Domain shuffling and the evolution of vertebrates

Takeshi Kawashima, Shuichi Kawashima, Chisaki Tanaka, et al.

Genome Res. 2009 19: 1393-1403 originally published online May 14, 2009
Access the most recent version at doi:10.1101/gr.087072.108

Letter

Domain shuffling and the evolution of vertebrates

Takeshi Kawashima,[1,2,3,9] Shuichi Kawashima,[4] Chisaki Tanaka,[5] Miho Murai,[6] Masahiko Yoneda,[6] Nicholas H. Putnam,[2,7] Daniel S. Rokhsar,[2,7] Minoru Kanehisa,[4,8] Nori Satoh,[1] and Hiroshi Wada[5,9]

[1]*Okinawa Institute of Science and Technology, Uruma, Okinawa 904-2234, Japan;* [2]*Department of Energy Joint Genome Institute, Walnut Creek, California 94598, USA;* [3]*Japanese Society for Promotion of Sciences, Tokyo 102-8471, Japan;* [4]*Human Genome Center, Institute of Medical Science, University of Tokyo, Tokyo 108-8639, Japan;* [5]*Graduate School of Life and Environmental Sciences, University of Tsukuba, Tsukuba 305-8572, Japan;* [6]*Department of Nursing & Health, School of Nursing & Health, Aichi Prefectural University, Nagoya 463-8502, Japan;* [7]*Center for Integrative Genomics, University of California, Berkeley, Berkeley, California 94720, USA;* [8]*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan*

The evolution of vertebrates has included a number of important events: the development of cartilage, the immune system, and complicated craniofacial structures. Here, we examine domain shuffling as one of the mechanisms that contributes novel genetic material required for vertebrate evolution. We mapped domain-shuffling events during the evolution of deuterostomes with a focus on how domain shuffling contributed to the evolution of vertebrate- and chordate-specific characteristics. We identified ~1000 new domain pairs in the vertebrate lineage, including ~100 that were shared by all seven of the vertebrate species examined. Some of these pairs occur in the protein components of vertebrate-specific structures, such as cartilage and the inner ear, suggesting that domain shuffling made a marked contribution to the evolution of vertebrate-specific characteristics. The evolutionary history of the domain pairs is traceable; for example, the Xlink domain of aggrecan, one of the major components of cartilage, was originally utilized as a functional domain of a surface molecule of blood cells in protochordate ancestors, and it was recruited by the protein of the matrix component of cartilage in the vertebrate ancestor. We also identified genes that were created as a result of domain shuffling in ancestral chordates. Some of these are involved in the functions of chordate structures, such as the endostyle, Reissner's fiber of the neural tube, and the notochord. Our analyses shed new light on the role of domain shuffling, especially in the evolution of vertebrates and chordates.

[Supplemental material is available online at www.genome.org.]

The question of how novel structures are created by changing genetic information is one of the most challenging issues in evolutionary biology. It is generally accepted that the morphological features of various multicellular animals are built on a common set of genes. Carroll et al. (2001) and Davidson (2006) proposed that the novel features emerged as a result of altered gene expression patterns. Novel genetic material has also contributed to the evolution of novel structures. Much attention has been focused on gene duplications as a mechanism for the evolution of novel genetic material. Particularly in the case of vertebrate evolution, whole genome duplications that occurred twice in ancestral vertebrates have been regarded as the main force driving the evolution of novel structures (Holland et al. 1994). As an additional mechanism, we focus here on domain shuffling (Chothia et al. 2003; Babushok et al. 2007). Several different molecular mechanisms for domain shuffling have been proposed. Since the domains are often correlated with exon boundaries, exon shuffling is believed to be one of the major forces driving domain shuffling (Liu and Grigoriev 2004). In addition, the introns at the boundaries of domains show a marked excess of symmetrical phase combinations, and, consequently, these domains may be inserted without changing the reading frame of ancestral genes (Kaessmann et al. 2002; Vibranovski et al. 2005). Some other mechanisms might have been involved in domain shuffling, such as the simple fusion of genes and recruitment of mobile elements (Babushok et al. 2007; Ekman et al. 2007).

In contrast to thorough investigation of the mechanistic aspects of domain shuffling, the contribution of the new genes created by domain shuffling to the evolution of phenotype has not been sufficiently explored. Because domain shuffling occurs more frequently in metazoan lineages than in unicellular organisms, suggesting that domain shuffling played an important role in the evolution of multicellularity (Patthy 2003; Ekman et al. 2007), we reasoned that domain shuffling also contributed to the elaboration of metazoan body plans. In this study, we comprehensively investigated the domain shuffling events during deuterostome evolution. The complete amphioxus genome sequence provides sufficient deuterostome genomic sequences (Dehal et al. 2002; Sea Urchin Genome Sequencing Consortium 2006; Putnam et al. 2008) to map the domain shuffling events accurately. We listed gene models that were created by domain shuffling in the ancestors of vertebrates, and examined how domain shuffling contributed to the evolution of new genes for vertebrate-specific characteristics. The contribution of domain shuffling was also examined in the evolution of chordates.

Results and Discussion

Detection of domain shuffling in deuterostome evolution

We identified pairs of protein domains found in single-gene models, including the human, mouse, rat, chicken, frog, pufferfish,

[9]**Corresponding authors.**
E-mail 98champ@msg.biglobe.ne.jp; fax 81-29-853-4671.
E-mail takeshik@oist.jp; fax 81-98-934-5622.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.087072.108.

zebrafish, ascidian, and amphioxus genomes. We took the order of the domains into account, so that the domain pair A and B (in order from the N terminus) was treated as a distinct pair from domains in the order B and A. This is because these domain pairs with different orders are more likely to have been gained by distinct domain shuffling events than by divergence from a common ancestral pair. We surveyed domain pairs found in all splicing variants in order to detect all domain pairs transcribed in the genomes. As outgroups, domain pairs from the proteomes of a sea urchin, fly, mosquito, gastropod snail, two species of nematode, sea anemone, choanoflagellate, slime mold, yeast, and plant (*Arabidopsis*) were obtained (Supplemental Table 1). In the 20 proteomes, profile-HMM (HMMER) searches of the Pfam database (*E*-value cutoff of 10^{-3}; see Methods) recognized 4649 protein domains. Of these, 2727 were found paired with other domains. Approximately 2000–5500 domain pairs were found in the gene models of each species. For example, 4057 domain pairs were found in the human gene models, with 2251 found in ascidian and 5387 in amphioxus. From these datasets, we surveyed domain pairs shared by the major deuterostome taxa (Table 1). By this method, triplets of domains were detected as two or three domain pairs in one gene model (e.g., SNF2 histone linker PHD RING helicase or attractin in Table 2).

Table 1. Shared domain combinations and unique domains

	Class I combinations[a]	Class II combinations[b]	Unique domains[c]
VCBSO	0	1326	1532
VCBS_	0	20	2
VCB_O	0	355	116
VC_SO	0	129	25
V_BSO	0	458	161
_CBSO	0	15	0
VCB__	0	43	5
VC_S_	2	4	1
V_BS_	2	51	4
CBS	0	8	0
VC__O	0	143	11
V_B_O	0	496	83
_CB_O	0	28	2
V__SO	0	131	25
_C_SO	0	4	0
__BSO	0	92	10
VC___	1	34	3
V_B__	21	157	18
_CB__	0	35	0
V__S_	2	43	1
_C_S_	0	4	0
__BS_	2	71	1
V___O	0	439	118
_C__O	0	21	2
__B_O	0	242	25
___SO	0	44	3
V____	368	859	209
_C___	0	79	0
__B__	32	1933	7
___S_	10	455	2
____O	0	2951	361

[a]Number of shared domain combinations for which one of the domains was found only in that species.
[b]Number of shared domain combinations for which the domains were found in other species.
[c]Numbers of domains uniquely found in the species.
(V) Vertebrates; (C) *Ciona* (ascidian); (B) *Branchiostoma* (amphioxus); (S) *Strongylocentrotus* (sea urchin); (O) other species examined.

