Table S1:

Homeobox gene diversity in amphioxus, duplications in amphioxus or human and method for estimation of homeobox gene number in the chordate ancestor

Class	Gene number in amphioxus	Conserved families between amphioxus and human	Families duplicated in amphioxus	Families duplicated in human
ANTP	60	36	6	29
PRD	29*	20	1	12
CUT	4	2	0	2
LIM	7	6	1	6
PROS	1	1	0	1
SINE	3	3	0	3
TALE	9	6	1	5
ZF	5	4	0	4
CERS	1	1	0	1
POU	7	6	1	4
HNF	4	2	1	1
Other	3	0	0	0
	133*	87	11	68

**Includes Pax2/5/8 which has a partial homeobox, but excludes Pax1/9 and Pon that are Pax genes lacking a homeobox sequence*

ANTP class

36 shared gene families found in amphioxus and human: Hox1, Hox2, Hox3, Hox4, Hox5, Hox6-8, Hox9-15, Cdx, Nk6, Tlx, Hmx, Nk4, Gsx, Meox, Evx, Gbx, En, Nk1, Vax, Emx, Dbx, Barh, Barx, Lbx, Msx, Nk3, Nk2-2, Nk2-1, Dlx, Pdx, Noto, Hhex, Hlx, Bsx, Mnx, Vent (here we define Hox as seven gene families)

6 of these duplicated in amphioxus: Mnx, Evx, Emx, Vent, NK1, Hox9-15

29 retain duplicates in human: all seven Hox families, plus Cdx, Nk6, Tlx, Hmx, Nk4, Gsx, Mox, Evx, Gbx, En, Nk1, Vax, Emx, Dbx, Barh, Barx, Lbx, Msx, Nk3, Nk2-2, Nk2-1, Dlx.

Number of chordate ancestral genes = one per conserved gene family (except Hox6-8, Hox9-15 and possibly Dlx), plus the number of ancient gene families lost by vertebrates. In the absence of information to the contrary, we currently assume that the ancestral chordate Hox cluster had 14 genes. It is unclear if the last common ancestor of chordates (LCAC) had one or two Dlx genes. For the present calculation we assume one. Comparison to other invertebrate genomes reveals six additional ancient ANTP gene families present in amphioxus but lost in vertebrates: Nedx, ro, Bari, Msxlx, Abox and Nk7.

Hence, the number of ANTP class genes in LCAC was (approximately) 49

PRD class

20 gene families conserved to human: Pax2/5/8, Pax3/7, Pax4/6, Arx, Alx, Vsx, Isx, Dmbx, Drgx, Gsc, Otp, Otx, Phox, Prop, Pitx, Prrx, Rax, Shox, Uncxa, Hopx

1 of these duplicated in amphioxus: Uncxa

12 duplicated in human: Pax2/5/8, Pax3/7, Pax4/6, Alx, Vsx, Gsc, Otx, Phox, Pitx, Prrx, Rax, Shox (Pax1/9, which lacks a homeobox, is not included in the calculation).

Number of ancestral genes = one per conserved gene family (20), plus the Repo and Uncxb genes which are secondarily lost from vertebrates. (Pon which lacks a homeobox, is not included in the calculation)
Hence, the number of PRD class genes in LCAC was (minimally) 22.

CUT class
2 conserved gene families (human and amphioxus): Onecut, Cux
0 duplicated in amphioxus
2 duplicated in human: Onecut, Cux
Number of ancestral genes = one per conserved gene family, plus the presumed common ancestral gene to vertebrate Satb genes and invertebrate Compass genes. Hence, the number of CUT class genes in LCAC was (minimally) 3

LIM class
6 conserved gene families: Lhx2/9, Isl, Lhx1/5, Lhx3/4, Lhx6/8, Lmx
1 duplicated in amphioxus: Lhx2/9
6 duplicated in human: Lhx2/9, Isl, Lhx1/5, Lhx3/4, Lhx6/8, Lmx
Number of ancestral genes = one per conserved gene family. Hence, the number of LIM class homeobox genes in LCAC was (minimally) 6

PROS class
1 conserved gene families: Prox
0 duplicated in amphioxus
1 duplicated in human: Prox
Number of ancestral genes = one per conserved gene family. Hence, the number of PROS class genes in LCAC was (minimally) 1

SINE class
3 conserved gene families: Six1/2, Six3/6, Six4/5
0 duplicated in amphioxus
3 duplicated in human: Six1/2, Six3/6, Six4/5
Number of ancestral genes = one per conserved gene family. Hence, the number of SINE class genes in LCAC was (minimally) 3

TALE class
6 conserved gene families: Iro, Meis, Mkx, Pbx, Pknox, Tgif
1 duplicated in amphioxus: Iro
5 duplicated in human: Iro, Meis, Pbx, Pknox, Tgif
Number of ancestral genes = one per conserved gene family except possibly Iro. We currently infer that there was a single Iro gene in the last common ancestor of chordates (LCAC), but it is possible that there was a pair. Hence, the number of TALE class genes in LCAC was (minimally) 6.

ZF class
4 conserved gene families: Zfhx, Zeb, Tshx, Zhx
0 duplicated in amphioxus:
4 duplicated in human: Zfhx, Zeb, Tshx, Zhx
Number of ancestral genes = one per conserved gene family. Hence, the number of ZF class genes in LCAC was (minimally) 4

CERS class
1 conserved gene families: Cers

0 duplicated in amphioxus:
1 duplicated in human: Cers
Number of ancestral genes = one per conserved gene family. Hence, the number of CERS class genes in LCAC was (minimally) 1

POU class
6 conserved gene families: Pou1, Pou2, Pou3, Pou4, Pou6, Hdx
1 duplicated in amphioxus: Pou3
4 duplicated in human: Pou2, Pou3, Pou4, Pou6
Hdx is classified here as a POU class gene, even though it lacks a POU-specific domain

Number of ancestral genes = one per conserved gene family. Hence, the number of POU class genes in LCAC was (minimally) 6

HNF class
2 conserved gene families: Hnf1, Hmbox
1 duplicated in amphioxus: Hmbox
1 duplicated in human: Hnf1
Number of ancestral genes = one per conserved gene family. Hence, the number of HNF class genes in LCAC was (minimally) 2

OTHER
No conserved gene families

Total number of homeobox genes in the last common ancestor of cephalochordates and vertebrates is therefore estimated to be 103.

Note: classification and nomenclature of human homeobox genes is from Holland et al. (2007).

Table S2: Opsin genes

Number in tree (Fig. S3)	Gene models	Scaffold	Best blastp hit	Score	Expectancy
87094	87094	158	Amphiop6	565	2e-159
110002	110002	726	Amphiop6	536	1e-150
110003	110003	726	Amphiop6	505	2e-141
201585	201585/ 175449	7/507	Amphiop6	200	1e-49
86640	86640/ 86644	153/153	Amphiop6	206	2e-51
86195	86195/ 86253	149/150	Opsin-4 *Rutilus*	223	1e-56
65960	65960/ 65959	9/9	AmphiMop	1234	0.0
84890	84890/ 185357	136/52	PRED: sim. to TMT opsin [*Danio rerio*]	202	2e-50
70446	70446/ 206170	28/28	PRED: sim. To TMT opsin [*Danio rerio*]	284	1e-74
70447	70447/ 206045	28/28	Amphiop5	286	1e-75
74631	74631/ 84894	52/136	Amphiop4	650	0.0
124039	124039/ 74630	136/52	Amphiop5	673	0.0
71561	71561/ 110962	34/818	Amphiop2	652	0.0
91094	91094/ 256798	205/816	Amphiop1	617	4e-175
91095	91095/ 91106	205/205	Amphiop1	515	2e-144
215180	215180	61	Amphiop1	258	1e-66
210643	210643	42	PRED: sim. to neuropsin [*Gallus gallus*]	236	2e-60
65045	65045/ 73626	6/45	PRED: opsin 5 iso. 2 [*Macaca mulatta*]	243	2e-62
94083	94083	247	PRED: hypothetical protein [*Danio rerio*]	192	2e-47
90832	90832/ 108075	202/607	Amphiop3	638	0.0

Table S2 shows gene models IDs for amphioxus sequences used in phylogenetic analysis. If two gene models represent the same protein sequence, most likely allelic forms, both gene models are listed. The opsin sequences used were derived from gene models predicted in *Branchiostoma floridae* genome v1.0. Incorrectly predicted or missing exons were found manually or by blastn search. The best NCBI blastp hit is shown with score and expectancy value.

