urchin genome, were two vascular endothelial growth factor (VEGF) receptor–like genes and a Tie1/2 receptor, all of which were expressed in adult coelomocytes. Many of these genes are homologs of important inflammatory regulators and growth factors in higher vertebrates, and these sea urchin homologs may have similar functions in regulating coelomocyte differentiation and recruitment.

Representatives of nearly all subclasses of important vertebrate hematopoietic and immune transcription factors were present in the sea urchin genome. Notably, the genome contained homologs of immune transcription factors that had not been identified previously outside of chordates, including PU.1/SpiB/SpiC, a member of the Ets subfamily, and a zinc finger gene with similarity to the Ikaros subfamily. Transcript prevalence measurements showed that PU.1, the Ikaros-like gene and homologs of Gata1/2/3, E2A/HEB/ITF2, and Stem Cell Leukemia (SCL) were all expressed at substantial levels in coelomocytes (41). This was consistent with the presence of conserved mechanisms of regulating gene expression among sea urchin coelomocytes and vertebrate blood cells.

ABC transporters

Many chemicals are removed from cells by efflux proteins known as ATP-binding cassette (ABC) or multidrug efflux transporters. *S. purpuratus* has 65 ABC transporter genes in the eight major subfamilies of these genes [ABC A to H; (42)]. The ABCC family of multidrug transporters is about 25% larger than in other deuterostome genomes with at least 30 genes in this family (nearly half of the sea urchin ABC transporters), and 25 of these 30 genes showed substantial mRNA expression in eggs, embryos, or larvae. Much of the expansion is in the Sp-ABCC5 and Sp-ABCC9 families, whereas orthologs of the vertebrate gene ABCC2 (also called MRP2) are absent. Because the ABCC family is known to generally transport more hydrophilic compounds than other transporter families, such as the ABCB genes, sea urchins may have increased need for transport of these compounds. ABCC efflux activity has been described in sea urchin embryos and, consistent with the genomic expansion of the ABCC family, the major activity in early embryos ensues from an ABCC-like efflux mechanism.

Cytochrome P-450 monooxygenase (CYP)

Enzymes in the CYP1, CYP2, CYP3, and CYP4 families carry out oxidative biotransformation of chemicals to more hydrophilic products. The sea urchin has 120 CYP genes, and those related to CYP gene families 1 to 4 constitute 80% of the total, which suggests that there has been selective pressure to expand functionality in these gene families (42). Eleven CYP1-like genes are present in the sea urchin genome, more than twice the number in chordates. CYP2-like and CYP3-like genes are also present at greater numbers than in other deuterostomes. In addition to the CYPs in families 1 to 4, the sea urchin genome contains homologs of proteins involved in developmental patterning (CYP26), cholesterol synthesis (CYP51), and metabolism (CYP27, CYP46). Homologs of some CYPs with endogenous functions in vertebrates were not found; however, (CYP19, androgen aromatase; CYP8, prostacyclin synthase; CYP11, pregnenolone synthase; CYP7, cholesterol-7α-hydroxylase). These CYP genes in concert with additional expanded defensive gene families represent a large diversification of defense gene families by the sea urchin relative to mammals (42).

Oxidative defense and metal-complexing proteins

The metal-complexing proteins include three metallothionein genes and three homologs of phytochelatin synthase genes. Genes for antioxidant proteins include three superoxide dismutase (SOD) genes and a gene encoding ovoperoxidase (an unusual peroxidase with SOD-like activity), along with one catalase, four glutathione peroxidase, and at least three thioredoxin peroxidase genes. Reactive oxygen detoxification genes may be important in conferring the long life-span of sea urchins, because oxidative damage is thought to be a major factor in aging.

Diversity and conservation in xenobiotic signaling

The diversity of genes encoding xenobiotic-sensing transcription factors that regulate biotransformation enzymes and transporters was similar to other invertebrate genomes, but in most cases lower than vertebrates. For example, the sea urchin genome encoded a single predicted CNC-bZIP protein homologous to the four human CNC-bZIP proteins involved in the response to oxidative stress. There were two sea urchin homologs of the aryl hydrocarbon receptor (AHR), which in vertebrates mediates the transcriptional response to polynuclear and halogenated aromatic hydrocarbons and, in both protostomes and deuterostomes, also regulates specific developmental processes (43–45). One of the sea urchin AHR homologs was more closely related to the vertebrate AHR; the other shared greatest sequence identity with the *Drosophila* AHR homolog *spineless*. Sea urchins also

had two genes encoding hypoxia-inducible factors (HIFα subunits), which regulate adaptive responses to hypoxia, and a gene encoding ARNT, a PAS protein that is a dimerization partner for both AHRs and HIFs.

Strongylocentrotus purpuratus has 32 nuclear receptor (NR) genes (20), two-thirds the number in humans, including several with potential roles in chemical defense (42). The sea urchin genome also contains two peroxisome proliferator–activated receptor (PPAR, NR1C) homologs and an NR1H gene coorthologous to both liver X receptor (LXR) and farnesoid X receptor (FXR) (42). Genes homologous to the vertebrate xenobiotic sensor NR1I genes [pregnane X receptor, PXR; constitutive androstane receptor, CAR (46)] are absent, although three NR1H-related genes were found, which possibly form a new subfamily of genes involved in xenobiotic sensing.

Many of the defense genes are expressed during development (10, 42), which suggests that they have dual roles in chemical defense and in developmental signaling. In several cases (CYPs, AHR, NF-E2), the evolution of pathways for chemical defense may have involved recruitment from developmental signaling pathways (42).

