Sea anemone genome reveals the gene repertoire and genomic organization of the eumetazoan ancestor

92697

Abstract

Nematostella vectensis

Introduction

1, 2

3

4-

6

7, 8

e.g. *2*

9 *Hydra*

e.g. *Charnodiscus* *10*

11, 12 *13*

Nematostella vectensis

14, 15

16

Nematostella

14, 17

14, 18, 19 *Nematostella*

Nematostella

15, 20-

25

Nematostella

15

Nematostella

26

i.e.

27-29

i.e.

30

Nematostella

Nematostella

Genome Sequencing and Assembly

Nematostella *31*

32

Nematostella

32

32

32

33

32

7 *8* *Populus* *34*

35

Nematostella

Nematostella

Nematostella

Nematostella

36

37

Nematostella gene set

Nematostella *bona fide*

ab initio

32

32

32

The ancestral eumetazoan gene set

Nematostella

(i.e.

38

Nematostella

Drosophila *C. elegans*

Ciona intestinalis

Nematostella

Nematostella

Drosophila *C. elegans*

i.e.

Nematostella *i.e. Drosophila* *C. elegans*

Nematostella

Drosophila

C. elegans

Nematostella

Nematostella

32

Nematostella *Hydra*

Nematostella Hydra

Hydra

39

C. elegans *40, 41*

e.g. *42*

Mnemiopsis leidyii

32

Nematostella

32

Nematostella

43

et al. *42*

32

Nematostella

32

Arabidopsis

Cryptococcus neoformans *32*

46

46

Nematostella *Arabidopsis,*

Cryptococcus

47

et al *Platynereis*

Conservation of ancient eumetazoan linkage groups

48, 49

e.g. *50*

Nematostella

Nematostella

51

52

32

52

48

Nematostella

Nematostella

32

Nematostella

i.e.

Nematostella

32

Nematostella

Nematostella *vice versa*

Nematostella

49

53

Nematostella

32

Nematostella

54-56

Nematostella *Nematostella*

56-58

bona fide

Nematostella *54, 56, 59*

Origins of eumetazoan genes

i.e.

32

60

32

e.g. *61*

de novo

62

i.e.

Eumetazoan networks and pathways

32

Nematostella

63

Signaling Pathways

64

Nematostella

24, 27, 28, 62, 65, 66

e.g.

e.g.

e.g. *e.g.*

Emergence of the neuromuscular system

e.g.

e.g.

e.g.

e.g.

e.g.

e.g.

Concluding remarks

Nematostella

Hydra

Nematostella

FIGURE CAPTIONS

Figure 1. ***Nematostella*** development and anatomy.

Nematostella

Figure 2. Bayesian phylogeny of metazoa.

Ciona intestinalis *Takifugu rubripes* *Xenopus tropicalis*

Lottia gigantea *Drosophila melanogaster* *Caenorhabditis elegans*

Hydra magnipapillata *, Nematostella* *Reniera sp. JGI-2005* *Monosiga*

brevicollis *Saccharomyces cerevisiae*

Figure 3. Patterns of intron evolution in eukaryotes.

Arabidopsis thaliana *Crytococcus neoformans* *C. elegans* *D. melanogaster* *Ciona intestinalis* *Homo sapiens* *Nematostella* *32*

Figure 4. Conserved synteny between the human and anemone genomes.

Nematostella

32

Nematostella

Nematostella

Nematostella

Nematostella

Nematostella

Figure 5. Origins of eumetazoan genes.

References

Generelle Morphologie der Organismen

Animal Evolution, 2nd edition

On the Origin of Phyla

et al. Science 287

Science 282

Science 282

et al. Science 298

et al. Science 314

Proc Natl Acad

Sci U S A 89

Paleontology 36

Lethaia 22

Nature 361

et al. Dev Biol 248

Biological Bulletin 182

et al. Bioessays 27

Estuaries 17

Dev Genes Evol 212

Dev Genes Evol 216

Gastrulation: From Cells to Embryos

Development 131

Evol Dev 5

Dev Biol 275

Dev Genes Evol 212

et al. Nature 426

Evol Dev 7

Mol Phylogenet Evol 24

et al. Proc Natl Acad Sci U S A 103

Curr Biol 16

et al. Dev Biol 296

London: The Ray Society II

Genome Res 7

et al. Nucleic Acids Res 34

et al. Science 313

et al. Science 297

Mol Ecol 13

submitted

Science 278

et al. Proc Natl Acad Sci U S A 97

Genome Res 14

et al. Science 311

Proc Natl Acad Sci U S A

101

et al. Proc Natl Acad Sci U S A 101

Genome Informatics 17

et al. Science 310

Curr Biol 13

Pac Symp Biocomput

et al. Genome Res 10

et al. Nature 431

Nature 444

Nat Genet 31

Genome Res 15

J Struct Funct Genomics 3

Science 304

Curr Biol 16

et al. Genome Biol 7

et al. Nature 442

J Exp Zoolog B Mol Dev Evol

Evol Dev 1

BMC Evol Biol 4

Protein Evolution

et al. *Nature* 433

Int Comp Biol 43

Nat Rev Genet 4

et al. *Trends Genet* 21

Semin Cell Dev Biol 17

Nature 424

Acknowledgements

Nematostella

Figure 1.

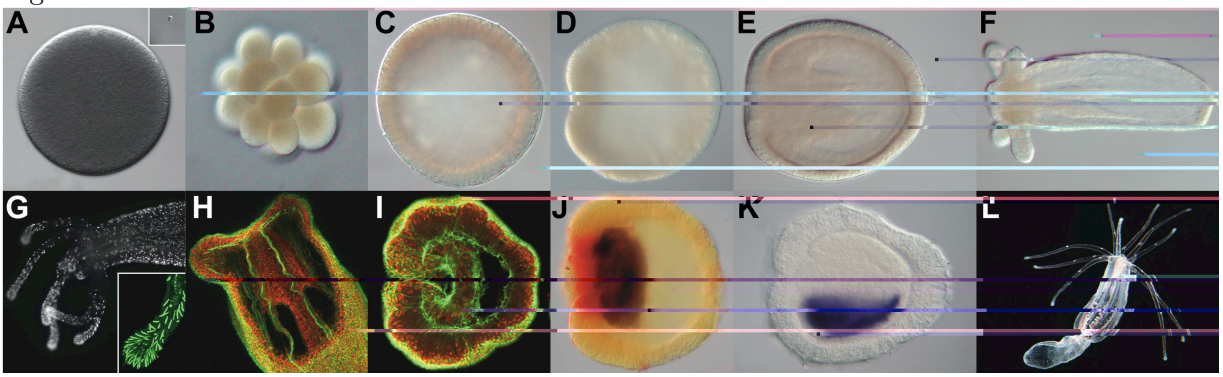


Figure 2a.