Table S3. Numbers of neural crest gene network homologs in amphioxus and human and summary of ectodermal expression patterns. Orthology of genes in a family was established by constructing phylogenetic trees using the Neighbor-Joining method. For details see Yu et al. 2008). In every case in which there is more than one amphioxus gene in a given family, phylogenetic analysis indicates they are the result of lineage-specific duplication events. *ColA expression is seen in the neural tube and pharyngeal mesoderm of amphioxus, though it is unclear if this expression can be considered homologous to neural crest-derived cartilage expression in vertebrates (Meulemans and Bronner-Fraser 2007).

Neural plate border induction signals	Amphioxus	Human	NJ bootstrap values	Conserved Ectodermal expression
BMP2/4	1	2	98	yes
Wnt1	1	1	(a)	no
Wnt3	1	2	(a)	yes
Wnt6	1	1	(a)	no
Wnt7	1	2	(a)	no
Wnt8	1	2	(a)	yes
FGF8/17/18	1	3	100	yes
Notch	1	4	(b)	yes
Delta	1	3	83	yes
Neural plate border specifiers				
Dlx	1	6	(c)	yes
Msx	1	2	(d)	yes
Zic	1	5	(e)	yes
Pax3/7	1	2	(f)	yes

Neural patterning/differentiation genes	Amphioxus	Human	NJ bootstrap values	Conserved Ectodermal expression
SoxB1	3	3	61	Yes
SoxB2	1	2	79	yes
NeuroD	1	4	99	unk
Neurogenin	1	3	98	yes
Achaete-Scute	2	4	90	unk
Beta3	1	2	98	unk
Olig	3	3	65	yes
Islet	1	2	(g)	yes
Hu/Elav	1	4	(h)	yes

Neural crest Specifiers	Amphioxus	Human	NJ bootstrap	Conserved ectodermal expression
Snail	1	2	(i)	yes
FoxD	1	5	(j)	no
Twist	1	2	(k)	no
Myc	1	4	82	no
Id	1	4	(l)	No
tfap2	1	5	(m)	no
SoxE	1	3	73	no

Neural crest effector genes	Amphioxus	Human	NJ bootstrap	conserved ectodermal expression
Mitf	1	4	100	yes, pigment cells
Tyrosinase	1	1	99	yes, pigment cells
Trp	2	2	99	yes, pigment cells
RhoA/B/C	2	3	98	unk
cRet	1	1	100	unk
Erbb	1	4	100	unk
ColA	1	5	(n)	unk*
cKit	0	1	N/A	N/A
Myelin protein P0	0	1	N/A	N/A

References:
a. Schubert et al. 2001; b. Holland et al. 2001; c. Holland et al. 1996; d. Sharman et al. 1999; e. Gostling and Shimeld 2003. f. Holland et al. 1999; g. Jackman et al. 2000; h. Benito-Gutierrez et al. 2005; i. Langeland et al. 1998; j. Yu et al. 2002; k. Yasui et al. 1998; l. Meulemans et al. 2003; m. Meulemans and Bronner-Fraser 2002; n. Zhang and Cohn 2006.

Table S4: Endocrine genes

Phylogenetic trees were calculated using the Maximum Likelihood (ML) method and the percentages were derived from these phylogenies in 100 bootstrap replicates. The ML bootstrap percentages are given for the node uniting amphioxus sequences with their relevant monophyletic orthologs. If no outgroup is available, bootstrap support for the amphioxus sequences could not be determined.

	Human	Amphioxus (0 = not found)	ML bootstrap percentages
Reproduction			
Kisspeptin	1	0	-
Kisspeptin Receptor GPR54	1	18	85
GnRH (Gonadotropin-Releasing Hormone)	2	0	-
GnRH Receptor	1	3	88, 98
Gonadotropins (LH, FSH, hCG)	3	0	-
Gonadotropin Glycoprotein Receptor	2	1†	100
Estrogen Receptor	2	1	74
Steroid Receptor	4	1*	55
Adrenal Axis			
CRH (Corticotropin-Releasing Hormone)	1	0	-
CRH Receptor	2	5	80
POMC (Proopiomelanocortin)	1	0	-
ACTH (Adrenocorticotropin) Receptor	1	0	-
Glucocorticoid Steroid Receptor	1	1*	55
Thyroid Axis and Hormone Metabolism			
TRH (Thyrotropin-Releasing Hormone)	1	1	‡
TRH Receptor	1	0	-
TSH (Thyrotropin) Subunit Beta	1	0	-
Thyrostimulin (GPA2/GPB5)	1α/1β	2α/1β	59/71
Thyrostimulin Glycoprotein Receptor	1	1†	100
Thyroid Hormone Receptor (TR)	2	1	99
Retinoid X Receptor (RXR)	3	1	100
SIS (Sodium/Iodide Symporter)	2	7	56
Thyroglobulin	1	0	-
Peroxidase (including Thyroid Peroxidase)	4 (1 Thyroid Peroxidase)	4	77
Deiodinase	3	5	

Transthyretin	1	0	-
Serpins (including TBG and CBG)	13 (1 TBG and 1 CBG)	6	100
CTHBP	2	1	100

Growth

GHRH (Growth Hormone Releasing Hormone)	1	0	-
GHRH Receptor	1	0	-
Somatostatin	1	0	-
Somatostatin Receptor	5	12	13
GH (Growth Hormone)	1	0	-
GH Receptor	1	0	-
Insulin/IGF/Relaxin	3	6	‡
Insulin/IGF Receptor	2	1	91
LGR7 and 8 (Leucine-rich Repeat GPCR)	2	6	83
Ghrelin	1	0	-
Ghrelin Receptor GHSR	1	2	89

Other Hormones

Amphitocin (Vasotocin-like)	1	1	79
Calcitonin-like	2	1	100
Activinβ/Inhibinβ	1	1	96
NPY/PPY/PP-like Family	3	1	100
Stanniocalcin	2	2	98

Other Receptors

Activin Receptor (Type 1)	2	1	42
Amphitocin (VP/Oxy) Receptor	4	2	73
Calcitonin-like Receptor (Secretin Family Rs)	2	1	82
Parathyroid Hormone Receptor (Secretin Family Rs)	2	4	93
Other Secretin Receptors	8	5	62

Steroid Hormone Metabolism

5α Reductase	2	2	71
3β HSD	3	6	100
SDR (17β HSD, 11β HSD and RDH group)	31	107	

MFE-2 (17β HSD activity)	1	1	100
3-Keto-Steroid Reductase (17β HSD activity)	1	1	
Aldo-Keto Reductase (AKR1C-like) (17β HSD activity)	4	0	-
Aromatase (Cyp19)	1	2	98
11β Hydroxylase (Cyp11)	3	3	72
17α Hydroxylase (Cyp17)	1	1	100
21-Steroid Hydroxylase (Cyp21)	1	0	-
StAR	1	1	98
Serpins (including CBG and TBG)	13 (1 CBG and 1 TBG)	6	100

Retinoic Acid Signaling and Metabolism

Retinoic Acid Receptor (RAR)	3	1	100
Retinoid X Receptor (RXR)	3	1	100
CRABP	2	0	-
CRBP	4	1	93
IRBP	1	0	-
ALDH1	3	6	100
ALDH8 (=ALDH12)	1	1	100
ADH	7	1	95
SDR (RDH and 17β HSD, 11β HSD group)	31	107	
Aldo-Keto Reductase (AKR1B-like)	2	0	-
LRAT	5	4	98
ARAT	1	1	81
BcoxI and BcoxII and RPE65	3	4	100
Retinyl Ester Hydrolase (REH)	5	6	67
Cyp26	3	3	86

[†]The amphioxus glycoprotein hormone receptor homologue is basal to vertebrate thyrostimulin receptor and gonadotropin receptors in phylogenetic analysis, but has the highest sequence similarity with vertebrate thyrostimulin receptor

*The amphioxus steroid receptor homologue is basal to the four vertebrate steroid receptors

[‡]Phylogenetic clustering of small hormones like TRH (3 amino acids) and insulin/IGF/relaxins do not generate strong statistical support; these bootstrap values are not included.

Table S5. Gene models near INSL genes in the *Branchiostoma floridae*. Each gene model is displayed with its human ortholog/s. The location of the human orthologs is indicated in relation to the insulin/IGF- and insulin-like/relaxin (INSL/RLN) paralogons. "x" denotes absence of a recognizable human homolog. Scaffold organization is according to *Branchiostoma floridae* JGI database v.1.0 May 2006 release. Note: Version 1.0 of the *B. floridae* genome includes both alleles. The folling amphioxus gene models are probably allelic: fgenesh2_pg.scaffold_302000049 and fgenesh2_pg.scaffold_59000102, estExt_fgenesh2_pg.C_410021 and estExt_fgenesh2_pg.C_590098, fgenesh2_pg.scaffold_50000083 and fgenesh2_pg.scaffold_674000009. Therefore, there are six distinct insulin-like genes in the amphioxus genome.