Nervous System

The echinoderm nervous system is the least well studied of all the major metazoan phyla. For a number of technical reasons, the structure and function of echinoderm nerves have been neglected. Analysis of the sea urchin genome has enabled an unprecedented glimpse into the neural and sensory functions and has revealed several novel molecular approaches to the study of echinoderm nervous systems (Table 3).

The nervous systems of echinoderm larvae and adults are dispersed, but they are not simple nerve nets. This organization differs from both vertebrates, which do not have a dispersed nervous system, and hemichordates, which do have nerve nets (47). Adult sea urchins have thousands of appendages, each with sensory neurons, ganglia, and motor neurons arranged in local reflex arcs. These peripheral appendages are connected to each other and to radial nerves, which provide overall control and coordination (47, 48).

Nearly all of the genes encoding known neurogenic transcription factors are present in the sea urchin genome, and several are expressed in neurogenic domains before gastrulation, which indicates that they may operate near the top of a conserved neural gene regulatory network (47). Axon guidance molecules known from other metazoans are also expressed in the developing embryo. Unexpectedly, genes encoding the neurotrophin-Trk receptor system that were thought to be vertebrate-specific because they were not found in *Ciona*, are present in sea urchin, which suggests a deuterostome origin and a potential loss in urochordates.

The genes required to construct neurons and to transmit signals are present, but the repertoire of neural genes and the initial characterization of expression of a number of them led to unexpected and surprising conclusions. There appear to be no genes encoding gap junction proteins, which suggests that communication among neurons depends on chemical synapses without ionic coupling. The repertoire of sea urchin neurotransmitters is large, but melatonin and adrenalin are lacking, as they are in ascidians (4, 47). Cannabinoid, lysophospholipid, and melanocortin receptors are not present in urchins, but orthologs were found in ascidians (4, 47). In contrast, some sets of genes thought to be chordate-specific have sea urchin orthologs, for example, insulin and insulin-like growth factors (IGFs) that are more similar to their chordate counterparts than those of other invertebrates (47). Overall, the genome contains representatives of all five large superfamilies of GPCRs, including those that mediate signals from neuropeptides and peptide hormones. Both the secretin and rhodopsin superfamilies display marked lineage-specific expansions (13, 47).

Sensory systems

There were 200 to 700 putative chemosensory genes that formed large clusters and lacked introns, which are features of chemosensory genes in vertebrates, but not in *Caenorhabditis elegans* and *Drosophila melanogaster*. Many of these genes encoded amino acid motifs that were characteristic of vertebrate chemosensory and odorant receptors (13, 47). Sea urchins had an elaborate collection of photoreceptor genes that quite surprisingly appeared to be expressed in tube feet (13, 47). These included many genes encoding transcription factors regulating retinal development and a photorhodopsin gene.

Human Usher syndromes are genetic diseases affecting hearing, balance, and retinitis pigmentosa (retinal photoreceptor degeneration). Most of the genes involved have been identified, and they encode a set of membrane and cytoskeletal proteins that form an interacting network that controls the arrangement of mechanosensory stereocilia in hair cells of the mammalian ear. Many or all of the proteins play some roles in photoreceptor organization and/or maintenance. Orthologs of virtually the entire set of membrane and cytoskeletal proteins of the Usher syndrome network were found in the sea urchin genome. These include the very large membrane proteins, usherin and VLGR-1 and large cadherins (Cadh23 and possibly Pcad15), all of which participate in forming links between stereocilia in mammalian hair cells, as well as myosin 7 and 15, two PDZ proteins (harmonin and whirlin) and another adaptor protein (SANS), which participate in linking these membrane proteins to the cytoskeleton. In addition, two membrane transporters, NBC (a candidate Usher syndrome target known to interact with harmonin) and TrpA1 (the mechanosensory channel connected to the tip links containing cadherin 23), have orthologs in the sea urchin genome. Sea urchins do not have ears or eyes, so they must deploy these proteins in other sensory processes. Sea urchins respond to light, touch, and displacement and probably use some of same sensory genes used by vertebrates.

The Echinoderm Adhesome

The *S. purpuratus* genome contained representatives of all the standard metazoan adhesion receptors (table S7), but the emphasis on different classes of receptors differed substantially from that used by vertebrates. The integrin family was intermediate in size between those of protostomes and vertebrates—several chordate-specific expansions of the integrin repertoire were absent, and there were some expansions unique (so far) to echinoderms. The cadherin repertoire was also small relative to vertebrates (a dozen or so instead of over a hundred), and many chordate-specific expansions were missing. Specialized large cadherins shared by protostomes and vertebrates were present, as well as some specialized large cadherins previously thought to be chordate-specific, but overall, the cadherin repertoire was more invertebrate than vertebrate in character. Sea urchins lacked the integrins and cadherins that link to intermediate filaments in vertebrates.

In contrast, sea urchins had large repertoires of adhesion molecules containing immunoglobulin superfamily, fibronectin type 3 repeat (FN3), epidermal growth factor (EGF), and LRR repeats. In addition to the expansion of TLRs and NLRs mentioned above, there are large expansions of other LRR receptor families, including GPCRs (32). The key neural adhesion systems involved in regulating axonal outgrowth were present (netrin/Unc5/DCC; Slit/Robo; and semaphorins/plexins), as were adhesion molecules involved in synaptogenesis (Agrin/MUSK; and neurexin/neuroligins). This was not surprising because these molecules were known in both protostomes and vertebrates. However, structurally, the synapses of echinoderms are unusual because there are no direct synaptic contacts (49). Some of them were expressed in sea urchin embryos before there are any neurons, suggesting that they may have other roles as well.

