The amphioxus genome illuminates vertebrate origins and cephalochordate biology


1–35[A complete list of author affiliations appears at the end of the paper before the Acknowledgments section.]

Cephalochordates, urochordates, and vertebrates evolved from a common ancestor over 520 million years ago. To improve our understanding of chordate evolution and the origin of vertebrates, we intensively searched for particular genes, gene families, and conserved noncoding elements in the sequenced genome of the cephalochordate Branchiostoma floridae, commonly called amphioxus or lancelets. Special attention was given to homeobox genes, opsin genes, genes involved in neural crest development, nuclear receptor genes, genes encoding components of the endocrine and immune systems, and conserved cis-regulatory enhancers. The amphioxus genome contains a basic set of chordate genes involved in development and cell signaling, including a fifteenth Hox gene. This set includes many genes that were co-opted in vertebrates for new roles in neural crest development and adaptive immunity. However, where amphioxus has a single gene, vertebrates often have two, three, or four paralogs derived from two whole-genome duplication events. In addition, several transcriptional enhancers are conserved between amphioxus and vertebrates—a very wide phylogenetic distance. In contrast, urochordate genomes have lost many genes, including a diversity of homeobox families and genes involved in steroid hormone function. The amphioxus genome also exhibits derived features, including duplications of opsins and genes proposed to function in innate immunity and endocrine systems. Our results indicate that the amphioxus genome is elemental to an understanding of the biology and evolution of nonchordate deuterostomes, invertebrate chordates, and vertebrates.

[Supplemental material is available online at www.genome.org.]
amphioxus diverged before the origin of vertebrates and could shed light on the prevertebrate condition. For example, the comparative anatomist E.S. Goodrich wrote, “Amphioxus is doubtless in some respects a very specialized animal . . . yet it preserves many primitive characters” (Gaskell et al. 1910). Gradually, this view became widely accepted, and has since been supported by modern studies. Recent molecular phylogenetic studies now position cephalochordates as the basal group within the phylum Chordata, with vertebrates and urochordates diverging later (Bourlat et al. 2006; Delsuc et al. 2006).

The genomes of several invertebrates have now been sequenced, including those of protostomes, such as nematodes (Caenorhabditis elegans Sequencing Consortium 1998; Stein et al. 2003) and insects (Adams et al. 2000; Mongin et al. 2004; The Honeybee Genome Sequencing Consortium 2006), and deuterostomes, such as the urochordate Ciona intestinalis (Dehal et al. 2002) and the sea urchin Strongylocentrotus purpuratus (Sodergren et al. 2006). Many vertebrate genome sequences are also now available, including those of humans (International Human Genome Sequencing Consortium 2001; Venter et al. 2001), mice (Waterston et al. 2002), chimpanzee (The Chimpanzee Sequencing and Analysis Consortium 2005), and several species of fish (Aparicio et al. 2002; Jaillon et al. 2004). Against this background of genomic data, the sequencing of the 520 megabase genome of the amphioxus Branchiostoma floridae provides crucial insights into the basal chordate condition (Putnam et al. 2008). The amphioxus genome exhibits considerable synteny with the human genome, but lacks the whole-genome duplications characteristic of vertebrates (Dehal and Boore 2005). Moreover, amphioxus, although very vertebrate-like, lacks several key vertebrate innovations including neural crest, bone, and an endoskeleton (Holland et al. 1996; Shimeld and Holland 2000; Meulemans and Bronner-Fraser 2004). Here, we report an analysis of the amphioxus genome that addresses the evolution of homeobox genes, the neural crest gene network, opsins, endocrine genes, immune genes, and transcriptional enhancers. We show that, in many respects, amphioxus reflects the primitive prevertebrate condition, yet it also exhibits uniquely specialized features that arose in the half a billion years after its divergence from the rest of the chordate lineage.

Results and Discussion

Homeobox genes

Hox genes encode a family of transcription factors that are typically arranged in chromosomal clusters and are involved in rostrocaudal patterning of most bilaterally symmetrical animals. The initial description of the Hox gene cluster of amphioxus, made over a decade ago, revitalized debate concerning the prevertebrate nature of cephalochordates (García-Fernández and Holland 1994). In contrast to the multiple Hox gene clusters of all vertebrate species examined, amphioxus possesses a single cluster. Initial studies revealed an organization remarkably like the inferred preduplication condition that must have existed in the long-extinct ancestor of all vertebrates. The amphioxus Hox gene cluster is the most prototypical deuterostome Hox cluster known, being intact, ordered, and containing the 14 Hox genes described to date (Garcia-Fernández and Holland 1994; Ferrier et al. 2000). Sequencing of the amphioxus genome allowed us to examine the region between the 14th Hox gene, Amphihox14, and the neighboring pair of Evx homeobox genes. This revealed a 15th gene, Amphihox15, which does not group strongly with any particular paralogy group in phylogenetic analysis (Fig. 2A; Supplemental Figs. S1, S2). Authenticity of this gene was confirmed by PCR amplification from genomic DNA of B. floridae with gene-specific primers. This DNA fragment was then used to isolate larger genomic clones from a cosmid library (MPCMg117J0736 and MPCMg117J2469). End sequences from these cosmids are in the expected places in the genome assembly. In addition, two-color chromosomal Fluorescence In Situ Hybridization (FISH) confirmed the tight linkage between Amphihox15 and the rest of the Hox gene cluster (Fig. 2B–E). Thus, the complete amphioxus Hox gene cluster comprises 15 genes spanning 470 kb.

Following discovery of the 14th Hox gene in amphioxus, Hox14 genes were also described in coelacanth and horn shark (Powers and Amemiya 2004). Hox14, however, has been secondarily lost from all tetrapod and teleost fish species examined to date. It now becomes pertinent to ask whether any extant vertebrates, or any deuterostomes other than amphioxus, possess a Hox15 gene, and whether the ancestral condition for deuterostomes evaded the preduplication condition that must have existed in the long-extinct ancestor of all vertebrates.