Scaffold 59 (entire scaffold	**JGI-gene model ID**	**INSL/RLN paralogon**	**Outside insulin-RLN paralogous regions**	**No hits in the human genome**
fgenesh2_pg.scaffold_59000001				x
estExt_gwp.C_590002	CCT2 12q15			
e_gw.59.61.1	DEPDC1 1p31.2+LOC91614 11p13	DEPDC1 1p31.2+LOC91614 11p13		
gw.59.95.1			MGC33365 3q24	
e_gw.59.94.1			MGC33365 3q24	
fgenesh2_pg.scaffold_59000006	C11orf10 11q12.2			
fgenesh2_pg.scaffold_59000007	ACTR6 12q23.1			
e_gw.59.63.1	TCP11L1 11p13			
fgenesh2_pm.scaffold_59000003	ZNF289 11p11.2+ARFGAP3 22q13.2			
fgenesh2_pg.scaffold_59000010			HRH1 3p25.3	
estExt_fgenesh2_pg.C_590011			MGC21644 5q32	
e_gw.59.206.1	PTPRO 12p12.3			
e_gw.59.44.1	PTPRB 12q15			
e_gw.59.49.1	NFASC 1q32.1			
e_gw.59.57.1			FN1 2q35	
fgenesh2_pg.scaffold_59000015	MANSC1 12p12.3			
e_gw.59.93.1			CSMD1 8p23.2+CSMD3 8q23.3	
e_gw.59.178.1			CSMD3 8q23.3	
e_gw.59.17.1		EGFL3 1p36.32		
e_gw.59.198.1	GABPB2 15q21.2	GABPB2 15q21.2		
estExt_gwp.C_590032		AKR1A1 1p34.1+AKR1B1 7q33		
e_gw.59.107.1		AKR1A1 1p34.1+AKR1B1 7q33		
estExt_gwp.C_590039			FLJ16237 7p21.1	
fgenesh2_pg.scaffold_59000021	EEA1 12q22			
fgenesh2_pg.scaffold_59000022	GABPB2 15q21.2	GBP2,-3,-6 1p22.2		
fgenesh2_pg.scaffold_59000023	GBP4,-6 1p22.2+GBP1,-3 1p22.2	GBP4,-6 1p22.2+GBP1,-3 1p22.2		
fgenesh2_pg.scaffold_59000024	GBP3,-4 1p22.2	GBP3,-4 1p22.2		
fgenesh2_pg.scaffold_59000025			AKR1D1 7q34	
fgenesh2_pg.scaffold_59000026				X

fgenesh2_pg.scaffold_59000027				X
fgenesh2_pg.scaffold_59000028				X
gw.59.145.1			CSMD1 8p23.2+CSMD3 8q23.3	
e_gw.59.84.1		EGFL3 1p36.32		
e_gw.59.80.1		EGFL3 1p36.32		
estExt_gwp.C_590049		AKR1A1 1p34.1+ARKR1B1 7q33		
fgenesh2_pg.scaffold_59000032		AKR1A1 1p34.1+ARKR1B1 7q33		
fgenesh2_pg.scaffold_59000033		MXR8 1p36.33		
e_gw.59.102.1			FLJ16237 7p21	
fgenesh2_pg.scaffold_59000035			RBPBP6 16p12.1	
fgenesh2_pg.scaffold_59000037	GBP6 1p22.2	GBP6 1p22.2		
fgenesh2_pg.scaffold_59000038			AKR1B1 7q33	
estExt_fgenesh2_pg.C_590039	AKR1A1 1p34.1			
fgenesh2_pg.scaffold_59000040	-			
fgenesh2_pg.scaffold_59000041	ZNF135 19q13.43			
fgenesh2_pg.scaffold_59000042	FLJ21839 2p23.3			
fgenesh2_pg.scaffold_59000043		TTRAP 6p22.2		
e_gw.59.31.1			EHHADH 3q27.2	
e_gw.59.35.1	MAPK8IP1 11p11.2+MAPK8IP2 22q13.33			
estExt_fgenesh2_pg.C_590046			KIF9 3p21.31	
estExt_fgenesh2_pg.C_590047	ATP5F1 1p13.2	ATP5F1 1p13.2		
estExt_fgenesh2_pg.C_590048	WDR77 1p13.2	WDR77 1p13.2		
e_gw.59.143.1			GALR118q23+GALR2 17q25.1	
estExt_fgenesh2_pg.C_590051				X
fgenesh2_pg.scaffold_59000052	PPP1R12A 12q21.2			
fgenesh2_pm.scaffold_59000008	PPP1R12A 12q21.2			
fgenesh2_pg.scaffold_59000054			FLJ25415 2q32.2	
e_gw.59.100.1			AATF 17q12	
e_gw.59.126.1		RASEF 9q21.32		
fgenesh2_pg.scaffold_59000057		FLJ16636 9q22.33		
fgenesh2_pg.scaffold_59000058				X
fgenesh2_pg.scaffold_59000059				X
estExt_gwp.C_590082	AKR1A1 1p34.1+AKR1B1 7q33			
e_gw.59.103.1	AKR1A1 1p34.1+AKR1B1 7q33			
fgenesh2_pg.scaffold_59000062	GBP1+GBP4 1p22.2	GBP1+GBP4 1p22.2		
gw.59.177.1	SELP 1q24.2+CSMD3 8q23.3	SELP 1q24.2+CSMD3 8q23.3		
e_gw.59.90.1				
e_gw.59.162.1			CSMD3 8q23.3	
fgenesh2_pg.scaffold_59000065	RBM23 14q11.2			
fgenesh2_pg.scaffold_59000066	C1orf69 1q42.13			
fgenesh2_pg.scaffold_59000067	DEAF1 11p15.5			
gw.59.34.1		EGFL3 1p36.32		
fgenesh2_pg.scaffold_59000069	FLJ40198 15q11.2	FLJ40198 15q11.2		

fgenesh2_pg.scaffold_59000070				X
fgenesh2_pg.scaffold_59000071	C1orf96 1q42.13			
fgenesh2_pg.scaffold_59000072	DEAF 11p15.5			
fgenesh2_pg.scaffold_59000073			NSUN7 4p14	
e_gw.59.199.1			PRKCB1 16p12.1	
e_gw.59.187.1	NINJ2 12p13.33+NINJ1 9q22.31	NINJ2 12p13.33+NINJ1 9q22.31		
fgenesh2_pg.scaffold_59000076		NINJ1 9q22.31		
estExt_fgenesh2_pg.C_590077	NINJ2 12p13.33			
fgenesh2_pg.scaffold_59000078				X
e_gw.59.106.1		TBN 6p21.1		
e_gw.59.14.1			PIGZ 3q29	
e_gw.59.36.1		LRRN6C 9p21.2		
fgenesh2_pg.scaffold_59000083			SRRM2 16p13.3	
fgenesh2_pg.scaffold_59000084				X
fgenesh2_pg.scaffold_59000085	PLCB3 11q13.1			
e_gw.59.203.1	GABPB2 15q21.2	GABPB2 15q21.2		
e_gw.59.33.1			CYP4V2 4q35.1	
e_gw.59.135.1				X
fgenesh2_pg.scaffold_59000089			DNAH1 3p21.1	
fgenesh2_pg.scaffold_59000090			HAVCR1 5q33.3	
fgenesh2_pg.scaffold_59000091				X
fgenesh2_pg.scaffold_59000092			THOC1 18p11.32	
fgenesh2_pg.scaffold_59000093				X
fgenesh2_pg.scaffold_59000094	SPON1 11p15.2			
estExt_fgenesh2_pg.C_590095	SPON1 11p15.2			X
fgenesh2_pg.scaffold_59000096				
estExt_fgenesh2_pg.C_590097			THOC1 18p11.32	
estExt_fgenesh2_pg.C_590098	INSL	INSL		
fgenesh2_pg.scaffold_59000099				X
e_gw.59.50.1			COLEC10 8q24.12+SFTPD 10q22.3	
estExt_fgenesh2_pg.C_590101	HAPLN2 1q23.1	HAPLN2 1q23.1		
fgenesh2_pg.scaffold_59000102	INSL	INSL		
estExt_fgenesh2_pg.C_590103			MPHOSPH1 10q23.31	
fgenesh2_pg.scaffold_59000104	TH 11p15.5			
e_gw.59.48.1	DUSP19 2q32.1			
e_gw.59.47.1	ASCL1 12q23.2+ASCL2 11p15.5			
fgenesh2_pg.scaffold_59000107			C6orf111 6q16.3	
fgenesh2_pg.scaffold_59000108				X
e_gw.59.147.1	GAL3ST3 11q13.1+GAL3ST1 22q12.2			
fgenesh2_pm.scaffold_59000012	ZNF678 1q42.13			
gw.59.75.1	ZNF678 1q42.13			
e_gw.59.179.1	ZNF678 1q42.13			