The basic metazoan basement membrane extracellular matrix (ECM) tool kit was present—two alpha-IV collagen genes, perlecan, laminin subunits, nidogen, and collagen XV/XVIII. There did not appear to be much, if any, expansion of these gene families, as is found in vertebrates, which suggests that there is less diversity among basement membranes. Quite a few ECM proteins present in chordates, but not protostomes, were also missing in sea urchins, including fibronectins, tenascins, von Willebrand factor, vitronectin, most vertebrate-type matrix proteoglycans, and complex VWA/FN3 collagens among others (32). Absence of these genes may be related to the absences of neural crest migration, a high shear endothelial-lined vasculature and, of course, cartilage and bone.

In addition to the components of Usher syndromes mentioned above, it was surprising to find a clear ortholog of reelin, a large ECM protein involved in establishing the layered organization of neurons in the vertebrate cerebral cortex. Reelin is mutated in the *reeler* mouse, and mutations in the *reeler* gene in humans have been associated with Norman-Roberts-type lissencephaly syndrome. Reelin has a unique domain composition and organization (Reeler, EGF, BNR) that has not been found outside chordates, but the sea urchin genome included a very good homolog of reelin. Receptors for reelin are believed to include low-density lipoprotein receptor–related proteins (LRPs), and there are a number of these receptors in *S. purpuratus* although it is as yet unclear whether they are reelin receptors, lipoprotein receptors, or something else. Similar receptors are also involved in human disease (atherosclerosis).

Biomineralization Genes

Among the deuterostomes, only echinoderms and vertebrates produce extensive skeletons. The possible evolutionary relations between biomineralization processes in these two groups have been controversial. Analysis of the *S. purpuratus* genome revealed major differences in the proteins that mediate biomineralization in echinoderms and vertebrates (50). First, there were few sea urchin counterparts of extracellular proteins that mediate biomineral deposition in vertebrates. For example, in vertebrates, an important class of proteins involved in biomineralization is the family of secreted, calcium-binding phosphoproteins, or SCPPs. Sea urchins did not have counterparts of SCPP genes, which supports the hypothesis that this family arose via a series of gene duplications after the echinoderm-chordate divergence (51). Second, almost all of the proteins that have been directly implicated in the control of biomineralization in sea urchins were specific to that clade. The echinoderm skeleton consists of magnesium calcite (as distinct from the calcium phosphate skeletons of vertebrates) in which is occluded many secreted matrix proteins. The sea urchin spicule matrix proteins were encoded by a family of 16 genes that are organized in small clusters and likely proliferated by gene duplication. Counterparts of sea urchin spicule matrix genes were not found in vertebrates, amphioxus, or ascidians. Likewise, other genes that have been implicated in biomineralization in sea urchins, including genes that encode the transmembrane protein P16 and MSP130, a glycosylphosphatidylinositol-linked glycoprotein, were members of small clusters of closely related genes without apparent homologs in other deuterostomes. The members of all three of these sea urchin–specific gene families were expressed specifically by the biomineral-forming cells of the embryo, the primary mesenchyme cells [see (50)]. As a whole, these findings highlighted substantial differences in the primary sequences of the proteins that mediate biomineralization in echinoderms and vertebrates.

Cytoskeletal genes

In addition to identifying genes for all previously known *S. purpuratus* actins and tubulins, one δ- and two ε-tubulin genes were found (52). Newly identified motor protein genes include members of four more classes of myosin, and eight more families of kinesins. The

first dynein cloned and sequenced was from sea urchin, and although most *S. purpuratus* dynein heavy chain genes mapped one-to-one to mammalian homologs, Sp-DNAH9 mapped one-to-three, as it was equidistant between the closely similar mammalian genes DNAH9, DNA11, and DNAH17 (52).

Conclusions

Our estimate of 23,300 genes is similar to estimates for vertebrates, despite the fact that two whole-genome duplications are believed to have occurred in the chordate lineage after divergence from the lineage leading to the echinoderms (25–27). From the analysis presented here, it seems likely that many mechanisms shaped the final genetic content of these genomes. On the one hand, there are cases of gene families that are expanded in vertebrates compared with sea urchin, including examples of the expected 4:1 ratio from two duplications (15). However other patterns are also found. The nuclear receptor family is only slightly reduced in sea urchin compared with that of humans, which suggests gene loss followed the vertebrate duplications. The unprecedented expansions of innate immune system diversity contrast sharply with the much smaller sets of counterparts that are present in the sequenced genomes of protostomes, *Ciona*, and vertebrates, an example of independent expansion in the sea urchin, whereas the GTPases described here have expanded in sea urchin to about the same numbers as in vertebrates. Thus, whereas the duplications of the chordate lineage were a contributor to the increased complexity of vertebrates, regional expansions clearly play a large role in the evolution of these animals.

The refinement of the inventory of vertebrate-specific or protostome-specific genes likewise benefits from the sea urchin genome. Many more human genes have shared ancestry across the deuterostomes, and in fact, bilaterian genes are more broadly shared than had been inferred from comparison of the previously limited genome sequences. The new biological niche sampled by the sea urchin genome provides not only a clearer view of the deuterostome and bilaterian ancestor, but has also provided a number of surprises. The finding of sea urchin homologs for sensory proteins related to vision and hearing in humans may lead to interesting new concepts of perception, and the extraordinary organization of the sea urchin immune system is different from any animal yet studied. From a practical standpoint, the sea urchin may be a treasure trove. Because of the many pathways shared by sea urchin and human, the sea urchin genome includes a large number of human disease gene orthologs. Many of the genes described in the preceding sections fall into this category (see tables S8 and S9) and cover a surprising diversity of systems such as nervous, endocrine, and blood systems, as well as muscle and skeleton, as exemplified by the Huntington and muscular dystrophy genes. Continued exploration of the sea urchin immune system is expected to uncover additional variations for protection against pathogens. The immense diversity of pathogen-binding motifs encoded in the sea urchin genome provides an invaluable resource for antimicrobial applications and the identification of new deuterostome immune functions with direct relevance to human health. These exciting possibilities show that much biodiversity is yet to be uncovered by sampling additional evolutionary branches of the tree of life.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