Figure 1. Juvenile amphioxus (Branchiostoma floridae). This specimen has not yet developed gonads. Anterior to left. (C) Buccal cirri surrounding mouth; (G) pharyngeal gill slits; (A) atrial cavity; (D) digestive diverticulum. Scale bar, 0.5 mm.

Figure 2. The amphioxus genome contains 15 Hox genes. (A) Genomic organization of the 470-kb amphioxus Hox gene cluster and linked Evx gene pair; beneath the horizontal line are shown the two Hox15 cosmid clones analyzed. Black vertical lines represent small sequence gaps; asterisks indicate genes with an intron within the homeobox; arrows denote transcriptional orientation. FISH to amphioxus interphase nucleus (B) and metaphase chromosomes (C–E). (Red signal) Cosmid MPCMg117J0736 (Amphihox15); (green signal) cosmid MPCMg117J2469 (Amphihox12 and Amphihox13). Methods as described in Castro and Holland (2002).
stomes was a Hox cluster with 15 genes adjacent to an Evx family gene.

There are many other homeobox genes beside Hox genes, and most have roles in development and cell differentiation (Bürglin 2005). Determining their patterns of gene duplication and divergence can contribute to understanding the evolution of numerous developmental processes. We identified a total of 133 homeobox genes in the draft genome sequence of B. floridæ (Supplemental Table S1). This figure represents a minimal estimate, because genome sequence coverage is not 100% complete and the genome possesses a high incidence of polymorphism, leading to potential confusion between loci and alleles. In our calculation of gene number, we include Pax2/5/8, which has a partial homeobox sequence, but exclude Pax1/9 and Pon, which lack homeobox sequences, but are likely derived from homeobox genes. The same rule was followed in a recent survey of the human genome, which revealed 235 human homeobox genes (and over 65 pseudogenes) (Holland et al. 2007).

Comparison to human and tunicate genomes suggests that the last common ancestor of the chordates possessed 103 homeobox genes (Supplemental Table S1). At the level of the gene family, 87 conserved homeobox gene families are shared between amphioxus and human; these have undergone very different patterns of gene duplication in the two lineages. Of these 87 families, 68 are represented by duplicate versions in the human genome, compatible with expansion by genome duplication, followed by gene loss. In contrast, just 11 are uniquely duplicated in amphioxus. These, which are usually represented as tandem duplicates, include Mnx, Hmbox, Unx, Pou3, Lhx2/9, Hox9-15, and Iro, in addition to the previously recognized duplications of Evx, Emx, Nkx1, and Vent genes (Ferrier et al. 2001; Minguillón et al. 2002; Luke et al. 2003). There is also a duplication of the amphioxus Ndx gene, which is lost in vertebrates (this name was proposed for CG13424 to reflect its genomic position: Next to Dll homeobog). It is striking that amphioxus retains many ancient bilaterian homeobox genes that were secondarily lost either on the Olfactores (tunicates + vertebrates) lineage, the vertebrate lineage, or the tunicate lineage (Fig. 3); we found no examples of homeobox gene loss in the cephalochordate lineage.

Expansion of opsin genes in amphioxus

Opsins are seven-pass transmembrane proteins widespread across animal phyla, typically mediating the first step in phototransduction cascades. Four opsin types are thought to have been present at the base of bilaterians: rhodopsin (r-opsins), ciliary (c-opsin), Go-opsins, and peropsin/RGR (Arendt et al. 2004; Terakita 2005, Raible et al. 2006). We have identified 20 different opsin genes in amphioxus representing all four groups (Supplemental Fig. S3; Supplemental Table S2). There are four r-opsin genes, three c-opsin, two in amphioxus or may have arisen on the Olfactores lineage. Four opsin genes from the Go-opsin family were identified, two in tandem orientation in the genome. There is one amphioxus peropsin/RGR gene, plus three putative neuroopsins related to vertebrate Opn5 that push back the date of origin of this group (Tarttelin et al. 2003). The expansion of three opsin subfamilies in the cephalochordate lineage (Amphitop6, encephalopsin/TMT-opsins, and Go-opsins) may relate to the diversity of photoreceptor organs in amphioxus, including the frontal eye and the lamellar body, both of which are ciliary photoreceptors, and the organs of Hesse and Joseph cells, which are both rhodallomer photoreceptors (Lacalli 2004).

Genes of the vertebrate neural crest gene network

Neural crest cells are a migratory cell population that forms at the neural plate border, migrates extensively, and differentiates into numerous differentiated cell types, including connective tissue of the craniofacial skeleton, much of the peripheral nervous system, and pigment cells (Meulemans and Bronner-Fraser 2004). These cells are responsible for many complex structures in the vertebrate head that are absent in amphioxus and tunicates. When we examined the amphioxus genome sequence to determine whether any key components of the vertebrate neural crest regulatory network are missing from amphioxus (Supplemental Table S3), we found that amphioxus has all the major transcription factors deployed in the vertebrate neural crest regulatory network. These include the epidermal ectoderm markers Dlx and AP2, the neural

![Figure 3. Homeobox gene loss in chordates. Homeobox gene loss has been extensive during the evolution of the Olfactores (vertebrates plus tunicates), yet in amphioxus, no homeobox gene loss has occurred since the last common ancestor of the chordates (LCAC). (Blue) ANTP class; (red) PRD class; (green) other classes. Gene losses along the tunicate branch occurred before the last common ancestor of Ciona and Oikopleura. Photo of Goodrich reproduced by permission of Department of Zoology, University of Oxford.](image-url)
ectoderm markers SoxB, Zic, and islet, genes encoding neural plate border specifiers (Pax3/7 and Msx), and neural crest specifiers (Snail, SoxE, AP2, Twist, Id, FoxD, Myc), as well as downstream mediators of neural crest migration and differentiation, notably, Rho, cRet, ErbB3, Mitf, tyrosinase, and tyrosinase-related genes. Hence, the evolution of neural crest in vertebrates cannot be explained by the gain of a new regulatory gene, but must relate to the acquisition of new functions by pre-existing genes. In contrast, two genes involved at later stages of neural development seem to be absent from amphioxus; these are homologs of c-Kit, a receptor tyrosine kinase essential for migration, survival, and differentiation of neural crest-derived melanocytes and of the gene for myelin protein P0, consistent with the notion that the glial myelin sheath is a vertebrate innovation. Interestingly, many of these neural crest specifier genes have multiple paralogs in vertebrates, presumably due to large-scale genome duplication events (Putnam et al. 2008), whereas amphioxus usually contains a single copy. These observations raise the possibility that neofunctionalization of some paralogs may have facilitated co-option of genes into a neural crest gene regulatory network.