fgenesh2_pg.scaffold_59000113	ZNF678 1q42.13			
e_gw.59.180.1	ZNF678 1q42.13			
gw.59.132.1	ZNF678 1q42.13			
gw.59.122.1				X
fgenesh2_pg.scaffold_59000117	ZNF665 19q13.42			

scaffold_302

fgenesh2_pg.scaffold_302000029	(Type 1 envelope glycoprotein J from Equid herpesvirus 1-if this sequence is blasted, it is mapped to human 11p15.5)			
fgenesh2_pg.scaffold_302000032				X
estExt_GenewiseH_1.C_3020024	SNRPEL1 9p24.1+EAW69166 19p13.3			
e_gw.302.47.1	TRIM3 11p15.4+TRIM25 17q23.2+TRIM2 4q31.3			
gw.302.43.1			TRIM59 3q26.1	
fgenesh2_pg.scaffold_302000040	GAL3ST3 11q13.1+GAL3ST2 22q12.2			
fgenesh2_pg.scaffold_302000041				X
fgenesh2_pg.scaffold_302000042			TTN 2q31.2	
estExt_fgenesh2_pg.C_3020042				X
e_gw.302.22.1	ASCL1 12q23.2+ASCL2 11p15.5 (co-orthologs)			
e_gw.302.8.1	SSH3 11q13.2			
e_gw.302.1.1	TH 11p15.5			
fgenesh2_pg.scaffold_302000049	INSL			

scaffold_41

fgenesh2_pg.scaffold_41000006				X
fgenesh2_pg.scaffold_41000007				X
e_gw.41.128.1				X
fgenesh2_pm.scaffold_41000001			CYP4V2 4q35.1	
e_gw.41.29.1			CYP4V2 4q35.1	
fgenesh2_pg.scaffold_41000011	C1QTNF4 11p11.2+C1QTNF3 5p13.3+C1QL2 2q14.2			
e_gw.41.169.1	MSCP 1q21.3	MSCP 1q21.3		
fgenesh2_pg.scaffold_41000013		NM_020702 9p13.1		
e_gw.41.216.1	ZNFs 19q13.12			
fgenesh2_pg.scaffold_41000015				X
fgenesh2_pg.scaffold_41000016				X
estExt_fgenesh2_pg.C_410017				X
fgenesh2_pg.scaffold_41000018			FBLN1 22q13.31	

estExt_fgenesh2_pg.C_410019			COLEC10 8q24.12	
estExt_fgenesh2_pg.C_410020	FLJ44817 14q24.1			
estExt_fgenesh2_pg.C_410021	INSL	INSL		
fgenesh2_pg.scaffold_41000022	INSL	INSL		
fgenesh2_pg.scaffold_41000023	ENSG00000196931 (LOC91664) 19q13.42			
estExt_fgenesh2_pg.C_410024			REST 4q12	
e_gw.41.15.1	ZNF567 19q13.12			
e_gw.41.50.1	ZNF84 12q24.33			
fgenesh2_pg.scaffold_41000028			ZNF182 Xp11.23	
e_gw.41.225.1	HKR1 19q13.12			
estExt_fgenesh2_pg.C_410030		TUBB2C 9q34.3		
estExt_fgenesh2_pg.C_410031		TUBB2C 9q34.3		
fgenesh2_pg.scaffold_41000032	SLC17A6 11p14.3+SLC17A8 12q23.1+SLC17A7 19q13.33			
fgenesh2_pg.scaffold_41000033	HRNR 1q21.3+RP1-14N1.3 (ENSG00000143520) 1q21.3+DSPP 4q22.1	HRNR 1q21.3+ENSG00000143520 1q21.3+DSPP 4q22.1		
fgenesh2_pg.scaffold_41000034	ZNF84 12q24.33			
gw.41.157.1			FGFR1 8p12	
gw.41.200.1		FIBCD1 9q34.13		
fgenesh2_pg.scaffold_41000036	LRIG2 1p13.2	LRIG2 1p13.2		
e_gw.41.86.1			GPR103 4q27+ CCKAR 4p15.2	

scaffold 372

gw.372.19.1			PKD1L2 16q23.2	
gw.372.23.1	PKD1 16p13.3			
gw.372.14.1	NNMT 11q23.2+PNMT 17q12+INMT 7p14.3			
fgenesh2_pg.scaffold_372000003				X
fgenesh2_pg.scaffold_372000004			NLGN2 17p13.1+CES2 16q22.1	
fgenesh2_pg.scaffold_372000005	INSL	INSL		
fgenesh2_pg.scaffold_372000006	Similar to insulin B-domain (fragment)	Similar to insulin B-domain (fragment)		
e_gw.372.1.1			ENSG00000106121 7p14.3	
fgenesh2_pg.scaffold_372000008			OPN4 10q23.2	
fgenesh2_pg.scaffold_372000009			RAMP3 7p13	
estExt_fgenesh2_pg.C_3720010			ZSWIM2 2q32.1	
e_gw.372.6.1			C7orf46 7p15.3	
fgenesh2_pg.scaffold_372000012			ENSG00000154978 7p11.2	
estExt_fgenesh2_pg.C_3720013			PARP14 3q21.1	
fgenesh2_pg.scaffold_372000014				X
fgenesh2_pg.scaffold_372000015				X

estExt_fgenesh2_pg.C_3720016				X
fgenesh2_pg.scaffold_372000017				X
fgenesh2_pg.scaffold_372000018			CRHR1 17q21.31	
scaffold 73				
fgenesh2_pg.scaffold_73000020	CAPS2 12q21.1			
e_gw.73.277.1	TRIM3 11p15.4+TRIM2 4q31.3			
fgenesh2_pg.scaffold_73000023	MYH7 14q11.2			
fgenesh2_pg.scaffold_73000024	ZNF299 19q13.31			
e_gw.73.262.1		SNRPE 1q32.1+SNRPEL1 9p24.1		
estExt_fgenesh2_pg.C_730025	TMEM56 1p21.3	TMEM56 1p21.3		
gw.73.232.1	ZNF672 1q44			
fgenesh2_pg.scaffold_73000027				X
fgenesh2_pg.scaffold_73000028		KIAA1543 (ENSG00000076826) 19p13.2		
fgenesh2_pg.scaffold_73000029			RBM28 7q32.1	
e_gw.73.9.1	ZNF135 19q13.43			
fgenesh2_pg.scaffold_73000031			RBM28 7q32.1	
fgenesh2_pg.scaffold_73000032				X
fgenesh2_pg.scaffold_73000033	CCDC131 12q21.1			
fgenesh2_pg.scaffold_73000034	INSL	INSL		
e_gw.73.317.1	MATN1 1p35.2+VIT 2p22.2			
e_gw.73.315.1	MATN1 1p35.2+VIT 2p22.2			
e_gw.73.114.1		ZNF782 9q22.33		
e_gw.73.211.1			RC74 8p21.1	
fgenesh2_pg.scaffold_73000039				X
fgenesh2_pg.scaffold_73000040			HIPK2 7q34	
fgenesh2_pg.scaffold_73000041	NLRP3 1q44			
fgenesh2_pg.scaffold_73000042			NOD27 16q13	
fgenesh2_pg.scaffold_73000043			NOD27 16q13	
gw.73.257.1			NOD27 16q13	
fgenesh2_pg.scaffold_73000045			NOD27 16q13	
fgenesh2_pg.scaffold_73000046		MUC16 19p13.2		
fgenesh2_pg.scaffold_73000047				X