54. We gratefully acknowledge the following support: BCM-HGSC, National Human Genome Research Institute (NIH) grant 5 U54 HG003273; Naples Workshop, Stazione Zoologica Naples and the Network of Excellence “Marine Genomics Europe” (GOCE-04-505403); M. Elphick, Biotechnology and Biological Sciences Research

Council (BBSRC), UK, grant S19916; J. Rast laboratory, Natural Sciences and Engineering Research Council (NSERC) of Canada, Canadian Institutes of Health Research (CIHR), and the Uehara Memorial Foundation; J. A. Coffman, **Mount Desert Island Biological Laboratory** (MDIBL), NIH grant GM070840; M. C. Thorndyke, K. H. Wilson, F. Hallböök, R. P. Olinski, Swedish Science Research Council, Network of Excellence Marine Genomics Europe (GOCE-04-505403), European Union Research Training Networks FP5 Trophic Neurogenome HPRN-ct-2002-00263, and the Royal Swedish Academy of Sciences, STINT; E. H. Davidson, R. A. Cameron, the Center for Computational Regulatory Genomics (E. H. Davidson, principal investigator) was supported by the NIH grant RR-15044, NSF IOB-0212869, and the Beckman Institute; also, support for the E. H. Davidson laboratory is from NIH grants HD-37105 and GM61005 and U.S. Department of Energy (DOE) grant DE-FG02-03ER63584; P. Oliveri, Camilla Chandler Frost Fellowship; G. M. Wessel laboratory supported by NSF IOB-0620607 and NIH grant R01 HD028152; B. Brandhorst, K. Bergeron, and N. Chen, NSERC; K. R. Foltz, NSF, IBN-0415581; M. Hahn, NIH grant R01ES006272; D. Burgess, NIH grant GM058231; L. C. Smith, NSF (MCB-0424235); R. O. Hynes, Howard Hughes Medical Institute and National Cancer Institute (NCI) (MIT Cancer Center core grant P30-CA14051); D. McClay, NIH grants GM61464, HD039948, and HD14483; V. D. Vacquier (group leader), G. W. Moy, H. J. Gunaratne, M. Kinukawa, M. Nomura, A. T. Neill, and Y.-H. Su, NIH grant R37-HD12896; R. D. Burke, NSERC and CIHR; L. M. Angerer, National Institute of Dental and Craniofacial Research (NIDCR), R. C. Angerer (NIDCR), Z. Wei (NIDCR), G. Humphrey, National Institute of Child Health and Human Development (NICHD), M. Landrum, National Center for Biotechnology Information (NCBI), O. Ermolaeva (NCBI), P. Kitts (NCBI), K. Pruitt (NCBI), V. Sapojnikov (NCBI), A. Souvorov (NCBI), W. Hiavina (NCBI), S. Fugmann, National Institute on Aging (NIA), M. Dean, National Cancer Institute–Frederick (NCIFCRF) Intramural Research Program of the NIH; P. Cormier, Association pour la Recherche contre la Cancer (ARC), France, grants 4247 and 3507 to P.C., Ligue Nationale contre le Cancer to P.C., Conseil Régional de Bretagne and Conseil Général du Finistère; W. H. Klein, National Eye Institute, NIH grant EY11930, NICHD HD66219, and the Robert A. Welch Foundation (G-0010); N. Adams, NSF grant IBN 0417003 and the Department of the Navy, Office of Naval Research, under Award N00014-05-1-0855; D. Epel, NSF 0417225; A. Hamdoun, F32-HD47136; C. Byrum, American Heart Association grant 0420074Z; K. Walton, U.S. Army Medical Research and Materiel Command grant W81XWH-04-1-0324; J. Stegeman, NIH 2P42 ESO7381; and J. Goldstone, NIH F32 ESO12794.

References and Notes

1. Gibbs RA, et al. Nature. 2004; 428:493. [PubMed: 15057822]
2. Cai WW, Chen R, Gibbs RA, Bradley A. Genome Res. 2001; 11:1619. [PubMed: 11591638]
3. Britten RJ, Cetta A, Davidson EH. Cell. 1978; 15:1175. [PubMed: 728997]
4. Dehal P, et al. Science. 2002; 298:2157. [PubMed: 12481130]
5. Vinson JP, et al. Genome Res. 2005; 15:1127. [PubMed: 16077012]
6. Hinegardner RT. Anal. Biochem. 1971; 39:197. [PubMed: 5544593]
7. Elsik CG, Mackey AJ, Reese JT, Milshina NV, Roos DS, Weinstock GM. Genome Biol. in press.
8. Sea Urchin Genome Project. (http://sugp.caltech.edu/resources/annotation.php)
9. Genboree. (www.genboree.org)
10. Samanta M, et al. Science. 2006; 314:960. [PubMed: 17095694]
11. Sullivan JC, et al. Nucleic Acids Res. 2006; 34:D495. [PubMed: 16381919]
12. Materna SC, Howard-Ashby M, Gray RF, Davidson EH. Dev. Biol. 10.1016/j.ydbio.2006.08.032, in press.
13. Raible F, et al. Dev. Biol. in press.
14. Grobben K. Verh. Zool. Bot. Ges. Wien. 1908; 58:491.
15. Davidson, EH. Gene Regulatory Networks in Development and Evolution. Academic Press/ Elsevier; San Diego, CA: 2006.
16. Davidson EH, et al. Science. 2002; 295:1669. [PubMed: 11872831]
17. Davidson EH, et al. Dev. Biol. 2002; 246:162. [PubMed: 12027441]
18. Wei Z, Angerer RC, Angerer LM. Dev. Biol. 10.1016/j.ydbio.2006.08.034, in press.
19. Erwin DH, Davidson EH. Development. 2002; 129:3021. [PubMed: 12070079]
20. Howard-Ashby M, Brown CT, Materna SC, Chen L, Davidson EH. Dev. Biol. in press.
21. Manning G, Plowman GD, Hunter T, Sudarsanam S. Trends Biochem. Sci. 2002; 27:514. [PubMed: 12368087]
22. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. Science. 2002; 298:1912. [PubMed: 12471243]
23. Bradham C, et al. Dev. Biol. 10.1016/j.ydbio.2006.08.074, in press.