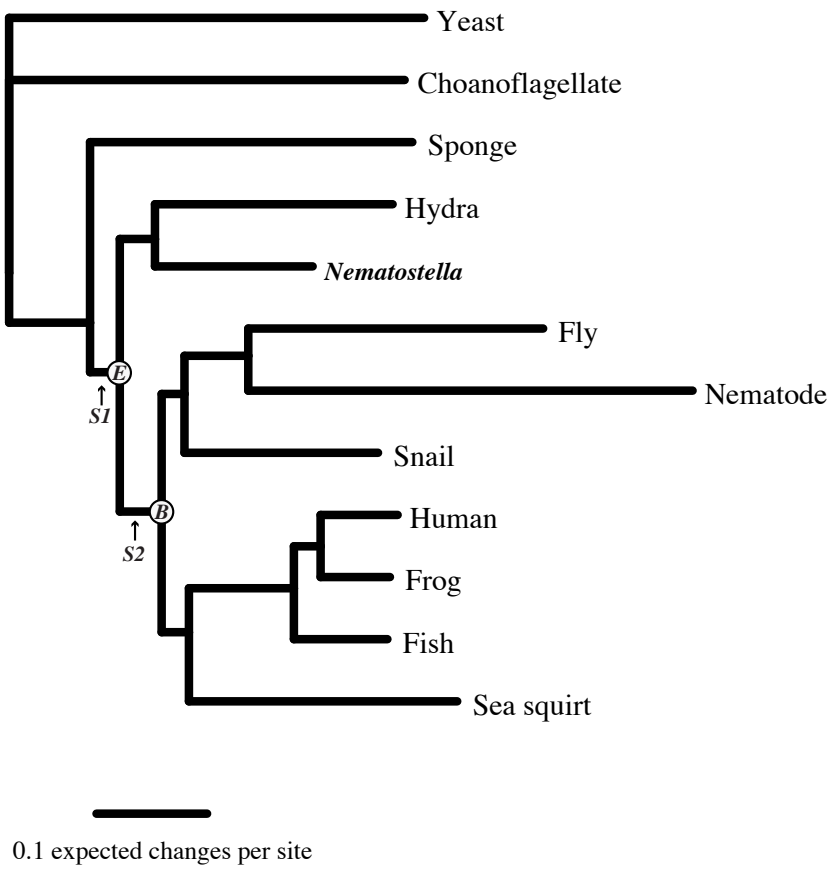

Yeast
Choanoflagellate
Sponge
Hydra
Nematostella
Fly
Nematode
Snail
Human
Frog
Fish
Sea squirt
E
S1
B
S2
0.1 expected changes per site

Figure 2b.

	Eumetazoan Stem	Bilaterian Stem
New genes originated	1148	662
New genes created through gene family expansion	1470	320
Reconstructed genes of the recent common ancestor	7766	8748

Figure 3.

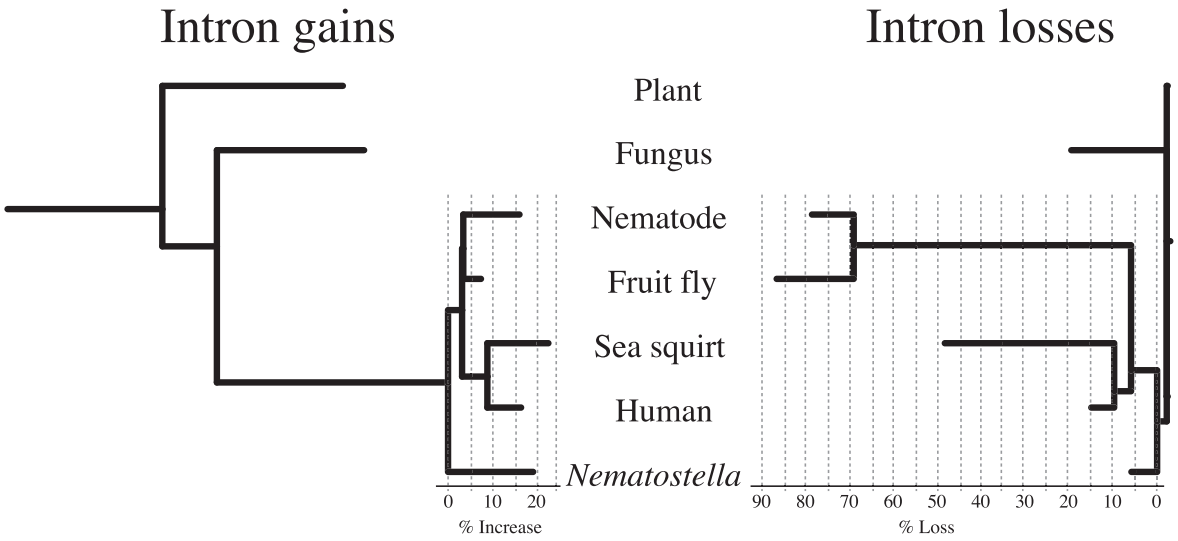

Intron gains
Intron losses
Plant
Fungus
Nematode
Fruit fly
Sea squirt
Human
Nematostella
0 10 20
% Increase
90 80 70 60 50 40 30 20 10 0
% Loss

Figure 3b.

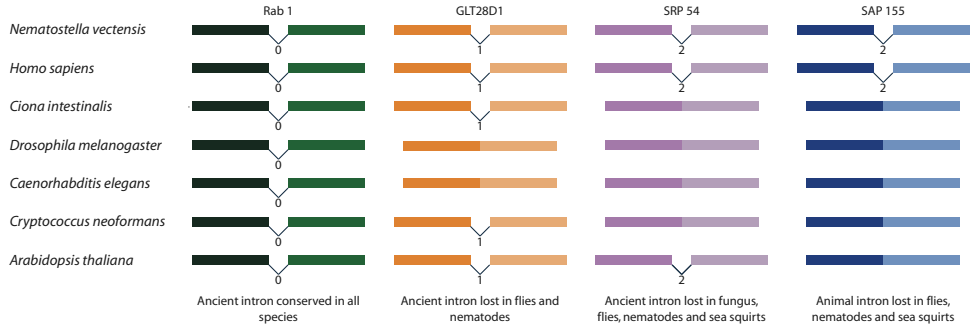

Rab 1
GLT28D1
SRP 54
SAP 155
Nematostella vectensis
Homo sapiens
Ciona intestinalis
Drosophila melanogaster
Caenorhabditis elegans
Cryptococcus neoformans
Arabidopsis thaliana
0
1
2
Ancient intron conserved in all species
Ancient intron lost in flies and nematodes
Ancient intron lost in fungus, flies, nematodes and sea squirts
Animal intron lost in flies, nematodes and sea squirts

Figure 4a.

Figure 4b.

	Nematostella scaffold:	3	5	46	26	53	61	44	144	7	74	18	88	52	42	156	89	10	8	34	118	91	191
Human Chromosome Segment:	2q11.2-35	25	32	17	16	17	9	12	7	1	2			2	1		1	1	1	2			
	12q12-14.3	16	14	9	5	8	3	6	5			1								1	1		
	17q12-21.32	12	8	4	10	6	4	3	1			1	1		1				1				
	7p11.2-21.3	4	10	3	3	2	7	1	2			1	1					2	1	1	1	1	
	10p11.22-13	8	6	1	1	2	1	4	1														
	14q12-32.33	10	3	2	4	5	3	3	2	23	12	13	11	17	11	9	8	1	2			1	
	11q12.1-13.1	4					2	2	1	12	7	1	6	6	1		4			2			
	1q32.2-44	6	4		1	1	2		1	11	6	6	3	6	2	2	1						1
	19q13.11-13.33	4	1	2	2	1	2	1		5	8	8	4	6	2		3	2	3		1	1	
	2p13.2-24.3	5	2	2	1	2	1	2	1	8	5	10	5	3	1	5	3	2		1			
	17q23.3-25.3	1	2						1	2		2	1					19	10	12	8	7	3
	16p11.2-13.3	1			1	1						1	1	1				17	19	9	5	6	6
	7p22.1-22.3											1		1	1			6	3	3	5	2	
	17p11.2-13.1				1						1			1			1	6	2		1	3	

Figure 4c.