We propose that migrating neural crest cells emerged after the split of amphioxus from tunicates and vertebrates. Therefore, it will be important to further examine neural crest-like cells in tunicates (Jeffery et al. 2004) and determine to what extent the genetic mechanisms for their specification, migration, and differentiation are evolutionarily conserved. Such studies in amphioxus and tunicates should help determine to what extent the glial myelin sheath is a vertebrate innovation. Interestingly, many of these neural crest specifier genes have multiple paralogs in vertebrates, presumably due to large-scale genome duplication events (Putnam et al. 2008), whereas amphioxus usually contains a single copy. These observations raise the possibility that neofunctionalization of some paralogs may have facilitated co-option of genes into a neural crest gene regulatory network.

We propose that migrating neural crest cells emerged after the split of amphioxus from tunicates and vertebrates. Therefore, it will be important to further examine neural crest-like cells in tunicates (Jeffery et al. 2004) and determine to what extent the genetic mechanisms for their specification, migration, and differentiation are evolutionarily conserved. Such studies in amphioxus and tunicates should help determine what constitutes the primitive genetic subnetwork for neural crest formation during chordate evolution.

Roots of the vertebrate endocrine system

The primary vertebrate endocrine organs (pineal gland, ovary, testis, endostyle/thyroid, brain, gut, and pituitary) have structurally simpler homologs in amphioxus. To determine which components of the endocrine system were present in the chordate ancestor, we searched the amphioxus genome for homologs of genes involved in the synthesis, secretion, and reception of endocrine signals. The relationships of these amphioxus genes to related genes in other organisms were determined by phylogenetic analyses. Support values for orthology relationships are listed in Supplemental Table S4.

Pituitary control of reproduction—A vertebrate innovation

Analysis of the amphioxus genome indicates considerable potential for neuroendocrine control of reproduction (Fig. 4A). For example, there are more than 10 homologs of the kisspeptin receptor (GPR54, also known as KISS1R), which in vertebrates is expressed in the hypothalamus and regulates release of gonadotropin-releasing hormone (GnRH). Both amphioxus and Ciona (Tello et al. 2005) have three GnRH receptors (Supplemental Table S4), which are functionally active in vitro. The presence of these receptors suggests that amphioxus also possesses the kisspeptin and GnRH peptides, even though we could not find them in the amphioxus genome; such short sequences are not readily identified by sequence analysis. Amphioxus also has a gene for thyrotropin stimulator, which may be ancestral to three of the pituitary hormones (follicle stimulating hormone, FSH, luteinizing hormone, LH, and thyroid stimulating hormone, TSH) (Sudo et al. 2005). However, amphioxus does not appear to have homologs of several vertebrate pituitary hormones [i.e., growth hormone (GH), adrenocorticotropic hormone (ACTH), and prolactin]. Not only were they not detected in the amphioxus genome, but neither were their receptors, suggesting that the reported presence of prolactin immunoreactivity in Hatschek’s pit (amphioxus pituitary homolog) should be reexamined (Weng et al. 2006). Since genes for ACTH, GH, and prolactin are also missing in the Ciona and sea urchin genomes, these hormones may be vertebrate novelties.

Gonad-expressed sex steroids comprise the final step in neuroendocrine control of vertebrate reproduction. Amphioxus has two genes for steroid receptors, one more closely related to vertebrate estrogen receptors (ER) and one closer to steroid receptors (SR), plus a complete set of genes for sex steroid-synthesizing enzymes: STAR, CYP11, CYP17, CYP19, 17βHSD, 3βHSD, and 5α-reductase (Supplemental Table S4). In contrast, Ciona lacks genes for steroid hormone receptors and has only a few for enzymes that synthesize steroid hormones (5α-reductase, 3βHSD, 17βHSD), suggesting tunicate-specific losses (Campbell et al. 2004). Thus, except for the pituitary hormones, amphioxus has a relatively complete endocrine pathway for control of reproduction. It may be that neuroendocrine GnRH activates the gonads without pituitary amplification. Indirect support for this hypothesis is that in vertebrates, GnRH, when synthesized locally, can act directly on the gonads (Leung et al. 2003).

Components of the stress response

The stress response in vertebrates depends on the adrenal medulla for epinephrine and on the adrenal cortex for cortisol. Adrenal cortex function in vertebrates is controlled by the brain (corticotropin-releasing hormone, CRH) and pituitary ACTH. Although amphioxus lacks a discrete adrenal gland, it possesses genes for catecholamine synthesis and for adrenergic receptors, suggesting that some cells may synthesize epinephrine. However, most of the stress-response genes (CRH, ACTH, ACTH receptor, cortisol and its receptor) are missing in amphioxus and have not been found in either Ciona or sea urchins. In addition, amphioxus also lacks a gene for the enzyme 21-hydroxylase (CYP21), necessary to synthesize cortisol precursors (Supplemental Fig. S5). The only component present in all three invertebrates is similar to the CRH receptor. However, since this receptor is also structurally related to the diuretic hormone receptor in Drosophila (Hauser et al. 2006), functional studies will be required to determine whether or not it functions in the stress pathway in amphioxus. In summary, the cortisol pathway most likely did not exist in the ancestor of chordates, but an epinephrine stress response did.