24. Byrum C, et al. Dev. Biol. in press.
25. Dehal P, Boore JL. PLoS Biol. 2005; 3:e314. [PubMed: 16128622]
26. Gu X, Wang Y, Gu J. Nat. Genet. 2002; 31:205. [PubMed: 12032571]
27. McLysaght A, Hokamp K, Wolfe KH. Nat. Genet. 2002; 31:200. [PubMed: 12032567]
28. Beane W, Voronina E, Wessel GM, McClay DR. Dev. Biol. 10.1016/j.ydbio.2006.08.046, in press.
29. Kusserow A, et al. Nature. 2005; 433:156. [PubMed: 15650739]
30. Croce J, et al. Dev. Biol. in press.
31. The Wnt homepage. (www.stanford.edu/~rnusse/wntwindow.html)
32. Whittaker CA, et al. Dev. Biol. 10.1016/j.ydbio.2006.07.044, in press.
33. Pancer Z. Proc. Natl. Acad. Sci. U.S.A. 2000; 97:13156. [PubMed: 11069281]
34. Pancer Z, Rast JP, Davidson EH. Immunogenetics. 1999; 49:773. [PubMed: 10398804]
35. Akira S, Uematsu S, Takeuchi O. Cell. 2006; 124:783. [PubMed: 16497588]
36. Fugmann SD, Messier C, Novack LA, Cameron RA, Rast JP. Proc. Natl. Acad. Sci. U.S.A. 2006; 103:3728. [PubMed: 16505374]
37. Kapitonov VV, Jurka J. PLoS Biol. 2005; 3:e181. [PubMed: 15898832]
38. Rast JP, et al. Science. 2006; 314:952. [PubMed: 17095692]
39. Smith LC, Azumi K, Nonaka M. Immunopharmacology. 1999; 42:107. [PubMed: 10408372]
40. Smith LC, Clow LA, Terwilliger DP. Immunol. Rev. 2001; 180:16. [PubMed: 11414357]
41. Hibino T, et al. Dev. Biol. 10.1016/j.ydbio.2006.08.065, in press.
42. Goldstone J, et al. Dev. Biol. 10.1016/j.ydbio.2006.08.066, in press.
43. Qin H, Powell-Coffman JA. Dev. Biol. 2004; 270:64. [PubMed: 15136141]
44. Walisser JA, Glover E, Pande K, Liss AL, Bradfield CA. Proc. Natl. Acad. Sci. U.S.A. 2005; 102:17858. [PubMed: 16301529]
45. Duncan DM, Burgess EA, Duncan I. Genes Dev. 1998; 12:1290. [PubMed: 9573046]
46. Xie W, Evans RM. J. Biol. Chem. 2001; 276:37739. [PubMed: 11459851]
47. Burke RD, et al. Dev. Biol. 10.1016/j.ydbio.2006.08.007, in press.
48. Cobb, JLS. Nervous Systems of Invertebrates. Ali, MA., editor. Plenum; New York: 1987. p. 483-525.
49. Cobb JL, Pantreath VW. Tissue Cell. 1977; 9:125. [PubMed: 898171]
50. Livingston BT, et al. Dev. Biol. 10.1016/j.ydbio.2006.07.047, in press.
51. Kawasaki K, Suzuki T, Weiss KM. Proc. Natl. Acad. Sci. U.S.A. 2004; 101:11356. [PubMed: 15272073]
52. Morris RL, et al. Dev. Biol. 10.1016/j.ydbio.2006.08.052, in press.
53. Koonin EV, Aravind L. Cell Death Differ. 2002; 9:394. [PubMed: 11965492]

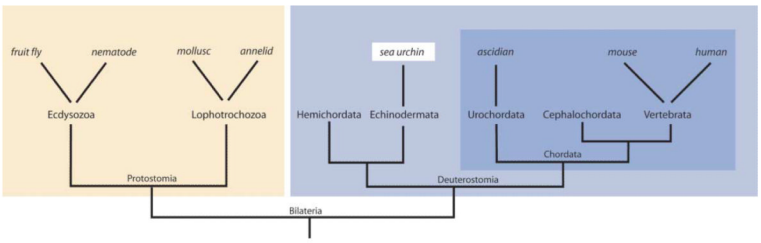


Fig. 1.

The phylogenetic position of the sea urchin relative to other model systems and humans. The chordates are shown on the darker blue background overlapping the deuterostomes as a whole on a lighter blue background. Organisms for which genome projects have been initiated or finished are shown across the top.