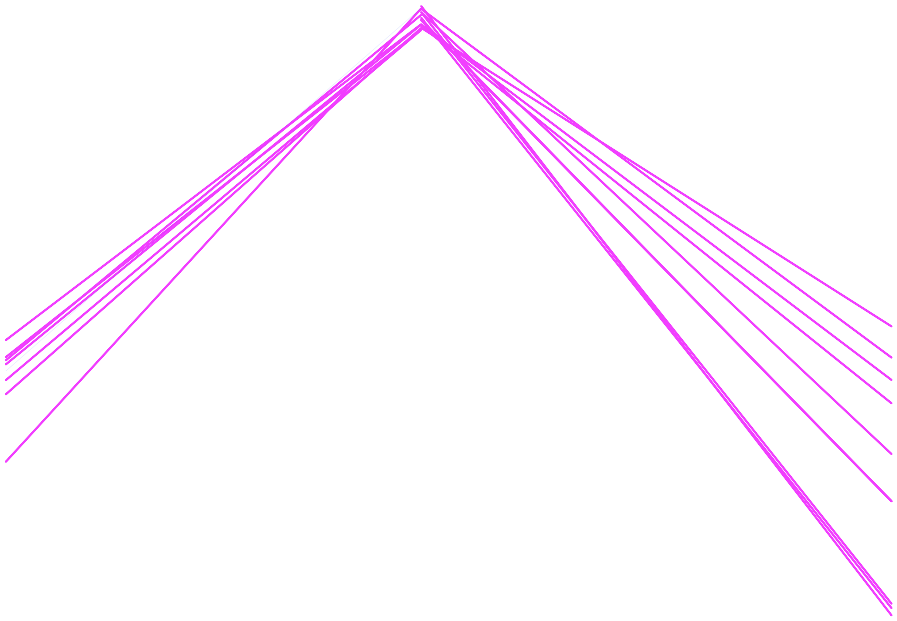
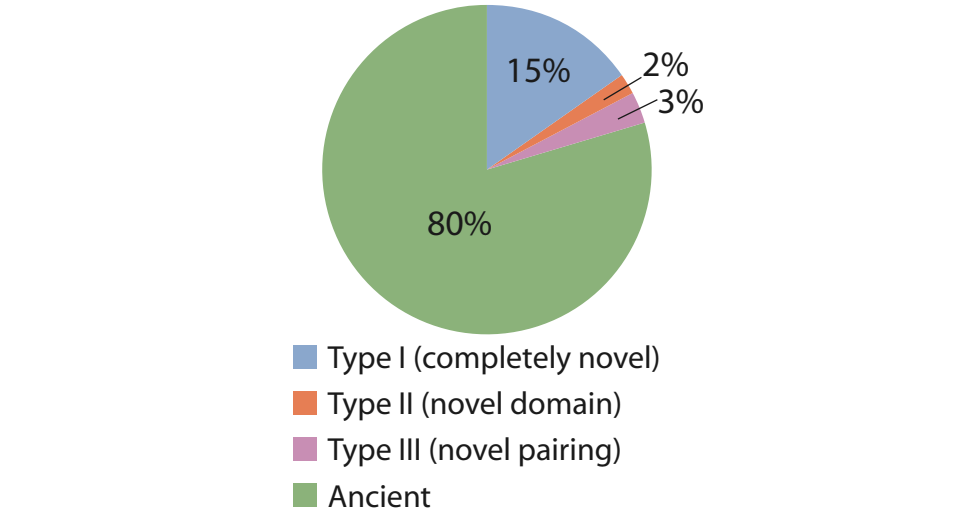


Figure 5a.



Type I Novelty: SMAD Family Proteins

Type II Novelty: Notch Proteins

Type III Novelty (Lim Homeodomain Proteins)

Figure 5b.

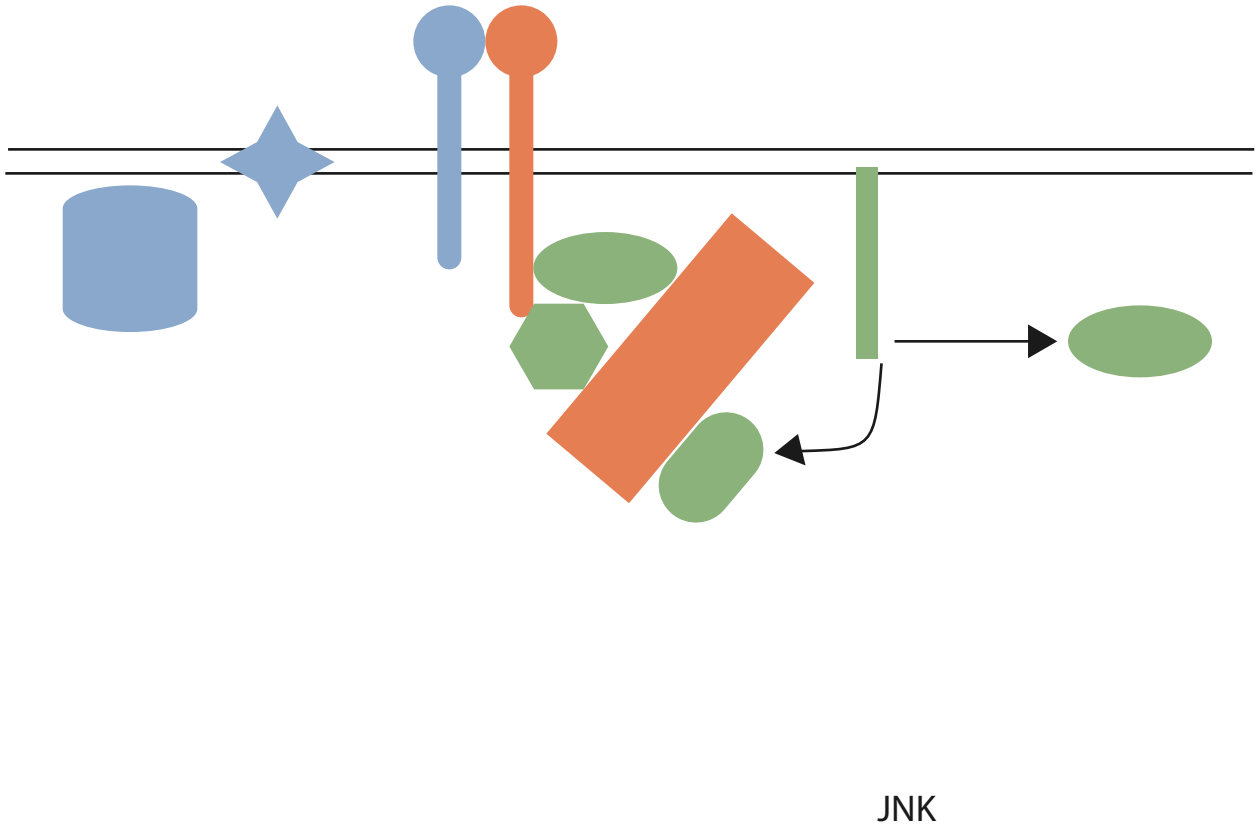


Figure 5c.