An ancient route for thyroid hormone production in amphioxus

The vertebrate thyroid gland sequesters iodine and secretes T3 and T4 hormones, which regulate metabolism, post-embryonic development, and metamorphosis in vertebrates. The amphioxus thyroid homolog, the endostyle, is primarily involved in mucous production for feeding, but it does incorporate iodine (Fredriksson et al. 1984), suggesting that it also may synthesize thyroid hormones. This suggestion has gained support from chromatography experiments with radiolabeled iodine (Covelli et al. 1960; Tong et al. 1962) from immunocytochemistry analyses with antibodies to vertebrate hormones (Fang and Weng 2002), and from the induction of amphibian metamorphosis by implantation of dried amphioxus endostyles (Sembrat 1953). The
amphioxus genome sequence further strengthens this claim, since we identified a thyroid hormone receptor (TR) gene and several genes for components of the thyroid hormone production system: a sodium iodide symporter (SIS), a thyroid peroxidase (TPO), deiodinases that convert T4 to active T3, and a binding protein (CTHBP) for cytosolic transport of TH (Fig. 4B). In contrast, genes for thyroglobulin and a thyroid-releasing hormone receptor (TRH) are missing. In vertebrates, hormones secreted by the brain and pituitary control production of T3 and T4. Although a gene for the pituitary thyroid-stimulating hormone is not found in the amphioxus genome, as noted above, there are genes for thyrostimulin (GPA2/GPB5) and its receptor, suggesting that a pathway for metabolic control was present in the ancestral chordate.

**Insulin family members**

The insulin superfamily includes insulin, insulin-like growth factors (IGFs), and relaxins. In vertebrates, insulin is produced by the pancreas, a gut derivative, and is essential for regulating glucose metabolism. IGF is critical for growth, and relaxins are associated with parturition among other roles. Amphioxus lacks a morphologically defined pancreas, but particular cells in the gut express an insulin-like peptide (Chan et al. 1990; Holland et al. 1997). In addition to this gene, we identified five other insulin/relaxin-related genes and their putative receptors in the amphioxus genome. Three of the insulin/relaxin-related genes are linked in the genome. Phylogenetic analysis of neighboring genes revealed synteny of four of the six insulin-like peptides with human paralogons hosting the insulin/IGF and relaxin-type genes (Supplemental Fig. S6; Supplemental Table S5). The organization of the corresponding genomic region in *Ciona* is similar (Olinski et al. 2006), indicating that insulin and relaxin-type precursor genes were positioned in one chromosomal segment in the ancestral chordate. Genes for other protein hormones present in amphioxus are shown in Supplemental Table S4.

**Nuclear receptors**

Nuclear receptors (NRs) bind a wide range of hormones and transduce their signals intracellularly. Amphioxus possesses genes for a total of 33 nuclear receptors. Eighteen have clear orthologs among the 48 human NRs (Supplemental Table S6). These genes include two related to the estrogen and steroid receptors (ER, SR) and one each for thyroid hormone receptor (TR) and retinoic acid receptor (RAR). Moreover, the ligand-binding and the DNA-binding domains (CEGCNK) of amphioxus SR and ER are more similar to the ER type than to the classical steroid type. A comparison of amphioxus, sea urchin, and *Ciona* nuclear receptors suggests a loss of several receptors in tunicates, including genes related to the ER and SR receptors and to the NR2E genes (TLL, PNR, and a novel duplicated amphioxus gene). TLL and PNR are involved in the development of the vertebrate brain and retina, and all three NRs may be important in amphioxus with its frontal eye, pineal eye, and multiple photoreceptors.

*Figure 4.* Predicted amphioxus endocrine pathways in which a red box identifies homologs found in the amphioxus genome. (A) Reproductive axis in which missing components include classical pituitary hormones (LH, luteinizing hormone; FSH, follicle stimulating hormone; Prl, prolactin) and their receptors. GPR54: glycoprotein receptor 54 (kisspeptin receptor); GnRH: gonadotropin-releasing hormone; E: estrogen; ER: estrogen receptor; SR: steroid receptor. (B) Thyroid metabolism showing that the thyroglobulin gene could not be identified in the amphioxus genome (hence, crossed out). CTHBP: cytosolic thyroid hormone binding protein; SIS: sodium iodide symporter; Tg: thyroglobulin; T2: diiodothyronine; T4: thyroxine; TR: thyroid hormone receptor; RXR: retinoid X receptor; TRE: thyroid hormone response element. (C) Retinoid metabolism in which CRABP and AKR genes could not be identified in the amphioxus genome (hence, crossed out). ADH: alcohol dehydrogenase; AKR: aldo-keto reductase; ARAT: acylCoA:retinol acyltransferase; BCOX: beta,beta-carotene-15,15'-oxygenase; CRABP: cellular retinoic acid binding protein; CRBP: cellular retinol binding protein; CYP26: cytochrome P450–26; LRAT: lecithin:retinol acyltransferase; RALDH: retinaldehyde dehydrogenase; RDH: retinol dehydrogenase; REH: retinyl ester hydrolase; RPE65: retinal pigment epithelium-specific protein 65 kDa; RAR: retinoic acid receptor; RXR: retinoid X receptor; RARE: retinoic acid response element.
The nuclear receptors most studied in amphioxus are RAR and RXR, which mediate the roles of retinoic acid during amphioxus development through an RAR:RXR heterodimer (Escriva et al. 2002; Yagi et al. 2003; Schubert et al. 2006). Not surprisingly, amphioxus also possesses genes for retinoic acid synthesis, storage, and degradation (Fig. 4C). These include an ortholog of cellular retinol binding protein (CRBP), but not of its retinoic acid equivalent (CRABP). Although a gene named CRABP has been described in amphioxus (Jackman et al. 2004), it does not group robustly with vertebrate CRABP. Thus, amphioxus lacks some genes involved in retinoid metabolism in vertebrates (such as CRABP), but others have duplicated independently (e.g., ALDH1s/RALDHs and REHs) (Supplemental Table S6).