Based on the table of shared domain pairs (Table 1), we mapped the gained and lost domain pairs on a phylogenetic tree that had been previously deduced by using concatenated amino acid sequences (Fig. 1A; Putnam et al. 2008). If more than two taxa shared domain pairs, we regarded the pairs as having been acquired in the last common ancestor. If pairs were shared in only some of the descendent taxa, we assumed that those pairs were lost in the rest of the descendent taxa. This method to reconstruct the gain/loss of domain pairs does not follow the maximum parsimony principle. Rather, to filter domain pairs that are more likely to have been achieved by independent domain shuffling events, we examined the overall domain structures of the gene models and defined a convergent index (CI) (see Methods). Although these independent acquisitions were thought to be rare (Gough 2005), we detected several domain pairs that are likely to have been achieved independently. By filtering domain pairs of $CI \geq 0.5$, we found that 15 pairs that were shared by the sea urchin and some chordates had rather different domain architectures (Supplemental Fig. 1). Therefore, these domain pairs were more likely to have been acquired by convergence. Among 269 domain pairs that were assumed to have been gained in the ancestors of chordates, 13 pairs showed a $CI \geq 0.5$, and thus were suggested to have evolved convergently. No domain pairs shared by ascidians and vertebrates, and new pairs acquired in the vertebrate lineage, showed a $CI \geq 0.5$.

After filtering domain pairs of $CI < 0.5$, 209 pairs were assumed to have been acquired in the common ancestors of deuterostomes, 256 pairs in the common ancestors of chordates, and 35 in the common ancestors of ascidian and vertebrates. Regarding the gained and lost pairs in each deuterostome taxon, it is notable that many of the new pairs (1965 pairs) were acquired in the amphioxus lineage, which exceeded the number observed in the vertebrate lineage. Because some ambiguity remains in the gene models predicted from the genome sequences and EST sequences, some of the predicted gene models may combine two distinct transcripts. If these misassigned gene models are abundant in the amphioxus genome, the number of amphioxus-specific domain pairs may be artifactually high. However, we think it is not the case here, because after filtering of the gene models by EST evidence (gene models supported by stretches of EST sequences and thus proven presence of the transcripts), 847 amphioxus-specific domain pairs were still found, while only 35,280 Pfam domains were recognized in the filtered gene models compared with 121,370 domains recognized in all amphioxus gene models (Supplemental Table 2). Therefore, we think this observation is unlikely to be an artifact of gene modeling. In amphioxus-specific domain pairs, Lectin_C, LRR_1, TPR_1, and TPR_2 occurred most frequently together with other domains, which also frequently were involved in domain shuffling in other lineages. On the other hand, some domains such as Death and F5_F8_type_C were more frequently involved in amphioxus compared with other lineages. Death is involved in apoptosis, and F5_F8_type_C is found in membrane protein or coagulation factors. This is consistent with previous reports that the gene repertoires for apoptosis and the innate immune system are expanded in amphioxus (Holland et al. 2008; Huang et al. 2008). Conversely, many of the pairs were lost in the ascidian lineage, which is not surprising considering that extensive gene loss has occurred in that lineage (Dehal et al. 2002). To correlate genomic complexity with phenotypic variation, it is worth noting that amphioxus species, which have more or less retained their ancestral morphology, acquired several new domain pairs, whereas

Table 2. Class II domain combinations specific to vertebrates

Human genes	Accession no.[a]	Domain 1		Domain 2		−3[b]	−5	−10	−15	−20	Other examples of human genes[a]
A. Function that many kinds of cells use											
Baculoviral IAP repeat-containing 3	Q13489	PF00653	BIR	PF00619	CARD	+	+	+	+	+	Q13075, Q13490
		PF00619	CARD	PF00097	zf-C3HC4	+	−	−	−	−	Q13490
Meprin A, alpha (PABA peptide hydrolase)	Q16819	PF00917	MATH	PF00008	EGF	+	+	+	−	−	
Occludin	Q16625	PF01284	MARVEL	PF07303	Occludin_ELL	+	+	+	+	+	NP_001033692
SNF2 histone linker PHD RING helicase	NP_001036148	PF00538	Linker_histone	PF00271	Helicase_C	+	−	−	−	−	
		PF00538	Linker_histone	PF00628	PHD	+	+	−	−	−	
Ubiquitin-conjugating enzyme E2B	NP_060784	PF05743	UEV	PF00056	Ldh_1_N	+	+	+	+	−	
Protein S (alpha)	P07225	PF00594	Gla	PF00054	Laminin_G_1	+	+	+	+	−	Q14393
Protein Z, vitamin K-dependent plasma glycoprotein	P22891	PF00594	Gla	PF00051	Kringle	+	+	+	+	−	
		PF00594	Gla	PF00089	Trypsin	+	+	+	+	+	P04070, P22891, P08709
B. Cell–cell communication											
Signaling receptors and ligands											
Agrin	O00468	PF00053	Laminin_EGF	PF01390	SEA	+	+	+	+	+	
Amyloid beta (A4) precursor protein	P05067	PF02177	A4_EXTRA	PF00014	Kunitz_BPTI	+	+	+	+	+	Q06481
Anthrax toxin receptor 1	Q9H6X2	PF00092	VWA	PF05587	Anth_Ig	+	+	+	+	+	P58335
		PF00092	VWA	PF05586	Ant_C	+	+	+	+	+	
Attractin	O75882	PF07646	Kelch_2	PF00059	Lectin_C	+	+	−	−	−	NP_997186
		PF00059	Lectin_C	PF01437	PSI	+	+	−	−	−	NP_997186
		PF01344	Kelch_1	PF00059	Lectin_C	+	+	−	−	−	NP_997186
		PF01437	PSI	PF00059	Lectin_C	+	+	−	−	−	NP_997186
Carboxypeptidase X (M14 family), member 2	Q8N436	PF00754	F5_F8_type_C	PF00246	Peptidase_M14	+	+	+	+	+	NP_001120, Q96SM3
Complement component 4A	NP_009224	PF01835	A2M_N	PF01821	ANATO	+	+	+	+	+	P01031, P01024, NM_001013640
Complement component 7	P10643	PF01823	MACPF	PF00084	Sushi	+	+	+	+	+	P13671
Discoidin, CUB and LCCL domain containing 1	Q8N8Z6	PF00431	CUB	PF03815	LCCL	+	+	+	+	+	Q96PD2
Growth-arrest-specific 6	Q14393	PF00594	Gla	PF02210	Laminin_G_2	+	+	+	+	−	P07225
Protease, serine, 7 (enterokinase)	P98073	PF00629	MAM	PF00089	Trypsin	+	+	+	+	+	
Integrin, beta 4	P16144	PF03160	Calx-beta	PF00041	fn3	+	+	+	+	+	
		PF00362	Integrin_beta	PF03160	Calx-beta	+	+	+	+	+	
		PF07965	Integrin_B_tail	PF03160	Calx-beta	+	+	+	+	+	
		PF07974	EGF_2	PF03160	Calx-beta	+	+	+	−	−	
Serine peptidase inhibitor, Kunitz type 1	O43278	PF07502	MANEC	PF00014	Kunitz_BPTI	+	+	+	+	+	
		PF00014	Kunitz_BPTI	PF00057	Ldl_recept_a	+	+	+	−	−	
Latrophilin1	O94910	PF02191	OLF	PF00002	7tm_2	+	+	+	+	+	O95490, Q9HAR2
Leucine-rich, glioma-inactivated 1	O95970	PF00560	LRR_1	PF03736	EPTP	+	+	−	−	−	Q8N0V4, Q8N145, Q8N135
Protein tyrosine phosphatase, receptor type, G	P23470	PF00194	Carb_anhydrase	PF00102	Y_phosphatase	+	+	+	+	+	P23471
		PF00194	Carb_anhydrase	PF00041	fn3	+	+	+	+	−	P23471
HtrA serine peptidase 1	Q92743	PF00050	Kazal_1	PF00595	PDZ	+	+	−	−	−	P83110
		PF00050	Kazal_1	PF00089	Trypsin	+	+	−	−	−	P83110
		PF07648	Kazal_2	PF00089	Trypsin	+	+	+	−	−	P83110, P83105, P05156
		PF00089	Trypsin	PF00595	PDZ	+	+	+	−	−	P83110
		PF07648	Kazal_2	PF00595	PDZ	+	+	+	−	−	P83110
Stabilin 1	Q9NY15	PF02469	Fasciclin	PF00193	Xlink	+	+	+	+	+	Q8WWQ8
von Willebrand factor	P04275	PF00093	VWC	PF00092	VWA	+	+	+	+	+	