Pathway	Type I Novelty	Type II Novelty	Type III Novelty	Ancient Gene
Integrin signaling	Integrin-alpha; caveolin	Collagen; Integrin-beta; Fak; Jun	Calpain	talin,; vinculin; paxillin; Ras; Grb2; SoS; Rap; ERK; MEK, Crk
Wnt signaling	Wnt; secreted frzzled related factors; frizzled; strabismus/van gogh	Dickkopf; arrow; dishevelled; axin		Beta-catenin; GSK3; APC; TCF/LEF; groucho
TGF-beta signaling	dpp/BMP; activin (nodal, nodal-related); gremlin; chordin; follistatin; R-SMAD; I-SMAD; co-SMAD	Type I receptors: TGFBR1, BMPR1A; ATF/JunB; snoN	Tolloid/BMP1	Type II receptors: ACVR2, BMPR2
Notch signaling	Numb; hairy/E(spl)	notch		Jagged; deltex; fringe; presenilin; ADAM10; nicastrin; furin; Aph1; PEN2; mastermind
Ephrin signaling		Ephrin; Fak	Eph (receptor)	Abl/SYK
Insulin signaling	insulin	insulin receptor substrate; phosphoinositide-3-kinase, catalytic	Insulin receptor/IGFR; phosphoinositide-3-kinase, class 2	phosphoinositide-3-kinase, class 3; phosphoinositide-3-kinase, regulatory subunit; 3-phosphoinositide dependent protein kinase- 1; PTEN
FGF signaling	FGF; Shc	Raf homolog serine/threonine-protein kinase; Ras GTPase activating protein	FGFR; RAS protein activator; phospholipase C, gamma; phosphoinositide-3-kinase, class 2; Protein kinase C iota	MAPK; phosphoinositide-3-kinase, class 3; Grb2; Protein kinase C; SoS; Rac
Cytokine signaling	inositol 1,4,5-triphosphate receptor; SOCS; arrestin; guanine nucleotide binding protein (G protein); gamma, regulator of G-protein signalling; REL/NFKB; NFAT	Adenylate cyclase 5/6; STAT5. ATF/Jun	CDC42 binding protein kinase	MAPK; Rho kinase; Rho

Figure 5d.

Process	Type I Novelty	Type II Novelty	Type III Novelty	Ancient Gene
neurogenesis	Hes, Gcm, Ephrin, netrin, semaphorin, dachsund, ski oncogene	notch, NGFR, Dsh, Arx, CREB/ATF, neuralized	neuropilin, Lhx, EPH receptor,	single-minded/HIF, achaete-scute, elav, Emx, Otp, Jagged, Deltex, Irx, Gli, Otx/Phox, stonal/neuroD/neuroG, reticulon
synaptic transmission	nitric oxide synthase (neuronal) adapter protein, DOPA-beta monoxygenase, calcium channel voltage dependent beta, syntrophin, synaptophysin, dystrophin, potassium large conductance calcium-activated channel, subfamily M, beta	cholinergic receptor nicotinic, neurexin	K-voltage gated channel, discs large	glutamate receptor, synaptotagmin, intersectin, synapsin, neuroligin/CES, syntaxin, glutamate transporter
ECM	netrin, dermatopontin, semaphorin, glypican, stereocilin	collagen, spondin, laminin,	nidogen, stabilin, neuropilin, matrix metalloprotease, thrombospondin	leprecan, microfibrillar associated protein
cell junction	par-6	tight junction protein		salvador
muscle contraction	voltage dependent calcium channel beta, beta-sarcoglycan, beta-dystrobrevin	cholinergic receptor nicotinic, nebulin, tropomyosin, calponin/transgelin	voltage dependent calcium channel alpha2/delta subunit, inositol triphophate receptor, calcium activated potassium channel slowpoke	phohorylase kinase, myosin light chain cytoplasmic, calcium channel alpha subunit, cGMP dependent protein kinase, calcium/calmodulin dependent kinase II, myosin regulatory light chain
Apoptosis	TNF5/10/11; Bcl2; BOK; GULP; engulfment adaptor PTB domain containing 1; CRADD; caspase 8/10; GULP1; growth arrest and DNA-damage-inducible; DNA fragmentation factor 40 kDa subunit ; Interleukin enhancer-binding factor 3; FMR	BIRC; CARD9/11	NGFR; SLIT-ROBO Rho GTPase activating protein; calpain	TNFRSR; TRAF; scavenger receptor class B; huntingtin interacting protein; programmed cell death 1/5; Bcl2-associated athanogene; Akt; SUMO; defender against cell death 1; apoptosis-inducing factor (AIF)-like mitochondrion-associated inducer of death; death-associated protein kinase
Transcription factors	L3MBT; T-Box; Nuclear hormone receptor; SMAD; dachsund; gcm; NFAT; nuclear respiratory factor; SNO and SKI family; sprouty; AP-2; onecut; MAF-related;	CBP/p300; ETO/MTG8/Nervy; groucho; Jun; Myt1; runt; STAT	hairless; nuclear protein 95; LIM homeobox; CCAAT enhancer binding; aryl hydrocarbon receptor related	zic; Gli; homeobox; bHLH; achaete-scute; sox; retinoblastoma binding potein 5/8; NFKB-related; Krueppel C2H2 type zinc finger; irx; Deltex; ataxin

Supporting Online Material

Supplement S1

Additional background information on ***Nematostella vectensis***.
The starlet sea anemone *Nematostella vectensis* (Family: Edwardsiidae) is a burrowing, brackish-water, solitary sea anemone with a worldwide distribution (*1, 2*). Self-sustaining laboratory cultures can be maintained year-round in artificial seawater, with daily feedings of brine shrimp (*3, 4*). While sexes are separate, they are not obviously morphologically distinguishable. *Nematostella* is unique among cnidarians in that it can be induced to spawn repeatedly on a regular cycle in the laboratory to produce large numbers of gametes that can be manipulated by simple in vitro fertilization methods (*4*). Development occurs via planula larvae that emerge from the jelly of the egg mass within two days at 20-25C (5-6 days at 18C) (*4*). Planulae are formed by gastrulation via invagination, and have an apical tuft at one end of the animal. A single planula larva is about 250 µm in length and consists of over 10,000 cells (Figure 3A-I). Metamorphosis into a four-tentacled juvenile polyp with two mesenteries (partitions that partially divide the gut and increase its surface area, also providing pouches for the production and storage of gametes) takes about a week, with sexual maturity reached in 3-4 months. Mature adults are hollow tubes typically 5-10 cm in length, with an open (oral) end encircled by 10-20 tentacles a few cm long, and a closed (aboral) end (Figure 2). The animals are carnivorous, capturing and consuming plankton, including small animals and their larvae, using tentacles and the characteristic stinging cells of cnidarians, which inject neurotoxin into prey.

Individual animals have been maintained in the laboratory for over fifteen years (C. Hand, private communication). Asexual reproduction can be induced by tying a fine thread around the body tube. Within a few days, the animal will separate into two individuals, producing both a new mouth and basal disc. As with other cnidarians, Nematostella possesses considerable regenerative abilities, reconstituting a complete and properly proportioned adult from only a part of the animal. Tentacles can also regrow when cut. It is not known how tentacle number or body tube length is regulated, either in regeneration or embryogenesis.

Table S1.1 contains a partial list of the merits of Nematostella as a model organism.

Figure 1 Methods
Nematocyst staining (Figure 1g): (Methods adapted from (*5*)) Juvenile and small adult Nematostella

and CH6 females. These parental strains – clones of which are widely available today in at least four laboratories and can be readily redistributed – are from the original colony established and maintained by Cadet Hand at the Bodega Bay Marine Laboratory in the early 1990's (*3*). Because commensals or symbionts have been reported for *Nematostella*, gametic or embryonic DNA is preferred to avoid contamination from symbionts and/or undigested food. DNA from the same preparation was used to create a BAC library, described below. Thanks to asexual reproduction, the haplotypes represented in the draft genome sequence and BAC library [see below] can be propagated indefinitely.