There are a few more examples for independent expansions of amphioxus endocrine genes. For example, genes in the LXR/FXR group, which control lipid/cholesterol metabolism in vertebrates have been duplicated in amphioxus as well as the genes for the 3BHSD enzymes, which are involved in vertebrate steroidogenesis, and for several SDR enzymes, which are important for processing and detoxification of a wide variety of biological compounds including steroids, alcohols, sugars, aromatic chemicals, and xenobiotics (Supplemental Tables S4, S6). These lineage-specific duplications in amphioxus of transcription factors (NRs of the LXR/FXR group) and enzymes (SDR, 3BHSD, ALDH1s/RALDHs, and REHs) involved in both metabolism and detoxification of various molecules represent an important example for a unique feature that has evolved specifically in the cephalochordate lineage. There are two (not mutually exclusive) evolutionary scenarios that might explain these amphioxus-specific duplications: (1) the signaling networks controlled by the molecules produced directly or indirectly by these duplicated genes have been extensively elaborated in the amphioxus lineage (for example steroid hormone synthesis) and/or (2) novel intracellular protection mechanisms against various kinds of toxins and xenobiotics derived from marine microorganisms or the surrounding environment might have evolved specifically in the amphioxus lineage, possibly as a result of the sedentary, filter feeding lifestyle of amphioxus adults.

Adaptive and innate immunity

Defense against pathogens is a basic requisite of all animals. The immune system, which performs this function, can be broadly divided into two separate, but closely integrated components: germ-line-encoded innate immunity and somatically modified adaptive immunity. Innate immunity in vertebrates relies on a relatively small number of pattern recognition receptors, notably the extracellular TOLL-like receptors (TLRs) and intracellular nucleotide-binding oligomerization domain-like receptors (NLRs) (Boraschi and Tagliabue 2006; Werts et al. 2006; Delbridge and O'Riordan 2007). TLRs consist of a TIR domain combined with leucine-rich repeats (LRRs); NLRs consist of a NACHT domain and LRRs (Martinon and Tschopp 2004; Inohara et al. 2005) (Fig. 5). A search of the amphioxus genome for homologs of innate immunity genes reveals the expansion of receptor gene families in an intriguing parallel to those discovered in the sea urchin genome (Rast et al. 2006) (Fig. 5; Supplemental Table S7). In the amphioxus genome, ab initio gene prediction yields 28

Figure 5. Expansion of innate immunity genes in amphioxus (right) compared with the human system (left). Novel domain architectures and (predicted) varied receptor conformations found in amphioxus proteins are expected to increase the specificity of innate immunity in amphioxus. We propose that some TLR-like proteins are localized in the endosome, as also seen in mammals. Domains: TIR (Toll/interleukin-1 receptor) domain, NACHT (NAIP, CIITA, HET-E and TP1), LRR (leucine-rich repeat), CARD (caspase recruitment domain), PYRIN (N-terminal domain of protein pyrin), DD (death domain), DED (death effector domain), SAM (sterile alpha motif) domain, Bir (Baculovirus Inhibitor of apoptosis repeat).
TLR gene models, compared with 10–15 in vertebrates, nine in Drosophila, and over 200 in sea urchin. The amphioxus number increases to 72 if additional ab initio models are considered. Variation in the number, length, and sequence of the LRR domains suggests that the repertoire of receptor shapes in amphioxus far exceeds that reported in other systems. Similar expansions are seen in several families of genes encoding proteins involved in apoptosis, a regulatory strategy in invertebrate host defense as well as in both innate and adaptive immunity in vertebrates.

In addition to TLRs and NLRs, other innate immunity genes are present in the amphioxus genome (Supplemental Tables S7–S9). The complement system of vertebrates is a highly sophisticated defense system that functions to opsonize pathogens through the covalent binding of the central component C3, through a unique intramolecular thioester bond (Law and Dodds 1997; Janssen et al. 2005). Most complement-system gene groups are highly conserved between amphioxus, Ciona, and vertebrates (Nonaka and Kimura 2006). In contrast, the sea urchin genome lacks the MASP/C1r/C1s and TCC (terminal complement component) genes. One unexpected novelty in the amphioxus genome is the presence of more than 39 C1q-like genes, of which 25 are encoded on a single scaffold, indicating an expanded complement-recognition repertoire (Supplemental Table S9).

Adaptive immunity has traditionally been viewed as a jawed vertebrate adaptation in which somatic reorganization of immunoglobulin (Ig) genes in B-lymphocytes or T-cell antigen receptor (TCR) genes in T-lymphocytes, results in the expression of cell surface receptors with unique molecular specificities. In lampreys and hagfish, genetic recombination of sequence elements encoding an LRR-containing receptor, termed a VLR (variable lymphocyte receptor), gives rise to an alternative form of adaptive immunity (Pancer and Cooper 2006). Beside the TLRs and NLRs mentioned above, the amphioxus genome contains at least 406 gene models predicted to encode proteins with multiple LRR modules that may play a role in host defense (Supplemental Table S10). Some of these encode cell surface proteins, with a single LRR domain capped by N- and C-terminal LRRs, a feature of VLRs. Some LRR-containing gene products also encode immunoglobulin (lg) domains. In jawed vertebrates, the rearrangement mechanism is dependent on the Recombination Activating Genes (RAG) 1 and 2, which are tightly linked in vertebrates. A homologous RAG1/2-like closely linked gene pair is present in the purple sea urchin genome; however, its function is unknown (Fugmann et al. 2006). Only a RAG1-like segment is identified in the amphioxus genome; the adjacent genomic sequences are unresolved, precluding accurate annotation. Many other gene products, mostly recruited from ancient DNA damage repair pathways, also function in the diversification of immune receptors in jawed vertebrates including XRCC4, DNA ligase IV, UNG, RAD18, Artemis, Cernunnos, Pol-δ, Pol-λ, Pol-β, REV1, and KU70/80 (Jung and Alt 2004; Litman et al. 2005). Homologs of these genes are present in both amphioxus and sea urchin; however, neither immunoglobulin nor TCR genes can be identified in either of these species, nor can homologs of MHCI and MHCII, the functional partners of TCR α/β (Supplemental Table S11). Terminal deoxynucleotidyltransferase (TdT), which catalyzes nontemplated addition of nucleotides at double-strand DNA breaks and accounts for a high degree of somatic variation in Igs and TCRs (Jung and Alt 2004), is a vertebrate homolog of DNA-polγ; our analyses reveal that amphioxus diverged before this duplication. Alternative processing likely gives rise to functionally distinct forms (data not shown). The amphioxus genome assembly does not support a model for full-length APOBEC (I-III) enzyme(s), which effects antiviral defenses in vertebrates or activation-induced cytidine deaminase (AICDA), which is integral to somatic hypermutation, Ig class switching, and gene conversion (Fugmann and Schatz 2002). However, only a short sequence exhibiting identity to APOBEC3 and two regions exhibiting weak sequence identity to the N- and C-terminal regions of human AICDA were identified. Although homologs of Ig and TCR have only been found in jawed vertebrates, highly polymorphic genes termed VCBP (variable region-containing chitin binding protein) have been identified in amphioxus and Ciona, but not in sea urchin (Cannon et al. 2002). Structures similar to the V (variable) regions of VCBPs are attractive candidates for ancestral forms of rearranging Ig-type immune receptors. Collectively, these data suggest that the ancestor of all chordates had a complex diversified innate immune system. Key effector and intermediate molecules, with roles in germ-line change and somatic variation associated with adaptive immunity were also present at the base of the chordates and were co-opted for adaptive immunity in the jawed vertebrate lineage and VLR diversification in jawless vertebrates (Litman et al. 2005; Nagawa et al. 2007; Rogozin et al. 2007).