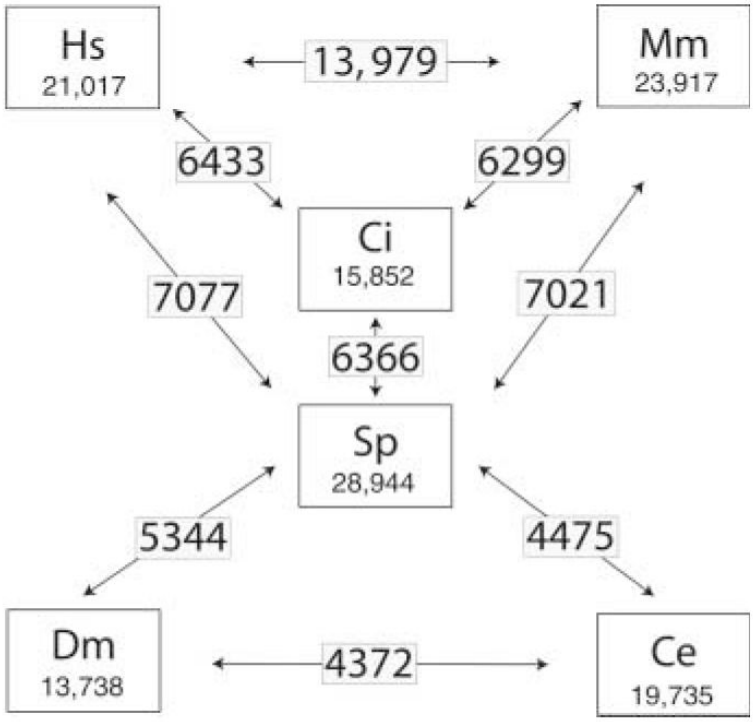


Fig. 2.

Orthologs among the Bilateria. The number of 1:1 orthologs captured by BLAST alignments at a match value of $e = 1 \times 10^{-6}$ in comparisons of sequenced genomes among the Bilateria. The number of orthologs is indicated in the boxes along the arrows, and the total number of International Protein Index database sequences is shown under the species symbol. Hs, *Homo sapiens*; Mm, *Mus musculus*; Ci, *Ciona intestinalis*; Sp, *S. purpuratus*; Dm, *Drosophila melanogaster*; Ce, *Caenorhabditis elegans*.

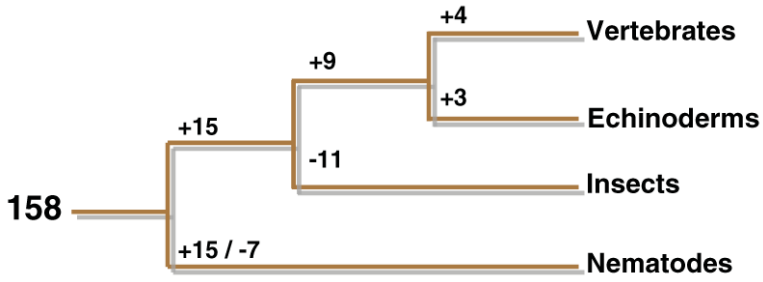


Fig. 3.

Protein kinase evolution: Invention and loss of protein kinase subfamilies in metazoan lineages. Deuterostomes share 9 protein kinase subfamilies absent from *C. elegans* and *Drosophila*, and the sea urchin has not lost any of the 158 metazoan primordial kinase classes, unlike insects or nematodes. [From (23)]

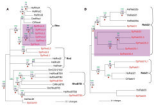
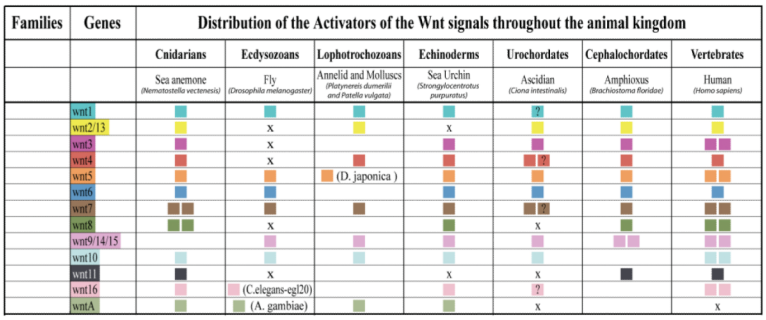


Fig. 4.

Partial phylogenies of the Rho (**A**) and the Rab families (**B**)ofsmallGTPases. The pink boxes highlight gene-specific duplications that increased sea urchin GTPase numbers, resulting in a complexity comparable to vertebrates. Numbers at each junction represent confidence values obtained via three independent phylogenetic methods [neighbor-joining (green), maximum parsimony (blue), and Bayesian (black)]; red stars indicate nodes retained by maximum likelihood. [From (28)]



Families	Genes	Distribution of the Activators of the Wnt signals throughout the animal kingdom						
		Cnidarians	Ecdysozoans	Lophotrochozoans	Echinoderms	Urochordates	Cephalochordates	Vertebrates
		Sea anemone *(Nematostella vectensis)*	Fly *(Drosophila melanogaster)*	Annelid and Molluscs *(Platynereis dumerilii and Patella vulgata)*	Sea Urchin *(Strongylocentrotus purpuratus)*	Ascidian *(Ciona intestinalis)*	Amphioxus *(Brachiostoma floridae)*	Human *(Homo sapiens)*
	wnt1					?		
	wnt2/13		x		x			
	wnt3		x					
	wnt4		x			?		
	wnt5			(D. japonica)				
	wnt6							
	wnt7					?		
	wnt8		x			x		
	wnt9/14/15							
	wnt10							
	wnt11		x		x	x		
	wnt16		(C.elegans-egl20)			?		
	wntA		(A. gambiae)			x		x

Fig. 5.