(*Continued*)

Table 2. Continued

Human genes	Accession no.[a]	Domain 1		Domain 2		−3[b]	−5	−10	−15	−20	Other examples of human genes[a]
Intracellular signal transduction											
Amyotrophic lateral sclerosis 2 (juvenile)	Q96Q42	PF00415	RCC1	PF02204	VPS9	+	+	−	−	−	
DDHD domain containing 2	NP_056029	PF02825	WWE	PF00536	SAM_1	+	+	−	−	−	
F-box protein 41	Q8TF61	PF00096	zf-C2H2	PF00646	F-box	+	−	−	−	−	
Obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF	Q5VST9	PF00041	fn3	PF00612	IQ	+	+	−	−	−	
Regulator of G-protein signaling 12	O14924	PF00595	PDZ	PF02188	GoLoco	+	+	−	−	−	
Rho-associated, coiled-coil containing protein kinase 2 (ROCK2)	O75116	PF02185	HR1	PF00169	PH	+	+	−	−	−	Q13464
		PF00069	Pkinase	PF02185	HR1	+	+	+	−	−	Q13464
		PF00433	Pkinase_C	PF02185	HR1	+	+	+	−	−	Q13464
Sel-1 suppressor of lin-12-like	Q9UBV2	PF00040	fn2	PF08238	Sel1	+	+	+	+	+	
Multimerin 2	Q9H8L6	PF07546	EMI	PF00386	C1q	+	+	+	+	+	Q9BXX0, Q9Y6C2, Q13201
ECM and cell adhesion											
Aggrecan	P16112	PF00047	ig	PF00193	Xlink	+	+	−	−	−	Q9GZV7, Q96S86, Q86UW8
		PF07686	V-set	PF00193	Xlink	+	+	+	+	−	Q9GZV7, Q96S86, Q86UW8
Cartilage acidic protein 1	Q9NQ79	PF07593	UnbV_ASPIC	PF07645	EGF_CA	+	+	+	+	+	
Coagulation factor C homolog, cochlin (*Limulus polyphemus*)	O43405	PF03815	LCCL	PF00092	VWA	+	+	+	+	+	
Collagen, type VI, alpha 3	P12111	PF01391	Collagen	PF00014	Kunitz_BPTI	+	+	+	+	+	Q02388
Fibronectin 1	P02751	PF00039	fn1	PF00041	fn3	+	+	+	+	+	
		PF00039	fn1	PF00040	fn2	+	+	+	+	+	Q04756
		PF00040	fn2	PF00041	fn3	+	+	+	+	+	
		PF00040	fn2	PF00039	fn1	+	+	+	+	+	P00748, Q04756
		PF00041	fn3	PF00039	fn1	+	+	+	+	+	
Fibulin 1	P23142	PF01821	ANATO	PF07645	EGF_CA	+	+	+	+	−	P98095
Nidogen 1	P14543	PF00086	Thyroglobulin_1	PF00058	Ldl_recept_b	+	+	+	+	+	Q14112
		PF07474	G2F	PF00086	Thyroglobulin_1	+	+	+	+	+	Q14112
		PF06119	NIDO	PF00086	Thyroglobulin_1	+	+	+	+	+	
		PF07645	EGF_CA	PF00086	Thyroglobulin_1	+	+	+	+	+	
Procollagen C-endopeptidase enhancer	Q15113	PF00431	CUB	PF01759	NTR	+	+	+	+	+	Q9UKZ9
Tectorin alpha	O75443	PF06119	NIDO	PF00100	Zona_pellucida	+	+	+	+	+	
		PF06119	NIDO	PF01826	TIL	+	+	+	+	+	
Vitronectin	P04004	PF01033	Somatomedin_B	PF00045	Hemopexin	+	+	+	+	+	Q92954
Matrix metallopeptidase 2	P08253	PF01471	PG_binding_1	PF00040	fn2	+	+	+	−	−	P14780
C. Transcription factors											
Arginine-glutamic acid dipeptide (RE) repeats	Q9P2R6	PF00249	Myb_DNA-binding	PF01448	ELM2	+	+	−	−	−	
		PF00320	GATA	PF01448	ELM2	+	+	−	−	−	
NACC family member 2, BEN and BTB (POZ) domain containing	NP_653254	PF00651	BTB	PF06665	DUF1172	+	+	+	+	+	NP_443108
Interleukin enhancer binding factor 3, 90kDa	Q12906	PF07528	DZF	PF00035	dsrm	+	+	+	+	+	

Gene models were classified according to Lee et al. (1999).
[a]NCBI Reference Sequence locus identifiers or UniProt IDs.
[b]Plus (+) and minus (−) indicate whether the combination was vertebrate-specific using the range of cutoff values; e.g., + in the −3 column indicates that the combination was vertebrate-specific at a cutoff *E*-value of 10^{-3}.