Conserved noncoding sequences

Although genome-scale comparisons between mammals and telost fish have revealed up to 3100 conserved noncoding sequences, most of which function as tissue-specific enhancers (Woolfe et al. 2005; Pennacchio et al. 2006), at greater phylogenetic distances, no conservation outside of coding sequences and microRNAs (Prochnik et al. 2007) has previously been identified. However, our comparison of the amphioxus and human genomes revealed a number of conserved noncoding elements (CNEs) near coding sequences, some of which directed tissue-specific expression in both amphioxus and the mouse. Comparisons of the amphioxus and human genomes yielded 27,880 aligned regions with more than 60% identity over 50 bp. However, most of these (99.7%) either overlapped transcribed or repetitive sequences in the human and/or amphioxus genomes or were not conserved with other vertebrate species and were, therefore, excluded from further analysis. We designated the remaining 77 alignments as putative chordate CNEs. Of these, 16 overlap with or are immediately adjacent to the 3’ or 5’ untranslated regions (UTRs) of human genes, and are, therefore, likely conserved UTRs. Four CNEs are adjacent to exons, and likely represent conserved splice enhancers. A single CNE overlapped a highly conserved microRNA (MIRN10B, adjacent to the human HOXD4 gene). The remaining 56 were of unattributable function. We tested the human orthologs of eight of these 56 CNEs for activity in a mouse enhancer assay (Pennacchio et al. 2006). Four reproducibly drove expression of a lacZ reporter gene in mice. One element, adjacent to the T-domain-containing transcriptional repressor TBX3 in human, drove expression in the limb and tail (4/9 embryos), which are two of the domains that normally express the gene (Chapman et al. 1996; Gibbon-Brown et al. 1996). A second element, adjacent to the dominant-negative helix-loop-helix inhibitor of DNA-binding transcription factor ID1, drove expression in the branchial arches, eyes, spine, midbrain, forebrain, and ear (7/8 embryos) (Supplemental Tables S12, S14, S15). The remaining two CNEs with enhancer activity are adjacent to the paralogous zinc finger transcription factors.
ZNFS03 and ZNF703 in human (Supplemental Fig. 8). Orthologs of these two genes occur in all sequenced vertebrate genomes, and correspond to a single locus in the amphioxus genome, amphioxus ZNFS03/703. The single amphioxus CNE and the two corresponding human sequences were tested for enhancer activity in both mouse and amphioxus (Fig. 6). LacZ reporter constructs of the conserved CNEs from all three genes showed tissue-specific expression in both amphioxus and the mouse (Fig. 6). Taken together, the computational and experimental results are consistent with shared ancestry of putative enhancer function, and suggest subfunctionalization of these noncoding elements following duplication in the vertebrate lineage. This represents the largest evolutionary distance (∼520 million years) identified to date, across which both the sequence and function of tissue-specific, cis-regulatory enhancers have been conserved.

Reconstructing the deuterostome ancestor

In the half-billion years since the amphioxus and vertebrate lineages split, the amphioxus genome has clearly undergone much less change than the genomes of other chordates. The ancestor of the Olfactores (tunicates + vertebrates) lost several regulatory genes, after which the vertebrate lineage underwent two whole-genome duplication events, while the tunicates experienced many additional gene losses. Moreover, while cis-regulatory enhancers are apparently not conserved between the Cliona and vertebrate genomes, amphioxus and humans share several. Thus, although there have been some independent gene duplications in the cephalochordate lineage (e.g., within opsin genes and innate immunity genes), the amphioxus genome appears to be closer than those of any other extant organism to the genome of the ancestral chordate.

Although amphioxus has no neural crest and, therefore, no homologs of vertebrate organs with major neural crest components, such as the adrenal gland, the last common ancestor of the chordates clearly possessed all of the known genes required for generating neural crest. What happened at the base of the vertebrates was evidently the co-option of pre-existing genes into the network specifying the cells at the edges of the neural plate and the evolution of new genes specific for some neural crest derivatives such as melanocytes. Similarly, amphioxus possesses homologs of vertebrate endocrine organs such as the thyroid and pituitary that are primarily epithelial derivatives (but not those with major contributions from neural crest), and these organs express the same upstream transcription factors as their vertebrate counterparts. The genes that are clearly vertebrate novelties include several coding for the downstream products of endocrine organs. A key vertebrate innovation is the adaptive immune system. In its absence, invertebrates, including amphioxus, rely on innate immunity for defense against pathogens. It is interesting that some gene families involved in innate immunity (e.g., Toll receptors) are greatly expanded in the amphioxus genome. Together with data from the fossil record (Chen et al. 1995) and embryology (Yu et al. 2007), our data indicate that the ancestral chordate was similar in many respects to a modern amphioxus: a small, free-living animal, with segmental muscles derived from somites, a notochord, a dorsal nerve cord, a robust innate immune system, and a basic set of endocrine functions.