Survey of the Wnt family of secreted signaling molecules in selected metazoans. Each square indicates a single Wnt gene identified either through genome analyses or independent studies, and squares with a question mark indicate uncertainty of the orthology. Letter X's represent absence of members of that subfamily in the corresponding annotated genome; empty spaces have been left for species for which genomic databases are not yet available. [From (30)]

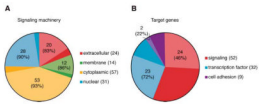
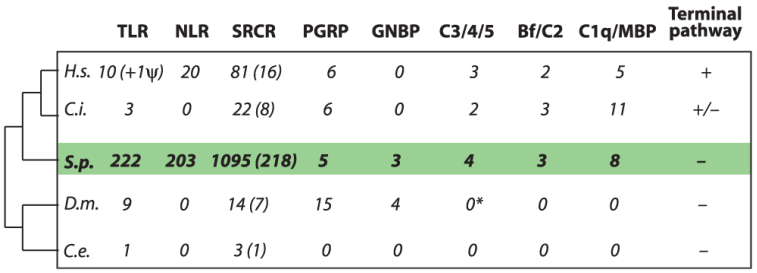


Fig. 6.

Presence of Wnt signaling machinery components (**A**) and target genes (**B**) in the *S. purpuratus* genome. (A) The 126 genes involved in the transduction of the Wnt signals have been separated into four categories from the extracellular compartment to the nucleus. Sea urchin homologs are identified by the lighter shade (indicated by both the number and the percentage of homologs that were identified within the chart); the total number of known genes is indicated in the chart legend. (B) The 93 reported Wnt targets have been divided into three categories: signaling molecules, transcription factors, and cell adhesion molecules. Colors and numbers are as in (A).



	TLR	NLR	SRCR	PGRP	GNBP	C3/4/5	Bf/C2	C1q/MBP	Terminal pathway
H.s.	*10 (+1ψ)*	*20*	*81 (16)*	*6*	*0*	*3*	*2*	*5*	*+*
C.i.	*3*	*0*	*22 (8)*	*6*	*0*	*2*	*3*	*11*	*+/–*
S.p.	***222***	***203***	***1095 (218)***	***5***	***3***	***4***	***3***	***8***	***–***
D.m.	*9*	*0*	*14 (7)*	*15*	*4*	*0**	*0*	*0*	*–*
C.e.	*1*	*0*	*3 (1)*	*0*	*0*	*0*	*0*	*0*	*–*

Fig. 7.

Gene families encoding important innate immune receptors and complement factors in animals with sequenced genomes. For some key receptor classes, gene numbers in the sea urchin exceeds other animals by more than an order of magnitude. Representative animals include *H.s.*, *Homo sapiens*; *C.i.*, *Ciona intestinalis*; *S.p. Strongylocentrotus purpuratus*; *D.m. Drosophila melanogaster*; and *C.e. Caenorhabditis elegans*. Indicated gene families include TLR, toll-like receptors; NLR, NACHT and leucine-rich repeat (LRR) domain–containing proteins similar to the vertebrate Nod/NALP genes; SRCR, Scavenger receptor cysteine-rich domain genes; PGRP, peptidoglycan recognition protein domain genes; and GNBP, Gram-negative binding proteins. C3/4/5, thioester proteins homologous to vertebrate C3, C4, and C5; Bf/C2, complement factors homologous to vertebrate C2 and factor B; C1q/MBP, homologs of vertebrate lectin pathway receptors; and Terminal pathway, homologs of vertebrate C6, C7, C8, and C9. SRCR gene statistics are given as domain number/gene number for multiple SRCR-containing proteins (numbers for *C. intestinalis* includes all SRCR proteins). Asterisk in the *D. melanogaster* C3/4/5 column is meant to denote the presence of related thioester genes (TEPs) and a true C3/4/5 homolog from another arthropod. +/− for *C. intestinalis* Terminal pathway column indicates the presence of genes with similarity to C6 only (Nonaka and Yoshizaki 2004). Phylogenetic relations among species are indicated by a cladogram at the left.

Table 1

Unique aspects of gene family distribution in sea urchin: Selected examples of the frequency of Interpro domains in the proteome of selected species. ID is the identification number used in the INTERPRO database; the second column shows the name given to the domain or motif family in the database. Species abbreviations: Sp, *Strongylocentrotus purpuratus*; Mm, *Mus musculus*; Ci, *Ciona intestinalis*; Dm, *Drosophila melanogaster*; Ce, *Caenorhabditis elegans*.

ID	Name	Species, total number (percentage of total matches)				
		Sp	Mm	Ci	Dm	Ce
IPR001190	Speract/scavenger receptor	361 (1.79)	14 (0.08)	1 (0.01)	2 (0.02)	0 (0.00)
IPR000157	TIR	248 (1.23)	22 (0.12)	9 (0.09)	9 (0.09)	2 (0.02)
IPR011029	DEATH-like	172 (0.85)	8 (0.05)	19 (0.18)	5 (0.05)	1 (0.01)
IPR007111	NACHT nucleoside triphosphatase	135 (0.67)	16 (0.09)	28 (0.27)	0 (0.00)	0 (0.00)
IPR011044	Quinoprotein amine dehydrogenase, β chain–like	122 (0.60)	7 (0.04)	15 (0.15)	5 (0.05)	6 (0.05)
IPR000558	Histone H2B	110 (0.54)	14 (0.08)	2 (0.02)	100 (1.00)	17 (0.13)
IPR001951	Histone H4	93 (0.46)	7 (0.04)	0 (0.00)	101 (1.01)	16 (0.12)
IPR002119	Histone H2A	87 (0.43)	24 (0.14)	2 (0.02)	104 (1.04)	19 (0.14)
IPR008042	Retrotransposon, Pao	76 (0.38)	0 (0.00)	0 (0.00)	0 (0.00)	6 (0.05)
IPR000164	Histone H3	72 (0.36)	17 (0.10)	5 (0.05)	103 (1.03)	22 (0.17)

Table 2

Distribution among sequenced animal genomes of various Pfam domains associated with selected aspects of eukaryotic cell physiology. In *S. purpuratus*, the number of annotated genes is listed; the number in parentheses is the total number of models (including ones that were not annotated) predicted to contain the Pfam domain. For *Nematostella veciensis* (Nv), numbers were obtained by searching Stellabase (11).