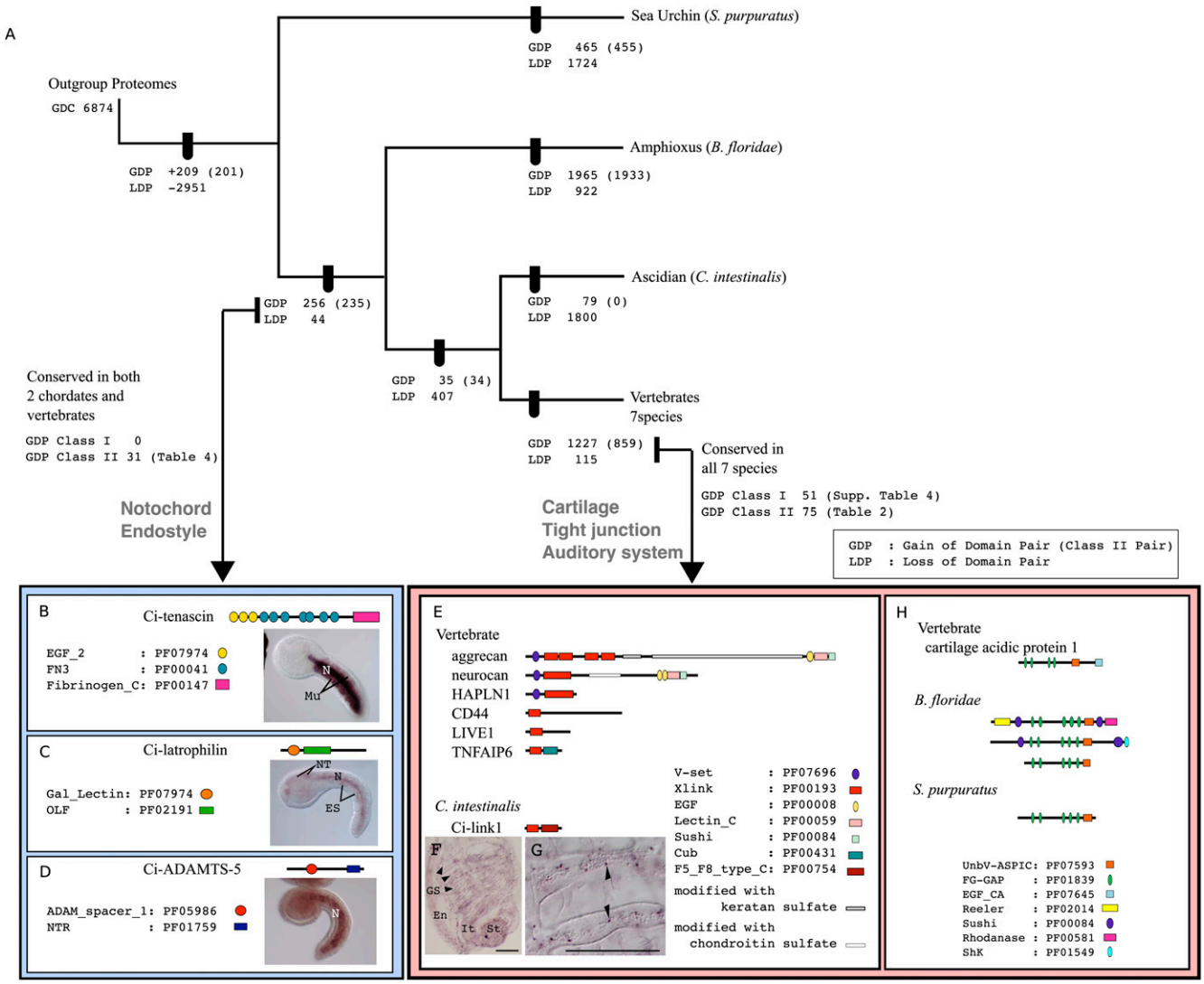

Figure 1. Domain shuffling during the evolution of deuterostomes. (*A*) Gain and loss of the domain pairs were mapped on the deuterostome phylogenetic tree constructed using molecular phylogenetic analyses of multiple gene sequences (Putnam et al. 2008). Numbers on each node present gains and loss of domain pairs after screening CI < 0.5 (see Methods for the definition of CI). Gains of Class II domain pairs are shown in brackets. (Arrows) Domain pairs acquired in the ancestral chordates, and those in ancestral vertebrates, that were examined in detail. Among 1227 pairs that were acquired in the ancestral vertebrates, 51 class I domain pairs and 75 class II domain pairs were conserved in seven vertebrate species examined. Some of these were involved in vertebrate-specific characters such as cartilage (*E–H*), tight junctions, and auditory systems (see text for details). Among 256 domain pairs unique to the chordates, 31 pairs were conserved in ascidian, amphioxus, and more than one species of vertebrates. Some of them were involved in chordate-specific characters such as the notochord (*B–D*) and endostyle (see text for details). (*B–D*) Tenascin, latrophilin, and *ADAMTS-5*, which were created by domain shuffling in ancestral chordates, are expressed in the notochords of ascidian embryos. Symbols are indicated with the domain ID and Pfam accession numbers (Finn et al. 2006). (*B*) *Ci*-tenascin is expressed in the ascidian notochord (N) and muscle cells (Mu). (*C*) *Ci*-latrophilin expression was detected in the notochord, neural tube (NT), and endodermal strand (ES) of ascidian larvae. (*D*) Expression of *Ci-ADAMTS-5* was restricted to the ascidian notochord. (*E*) Domain structures of the proteins that include an Xlink domain. (*F,G*) Expression of *Ci-link1* was observed in some of the blood cells from an ascidian juvenile. (Arrowheads) Some of the *Ci-link1*-positive cells. Scale bars, 50 μm. (En) Endostyle, (GS) gill slits, (St) stomach, (It) intestine. (*H*) Domain structures of cartilage acidic protein and proteins containing an ASPIC-and-UnbV domain encoded by amphioxus gene models.

many pairs were lost in the ascidian lineage, which has evolved extensive morphological variation.

Domain shuffling in the ancestral vertebrates

The vertebrate lineage acquired 1227 domain pairs. Of these, 137 pairs were conserved among all seven species, whereas the rest were restricted to certain taxa (Fig. 1A). Surveying the NCBI Conserved Domain Database (CDD) with each domain pair using CDART searches (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) identified 11 of the pairs in invertebrates such as bees, *Drosophila*, or *Caenorhabditis elegans* (Supplemental Table 3; some genes from *Drosophila* or other species in the CDD were missing from the proteome database because the gene models were incomplete). Of the remaining 126 pairs, 51 contained one or two domains found only in vertebrate genomes (Supplemental Table 4). These pairs, which we designated as class I domain pairs, therefore may not have evolved via domain shuffling, but instead

may have resulted from de novo establishment of new domain sequences. Alternatively, some domain sequences might be too divergent to be detected in outgroups, and may not be novel combinations. In contrast, the remaining 75 pairs consisted of two domains, both of which were found in invertebrates as well as in vertebrates (Table 2). We designate them as class II domain pairs, which certainly evolved via domain shuffling. Only two gene models contain both class I and class II domain pairs (latrophilin1 and amyloid beta [A4] protein precursor).

We first asked whether these class II domain pairs were acquired by exon shuffling events. We examined the exon–intron structures of 75 pairs that were found in 143 human gene models. In 24% of these gene models, the two domains are encoded by exons separated by one intron (referred to as "neighbor" in Fig. 2 and Supplemental Table 5), including those in which either domain is encoded by more than two exons. In 65%, the two domains were encoded by exons separated by two or more introns and one or more exons ("farther" in Fig. 2 and Supplemental Table 5). In only 11%, the two domains were not separated by any introns and, thus, there is an exon that codes at least part of both domains ("same" in Fig. 2, Supplemental Table 5). Although in some domains the boundary of the exon does not always correlate with the boundary of the domain, and the shuffling of these exons provides a stretch of amino acid sequences between the domains (Supplemental Table 5), these data suggest that most of the new domain pairs were acquired by exon shuffling. Supporting this idea, for more than half of the domains, the introns forming their boundaries have symmetrical phase combinations, and they can be inserted without disrupting the reading frame of the ancestral genes (Table 3). In addition, many of the domains are abutted by phase 1 introns at both of their boundaries, while phase 0 introns are dominant in the entire genomes of human and amphioxus (Fig. 3; Table 3; Supplemental Table 5). This is consistent with a previous study in which large numbers of introns participating in exon shuffling are phase 1 (Patthy 1999). Note also that while the novel domain is inserted at either the N or C terminus in most cases (Ekman et al. 2007), we found one clear example in which a new domain was inserted between the domains of the ancestral gene (Supplemental Fig. 2).

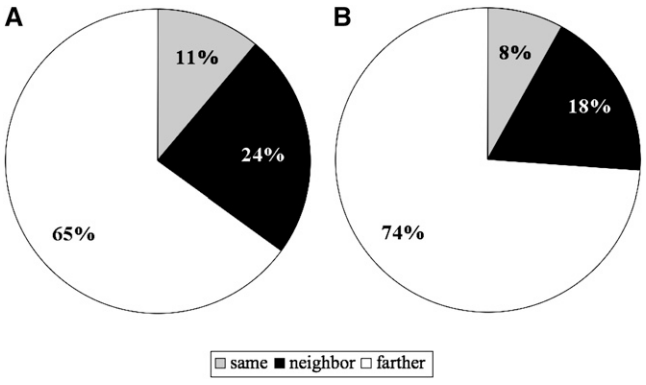

Figure 2. Position of the domain sequences relative to introns. (Neighbor) Domains are encoded by exons separated by one intron, (farther) domains are encoded by exons separated by one or more exons and two or more introns, (same) domains are encoded by the same exon. (*A*) Seventy-five class II pairs of the vertebrates (Table 2) were found in 143 gene models in the human genome. The relative position of the domain sequences in these gene models is summarized. Details of the data are given in Supplemental Table 5. (*B*) Thirty-one class II pairs of the chordates (Table 4) were found in 76 gene models in the human genome. The relative position of the domain sequences in these gene models is summarized. Details of the data are given in Supplemental Table 7.

Table 3. Intron phases of the domains

Intron phase pair-type	Vertebrate-specific Class II		Chordate-specific Class II	
	Observed	*P*-value	Observed	*P*-value
0-0	7	1.00	15	0.974
1-1	115	0	41	0
2-2	2	0.997	4	0.735
0-1	31	0.566	15	0.637
0-2	6	1.00	4	0.989
1-2	9	0.954	1	0.995
1-0	29	0.709	11	0.923
2-0	5	1.00	10	0.653
2-1	13	0.735	10	0.214
Total	217		111	

Statistical analyses were performed based on the observed intron phases in human genomes (Phase 0, 66,882; Phase 1, 46,696; Phase 2, 32,260; total: 145,838).