Looking deeper in time, we compared the genomes of amphioxus B. floridana, the tunicate C. intestinalis, and the echinoderm S. purpuratus. We found that the 814-mb sea urchin genome and the 520-mb amphioxus genome are more representative of that of the ancestral deuterostome than is the 155-mb genome of C. intestinalis, which has undergone considerable secondary gene loss. Echinoderms and hemichordates (the free-living enteropneusts plus the sessile pterobranchs) are sister groups, and together constitute the Ambulacaria, which, with the Xenoturbellida, constitute the nonchordate deuterostome lineage (Bourlat et al. 2006). Hemichordate genomes have not yet been se-
quenced, and it is uncertain whether enteropneusts or pterobranchs are basal within the hemichordates (Cameron et al. 2000). Moreover, *S. purpuratus* is derived within the echinoderms. Therefore, for gene families in which the amphioxus genome differs considerably from those of protostomes (e.g., TLF genes), it is difficult to infer the ancestral deuterostome condition. Nevertheless, by integrating the findings of molecular phylogenetics, genomic data from chordates and an echinoderm, and analyses of axial patterning systems in hemichordates, amphioxus, and vertebrates, we can evaluate alternative theories concerning deuterostome evolution. These data imply that the evolutionary scenarios of Garstang, Romer, and Berrill (for review, see Lacalli 2004), prevalent for much of the 20th century, are highly unlikely. The deuterostome ancestor was almost certainly not a sessile lophophorate-like animal that gave rise to an ancestral ascidian. In turn, this larval ascidian did not give rise to the ancestor of amphioxus and the vertebrates by paedomorphosis. It is much more likely that the ancestral deuterostome was a free-living, bilateral, worm-like animal with a series of pharyngeal gill slits. Enteropneust hemichordates, amphioxus, and even vertebrates retain body plans comparable to this inferred ancestral deuterostome, albeit each with its own specializations, while the body plans of echinoderms and tunicates are considerably more derived.

**Methods**

The starting point for these analyses was the draft assembly of the amphioxus genome sequence (http://genome.jgi-psf.org/Brafl1/Brafl1.home.html), in which initial gene and protein predictions had been made using the JGI annotation pipeline as described by Putnam et al. (2008). For each gene family analyzed in the present study, the assembly was searched using tBLASTn, and for each putative match, the deduced amphioxus protein sequences were retrieved. Gene annotations were corrected manually, taking into account additional cDNA information and alignment to homologs from other species. In the few cases when previous genomic clone data differed from genome sequence data, putative sequencing or assembly errors were annotated on the genome and corrected sequences used in further analysis. Gene models representing allelic pairs were identified using a combination of similarity of predicted peptide sequence and gene neighborhood context. For phylogenetic analysis, predicted amphioxus proteins were aligned with other family members and with outgroup proteins using ClustalW or MUSCLE, edited manually, and analyzed using neighbor-joining, maximum parsimony, and maximum likelihood algorithms, implemented using PHYLIP, MEGA 3.0, or PhyML. Branch support of the phylogenies was assessed by bootstrap analyses. Isolation of Hox15 cosmid clones and fluorescent in situ hybridization to chromosomes followed published protocols (Castro and Holland 2002).

**Computational detection of conserved noncoding sequences**

Whole-genome alignments of the amphioxus genome scaffold sequences and human genome reference sequence (NCBI v35) were performed with the VISTA pipeline infrastructure (Frazer et al. 2004) based on efficient combination of glocal and local alignment methods. We first obtained a map of large blocks of conserved synteny between the two species by applying ShuffleLAGAN global chaining algorithm (Brudno et al. 2003) to local alignments produced by translated BLAT (Kent 2002). Then, in each syntenic block, we applied Shuffle-LAGAN again to obtain a more fine-grained map of small-scale rearrangements, such as inversions. Finally, we used VISTA with a similarity cut-off of 60% identity over 50 bp to identify candidate-conserved non-coding sequences. The human genome coordinates of all alignments were used in the UCSC table browser to identify and remove: (1) elements overlapping known genes, human mRNAs, spliced ESTs, Ensembl genes, human repeats, simple repeats, retroposed genes, and high-copy number regions in the human genome (defined by more than one overlapping self-chain hit); (2) elements not overlapping the “most conserved” track in the UCSC genome browser; (3) elements not conserved between human and at least one of the fish, chicken, or frog genomes. The amphioxus genome coordinates of all remaining alignments were used to exclude elements with BLASTX alignments to proteins encoded by the amphioxus genome.

**Mouse transgenic enhancer assay**

Primers were designed to flank the conserved element by several hundred base pairs using primer32 (Supplemental Tables S12, S13). Enhancer element constructs were PCR amplified from human genomic DNA (BD Biosciences) and directionally cloned into the pENTR/D-TOPO vector (Invitrogen). All inserts were sequence validated and transferred into a Hsp68-LacZ vector33 encompassing a Gateway cassette using LR recombination (Invitrogen). Generation of transgenic mice and embryo staining were as previously described (Pennacchio et al. 2006) in accordance with protocols approved by the Lawrence Berkeley National Laboratory. Transgenic mouse DNA was prepared from yolk sacs that were carefully dissected from embryos, boiled for 5 min in lysis solution (50 mM Tris HCl at pH 8.0, 20 mM NaCl, 1 mM EDTA at pH 8.0, 1% SDS), and then screened by PCR with LacZ primers (LacZ-fwd: 5 prime-TTTCCATGGTGCACCTGCG; LacZ-rev: 5 prime-AACGCGTTCGTTCAAGCA) for positive transgenic animals. Images were obtained using a Leica MZ16 microscope and DC480 camera, cropped, and level adjusted with Adobe Photoshop. An element was classed as a positive enhancer if at least three independent transgenic embryos showed the same expression pattern.