CHORI BAC library

A Bacterial Artificial Chromosome (BAC) library was produced by Drs. Baoli Zhu and Pieter de Jong at the Children's Hospital Oakland Research Institute (CHORI). This library provides a ten-fold coverage of the genome. The average size of the inserts in the library is 168 kb. Funding for construction of the library was provided by a grant from the NSF (Robert Steele, PI, Ulrich Technau, Co-PI). The library is available through the CHORI BACPAC resource (deJong et al). More information can be found at http://bacpac.chori.org/library.php?id=219.

Whole Genome Shotgun (WGS) Sequencing and Assembly.

The genome of *Nematostella vectensis* was sequenced and assembled by whole genome shotgun (WGS) (*7*) as previously described (*8*). Briefly, genomic DNA prepared as described above was used to create shotgun libraries with inserts of approximately 3,000 bp, 6,500 bp and 35,000 bp. The libraries used, their mean insert sizes, and the numbers of reads sequenced are listed in Table S2.1. The shotgun reads were trimmed of low quality and vector-derived sequence, and assembled using JAZZ(*8, 9*). Approximately one third of the shotgun reads are composed entirely of high copy-number repeat sequences, and are therefore masked at the alignment stage of JAZZ, and therefor remain unassembled. Table S2.2 lists 10 abundant tandemly-repeated sequences in the shotgun dataset which together account for 32% of shotgun reads.

The assembled genome contains a total of 59,124 contiguous reconstructed sequences ("contigs") with a total length of 297 million base pairs (Mbp) and 10,804 "scaffolds", or reconstructed fragments of the genome that include gaps of unknown sequence, with a total length of 356 Mbp. Half of the contig sequence is contained in the largest 3,617 contigs, which are all at least 19,835 bp in length (N50). Half of the total scaffold sequence is contributed by the largest 181 scaffolds, which are each at least 472 Kbp in length.

Approximately 0.8% of positions in the assembly contain a polymorphic site (Figure S2.1), and we estimate that the mean pairwise variation between the four haplotypes represented in the libraries is 0.64 % (Figure S2.2).

Expressed sequence tag (EST) library preparation, sequencing, and assembly

A mixed stage cDNA library for *Nematostella* was prepared in the laboratory of Ulrich Technau, cloning polyA RNA from unfertilized eggs through metamorphosis into pSPORT 6.1. The library contains 56 million colony forming units (cfu) at a concentration of 4.7 million cfu/ml. The average insert size of the library is 1.96 kb, with greater than 99.5% recombinant, and an estimated 75% full length based on pilot sequencing. Of 1,152 sample sequences, 99.9% were passing, and 80% possessed significant BLASTX hits (E-value < 1E-5). 780 contigs were produced, with 680 single clones; the most abundant sequence was EF-1a, found in 3% of the sample, indicating that even without normalization this library

has a relatively low level of redundancy.

To enable the characterization of gene structures and to provide resources for further study, 88,704 cDNA clones from the library were end-sequenced to provide 146,095 expressed sequence tags (ESTs). The ESTs were clustered and assembled into 30,813 contigs via the JGI EST pipeline. Of these, 7,925 contigs were found to have a complete (start codon to stop codon) open reading frame (ORFs) of at least 450 bp. These putatively full-length EST contigs were aligned to the assembled WGS scaffolds using BLAT(*10*) (-maxIntron=100000 -extendThroughN).

To evaluate the completeness of the WGS assembly with respect to this collection of ESTs, we considered the number of putative full length EST contigs aligned to the genome at varying levels of completeness. For alignments of at least 95% sequence identity, 7,738 (97.6%) had an alignment spanning at least 25% of the length of the EST contig, 7,557 (95.4%) had an alignment spanning at least 75% of the length of the EST contig, and 7,193 (90.8%) had an alignment spanning at least 95% of the length of the EST contig. 138 of the 222 EST contigs that lacked an alignment over at least 50% of their length had an identifiable alignment to human refseq genes by BLASTP(*11*) (-e 1e-5), indicating that they are likely to represent *bona fide* protein-coding transcripts rather artifactual sequence. Others may be contaminants of the EST library, or novel genes.

839 (11.1%) of the EST contigs had alignments of at least 95% identity spanning at least 75% of their length with multiple locations in the assembly, indicating that up to approximately 10% of the non-repetitive genome may be represented redundantly in the assembly.

For *Mnemiopsis leidyi*, a cDNA library was created from total RNA prepared from gastrual stage embryos and reversed transcribed with oligo dT primers and the ZAP cDNA Synthesis Kit (Stratagene) by Kevin Pang and Mark Martindale. cDNA fragments with sizes ranging from ~500-2000 base pairs were cloned into pBluescript SK, and 15,360 paired clone end sequences were generated at JGI.

Repeat sequences reconstructed from unassembled WGS reads

Repeats were identified by assembling 16-mers (DNA sequences of length 16 bp) that frequently occurred in both ends of a sample of 50,000 fosmid clones from the ASYG library. Any 16-mers that occurred in both ends of at least 20 clones were used in the assemblies. The assemblies were performed using juggernaut.pl, a script developed for this purpose. tRNAScan-SE(*12*) was used to look for tRNAs and BLASTN(*11*) against nr and Repbase(*13*) to identify the 5S,18S,28S,U2,U6 RNAs, and two *Nematostella* transposons (see below). The five elements lacking notes are not identified by either of these methods.

The tandem array sizes are estimated by calculating the probability that a fosmid end matches the repeat given that its sister does. This probability can be used to estimate the expected array size (an average over multiple arrays in some cases) in terms of the mean fosmid length (37kb). These estimates depend on the assumptions of "normal" cloning behavior for these repetitive sequences.

10 families of tandemly repeated sequences were identified which occur in arrays longer than fosmid-length and account for 32% of the WGS data set. The key characteristics of these repeats are described in Table S2.2. See the file juggernaut.fasta for the complete sequences of these 10 elements.

Transposable elements in the sea anemone genome

Transposable elements (TEs) constitute more than 26% of the assembled sea anemone genome (Table S2.3) and belong to >500 families. These families are composed of a small number of copies (from 1 to ~5,000) and they all are relatively young: elements from the oldest families are less than 15% divergent from their consensus sequences and their ORFs coding for transposases, reverse transcriptases, and other transposon-specific proteins are not severely damaged by mutations.

In terms of their bulk contribution to the genome size, DNA transposons are fourfold more abundant than retrotransposons (Table S2.3). However, while different classes of anemone retrotransposons, including Gypsy, DIRS, Penelope, and CR1, are composed of more than 50-100 families each, different classes of autonomous DNA transposons are represented by just a few families. It appears that retrotransposition of retrotransposons, despite their high diversity, has not been as efficient as propagation of DNA transposons in the anemone genome.

The variety of different types of DNA transposons found in the anemone genome is the highest among eukaryotic species studied so far. Representatives of all reported superfamilies and groups of eukaryotic DNA transposons (*14-16*), excluding the Transib superfamily and the Mariner group of the Mariner superfamily, are present in the anemone genome. Even, En/Spm (also called CACTA) and transposons, which were believed to populate plants genomes only (*14*), reside in the anemone genome. While the anemone 10,632-bp EnSpm-1_NV and 9,347-bp EnSpm-2_NV transposons encode transposases (TPase) similar to the plant En/Spm TPase and are flanked by 3-bp targets site duplications typical for known En/Spm elements, their 5'-CACAG termini differ from the 5'-CACTA termini of the plant transposons.

Over 3% of the anemone genome is made of fossilized copies of self-synthesized Polinton DNA transposons whose transposition depends on the Polinton-encoded DNA polymerase and integrase (*17*). It makes *Nematostella* the first metazoan with Polintons constituting a substantial portion of the genome (*17*).