Among these class II domain pairs, some were obviously involved in the evolution of vertebrate-specific characteristics. As an example, aggrecan—the most abundant noncollagenous protein in cartilage—was created by domain shuffling in ancestral vertebrates, as has been previously suggested (Upholt et al. 1994; Valhmu et al. 1995; Patthy 2003). Aggrecan consists of five types of domains, and the combination of the immunoglobulin domain (V-set: PF07686; ig: PF00047) and extracellular link domain (Xlink: PF00193) was identified as a novel domain pair (Fig. 1E). Aggrecan is a heavily glycosylated protein. Repeated amino acid sequences that are modified with keratan sulfate and chondroitin sulfate moieties are located between the Xlink domain and the EGF domain (Fig. 1E), although these repeats were not recognized as protein motifs in the Pfam database. The Xlink domains are essential for binding with glycosaminoglycan hyaluronic acid. The molecular complex of hyaluronic acid and heavily glycosylated aggrecan provides the tensile strength that allows cartilage to absorb shock and resist compression in joints, and mutations in aggrecan cause severe cartilage defects and result in dwarfism (Watanabe and Yamada 1999).

To trace the evolutionary history of aggrecan genes, we examined each domain in the amphioxus and ascidian genomes. Since the V-set, ig, Lectin C, EGF, and Sushi domains are relatively abundant in both genomes, we traced the origin of the Xlink domain. Two ascidian gene models and 20 amphioxus gene models contained Xlink domains. One of the ascidian gene models (*Ci-link1*) encoded a single Xlink domain and an F5/8 type C domain (PF00754), and is expressed in some blood cells in juveniles (Fig. 1F,G). In vertebrates, the Xlink domain is found not only in aggrecan and its paralogs, but also in CD44, TNFAIP6 (also known as TSG6), LYVE1, and stabilin 1. These proteins contain single Xlink domains (Fig. 1E) and are involved in lymphocyte migration (Ponta et al. 2003; Kzhyshkowska et al. 2006). Therefore, the molecular structure and function of Ci-link1 more closely resemble those of these single-link proteins than those of aggrecan. These results suggest that the Xlink domain is a component of surface molecules on blood cells in protochordate ancestors. Moreover, the Xlink domain was combined with other domains, such as the V-set domain, in a vertebrate ancestor, and the protein was used as a component of cartilage—a unique vertebrate structure. Cartilage acidic protein 1 was also created by combining

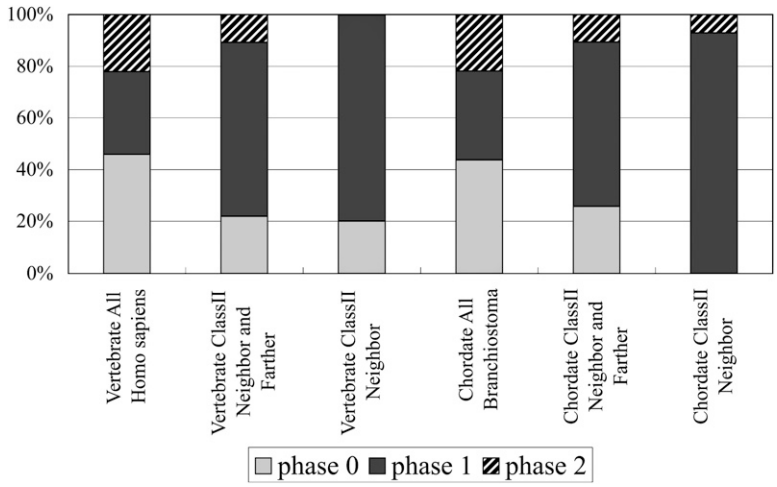

Figure 3. Phase of introns involved in domain shuffling. The intron phases of exons that encode the C terminus of domain 1 and that of the N terminus of domain 2 were counted here. Although the phase 0 intron is dominant in the gene models of *Homo* and *Branchiostoma*, the phase 1 intron is dominant in the class II pairs. The intron phase of the class II pairs was analyzed using the human gene models. Details of the data are given in Supplemental Tables 5 and 7.

domains, including the ASPIC-and-UnbV (UnbV_ASPIC: PF07593) and calcium-binding EGF domains (EGF_CA: PF07645; Fig. 1H).

Among the vertebrate-specific domain pairs, we identified two proteins that contribute to cell junctions. Occludin—a component of tight junctions—contains a domain pair of a membrane-associating domain (MARVEL: PF01284) and an occludin_ELL domain (PF07303) that is specific to vertebrates (Supplemental Fig. 3). The amphioxus and ascidian proteomes possess an occludin_ELL domain in the RNA polymerase II elongation factor ELL. Conversely, 17 amphioxus genes and four ascidian genes contain sequences encoding the MARVEL domain, and for each of these genes, the MARVEL domain was the only recognized functional domain. The structures of protochordate proteins containing the MARVEL domain were more similar to MAL and physins, which are involved in vesicle trafficking (Sánchez-Pulido et al. 2002). Therefore, a gene in an ancestral vertebrate acquired the sequence encoding the occludin_ELL domain, resulting in a new protein that functioned as a component of tight junctions, which are found only in vertebrates and ascidians together with the major tight junction component claudin. Interestingly, occludin is not an essential component of tight junctions, but rather has been suggested to be involved in the regulation of tight junctions by mediating extracellular cell adhesion signals (Saitou et al. 2000). Therefore, acquiring occludin as a component of tight junctions may have provided vertebrates with additional uses for these complexes. The gap junction protein, alpha 1, 43 kDa (connexin43), was also found to contain a class I vertebrate-specific domain pair, which was composed of the Connexin (PF00029) and Connexin43 (PF03508) domains (Supplemental Fig. 3). The Connexin43 domain, which is found only in vertebrates, serves as a regulatory domain for the gap junction protein alpha 1, 43 kDa (connexin43) through its phosphorylation status (Axelsen et al. 2006).

Another notable class of genes that contained sequences encoding vertebrate-specific domain pairs included those involved in the vertebrate auditory system. Tectorin-alpha is a major component of the tectorial membrane in the mammalian inner ear; in humans, a missense mutation in the ZP domain of tectorin alpha causes hearing loss in the range of 50–80 dB (Verhoeve et al. 1998). Furthermore, a mutation in the LCCL domain of cochlin, a novel gene created by domain shuffling in an ancestral vertebrate, causes deafness (Supplemental Fig. 3; Robertson et al. 2006).

Note also that genes involved in neurogenesis and immune responses were also identified. Amyloid beta (A4) precursor and agrin are involved in synaptic physiology and synapse formation (Banks et al. 2003; Kamenetz et al. 2003), stabilin 1 is a scavenger receptor (Kzhyshkowska et al. 2006), and attractin is essential for initial immune cell clustering (Duke-Cohan et al. 1998). Complements C6/7 and C3/4/5 were also identified due to their unique combinations of domains (Azumi et al. 2003). Fibronectin, nidogen, and vitronectin were recognized as vertebrate-specific extracellular matrix (ECM) proteins among the class II domain pairs (Patthy 2003).

Although they may not have been achieved by domain shuffling, it is worth noting that various transcription factors (HOXC9, HNF1A, CDX2, NFIC, and ZFY) categorized as class I have acquired vertebrate-specific transactivation domains. It has been suggested that the evolution of the transactivation domain was critical for the evolution of arthropod leg morphogenesis (Galant and Carroll 2002).

Domain shuffling in the common ancestors of chordates

Of the 256 pairs acquired by the ancestral chordates (the last common ancestors of amphioxus, ascidians, and vertebrates), 43 pairs were conserved between the ascidian, amphioxus, and at least some of the vertebrate species. After surveying each domain pair in the CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml), 12 were found in other invertebrates (Supplemental Table 6) and were omitted from further analysis. The remaining 31 pairs (Table 4) were found in proteins encoded by 14 types of vertebrate genes. Sequences coding for all of these domains were found in nonchordate genomes; therefore, there were no class I pairs. Thirty-one pairs were found in 76 human gene models. Among these gene models, 18% were encoded by exons separated by one intron (referred to as "neighbor" in Fig. 2 and Supplemental Table 7) and 74% by exons separated by two or more introns and one or more exons ("farther" in Fig. 2 and Supplemental Table 7). Only 8% were encoded by the same exon ("same" in Fig. 2 and Supplemental Table 7). Many of the domains involved in chordate-specific pairs were also separated by phase 1 symmetric introns, and so these domain pairs were likely achieved by exon shuffling (Fig. 3; Table 3). In most cases, we confirmed that the protochordate genes listed in Table 4 encode the cognates of human genes by checking the best-hit genes in a BLASTP search.