scaffold 50

e_gw.50.66.1		FNDC3A 13q14.2+FNDC3B 3q26.31		
fgenesh2_pg.scaffold_50000075		NMUR1 2q37.1		
fgenesh2_pg.scaffold_50000076		FNDC3A 13q14.2		
fgenesh2_pg.scaffold_50000077				X
fgenesh2_pg.scaffold_50000078		MAOA Xp11.3+MAOB Xp11.3		
estExt_fgenesh2_pg.C_500078				X
e_gw.50.91.1		ATP11A 13q34+ATP11B 3q26.33		
estExt_gwp.C_500103		TUBGCP3 13q34		
e_gw.50.63.1		ANKMY1 2q37.3		
estExt_fgenesh2_pg.C_500081		ZC3H13 13q14.13		
fgenesh2_pg.scaffold_50000083	INSL	INSL		
estExt_gwp.C_500107			MAOA Xp11.3+MAOB Xp11.3	
fgenesh2_pg.scaffold_50000085	PRG4 1q31.1	PRG4 1q31.1		
fgenesh2_pg.scaffold_50000086	TRIM2 4q31.3+TRIM3 11p15.5			
fgenesh2_pg.scaffold_50000087	SMYD3 1q44+SMYD2 1q41			
estExt_gwp.C_500110			ABCC4 13q32.1	
estExt_gwp.C_500111			GSR 8p12	
fgenesh2_pg.scaffold_50000090				X
fgenesh2_pg.scaffold_50000091				X
fgenesh2_pg.scaffold_50000092				
fgenesh2_pg.scaffold_50000093		PTCHD3 10p12.1		
fgenesh2_pg.scaffold_50000094	FOLH1 11p11.12			X
fgenesh2_pg.scaffold_50000095	HEPHL1 11q21+HEPH Xq12+CP 3q25.1			

scaffold 674 (entire scaffold)

e_gw.674.8.1				X
gw.674.20.1				
fgenesh2_pg.scaffold_674000002			FIS1 7q22.1	
e_gw.674.12.1			ABCC4 13q32.1	
fgenesh2_pg.scaffold_674000006	TMEM56 1p21.3	TMEM56 1p21.3		
fgenesh2_pg.scaffold_674000007	TRIM3 11p15.5+TRIM2 4q31.3			
estExt_fgenesh2_pg.C_6740008		MAOA Xp11.3+MAOB Xp11.3		
fgenesh2_pg.scaffold_674000009	INSL	INSL		
estExt_fgenesh2_pg.C_6740010		ZC3H13 13q14.13		
e_gw.674.2.1		ANKMY1 2q37.3		
e_gw.674.18.1		TUBGCP3 13q34		

Table S6: Nuclear hormone receptor (NR) genes

NR group	NR name in humans	NR name in *Drosophila*	NR name in amphioxus	ML bootstrap percentages
NR0A		KNI, KNRL, EG		
NR0B	SHP, DAX		NR0B	
NR1A	TRα, TRβ		TR	99
NR1B	RARα, RARβ, RARγ		RAR	100
NR1C	PPARα, PPARβ, PPARγ		PPAR	100
NR1D	REV-ERBα, REV-ERBβ	E75	REV-ERB	98
NR1E		E78		
NR1F	RORα, RORβ, RORγ	DHR3	ROR	100
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-1	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-2	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-3	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-4	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-5	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-6	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-7	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-8	94
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-9†	71
NR1H	LXRα, LXRβ, FXRα	ECR	NR1H-10†	71
NR1I	VDR, PXR, CAR			
NR1J		DHR96		
NR2A	HNF4α, HNF4γ	HNF4	HNF4	100
NR2B	RXRα, RXRβ, RXRγ	USP	RXR	100
NR2C	TR2, TR4		TR2/4	100
NR2D		DHR78		
NR2E	TLL	TLL, DSF	TLL	87
NR2E	PNR	PNR	PNR	100
NR2E			NR2E	84
NR2E		FAX1		
NR2F	COUPTFα, COUPTFβ, EAR2	SVP	COUPTF*	100
NR3A	ERα, ERβ		ER	74
NR3B	ERRα, ERRβ, ERRγ	ERR	ERR	98
NR3C	GR, MR, AR, PR		SR	55
NR4A	4A1, 4A2, 4A3	4A4	4A	100

NR5A NR5B	5A1, 5A2	5A3 5B1	5A 5B	100 92

NR6A	GCNF	GRF	GCNF	96

NRa			NRa	
NRb			NRb	
NRc			NRc†	

Total	**48 NRs in humans**	**21 NRs in *Drosophila***	**33 NRs in amphioxus**

Table S6: Nuclear hormone receptor (NR) genes
Phylogenetic trees were calculated using the Maximum Likelihood (ML) method and the percentages were derived from these phylogenies in 100 bootstrap replicates. The ML bootstrap percentages are given for the node uniting amphioxus sequences with their relevant monophyletic orthologs. If no outgroup is available, bootstrap support for the amphioxus sequences could not be determined.

†DNA Binding Domain only

*COUPTF has been cloned from amphioxus, but could not be identified in the genome sequence

Table S7. Innate immunity and apoptosis related protein families

Domain (Pfam ID)	*B. floridae* (amphioxus)	*H. sapiens* (human)	*S. purpuratus* (sea urchin)	*D. melanogaster* (fruit fly)	*C. elegans* (nematode)	*N. vectensis* (sea anemone)
TIR (PF01582)	134(125)	24(23)	244(216)	11(11)	2(2)	7(7)
NACHT (PF05729)	95(94)	23(22)	326(320)	0	0	45(43)
CARD (PF00619)	84(84)	23(22)	12(10)	1(0)	1(1)	8(8)
Death (PF00531)	139(136)	31(29)	87(82)	5(5)	2(2)	5(5)
DED (PF01335)	57(57)	8(8)	3(3)	0	0	9(9)
BIR (PF00653)	18(18)	7(7)	6(6)	4(4)	2(2)	3(3)
Caspase (PF00656)	53(53)	11(11)	42(42)	7(7)	5(5)	10(10)

The value in each domain category for each species is the total number of protein sequence hits in its corresponding genome protein set; the value in parentheses is the number confirmed by Pfam or CD-Search under the default threshold. Several rounds of PSITBLASTN searches were performed against each genome using related human domain sequences as seeds. Hits were mapped to the corresponding genome protein set in order to obtain the full-length protein sequences (instead of the partial sequences which got from PSITBLASTN search at the first stage). Numbers are approximate owing to both the diversity of domain sequences, draft nature of genome coverage, and limited experimental verification of gene predictions. Counts are also dependant on significance thresholds for gene prediction and sequence recognition tools used. Genome data for sea anemone was from JGI, sea urchin from Baylor College of Medicine Human Genome Sequencing Center and others from Ensembl.

Table S8. Prediction of the genes encoding interleukins/cytokines/TNFs and their receptors in the amphioxus genome.

Category	Gene	Number of genes	E-value
Interleukin[a]	IL1R1-like	1	0.0027
	IL8RA-like	1	e-16
	IL8RB-like	1	e-17
	IL9R-like	1	0.0001
	IL12A-like	1	0.01
	IL17-like	1	0.00015
	IL17R-like	1	0.01
	IL18R1-like	1	e-6
Cytokine[a]	KIT-like	5	e-44
	FLT3-like	5	0~e-39
	MDK-like	1	e-10
	PTN-like	1	e-21
TNF[b]	TNFSF10-like	16	e-12~e-31
	FASLG/TNFSF10/11-like	5	e-6~e-10
	TNFSF13/13B-like	6	e-4~e-11
TNFRSF[c]	TNFRSF1-like	2	e-6~e-13
	TNFRSF2-like	2	e-10
	TNFRSF3-like	10	e-4~e-21
	TNFRSF4-like	1	e-6
	TNFRSF5-like	10	e-3~e-18
	TNFRSF9-like	2	e-4
	TNFRSF10a-like	1	e-3
	TNFRSF10c-like	9	e-4~e-9
	TNFRSF11b-like	7	e-13~e-23
	TNFRSF14-like	2	e-6~e-11

a. The amphioxus genome assembly was searched by BLAST using human genes as probes. Amphioxus genes exhibiting less than 0.01 E-value were collected.

b. The amphioxus genome assembly was searched by BLAST using human genes as probes. A total of 27 amphioxus genes exhibiting less than e-3 were identified. These genes were classified into three groups, TNFSF10-like, TNFSF10/11 (and/or FASLG)-like, and TNFSF13/13B-like, based on phylogenetic analysis. Domain analysis (HMMPfam, ProfileScan and HMMSmart) predicted a TNF domain in all of these genes.

c. BLAST searches of the amphioxus genome assembly retrieved a total of 60 TNFR-like genes exhibiting less than e-02 E-value to a human TNFR domain. Of these genes, HMMPfam and HMMSmart programs predicted at least one TNFR domain in 46 genes, which are listed in the Table. Of the 46 genes, 25 contained only a TNFR domain (a human CD40 type), and nine contained both TNFR and DEATH domains (a human TNFR □ type). The remaining 12 genes contained other domains in addition to the TNFR domain.