**Amphioxus transgenic enhancer assay**

AmPrimer were designed to flank the conserved element by several hundred base pairs using primer33, and an additional sequence containing a HindIII restriction site was added to the 5' end of each primer. Enhancer element constructs were PCR amplified from amphioxus (*B. floridae*) genomic DNA. Amplified fragments were digested using HindIII and ligated into HindIII digested pASCID-FoxD vector35 using a Fast-link DNA ligation solution (50 mM Tris HCl at pH 8.0, 20 mM NaCl, 1 mM EDTA at pH 8.0, 1% SDS), and then screened by PCR with LacZ primers (LacZ-fwd: 5 prime-TTTCCATGGTGCACCTGCG; LacZ-rev: 5 prime-AACGCGTTCGTTCAAGCA) for positive transgenic animals. Images were obtained using a Leica MZ16 microscope and DC480 camera, cropped, and level adjusted with Adobe Photoshop. An element was classed as a positive enhancer if at least three independent transgenic embryos showed the same expression pattern.

**List of Affiliations**

1Marine Biology Research Division, Scripps Institution of Oceanography, La Jolla, California 92093-0202, USA; 2Department of Genetics, Faculty of Biology, University of Barcelona, Barcelona E-08028, Spain; 3Division of Innovative Research, Creative Research Initiative “Sousset”, Hokkaido University, Sapporo 001-0201, Japan; 4U.S. Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA; 5Division of Biology 139-74, California Institute of Technology, Pasadena, California 91125, USA; 6Institut de Génomique Fonctionnelle de Lyon, CNRS UMR5242, UCB, ENS, INRA 1288, IFR128 BioSciences Lyon-Gerland Ecole Normale Supérieure de Lyon, 69364 Lyon.
Cedex 07, France; 7Department of Zoology, University of Oxford, South Parks Road, Oxford, OX1 3PS, UK; 8Department of Biologica, Università di Genova, viale Benedetto XV 5, 16132 Genova, Italy; 9H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida 33612 USA; 10Department of Molecular Genetics, All Children’s Hospital, St. Petersburg, Florida 33701 USA; 11The Gatty Marine Laboratory, University of St Andrews, St Andrews, Fife, KY16 8LB, Scotland; 12Department of Biology, Washington University in St Louis, St. Louis, Missouri 63130, USA; 13Dipartimento di Scienze Biomoleculari e Biotecnologie, Università di Milano, Milano, Italy; 14Burnham Institute for Medical Research, La Jolla, California 92037, USA; 15Unit of Developmental Neuroscience, Department of Neurosciences, Uppsala University, Uppsala, Sweden; 16Department of Biological Sciences, Graduate School of Science and Engineering, Tokyo Metropolitan University, Hachioji, Tokyo 192-0397, Japan; 17Department of Molecular Life Science, Tokai University School of Medicine, Isehara, Kanagawa 259-1193, Japan; 18Department of Pathology, Hokkaido University Graduate School of Medicine, Sapporo 060-8638, Japan; 19Center for Integrative Genomics, Department of Cell and Molecular Biology, University of California at Berkeley, Berkeley, California 94720, USA; 20Department of Biological Sciences, Graduate school of Science, The University of Tokyo, Tokyo 113-0033, Japan; 21Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Vídeňská 1083, 142 20 Prague 4, Czech Republic; 22Center for Advanced Marine Research, Ocean Research Institute, University of Tokyo, Nakano, Tokyo 164-8639, Japan; 23Department of Pediatrics, University of South Florida, St. Petersburg, Florida 33701 USA; 24Bioinformatics and Pattern Discovery Group, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA; 25Center of Marine Biotechnology, University of Maryland Biotechnology Institute, Baltimore, Maryland 21202 USA; 26Sunnybrook Research Institute and Department of Medical Biophysics, University of Toronto, Toronto, Ontario M4N 3M5, Canada; 27Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland; 28Department of Biology, University of Victoria, Victoria, B.C., V8W 3N5, Canada; 29Shimoda Marine Research Center, University of Tsukuba, S-10-1, Shimoda, Shizuoka, 415-0025 Japan; 30Department of Molecular Immunology, Institute of Development, Aging and Cancer, Tohoku University, Sendai 980-8575, Japan; 31Department of Zoology, Graduate School of Science, Kyushu University, Kyoto 606-8502, Japan; 32Department of Bioscience, Faculty of Bioscience, Nagahama Institute of Bio-Science and Technology, Nagahama, Shiga 526-0829, Japan; 33State Key Laboratory of Biocontrol, Department of Biochemistry, College of Life Sciences, Sun Yat-Sen (Zhongshan) University, Guangzhou, People’s Republic of China; 34Institute for Environmental and Gender-Specific Medicine, Juntendo University, Chiba 279-0021, Japan; 35Children’s Hospital of Oakland Research Institute, Oakland, California 94609, USA

Acknowledgments

This work was funded by grants from the National Science Foundation, USA (I.Z.H.), National Institutes of Health, USA (G.W.L.), the Wellcome Trust (P.W.H.), BBSRC, UK (D.E.K., T.B., P.W.H.), and MEXT, Japan (H.S., N.S.), Center for Applied Genomics MSMT and Academy of Sciences, Czech Republic (Z.K.), and the 21st Century and Global COEs at Kyoto University (N.S.), from Ministerio de Educación y Ciencia, Spain (J.G.-F.), MIUR Italy, FIRB 2001 BAU01WAFY (S.C., M.P.), JSPS, Japan (H.S.), and MENRT, CNRS and CRESCENDO, a European Union Integrated Project of FPs (V.L., M.S.).

References


Received October 26, 2007; accepted in revised form February 24, 2008.