Two unique domain pairs (the PF03098 An_peroxidase domain with Sushi or EGF) were encoded in the thyroid peroxidase gene, which is expressed in the endostyle or thyroid of chordates where it performs an essential role in iodide metabolism

Table 4. Domain combinations specific to chordates

Human gene	Accession no.[a]	Domain 1		Domain 2		−3[b]	−5	−10	−15	−20	*Branchiostoma* genes[c]	*Ciona* genes[c]	*Ciona* GC[d]	Expression[e]
Thyroid peroxidase	P07202	PF03098	An_peroxidase	PF00084	Sushi	+	+	+	–	–	fgenesh2_pg.scaffold_15000007 (3)	gw1.309.4.1 (1)	15898	Endostyle (Ogasawara et al. 1999)
		PF03098	An_peroxidase	PF00008	EGF	+	+	–	–	–	fgenesh2_pg.scaffold_15000007 (3)	gw1.309.4.1 (3)	15898	Endostyle (Ogasawara et al. 1999)
SCO-spondin homolog (Bos taurus)	Q8CG65	PF00094	VWD	PF00090	TSP_1	+	+	+	+	+	fgenesh2_pg.scaffold_343000009 (4)	fgenesh3_pg.C_scaffold_806000001 (1)	—	Neural tube (Supplemental Fig. 5)
von Willebrand factor	P04275	PF00094	VWD	PF00093	VWC	+	+	+	+	+	estExt_fgenesh2_pg.C_340142 (4)	estExt_fgenesh3_pg.C_550053 (2)	03415	Endostyle (Satou et al. 2005)
		PF01826	TIL	PF00093	VWC	+	+	+	+	+	estExt_fgenesh2_pg.C_340142 (4)	fgenesh3_pg.C_scaffold_26000037 (1)	15744	Testis, digestive gland (EST)
		PF00092	VWA	PF00094	VWD	+	+	+	+	+	fgenesh2_pg.scaffold_185000039 (2*)	fgenesh3_pg.C_chr_04q000231 (1)	33469	Heart (EST)
		PF01826	TIL	PF00092	VWA	+	+	+	+	+	fgenesh2_pg.scaffold_171000002 (5)	estExt_fgenesh3_pg.C_chr_07q0146 (1*)	—	Not detected
Complement component 7	P10643	PF00090	TSP_1	PF01823	MACPF	+	+	+	+	+	fgenesh2_pg.scaffold_105000080 (7)	estExt_fgenesh3_pg.C_chr_01q0717 (6)	14110	Blood (EST)
		PF00057	Ldl_recept_a	PF01823	MACPF	+	+	+	–	–	fgenesh2_pg.scaffold_105000080 (8)	estExt_fgenesh3_pg.C_chr_01q0717 (6)	14110	Blood (EST)
Surfactant protein D	P35247	PF01391	Collagen	PF00059	Lectin_C	+	+	+	+	+	estExt_fgenesh2_pg.C_2110004 (15*)	fgenesh3_pg.C_chr_05q000848 (2)	16667	Nervous system (Supplemental Fig. 5)
Tenascin N	Q9UQP3	PF00041	fn3	PF00147	Fibrinogen_C	+	+	+	+	+	estExt_fgenesh2_pg.C_1550085 (1)	gw1.09q.322.1 (1)	44683	Notochord, muscle (Fig. 1)
		PF07974	EGF_2	PF00147	Fibrinogen_C	+	+	+	+	+	estExt_fgenesh2_pg.C_1550085 (2)	gw1.09q.322.1 (1)	44683	Notochord, muscle (Fig. 1)
Latrophilin 1	O94910	PF02140	Gal_Lectin	PF02191	OLF	+	+	+	+	+	fgenesh2_pg.scaffold_156000051 (2)	fgenesh3_pg.C_chr_08q000436 (1)	—	Notochord, neural tube (Fig. 1)
ADAMTS-like 5	Q6ZMM2	PF05986	ADAM_spacer1	PF01759	NTR	+	+	–	–	–	fgenesh2_pg.scaffold_505000023 (1)	fgenesh3_pg.C_chr_04q000263 (1)	12258	Notochord (Fig. 1)
Insulin-like growth factor-binding protein 4	P22692	PF00219	IGFBP	PF00086	Thyroglobulin_1	+	+	+	+	+	estExt_fgenesh2_pg.C_1700020 (1)	estExt_fgenesh3_pg.C_chr_13q0261 (2)	01928	Adhesive palp
Polycystic kidney disease 1 (autosomal dominant)	P98161	PF00560	LRR_1	PF02010	REJ	+	+	–	–	–	fgenesh2_pg.scaffold_5000226 (1)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)

(*Continued*)

Table 4. Continued

Human gene	Accession no.[a]	Domain 1		Domain 2		−3[b]	−5	−10	−15	−20	*Branchiostoma* genes[c]	*Ciona* genes[c]	*Ciona* GC[d]	Expression[e]
		PF00560	LRR_1	PF01825	GPS	+	+	+	–	–	fgenesh2_pg.scaffold_5000226 (3*)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
		PF00560	LRR_1	PF00801	PKD	+	+	–	–	–	fgenesh2_pg.scaffold_5000226 (1)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
		PF00560	LRR_1	PF08016	PKD_channel	+	+	–	–	–	fgenesh2_pg.scaffold_5000226 (1)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
		PF00059	Lectin_C	PF08016	PKD_channel	+	+	+	+	+	fgenesh2_pg.scaffold_19000149 (6*)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
		PF00059	Lectin_C	PF00801	PKD	+	+	+	–	–	fgenesh2_pg.scaffold_218000066 (2*)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
		PF00560	LRR_1	PF01477	PLAT	+	+	–	–	–	fgenesh2_pg.scaffold_5000226 (1)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
		PF00560	LRR_1	PF00059	Lectin_C	+	+	+	+	+	fgenesh2_pg.scaffold_96000019 (2*)	estExt_fgenesh3_pg.C_chr_10q0087 (1)	31202	Egg (EST)
Adaptor-related protein complex 3, delta 1 subunit	O14617	PF01602	Adaptin_N	PF06375	BLVR	+	+	+	+	+	estExt_gwp.C_280211 (2)	estExt_genewise1.C_chr_05q0817 (1)	15999	Neural complex (EST)
		PF02985	HEAT	PF06375	BLVR	+	+	–	–	–	estExt_gwp.C_280211 (2)	estExt_genewise1.C_chr_05q0817 (1)	15999	Neural complex (EST)
Diacylglycerol kinase, delta 130 kDa	Q16760	PF00169	PH	PF00536	SAM_1	+	+	+	+	+	e_gw.81.48.1 (1)	gw1.08q.1141.1 (1)	11396	Heart, blood (EST)
		PF00169	PH	PF00609	DAGK_acc	+	+	+	+	+	e_gw.81.48.1 (1)	gw1.08q.1141.1 (1)	11396	Heart, blood (EST)
		PF00169	PH	PF00781	DAGK_cat	+	+	+	+	+	e_gw.81.48.1 (1)	gw1.08q.1141.1 (1)	11396	Heart, blood (EST)
		PF00169	PH	PF07647	SAM_2	+	+	+	+	+	e_gw.81.48.1 (1)	gw1.08q.1141.1 (1)	11396	Heart, blood (EST)
B-factor, preperdin[f]	NP_997631	PF00089	Trypsin	PF00084	Sushi	+	+	+	+	+	estExt_fgenesh2_pg.C_140131 (1*)	estExt_fgenesh3_pg.C_1280016 (1*)	06646	Endoderm, mesenchyme (Satou et al. 2005)
Meprin 1 beta[f]	Q61847	PF00629	VWD	PF07974	EGF_2	+	+	–	–	–	fgenesh2_pg.scaffold_132000059 (1*)	estExt_fgenesh3_pg.C_2000004 (1*)	16539	Heart, blood (EST)

[a]NCBI Reference Sequence locus identifiers or UniProt IDs.
[b]Plus (+) and minus (–) indicate whether the combination was chordate-specific using the range of cutoff *E*-values.
[c]Numbers shown with the gene models from *Branchiostoma* and *Ciona* indicate the numbers of gene models that contained the module combinations. Asterisks indicate that some of the gene models did not identify corresponding human genes after BLASTP searches; thus, these module combinations were possibly a result of convergence.
[d]*Ciona* GeneCollection (Satou et al. 2005).
[e]"EST" indicates that the expression pattern of the gene was confirmed using an EST from *C. intestinalis* (Satou et al. 2005).
[f]Gene models were not found using human proteomes, because the human gene model used here did not contain the module combinations.