Table S9. Complement genes in genomes of representative species.

Species	**C3/C4/C5**	**Bf/C2**	**MASP/C1r,s**	**TCC**	**C1q-like**
Homo sapiens	4	2	4	6	23
Gallus gallus	4	1	3	4	17
Xenopus tropicalis	3	6	5	4	14
Danio rerio	10	10	2	3	36
Branchiostoma floridae	3	2	5	9	39
Ciona intestinalis	2	3	4	11	2
Strongylocentrotus purpuratus	4	3	0	0	4
Drosophila melanogaster	0	0	0	0	0
Caenorhabditis elegans	0	0	0	0	0
Nematostella vectensis	2	1	1	0	0

BLAST searches of the amphioxus genome assembly used vertebrate sequences as queries. The pattern-based method was used to identify the gene models, which predict the same domain architecture as those of vertebrate complement components. The numbers of complement genes of bilaterian species were determined by a domain architecture-based search using SMART (http://smart.embl-heidelberg.de/). In contrast to the other complement components, which have a characteristic domain architecture unique to complement components, C1q has a simple domain architecture composed only of collagen and C1q domains. Only three of 23 human proteins with this domain architecture are complement components. Thus, these genes are termed C1q-like. *Nematostella* genes were identified by BLAST search of the *Nematostella vectensis* genome (http://genome.jgi-psf.org/Nemve1/home.html).

Table S10. Diversity of LRR-containing gene predictions

Gene Organization	Genes in Category	LRR Modules	Ig Modules	Membrane Bound	GPI Anchored	Secreted
Single LRR domain	230	2-23	0	141	10	79
Single LRR domain and Igs	147	3-23	1-4	124	3	20
Multiple LRR domains (2-3)[a]	8	12-39	0	1	0	7
Multiple LRR domains (2-3)[a] and Igs	3	16-28	1-2	0	0	3
LRRs and additional modules	18	2-16	1-2 (5)[b]	8	0	10

a. In parentheses the number of leucine-rich repeat (LRR) domains

b. In parentheses the number of immunoglobulin (Ig)-containing genes

Of 406 genes, over half encode proteins that contain a single LRR domain, but are devoid of Ig or other domains; the LRR domain, made up of 2-23 LRR modules, is frequently capped both by N- and C-terminal LRRs. Most proteins in this category (denoted single LRR domain in Table) are predicted to be cell surface proteins (66%), associated via a transmembrane domain or GPI-anchorage. These molecules reveal structural similarity to the GPI-anchored variable lymphocyte receptors (VLR) of jawless fish [56]. Proteins in the second largest category (single LRR domain and Igs) consist of single domain LRR modules and one to four C-terminal Ig domains. The general organization is similar to that of the first category and ~86% of the LRR-Ig genes are predicted to encode surface proteins. In addition, there are 11 LRR-containing genes with 2-3 LRR domains, each consisting of several LRR modules capped by N- and C-terminal LRRs; eight of these encode proteins with no Ig domains (multiple LRR domains) and the remaining three encode C-terminal Ig domains (multiple LRR domains and Igs). The last category includes 18 genes consisting of LRRs as well as various additional protein domains, of which five also include Ig.

Table S11. Amphioxus genes with homologous copies in the human MHC region

Category	Gene	Number of genes	E-value
MHC	ABCF1/2/3	2	0~1e-163
	BAT1-like	2	1e-174~1e-142
	BAT5-like	2	0~1e-177
	BRD2/3/4	1	0
	C3/4/5-like	1	1-e-119
	CSNK2B	2	1e-127
	DDR1-like	2	0
	DOM3Z	2	1e-111~1e-108
	FLOT1	2	1e-155~1e-132
	GABBR1	1	1e-170
	GNL1	2	1e-122~1e-117
	GTF2H4	2	1e-179~1e-178
	HSPA1/2/5/6/8	2	0
	MSH5	1	1e-110
	NEU1-like	6	1e-111~1e-102
	PSMB5/8	1	1e-107
	RING1-like	1	1e-104
	RXRA/B/G	2	1e-159
	SKIV2L	1	1e-161
	SLC44A-like	2	1e-167~1e-161
	TAP1/2-like	4	0~3e-154
	TN-like	1	1e-150
	TUBB-like	7	0
	VARS2	2	0
	VARS2L	2	1e-161
	VPS52	2	0
	WDR46-like	2	1e-164

BLAST searches of the amphioxus genome assembly used 148 genes located in the human HLA regions as queries. These searches revealed 32 conserved genes (BLASTp score: P-value <e-100), 55 moderately conserved genes (e-10< P-value< e-100) and 61 non-conserved genes (e-10< P-value). Phylogenetic analysis of 30 conserved genes, excluding DHX16 and NOTCH4, were consistent with previous known evolutionary relationships. Genes exist on 36 scaffolds and reflect 13 gene duplication or haplotypic differences. Scaffold 11 has a genome structure (GNL1 - GTF2H4 - VARS2L - BAT1-like - BAT5-like) that is conserved relative to the human HLA

region. The amphioxus genome has at least four TAP1/2-like genes and one PSMB5/8 gene. However, MHC I, MHC II and the tapasin gene, which encodes the TAP binding protein, were not identified, suggesting that the MHC antigen presentation system was established after speciation of amphioxus. Six NEU1-like genes were identified in amphioxus. The up-regulation of Neu1 expression is important for the primary function of macrophages and establishes the link between Neu1 and the cellular immune response, suggesting that these genes could play a role in non-self recognition.

Supplemental Table S12. Conserved non-coding Elements tested for Enhancer Activity in Mouse

A. Human sequences tested

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near ZNF703 chr8:37652068-37652163 CACCTGCTTAGCAAAGCC
PCR Product Sequence:
GCATTGCTTAGCAAAGCCCTGAACTCTGCCTCCCACCAACACCCCCAAGGCTAAGTGA
GGTCAGGGGTGCTGACATTGGCCGAGTGAGTTCTTATGTCACCTGCCTCAACAACTTT
GGTTCTCCGCTTCCAATTATTTCTATCTTGGTCCATTTTCCCAGCTCTTCCTCCATCCTC
CCTCCTTCCTGGGGACTCTGGGAAGTGCCTGAGTGTGGCCTTGCAGGTTGCGCCACCG
CTTTCATCCCAAGTTTGATGTGACAAGCCTGATAGGCGTTGATTTACTTACAGACTGAT
GGGCTTTTAATTGAGCACGCCATCCTAGTCACTTCACCTCACCCGGCTCTGTAGCAAA
GGAACACATTTTCTGTAACTCAGAAATGGCCTTTTGTCTGCTGCCAAGGAAGGCAGCC
CCAGGCCCTTGATGTGCAAAATGGCTACTTTGTTCTGACCTCCTGTGCTG

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near SOX21 chr13:94156903-94157110 CACCCTAGGTATCCCCCA
PCR Product Sequence:
GCCCCTAGGTATCCCCCAGCCCCTACCTTTATTAGGAGGCCCGGGTTTCCCCCATTAA
ACTTATATTCATTCATTCATTCATTCAAGCCAGAGCCTCAGGACTGGCTCCCGGGAAT
GTTTGAAGATCATCGCGTATCTTGGAATTGCCAGGCTTCCTGCTCCACAGAATTACTG
CTACCCTATATTATCCCAACCTGTTCAATTTCTATCTGCCGCAGACAATGCTAATAATG
GCAATTATTGCTTCTATAATTCAGCCCTTGCAAAAGCATATCTATTCAAAGCCCTTGCA
AAAAGCATATCTATTCAAAGGGCATTTAGTCCATTTCTTTCTTTCCAGTTTTGGTCACC
CATTTTATTTATTTATGCGCCTGCTCCTATTCAGGGGCTCATGTATTTTTTATTATGTTG
TTTCACTGTAAAAAAACCTTGTGAAAAGAAACGCCTCACAATGGACACAGAAGAAGT
GCACAATGCTAACAGTTTTCCAAATACCACTGATTCTCGACTGTCTAAACGCTCCCAC
TGAAGGACAAAAAAAAAAAAAAGGATAAAAAAAGCTCTTTTCACCATTTCCCTTTTTT
TCCGAGAAAGAAAGGAGGAGAGATTAATAAAGTTTGGACGGCTCCCTCGGCCTCCGC
TGGGCCGCGGGCGCAAGAAAGACGCGGCGACCGTCCCTATCTGCTCCGGGGAGAGTG
GAATCTTCTCTCCATCTTCTGCGCCCCGAGCCTGGGGTGAGAGGGAGATCCGACAGCG
TCCCCAGCACCAGTTCTTACTGCTTACACAGCGC