Remarkably, the sea anemone genome is a safe haven for unusual transposons that have never been seen before. For instance, Troyka, a novel type of LTR retrotransposons distantly related to the Gypsy superfamily, is characterized by 3-bp target site duplications (TSDs), while all known LTR retrotransposons, including retroviruses, are defined by 4-6 bp TSDs (*14*). Among DNA transposons, the hAT superfamily is well-known for TSDs that are always 8 bp long (*14*). However, the sea anemone genome, in addition to the canonical hAT transposons contains two novel groups, hAT5 and hAT6, characterized by 5- and 6-bp TSDs, respectively. Importantly, using reverse transcriptase/integrase and transposase encoded by the anemone Troyka, hAT5, and hAT6 transposons as queries in TBLASTN searches against GenBank DNA sequences, we found that proteins closest to the queries (>30% protein identity) are encoded by TEs characterized by the same unusual lengths of TSDs. For instance, Troyka retrotransposons are present also in sea urchin, and the hAT5 and hAT6 transposons are wide spread in sea urchin, sea squirts and lancelet.

The anemone genome is also populated by a novel superfamily of eukaryotic "cut and paste" DNA

transposons, called IS4EU, characterized by their TPase distantly related to the bacterial IS4 TPase. Following identification of the IS4EU TEs in the anemone genome, members of this superfamily have been also found in other species, including lancelet.

Analyzing anemone TEs, we have also advanced in our understanding of evolution of non-LTR retrotransposons (Fig. S2.1). For instance, the anemone genome harbors two families of Tx1-like non-LTR retrotransposons, Tx1-1_NV and Tx1-2_NV, inserted in 5S rRNA and U2 smRNA, respectively, at target sites identical to those of different Tx1 elements in fish (*18*), frog and lancelet. We suggest that Tx1-like elements form a novel clade of non-LTR retrotransposons differing from the L1 clade elements by the strong target-site specificity.

RTE is another clade of non-LTR retrotransposons first described a few years ago (*14, 19*). All known RTE elements, including those in plants, insects, nematodes, and vertebrates, contain only one ORF and are characterized by extremely frequent 5' truncations of the RTE elements during their retrotransposition. Here, we show that the anemone genome contains several families of RTE-like elements, RTEX in Fig. S2.1, which are longer than canonical RTE elements and contain an additional ORF at their 5' terminal portion that codes for the esterase domain, analogously to elements from the CR1/L2 clade (*20*).

Transposable Element Analysis Methods

Transposable elements were identified using WU-BLAST (http://blast.wustl.edu) and its implementation in CENSOR (http://girinst.org/censor/). First, we detected all fragments of the anemone genome coding for proteins similar to transposases, reverse transcriptases, and DNA polymerases representing all known classes of TEs. The detected DNA sequences have been clustered based on their pairwise identities by using BLASTclust (standalone NCBI BLAST(*11*)). Each cluster has been treated as a potential family of TEs described by its consensus sequence. The consensus sequences were built automatically based on multiple alignments of the cluster sequences expanded in both directions and manually modified based on structural characteristics of known TEs. Using WU-BLAST/CENSOR we identified fragments of the anemone genome similar to the consensus sequences that were considered as copies of TEs. Second, given the identified consensus sequences, we detected automatically insertions longer than 50-bp present in the identified copies of the protein-coding TEs. The insertions have been treated as potential TEs, clustered based on their pairwise DNA identities and replaced by their consensus sequences built for each cluster. After manual refinements of the consensus sequences, the identified families of TEs were classified based on their structural hallmarks, including target site duplications, terminal repeats, encoded proteins and similarities to TEs classified previously. Identified TEs are deposited in Repbase (*13*).

Supplement S3

Gene prediction and quality control

The genome of *Nematostella vectensis* includes 27,273 predicted gene models built using the JGI Annotation Pipeline, described below. The genomic sequence, predicted genes and annotations of *Nematostella*, together with available evidence, are available at the JGI Genome Portal (www.jgi.doe.gov/Nematostella)

The JGI Annotation Pipeline was used for annotation of the v1.0 *Nematostella* assembly described here. The pipeline includes the following annotation steps: (1) repeat masking, (2) mapping ESTs, full length cDNAs, and putative full length genes, (3) gene prediction using several methods, (4) protein

annotation using several methods, and (5) combining gene predictions into a non redundant representative set of gene models, which are subject to genome-scale analysis.

Transposons were masked in the *Nematostella* assembly using RepeatMasker (*21*) tools and a custom library of manually curated repeats (available upon request from V. Kapitonov). 146,095 ESTs were clustered into 30,813 consensus sequences and both individual ESTs and consensus sequences were mapped onto genome assembly using BLAT(*10*).

Gene predictors used for annotation of *Nematostella* v1.0 included *ab initio* FGENESH (*22*), homology-based FGENESH+ (*22*), homology-based GENEWISE (*23*), and EST-based ESTEXT (Grigoriev, unpublished).

A set of 1,678 genes derived from EST clusters with a putative full length ORF was directly mapped to the genomic sequence to build gene models. FGENESH was trained on this set to achieve sensitivity and specificity of 81% and 80%, respectively. To generate homology-based gene models, proteins from the NCBI NR database were aligned against genomic sequence using BlastX(*11*). High quality seed proteins were then used to build models using FEGENESH+ and GENEWISE. GENEWISE gene models were then filtered to remove models with frameshifts and internal stop-codons and extended to include start and stop codons where possible. FGENESH, FGENESH+ and GENEWISE gene models were then processed using ESTEXT to correct them according to splicing patterns observed in available ESTs and to extend 3' and 5' UTR of the genes.

All gene models were annotated by homology to other proteins from NCBI NR, SwissProt and KEGG databases. Using InterproScan (*24*) we predicted proteins domains. Using both these sources of information, annotation of each protein was mapped to the terms of Gene Ontology (*25*), KOG clusters of orthologs (*26*), and mapped to KEGG pathways (*27*).

The large set of all predicted models was reduced to a non-redundant set of 27,273 representative models (Filtered Models), where every locus is described by a single best gene model according to the criteria of homology and EST support. For this set of representative gene models we assigned GO (*25*) terms to 12,786 proteins, 16,625 (78%) proteins to KOG clusters (*26*), and 695 distinct EC numbers were assigned to 2,822 proteins mapped to KEGG pathways (*27*). Table S3.1 summarizes the set of predicted genes.

The data are available from JGI Genome Portal (www.jgi.doe.gov/Nematostella) and from the GenBank under accession numbers XXXXXXXXXXXX

Gene Content

Human Genes Sharing Ancestry with Nematostella Genes

To determine the number of genes in the *Nematostella* genome, we estimated how many of the 27,273 predicted gene models represent unique genes in the genome, as opposed to spurious gene predictions, fragmentary gene models, pseudogenes or unrecognized transposable element sequence. First, the *Nematostella* gene models were divided into categories based on the quality of their hits to the human proteome. Specifically we define the "best C-value", for each *Nematostella* gene, to be the ratio of the BLAST score of its best hit to the human genome to the highest BLAST score of the best-hitting human gene to any *Nematostella* gene. The number of genes with best C-value greater than or equal to Cmin, for Cmin from 0 to 1, is plotted in Figure S3.1 for two choices of BLAST e-value threshold. This value is by construction equal to 1 for genes with a mutual best, and the human and nematostella curves converge at Cmin=1 for each choise of e-value. At the opposite extreme of Cmin=0, the curves reach the total number of genes with detectable alignment in the other genome.