Process	Domain	PFAM no.	Sp	Hs	Dm	Ce	Nv
Cell cycle control	Cyclin_N	PF00134	15 (17)	21	11	7	7
	Cyclin_C	PF02984	7 (8)	12	4	5	4
	E2F_TDP	PF02319	3 (5)	11	3	4	3
	RB_A	PF01858	2	3	2	1	0
	RB_B	PF01857	2	3	2	1	0
	P53	PF00870	1	3	1	1	0
	Cullin	PF00888	7	9	8	7	4
	Skp1	PF01466	1	3	5	21	1
Histone metabolism	Histone*	PF00125	49	75	8	?	?
	Linker histone*	PF00538	5	8	2	?	?
	Nucleo-plasmin	PF03066	2	5	2	0	1
	NAP	PF00956	2	24	4	2	0
	HDAC	PF00850	8	11	5	8	3
	D0T1	PF08123	1	1	1	6	1
RNA metabolism	RRM_1	PF00076	140 (178)	245	126	99	41
	TUDOR	PF00567	15	13	15	8	7
	DEAD	PF00270	93 (125)	78	56	65	27
	LSM	PF01423	17	21	17	18	4
	KH-1	PF00013	28 (31)	36	28	28	5
	DSRM	PF00035	14 (15)	21	14	13	8
	3′-5′-Exo-nuclease	PF01612	13 (15)	5	5	9	5
	Exonuc_X-T	PF00929	9 (11)	15	7	10	5
Apoptosis†	Caspase	PF00656	31 (33)	14	7	4	5
	BIR	PF00653	4 (7)	8	4	1	4
	Bcl-2	PF00452	10	11	2	1	7

Process	Domain	PFAM no.	Sp	Hs	Dm	Ce	Nv
	TNFR_c6	PF00020	8 (9)	8	1 (no DD)	1 (no DD)	2
	NACHT	PF05729	129 (145)	18	1	1	2
	NB-ARC	PF00931	3	1	1	1	0
	DEATH	PF00531	47 (101)	30	9	6	6
	DED	PF01335	4 (5)	7	1	0	5
	CARD	PF00619	5 (10)	20	1	0	8

Complexity intermediate between that in vertebrates and protostome invertebrate model organisms

Complexity greater than that found in other model organisms

Complexity lower than that found in other model organisms

*Numbers of histone genes refer to distinct core or linker histone genes, as opposed to total gene number as a result of large tandemly repeated arrays (e.g., ~400 clusters of early histone arrays in sea urchin, 100 copies of a tandem array in *Drosophila*, with each array containing a gene for the four core and one H1 histone).

†Numbers for Hs, Dm, and Ce obtained from (53).

Table 3

Genomic insights into sea urchin neurobiology.

Neural process	Revelations from the genome	Genes
Neural development	Neurogenic ectoderm is specified in early embryonic development.	Sp-Achaete-scute, Sp-homeobrain, Sp-Rx (retinal anterior homeobox), Sp-Zic2
Synapse structure and function	Echinoderm synapses are structurally unusual, despite the presence of many genes encoding proteins involved in synapse function.	Sp-Neurolignin, Sp-neurexin, Sp-agrin, Sp-MUSK, Sp-thrombospondin, Sp-Rim2, Sp-Rab3, exocyst complex, Snares, SM, synaptotagmins
Electrical signaling and coupling	Neurons have ion channel proteins, but lack electrical coupling via gap junctions.	Voltage-gated K^+, Ca^{2+}, and Na^+ channels, but no connexins or pannexins/innexins
Neurotransmitter/neuromodulatory diversity	Neurons use the same neurotransmitters as vertebrates, but lack melatonin and adrenalin.	Enzymes involved in synthesis, transport, reception, and hydrolysis of serotonin, dopamine, noradrenaline, g-aminobutryic acid (GABA), histamine, acetycholine, glycine, and nitric oxide
GPCR signaling	Identification of GPCRs that are unique to chordates and identification of expanded GPCR families.	Orthologs of vertebrate cannabinoid, lysophospholipid, and melanocortin receptors are absent; 162 secretin receptor-like genes
Peptide signaling	G protein–coupled peptide receptors indicate diversity in peptide signaling systems, but only a few sea urchin neuropeptides or peptide hormones identified.	37 G protein–coupled peptide receptors. Precursors for SALMFamides, NGFFFamide, and a vasotocin-like peptide
Neurotrophins	Neurotrophins and neurotrophin receptors are not unique to chordates.	Sp-Neurotrophin, Sp-Trk, Sp-p75NTR, ependymins
Insulin and IGFs	More similar to vertebrate forms than invertebrate insulin-like molecules.	Sp-IGF1, SpIGF2
Chemosensory functions	A large family of predicted chemoreceptor genes, some expressed in tube feet or pedicellariae, indicates a complex chemosensory system.	Over 600 genes encoding putative G protein–coupled chemoreceptors, many tandemly repeated and lacking introns
Photoreception functions	Genes associated with photoreception are expressed in tube feet.	Photorhodopsins, Sp-Pax6, retinal transcription factors
Mechanosensory functions	Orthologs of vertebrate mechanosensory genes are present.	Sp-Usherin, Sp-VLGR-1, Sp-cadherins, Sp-myosin 7, Sp-myosin 15, Sp-harmonin, Sp-whirlin, Sp-NBC, Sp-TrpA1