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near SIX1 chr14:60191711-60191798 CACCGTAAGGAGGCCCA
PCR product sequence:
AATGGTAAGGAGGCCCAGTGACAAGGAAAGCAGACATTTAGCACCCCTCCCCATCT
CTTCACTTGAATTCTTACAATAAGCTTCGTTAGCGTGATTCACTGACTATCTGGTTTC
CTCAGGTAAACCTGCACTTACATATAAATCCTAGTAATTTGAAACTTCTCCCTTAATG
CGATGGCAATTGGAGTCCCTGATGACTGCGGCTCGGGTTTCTTGTAATGGCTCTACC
ACATTTTCCTGGACCTCCTTCGCCACAGCAGGCTGGGCGGCTGGGGAGAGCAGACG
TCTGCACGCGGGGCGCCGCAACCCGAAAACCTTTCGCTACTCTCTTGGGACTCTCCA
GTTATCGAGCGTTCCCACCCTACAGCGTCCGCCCAGGCGCAGGCCTGCCCCTAGTTC
AGGAGTGAACCC

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near ZNF503 chr10:76835035-76835166 CACCGTCAGCGTGCGTT

PCR Product Sequence:
CGTGGTCAGCGTGCGTTCGCTTTTCTCAAAACTCCACTCCGAGGGTCCGAGCGCATGG
GGGCGCCTGGGGGCCTGCGCGCTCTGCGCCAGGCCTAAAGCCTCCAACCCCCAGCAG
CAGTTGGCGTGGATGTCCTGCGGATTTTATTTGCAAACAATGGAAATTGAACTATTCA
CTGTCGAGATGGCTAATCAGTATTGAGGCCGGAGGGCAGGCGGGCCGCCTTTCACCG
CCGCTTCCCTTTCTCCCCTATTAATCTCCCTCGAAGATAATCCGCCTGACAGGTCAGGC
CCGGCGAGCTGGAAAGCCTGGCCCAGAGCACCCAAGCACCCAGGTCCGCCCCACCTG
ACTGGGAGCAAGGCCCAGATTCGCGGGGCCAGTAGGGGACAGTCCCGCGCGGCCCTC
TCCCTTCTTTTTGTTTTCTCTCCCGCGTCTTTCTGGGCTCCCCCATCTTTTTCCTTCATT
TCCTTTTCCTTTTCCCTTTTTTTTAC

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near ID1 chr20:29655769-29655845 CACCTACAGTGCCCGCTG

PCR Product Sequence:
TTCATACAGTGCCCGCTGTACAAAGTAAGAAATTAATAATGGTCACGTTTGTGATTAT
TATTTAAACACCAAGCAGATATTAAAGAGAGGTGGGGGGAGCGGTGAAGAAACCCCA
AGCGGCCCAAGCTGTGGGTCTGGGTTGGGAGACTCGCAGGTGTGGGGCACTAGCTAG
ACCAGTTTGTCGTCTCCATGGCGACCGCCCGCGCGGCGCCAGCCTGACAGTCCGTCCG
GGTTTTATGAATGGCGCCGTGGGAGGGAGACCCTGCTCTGAGGTCTTTGAGAAGAAA
ATTTAAAAAGCAGCCAAAAATGGGAAAAAACATTAAAAAACCCAGCCCCAAACTTAC
TAGACTTTCCCGAATTAATTGCTCCCACCCGGGAGGGATCTGGGTAGGCCCTCCGGGT
CTCAG

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near TBX3 chr12:113586552-113586623 CACCATGTCTTGGACCAT
GCCAATGTCTTGGACCATCCTTGCTTAGATGGAGGGAAATCTTGAAAACAATGTTCTT
TTATCTTTCCTGAGATGAAAGCTTCTGCATTTTGCCTTGTGCACCCGCGCAACCCCCAG
CAACCCCGAAACCCCAACGTGGCACCAAGGGAAGGGTGAAAATGAAGTGTTGAGAAT
TCCCTGAAGTCACCAGAGAGAATAGTGTGGTTCTGCGAACTCCGTCTGGCTCGCCATC
TCGCCCCTTTCTTTCCTCATAACACCCCCTCCCGCATCCCCTCCTCCCCACCTGAAGGG
TGCGAAAGGCCAGGCCATTTCCCGCCTTACACAGAAC

CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near TBX3 chr12:113587672-113587734 CACCTACTTGGGGTGGAA
CCACTACTTGGGGTGGAAGCAAAGTAGTTCTACAACCACACTGTGACCGACCTCCTC
CTCACCCACTCGGTGCCCCCAAGCGGCTCCGCGGGTGGACTGAGAGGGAGTTCTGC
CCCTCACACTGTCCTCCTCCCTCAATTGCCTATTTTTTGCCCATAGCTCTTCGAATCG
AAAATGTCGACATCTTGCTAATGGTCTGCAAACTTCCGCCAATTATGACTGACCTCC
CAGACTCGGCCCCAGGAGATAAACCCACTCTGGCGCCGGCCATGCGCTGGGTGATT
AATTTGCGAACAAACAAAAGCGGCCTGGTGGCCACTGCATTCCCGGCTATCTGCGGT
CTTTTTCTCTCTGCAGACCCCTCGATCCTCTTTCCTTCGGTCTCACACTCAACAAAAG
ACAGACTAGA
CNS tested CNS coordinates Forward Primer
Human sequence of aCNS near SIX1 chr14:60188067-60188154 CACCTGCACCCTGCAACA

PCR Product sequence:
ACAGTGCACCCTGCAACATACATTCTAAACTCATAAAGGGTTCCCACCTCAGGGCTG
CCACATGGGAGATTCCAAGTAGACATTTTTTTAAAAAAATGCATTTAATAGGTGTGC
ATGCTTTCAAAAGAGCCTCTGCCATATAGTATGCCCAATGACTTTCTAAACTGCCCA
GAAGGAATGCCTCTGCTTGAAACAGTCCTTATTGACATGGCAGCTGTGTGAGGACCG
GAACGGAGGCTTAAACCCGGAGTATGGGGTGTGCTTTGTAATTAACAATTTTGTATA
AGGAAATGAGGCATAGCCTGCACAAATAAGGGCAGGTATAAATTTACAGGACGTAC
TACATTGCCGTCATCTGGGAAGACGACGTGTAGAAGGTGTGATTGCGGGAACCCGG
GCAGAAAAGAGGTGCTACCTGCTATTTCGATCCCTCGGTTAGGGGAACCAATTGTTA
TAAATGTTAATTATTGGCTCGGATATATAAGGTTTCCCAAA

B. Amphioxus sequence tested

CNS tested	CNS coordinates	Forward Primer
An aCNS near amphioxus ZNF503/703	scaffold_9:3320559-3320655	CACCTCGCGTTTTTGTTT

PCR Product Sequence:
CACCTCGCGTTTTTGTTTGACAAGAAAAATTGTCCAAAAATTACAGATAATCAATTCT
TTACATTGTTTCCGGCCCTTGCCGGGCGGGTGCGAGGGGGAGGGCGCCCGGGCACCA
CCAGCGGTGCTGCGCCGCCCTTCTCCCCAAGAATCCTAGACAGCCGCCTGGCGCTGAA
CGGGCTAGTCCGCGAGGATTGACGCAGCCGGGAAGCGTAGTTTGTCGGCTGTGTTTGT
GTAGCGTGACCGGCTCTTTTATACGGTAGGTTGATGAATGGGCCCGACTCGCAGAAGC
CAAAGCGAACAAAACGAGCAGTTGTGATTAAAATTTTGCCCCGCCAAAAGAGACAGC
TTACGTCCTCGACTTGTTTGAAGTCGCCGCTCCTTAGGCGCTCAATTAAAAGCCTATCA
GTGTTTATATTTGGAACAATAACAGGCCGGACGACAAAGGACGCACACGGGGTAACC
AGAATCACACCAAATGTGTGAATGGCAGCCCTGGGGGCATTGGCGGGTACATGTGGT
TTTCATAAAGGCTGCTCCGGGCCTGTCAGGCCAGTCAGATACGCACGATTACCAACCA
CGATGATATATGGTGCCCTTTTTACACCGGACGTGCGTGAGAAGCCAGTTGAAAACGT
GTTCCCCGGGCGGGCAACGGTGCCGCGAAACACGCAAACACAAGAAGAGGACAACA
CGTTGATTGAAAGGCACTGAGTGAGCGCTAGAAACCCGCAACACACCAAACACTTCA
TTTCTTCTCTCCCCCAAAAGAAACATCTGCGAAGGACTGTTGA