If a species has undergone extensive "paralog-formation', for example by a genome duplication relative to the other, we will expect the curve for genes of the 'duplicated' species hitting genes of the 'unduplicated' species being above the vice versa, for ranges 0.8 <= Cmin <1, i.e. the 'co-orthologs' range, as we observe for human in the plot.

If the curve for a species does not flatten as Cmin -> 0 this means that there are many genes in that species having low best C-values, which is what we expect for pseudogenes and/or transposons where partial gene predictions have been made. For *Nematostella*, this curve shows a large excess, exceeding the human curve for values of Cmin > 0.5, while falling below human at high Cmin values. This type of reversal does not appear in human-*Drosophila*, human-*Caenorhabditis* , or *Drosophila*-*Caenorhabditis* comparisons (data not shown).

To asses whether the excess of gene models with low best C-value in *Nematostella* reflect the contribution of a large number of small, fragmentary models and pseudogenes, 60 *Nematostella* genes were subjected to a detailed manual review. Twenty genes were selected at random from the JGI *Nematostella* Filtered Models version 1.0 ("FM1.0 set") in each of the following categories:

1) BCV (best C-value to human) = 0, meaning no BLAST hit to human. 5486 of the FM1.0 set have BCV = 0.
2) 0 < BCV < 0.4. 4889 of the FM1.0 set.
3) BCV >= 0.4. 18274 of the FM1.0 set.

Manual review is by definition somewhat subjective, but using conservative criteria, i.e. avoiding dismissing too many genes, the results of the sampling indicate that about one third of all genes in the FM1.0 set could be expected to be rejected by manual reviews.

Category 1), 8 of the 20 were deemed "real genes", i.e. from the total number of genes with BCV = 0 we would expect ~ 0.4 * 5489 = 2194 genes to "pass manual scrutiny". Note that 15 of the 20 in this category have 1 or 2 exons.

Category 2). 10 of 20 were deemed real. 11 of the 20 have 1 or 2 exons. Predicted # genes to pass manual review: 4889 * 0.5 = 2445

Category 3). These are high BCV genes, 13 of which have BCV > 0.8. Here, 15 of the 20 are thought to be real genes. In some cases, it looked like two gene models should be merged, and I tried roughly to call a gene here every other time, to approximately get the right gene count. From the counts here, we would expect ~ 0.75 * 18274 = 13706 genes in this category.

Adding up these expected numbers gives us an estimate of 18,345 bona fide *Nematostella* genes. Even this may be an overestimate, since quite a few of the genes with lower c-values are at the edges of short scaffolds, and their other half may be picked up by another scaffold, causing 2 annotations for a single gene.

Additional observations on the ***Nematostella*** proteome

- The human genome has more genes with a mutual best hit in *Nematostella* than in the proteomes of *Ciona*, fruit flies or nematodes. (Figure S3.2)
- The *Nematostella* genome contains many protines with domain architectures (combinations of PFAM domains) that are shared exclusively with vertebrate genes. (Figure S3.3)
- Of the PFAM domains present in human, mouse, dog, chicken, frog and fugu, Nematostella has

more in common than any of *Ciona*, fruit fly, or nematode. (Figure S3.4)

- There are 5 large clusters of short proteins (around ~100aa), each comprising 55-74 members with weak similarity to hypothetical short ORFs from fungi (*28*)

- There are 242 clusters of tandemly duplicated genes, comprising 2-13 members, with annotated Pfam domains, which apparently were duplicated after split of bilateria

- There are 9 neurotoxins genes, with an anemone neurotoxin domain (PF0076) previously found only in the Cnidaria, but not previously in *Nematostella*, and 5 copies of green fluorescent protein (PF01353), originally found in jellyfish and predominantly found in Cnidaria.

- 16 Pfam domains previously exclusively found only in vertebrates, but not in other phyla of bilateria (or other eukaryotes), are present in Nematostella genome, including:

PF01500 - Keratin, high sulfur B2 protein
PF00040 - Fibronectin type II domain
PF06954 - Resistin
PF06990 - Galactose-3-O-sulfotransferase
PF05038 - Cytochrome b558 alpha-subunit

Lineage Specific Expansions

We identified 809 “recent” tandem expansions in the *Nematostella* genome, comprising 1,854 protein-coding genes. A similar algorithm applied to the ENSEMBL annotation of the human genome detected 504 recent expansions with 1,317 genes. The algorithm is as follows: first, all genes on chromosomes or scaffolds with three or more annotated genes were numbered in occurring order. From an all-against-all Smith-Waterman alignment of these peptides, all hits with greater than 60% identity and with at least 25 conserved four-fold degenerate codons were retained. This filtering step helps eliminate pseudogenes and spurious hits of low-complexity regions, and allows a divergence epoch estimate for the pair based on four-fold degenerate transversion frequency (4DTv)(*29*). Since our focus is on expansions specific to the nematostella lineage, we only consider hits with 4DTv < 0.2, i.e. 20% or less observed transversions at four-fold degenerate 3rd codon positions. Extrapolating from vertebrate calibrations, this corresponds to gene duplications no older than 150-200 million years. For comparison, human-mouse orthologs have typical 4DTv distances of ~ 0.15, and human-opossum have 4DTv ~ 0.26 (data not shown).

Next, the scaffolds were scanned for pairwise hits under the above criteria with no more than three unrelated genes separating them. This allows for intervening spurious gene models as well as small-scale inversions. Finally, all such pairs with one of the genes being within three genes of a member of another pair were clustered in a single-linkage fashion. To assess the probability of detecting tandem expansions by chance, we repeated this approach on versions of the human and nematostella gene sets in which the gene order had been randomly scrambled. We found a single spurious 2-member cluster in nematostella and four in human. Hence, we expect the false positive rate of this approach to be less than 1%.

In order to assess to what extent these relatively recent expansions have been retained by positive selection, and to compare the types of expansions found in *Nematostella* to those in vertebrates, we performed the following analysis: first, we scanned all of the genes in the human and *Nematostella* gene sets for PFAM-A domains using hmmpfam(*30*). We were able to assign one or more PFAM domains to 15,102 human genes and 12,202 *Nematostella* genes. We then formulated a neutral-evolution hypothesis that any gene has an equal probability of getting duplicated and fixed in the population. For genes with a certain domain we can then test the validity of this hypothesis by comparing the frequencies of such genes in the recent expansions to the overall frequency. For example, the number of recently created genes in *Nematostella* containing a PF000001 seven transmembrane family (rhodopsin family) domain is 33 (subtracting one “seed” member of each tandem cluster). Since 779 of the 12,202 *Nematostella* genes contain this domain, the expected number in the recently expanded set (with a total of 572 genes with PFAM domains) under the neutral hypothesis is 36.5 +- 5.8, where the binomial approximation has been used since the recent genes constitutes a small fraction of the total

genes in both species. Hence, in *Nematostella*, there is no evidence for recent selection for retention of new genes greated by tandem duplication with PF000001. In the human genome, on the other hand, 112 such genes are observed, with an expected value of 29 +- 5.3, consistent with a strong recent selective retention of such receptors (olfactory and visual) within vertebrates or mammals. Tables S3.3 (*Nematostella*) and S3.4 (human) show all PFAM domains found in at least four genes in recent tandem expansions, and with a frequency of at least 3 sigma above the expected frequency under the neutral hypothesis. In general, the gene families showing strong expansions along the two lineages are different. In addition to olfactory and taste receptors, the human genome shows strong recent preference of C2H2 zinc finger genes with a KRAB domain, keratin, and immune defence proteins. This newly acquired repertoire almost certainly plays a key role in defining vertebrates and mammals. Similarily, the genes listed in Table S3.3 can be hypothesized to play a significant role in distinguishing *Nematostella*. Note that this analysis is biased towards vertebrates, for which more domains have been characterized.