(Ogasawara et al. 1999). Since the endostyle/thyroid is a chordate-specific structure, it is likely that a gene with an essential role in the endostyle was created in ancestral chordates. The gene encoding the mucous protein VWF is also specific to chordates. *Ciona VWF* (*Ciona* gene collection ID 03415) has been reported to be expressed in the endostyle (Satou et al. 2005). vWF may be a component of mucus excreted from the endostyle, and originally functioned in the filter-feeding systems of chordates, which were subsequently used to form the blood-clotting machinery of vertebrates. The main component of Reissner's fiber, a chordate-specific structure that extends from the infundibular organ of the neural tube, is SCO-spondin (Gonçalves-Mendes et al. 2003), which was created by domain shuffling in chordates. Our results demonstrated that *Ciona* SCO-spondin was expressed in the neural tube (Supplemental Fig. 4).

The development of an innate immune system in chordates was facilitated by the expansion of protein repertoires. Our analysis recognized three domain pairs in two types of genes involved in innate immune systems as chordate-specific: complement component C6/7/8/9 and surfactant protein D (SP-D). *Ciona C6/7/8/9* is expressed in the blood cells of ascidians and is also likely involved in their innate immune system (Satou et al. 2002, 2003). In contrast, the expression of *Ciona SP-D* is detected only in the embryonic nervous system (Supplemental Fig. 4).

These examples indicate that, as in ancestral vertebrates, domain shuffling played important roles in the phenotypic evolution of ancestral chordates. In this study, we have taken advantage of the availability of the *Ciona* gene collection (http://ghost.zool.kyoto-u.ac.jp/indexr1.html; Satou et al. 2005) to examine the expression of *Ciona* genes using in situ hybridizations. One of the important structures acquired by ancestral chordates was the notochord, which had an essential role in the formation of the chordate body plan (Satoh 2003). Therefore, it is notable that three of the identified ascidian genes, tenascin, latrophilin, and the ADAMTS-like 5 homolog, are expressed in the notochord (Fig. 1B–D). *Ciona* tenascin is expressed in both the notochord and muscle (Fig. 1B), whereas vertebrate tenascin genes are expressed in a wide variety of cells, including the notochord and somites (Tongiorgi 1999). One of the original functions of tenascin may lie in the formation of the notochord. Latrophilin is thought to be involved in exocytosis (Lang et al. 1998), whereas ADAMTS-like protein 5 is a metalloprotease involved in ECM digestion (Tortorella et al. 2002). These proteins may have been essential for the evolution of notochord cells, which secrete the ECM that makes up the notochord sheath.

We made a list of domain pairs that were inferred to have been acquired by the common ancestors of vertebrates, as well as those acquired by common ancestors of chordates. These genes are likely to have gained new functional roles by acquiring new domains, and are likely to be involved in phenotypic evolution. Indeed, we have presented some examples showing that novel genes drove the evolution of the characteristic features of vertebrates or chordates.

Recently, comparative developmental biology revealed that the morphological features of various multicellular animals are built on a common set of genes, and morphological variation is attributable to the altered expression patterns of preexisting genes (Carroll et al. 2001). However, these studies are primarily concerned with transcription factors and signaling molecules, with less attention to structural proteins. We propose that, in addition to the regulatory change of preexisting genes, dramatic molecular evolutionary events such as the evolution of novel genes via domain shuffling have contributed to phenotypic evolution.

Methods

We used the Pfam database (Finn et al. 2006; Pfam_ls, release 20.0; http://pfam.sanger.ac.uk), which contains 8296 conserved domains. Protein entries matching conserved domains in Pfam were identified using HMMER searches (Eddy 1998) of a set of 20 species with complete genomic sequences. Details of the data sources are presented in Supplemental Table 1. In total, 4649 of 7677 Pfam domains were found in these proteomes (*E*-value cutoff of 10^{-3}). Some scripts for parsing the data and for making pairwise domain matrixes were written using Ruby (http://www.ruby-lang.org) and BioRuby (http://bioruby.org/).

Our analyses did not consider the number of domains in each protein; regardless of whether domain A was present with a single copy of domain B or multiple copies of domain B, the pair of domain A and domain B was simply scored as present. We did, however, consider the order of the domains. A domain pair of A and B with A closer to the N terminus was treated as distinct from a pair with B closer to the N terminus. We examined domain pairs specific to vertebrates or chordates using HMMER searches and *E*-value cutoffs from 10^{-3} to 10^{-20}, and decided to use *E*-value cutoffs of as high as 10^{-3} because some vertebrate-specific pairs were eliminated when the cutoff values were lower, as the amino acid sequences of some of the domains were less conserved in some species. For these cases, we still regarded the domain pairs as significant. In contrast, some domain pairs were eliminated when the cutoff values were higher, because some pairs became positive in an outgroup taxa. We did not regard these pairs as specific to vertebrates or chordates. Therefore, the specific domain pairs were identified using a cutoff value of 10^{-3}, and sequence conservation was assessed by checking whether the pairs were still specific at lower cutoff values.

The convergent index (CI) was calculated by comparing the domain architecture between two gene models. When we defined X as the number of specific domain(s) of gene model A, Y as that of gene model B, and Z as the number of common domains between the two gene models, then $CI = X \times Y/(Z + X \times Y)$. We compared the CI between all gene models that had the domain pairs. When the lowest CI was ≥ 0.5, we regarded the pairs as likely to have been acquired convergently (Supplemental Fig. 1).

We examined the expression of *Ciona* genes in the tailbud and young juvenile stages of all of the gene models listed in Table 4 except those whose expression had already been published. For those in which early expression was not detected, we list information on their expression from previously published studies. In situ hybridization was performed as described (Ogasawara et al. 2002; Satou et al. 2005).

Acknowledgments

We thank Seb Shimeld and Peter Holland for their critical reading of the manuscript and helpful comments. This work was supported by KAKENHI (Grant-in-Aid for Scientific Research) on Priority Areas "Comparative Genomics" from the Ministry of Education, Culture, Sports, Science and Technology of Japan to H.W. and N.S. T.K. was supported by JSPS Postdoctral Fellowships for Research Abroad. D.S.R. and N.H.P. were supported by CIG. D.S.R. was supported by R.A. Melon.

References

Axelsen LN, Stahlhut M, Mohammed S, Larsen BD, Nielsen MS, Holstein-Rathlou N-H, Andersen S, Jensen ON, Hennan JK, Kjølbye AL. 2006.