Supplemental Table S13: Conserved non-coding Elements tested for Enhancer Activity in Amphioxus

A. Human sequence tested

CNS tested	CNS coordinates	Forward Primer
human sequence of aCNS near ZNF703	chr8:37,651,531-37,652,521	GGCAAGCTTAGCAAAGC

Sequence of PCR Product
GCATTGCTTAGCAAAGCCCTGAACTCTGCCTCCCACCAACACCCCCAAGGCTAAGTG
AGGTCAGGGGTGCTGACATTGGCCGAGTGAGTTCTTATGTCACCTGCCTCAACAACT
TTGGTTCTCCGCTTCCAATTATTTCTATCTTGGTCCATTTTCCCAGCTCTTCCTCCATC
CTCCCTCCTTCCTGGGGACTCTGGGAAGTGCCTGAGTGTGGCCTTGCAGGTTGCGCC
ACCGCTTTCATCCCAAGTTTGATGTGACAAGCCTGATAGGCGTTGATTTACTTACAG
ACTGATGGGCTTTTAATTGAGCACGCCATCCTAGTCACTTCACCTCACCCGGCTCTGT
AGCAAAGGAACACATTTTCTGTAACTCAGAAATGGCCTTTTGTCTGCTGCCAAGGAA
GGCAGCCCCAGGCCCTTGATGTGCAAAATGGCTACTTTGTTCTGACCTCCTGTGCTG
AGTCAGCCT

CNS tested	CNS coordinates	Forward Primer
Human sequence of aCNS near ZNF503	chr10:76834989-76835164	GGCAAGCTTGTCAGCGT

Sequence of PCR Product
CGGGTCGTGGTCAGCGTGCGTTCGCTTTTCTCAAAACTCCACTCCGAGGGTCCGAGC
GCATGGGGGCGCCTGGGGGCCTGAATTACTCTGCGCCAGGCCTAAAGCCTCCAACC
CCCAGCAGCAGTTGGCGTGGATGTCCTGCGGATTTTATTTGCAAACAATATGCCTTG
AACTATTCACTGTCGAGATGGCTAATCAGTATTGAGGCCGGAGGGCAGGCGGGCCG
CCTTTCACCGCCGCTTCCCGCCTCCCCTATTAATCTCCCTCGAAGATAATCCGCCTGA
CAGGTCAGGCCCGGCGAGCTGGAAAGCCTGGCCCAGAGCAGCCTCTGCACCCAGGT
CCGCCCCACCTGACTGGGAGCAAGGCCCAGATTCGCGGGGCCAGTAGGGGACAGTC
CCGCGCGTCTTCCTCCCTTCTTTTTGTTTTCTCTCCCGCGTCTTTCTGGGCTCCCCCAT
CTTTTTCCTTCATTTCCTTTTCCTTTTCCCTTTT

B. Amphioxus sequence tested

CNS tested	CNS coordinates	Forward Primer
Amphioxus sequence near ZNF503/703	scaffold_9:3320559-3320655	GGCAAGCTTCGCGTTTTT

Sequence of PCR Product
GGCAAGCTTCGCGTTTTTGTTTGACAAGAAAAATTGTCCAAAAATTACAGATAATCA
ATTCTTTACATTGTTTCCGGCCCTTGCCACCCGGGTGCGAGGGGGAGGGCGCCCGGG
CACCACCAGCGGTGCTGCGCCGCCCTTCTCCCCAAGAATCCTAGACAGCGTTTTGGC
GCTGAACGGGCTAGTCCGCGAGGATTGACGCAGCCGGGAAGCGTAGTTTGTCGGCT
GTGTTTGTGTAGCGTGATAGTGCTCTTTTATACGGTAGGTTGATGAATGGGCCCGAC
TCGCAGAAGCCAAAGCGAACAAAACGAGCAGTTGTGATTAAATTGTTGCCCCGCCA
AAAGAGACAGCTTACGTCCTCGACTTGTTTGAAGTCGCCGCTCCTTAGGCGCTCAAT
TAAAAGCCTATACACGTTTATATTTGGAACAATAACAGGCCGGACGACAAAGGACG
CACACGGGGTAACCAGAATCACACCAAATGTGTGAATGCTGGCCCTGGGGGCATTG
GCGGGTACATGTGGTTTTCATAAAGGCTGCTCCGGGCCTGTCAGGCCAGTCAGATAC

GCACATCTCCCAACCACGATGATATATGGTGCCCTTTTTACACGGACGTGCGTGAGA
AGCCAGTTGAAAACGTGTTCCCCGGGCGGGTTTACGGTGCCGCGAAACACGCAAAC
ACAAGAAGAGGACAACACGTTGATTGAAAGGCACTGAGTGAGCGCTAGAAACCC
TCACACACCAAACACTTCATTTCTTCTCTCCCCCAAAAGAAACATCTGCGAAGGACT
GTTGAAGCTTGCC

Supplementary Table S14.

SUMMARY of EXPRESSION OF ELEMENTS TESTED FOR ENHANCER ACTIVITY IN MOUSE

CNS tested	Total number of embryos	Number of Blue embryos	Results
Human sequence of aCNS near ZNF703	69	9	Confirmed enhancer
Human sequence of aCNS near SOX21	14	1	
Human sequence of aCNS near SIX1 (16.1)	57	2	
Human sequence of aCNS near DLX3	51	5	
Human sequence of aCNS near ZNF503	63	7	Confirmed enhancer
Human sequence of aCNS near ID1	62	8	Confirmed enhancer
Human sequence of aCNS near TBX3 (12.1)	30	5	
Human sequence of aCNS near TBX3 (13.1)	47	9	Confirmed enhancer
Human sequence of aCNS near SIX1 (5.1)	48	9	
Amphioxus sequence of aCNS near ZNF503/703	42	9	Confirmed enhancer

Table S15. Expression in mouse embryos of the conserved transcriptional enhancers in the mouse ZNF503 and ZNF703 genes and the amphioxus ZNF503/703 gene

Expression in mouse embryos driven by enhancer elements (Number of embryos with pattern / Number of blue embryos)											
	Limb	**Tail**									
Human aCNS near TBX3 (13.1)	5 / 7	5 / 7									
	Branchial arch	**Eye**	**Spine**	**Mid brain**	**Fore brain**	**Ear**					
Human aCNS near ID1	7 / 7	7 / 7	7 / 7	7 / 7	6 / 7	7 / 7					
	Fore brain	**Mid brain**	**Hind brain**	**Eye**	**Branchial Arch**	**Ganglia**	**Limb**	**AER**	**Side**	**Nostril**	**Abdo men**
Human aCNS near ZNF503	4 / 6	6 / 6	2 / 6	6 / 6	6 / 6	3 / 6	6 / 6	5 / 6	0 / 6	6 / 6	6 / 6
Human aCNS near ZNF703	8 / 9	7 / 9	7 / 9	9 / 9	7 / 9	6 / 9	8 / 9	0 / 9	0 / 9	0 / 9	0 / 9
Amphioxus aCNS near ZNF503/703	0 / 10	0 / 10	8 / 10	0 / 10	9 / 10	0 / 10	0 / 10	0 / 10	7 / 10	0 / 10	0 / 9

Table S16. Expression in amphioxus embryos of the conserved transcriptional enhancers in the mouse ZNF503 and ZNF703 genes and the amphioxus ZNF503/703 gene

The percentage of amphioxus embryos expressing reporter constructs in a given tissue as a percentage of total expressing embryos								
CNE	ectoderm	somites	Noto muscle ≤ 4 cells	Noto muscle >4 cells	Noto Müller cells	CNS	endoderm	Total expressing /total embryos
Human ZNF503	31.8%	27%	0%	0%	9%	13.6%	9.1%	22/201
Human ZNF703	28%	20%	4.3%	1.1%	6.5%	2.2%	3.3%	92/294
Amph ZNF504/703	42.6%	17.8%	20.2%	10.9%	12.4%	6.2%	1.6%	129/506

Total expressing embryos include those with expression only in abnormal cells in gut lumen; indicating successful injection. Injected embryos that developed very abnormally are not included in total. Expression, as is typical for transient transgenics, is mosaic. The notochord (noto) consists of modified striated muscle cells and associated cells, the Müller cells.