Identification of ischemia-regulated phosphorylation sites in connexin43: A possible target for the antiarrhythmic peptide analogue rotigaptide (ZP123). *J Mol Cell Cardiol* **40:** 790–798.
Azumi K, De Santis R, De Tomaso A, Rigoutsos I, Yoshizaki F, Pinto M, Marino R, Shida K, Ikeda M, Ikeda M, et al. 2003. Genomic analysis of immunity in a Urochordate and the emergence of the vertebrate immune system: "Waiting for Godot." *Immunogenetics* **55:** 570–581.
Babushok DV, Ostertag EM, Kazazian HH Jr. 2007. Current topics in genome evolution: Molecular mechanisms of new gene formation. *Cell Mol Life Sci* **64:** 542–554.
Banks G, Choy P, Lavidis N, Noakes P. 2003. Neuromuscular synapses mediate motor axon branching and motoneuron survival during the embryonic period of programmed cell death. *Dev Biol* **257:** 71–84.
Carroll SB, Greiner JK, Weatherbee SD. 2001. *From DNA to diversity*. Blackwell Science, Malden, MA.
Chothia C, Gough J, Vogel C, Teichmann S. 2003. Evolution of the protein repertoire. *Science* **300:** 1701–1703.
Davidson EH. 2006. *The regulatory genome*. Academic Press, San Diego, CA.
Dehal P, Satou Y, Campbell PK, Chapman J, Degnan B, De Tomaso A, Davidson B, Di Gregorio A, Gelpke M, Goodstein DM, et al. 2002. The draft genome of *Ciona intestinalis*: Insights into chordate and vertebrate origins. *Science* **298:** 2157–2167.
Duke-Cohan JS, Gu J, McLaughlin DF, Xu Y, Freeman GJ, Schlossman SF. 1998. Attractin (DPPT-L), a member of the CUB family of cell adhesion and guidance proteins, is secreted by activated human T lymphocytes and modulates immune cell interactions. *Proc Natl Acad Sci* **95:** 11336–11341.
Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14:** 755–763.
Ekman D, Bjorkund A, Elofsson A. 2007. Quantification of the elevated rate of domain rearrangements in Metazoa. *J Mol Biol* **372:** 1337–1348.
Finn RD, Mistry J, Schuster-Böckler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al. 2006. Pfam: Clans, web tools and services. *Nucleic Acids Res* **34:** D247–D251.
Galant R, Carroll SB. 2002. Evolution of a transcriptional repression domain in an insect Hox protein. *Nature* **415:** 910–913.
Gonçalves-Mendes N, Simon-Chazottes D, Creveaux I, Meiniel A, Guénet J-L, Meiniel R. 2003. Mouse SCO-spondin, a gene of the thrombospondin type 1 repeat (TSR) superfamily expressed in the brain. *Gene* **312:** 263–270.
Gough J. 2005. Convergent evolution of domain architectures (is rare). *Bioinformatics* **21:** 1464–1471.
Holland PW, Garcia-Fernandez J, Williams NA, Sidow A. 1994. Gene duplications and origins of vertebrate development. *Development* **1994 (Suppl.):** 125–133.
Holland, LZ, Abalat R, Azumi K, Benito-Gutiérrez E, Blow MJ, Bronner-Fraser M, Brunet F, Butts T, Candiani S, Dishaw LJ, et al. 2008. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18:** 1100–1111.
Huang S, Yuan S, Guo L, Yu Y, Li J, Wu T, Liu T, Yang M, Wu K, Liu H, et al. 2008. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res* **18:** 1112–1126.
Kaessmann H, Zollner S, Nekrutenko A, Li W. 2002. Signatures of domain shuffling in the human genome. *Genome Res* **12:** 1642–1650.
Kamenetz F, Tomita T, Hsieh H, Seabrook G, Borchelt D, Iwatsubo T, Sisodia S, Malinow R. 2003. APP processing and synaptic function. *Neuron* **37:** 925–937.
Kzhyshkowska J, Gratchev A, Goerdt S. 2006. Stabilin-1, a homeostatic scavenger receptor with multiple functions. *J Cell Mol Med* **10:** 635–649.
Lang J, Ushkaryov Y, Grasso A, Wollheim CB. 1998. Ca^{2+}-independent insulin exocytosis induced by α-latrotoxin requires latrophilin, a G protein-coupled receptor. *EMBO J* **17:** 648–657.
Lee YH, Huang GM, Cameron RA, Graham G, Davidson EH, Hood L, Britten RJ. 1999. EST analysis of gene expression in early cleavage-stage sea urchin embryos. *Development* **126:** 3857–3867.
Liu M, Grigoriev A. 2004. Protein domains correlate strongly with exons in multiple eukaryotic genomes: Evidence of exon shuffling? *Trends Genet* **20:** 399–403.
Ogasawara M, Lauro RD, Satoh N. 1999. Ascidian homologs of mammalian thyroid peroxidase genes are expressed in the thyroid-equivalent region of the endostyle. *J Exp Zool* **285:** 158–169.
Ogasawara M, Sasaki A, Metoki H, Shin-i T, Kohara Y, Satoh N, Satou Y. 2002. Gene expression profiles in young adult *Ciona intestinalis*. *Dev Genes Evol* **212:** 173–185.
Patthy L. 1999. Genome evolution and the evolution of exon-shuffling—A review. *Gene* **238:** 103–114.
Patthy L. 2003. Modular assembly of genes and the evolution of new functions. *Genetica* **118:** 217–231.
Ponta H, Sherman L, Herrlich PA. 2003. CD44: From adhesion molecules to signalling regulators. *Nat Rev Mol Cell Biol* **4:** 33–45.
Putnam NH, Butts T, Ferrier DEK, Furlong, RF, Hellsten, U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al. 2008. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453:** 1064–1071.
Robertson N, Hamaker S, Patriub V, Aster J, Morton C. 2006. Subcellular localisation, secretion, and post-translational processing of normal cochlin, and of mutants causing the sensorineural deafness and vestibular disorder, DFNA9. *J Med Genet* **40:** 479–486.
Saitou M, Furuse M, Sasaki H, Schulzke J-D, Fromm M, Takano H, Noda T, Tsukita S. 2000. Complex phenotype of mice lacking occludin, a component of tight junction strands. *Mol Biol Cell* **11:** 4131–4142.
Sánchez-Pulido L, Martín-Belmonte F, Valencia A, Alonso MA. 2002. MARVEL: A conserved domain involved in membrane apposition events. *Trends Biochem Sci* **27:** 599–601.
Satoh N. 2003. The ascidian tadpole larva: Comparative molecular development and genomics. *Nat Rev Genet* **4:** 285–295.
Satou Y, Yamada L, Mochizuki Y, Takatori N, Kawashima T, Sasaki A, Hamaguchi M, Awazu S, Yagi K, Sasakura Y, et al. 2002. A cDNA resource from the basal chordate *Ciona intestinalis*. *Genesis* **33:** 153–154.
Satou Y, Kawashima T, Kohara Y, Satoh N. 2003. Large scale EST analyses in *Ciona intestinalis*, its application as Northern blot analyses. *Dev Genes Evol* **213:** 314–318.
Satou Y, Kawashima T, Shoguchi E, Nakayama A, Satoh N. 2005. An integrated database of the ascidian, *Ciona intestinalis*: Towards functional genomics. *Zoolog Sci* **22:** 837–843.
Sea Urchin Genome Sequencing Consortium. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. *Science* **314:** 941–952.
Tongiorgi E. 1999. Tenascin-C expression in the trunk of wild-type, cyclops and floating head zebrafish embryos. *Brain Res Bull* **48:** 79–88.
Tortorella MD, Liu R-Q, Burn T, Newton RC, Arner E. 2002. Characterization of human aggrecanase 2 (ADAM-TS5): Substrate specificity studies and comparison with aggrecanase 1 (ADAM-TS4). *Matrix Biol* **21:** 499–511.
Upholt WB, Chandrasekaran L, Tanzer ML. 1994. Molecular cloning and analysis of the protein modules of aggrecans. *EXS* **70:** 37–52.
Valhmu WB, Palmer GD, Rivers PA, Ebara S, Cheng JF, Fischer S, Ratcliffe A. 1995. Structure of the human aggrecan gene: Exon–intron organization and association with the protein domains. *Biochem J* **309:** 535–542.
Verhoeve K, Van Laer L, Kirschhofer K, Legan P, Hughes D, Schatteman I, Verstreken M, Van Hauwe P, Coucke P, Chen A, et al. 1998. Mutations in the human α-tectorin gene cause autosomal dominant non-syndromic hearing impairment. *Nat Genet* **19:** 60–62.
Vibranovski MD, NJ Sakabe, de Oliveira RS, de Souza SJ. 2005. Signs of ancient and modern exon-shuffling are correlated to the distributuon of ancient and modern domains along proteins. *J Mol Evol* **61:** 341–350.
Watanabe H, Yamada Y. 1999. Mice lacking link protein develop dwarfism and craniofacial abnormalities. *Nat Genet* **21:** 225–229.

Received September 23, 2008; accepted in revised form April 24, 2009.