Supplement S4

Construction and characterization of eumetazoan gene families

To understand gene creation and duplication we designed a phylogenetically informed clustering algorithm which produces clusters at the base (most distant in time) and tip (most recent point) of a given internal branch (stem) of the species tree. Each cluster is composed of a group of modern genes that are the offspring of one gene in the common ancestor. Our algorithm takes as input:
a) The genomes that have arisen as descendants from our stem of interest. These are our in-group genomes.
b) Other genomes which serve as phylogenetic out-groups.
c) Pairwise alignment scores for all pairs of genes in the in- and out-groups.
d) Any previous clusterings made of the in-group genomes we want to preserve.

From this data our algorithm operates as follows:
i) A graph is made where each node is an in-group gene. Edges are added if two genes are mutual best hits between species. Edges are also added if two genes are in any clusters in input (d).
ii) A single linkage clustering is done of the graph. This represents the clusters at the tip of our stem. The mutual best hits captures the likely orthologs between the organisms while the clusters passed in as input (d) captures the paralogs from the stems emanating from the tip of the current stem of interest.
iii) For each cluster made in (ii), the top *m* hits to the out-groups are found where *m* = twice the number of out-groups. This collection of out-group genes is called the potential blockers for this cluster.
iv) Two clusters from (ii) are merged if they share at least one potential blocker and for every potential blocker the genes with which it aligns are closer [by BLAST score] to each other than either is to its potential blocker. This gives us a set of clusters that existed at the base of our stem of interest.

Blastp was run using BLOSUM45, evalue cutoff 0.001, and filtering was turned off. Only the top 1500 hits were considered if more hits passed these criteria. The genomes used are as follows:

Xenopus tropicalis JGI v4.1

Takifugu rubripes JGI v4.0
Nematostella vectensis JGI V1.0 (this work)
Homo sapiens Ensembl build 38
Drosophila melanogaster Ensembl build 38
Caenorhabditis elegans Ensembl build 38
Arabidopsis thaliana From NCBI on 11/2005
Saccharomyces cerevisiae From genome-ftp.stanford.edu, version released on July 7, 2004
Dictyostelium discoideum From dictybase.org, Annotations released on 7/11/2005

Supplement S5

Phylogenetic analysis of metazoa

We compared predicted protein sequences from Nematostella to those from other metazoan and out-group genomes, and find that *Nematostella* genes are more similar to vertebrate genes than to fly and nematode genes using bayesian branch length estimation and an analysis of percent sequence identity. ((*31*) came to the same conclusion using ESTs and BLAST e-value to measure similarity.) Of the 7,766 ancestral metaozan gene clusters, 1,619 are composed of a single gene from each of the six representative metazoan genomes listed in Supplement S4: human, fish, frog, *Nematostella*, fruit fly and nematode. Starting with this set of apparently single-copy genes in these six genomes, we searched six additional complete or partial genome sequence data sets (of a tunicate, a gastropod mollusk, a hydrozoan cnidarian, a choanoflagellate, a sponge, and yeast), and a collection of ESTs from the ctenophore *Mnemiopsis leidyi* (see Table S5.1 for a list of data sources) for orthologous genes, making a total of twelve whole genome data sets, plus the EST-derived sequences from *Mnemiopsis*. For each additional genome, if a mutual-best hit existed to the human gene in the cluster, that gene was identified as an ortholog, and added to the cluster. We compared the results obtained with this set with those obtained using *Nematostella* rather than human as the anchor for identifying orthologs, and found that it did not change the results. By this method, 337 ortholog sets were identified that had one gene representing each of the twelve whole genome datasets. Only nine ortholog sets contained one gene from each of the twelve whole genomes plus a *Mnemiopsis* sequence.

We constructed two concatenated multiple sequence alignments from the identified orthologs: one with and one without the ctenophore sequence. In each case, multiple sequence alignments for each orthologous set were computed with MUSCLE(*32*), and well-aligned regions extracted with GBLOCKS(*33*) using conservative settings (all available sequences in an orthologous group were required to be well aligned at the start and the end of each extracted block: -b1=N -b2=N, where N is equal to the number of sequences in the alignment.). We constructed two concatenated multiple alignments for investigating metazoan phylogeny and relative rates of protein sequence evolution among the different lineages. The first (Alignment 1) excludes sequence from the *Mnemiopsis* ESTs, and includes only the 337 ortholog sets with representation from each of the other twelve genomes. The second (Alignment 2) was compiled from the multiple alignments including the *Mnemiopsis* data and includes all ortholog sets with twelve or thirteen members, plus all ortholog sets including a *Mnemiopsis* sequence.

Alignment 1 consists of 19,563 columns, with no missing data. This data matrix was analyzed using *mrbayes* version 3.1.2(*34, 35*), using a the WAG(*36*) model of protein evolution, a Gamma distribution of rate variation among sites, approximated by four rate categories, and a category for invariant sites. Multiple runs from different starting topologies all converged on the same topology, branch lengths and posterior probabilities for protein evolution model parameters within approximately 10,000 monte carlo iterations. The mean and variance of the posterior probabilities for total tree length, Gamma distribution shape parameter alpha and the fraction of invariable sites were 2.278 +- 0.001, 0.818 +-

0.001, and 0.2291 +- 0.0001, respectively. Figure S5.1 shows the consensus tree topology and branch lengths. All nodes were resolved as shown in 100% of the samples trees. The sequences of the genes used in Alignment 1 are available in FASTA format in S5.fasta.

Alignment 2 consists of 19,977 columns, however only 2272 columns contain *Mnemiopsis* sequence. To test whether this data could be used to shed light additional light on the phylogenetic relationships among cnidarians, ctenophores and bilateria, we submitted this dataset to a maximum likelihood analysis using the PHYLIP package's PROML program(*37*), and compared the likelihood scores of three topologies: ctenophores sister to cnidarians+bilaterians, ctenophores sister to bilaterians, and ctenophores sister to cnidarians. Of these, the first had the highest likelihood score, but it was not significantly better than the second in a Shimodaira-Hasegawa test. The branch lengths for the tree shown in Figure 2 were estimated using PROML, for the defined topology illustrated, with a trifurcation at the cnidarian/ctenophore/bilaterian divergence.

To make an extremely rough estimate of divergence time between bilaterians and cnidarians, we interpolated following Dawkins (*38*) between recent molecular clock estimates(*39*) of the timing of the protostome-deuterostome (95% confidence interval: 640-760 Mya) and choanoflagellate-metazoan (95% CI: 760-960 Mya) divergences. We see from Figure 2b that the cnidarian-bilaterian split lies ~30% of the way between these two nodes (adopting the midpoint rooting as shown), suggesting that the eumetazoan ancestor lived between 670 and 820 Mya.

Figure S5.2 shows a more direct way the greater similarity between human and Nematostella proteins than between human and fly/nematode proteins.

Supplement S6

Intron Splice Site Conservation

To study intron loss and gain in orthologous genes in multiple species, we first aligned the *Nematostella* gene set to the set of human ENSEMBL models (release 26.35.1) and to the TIGR release 5 of *Arabidopsis thaliana* genes. In 2,347 cases, a human gene was found to have a mutual best hit to both a *Nematostella* and an *Arabidopsis* gene, forming a tentative cluster of orthologous genes to be studied further.

Gene models are often incomplete in the 5' ends and may have have poorly determined splice sites, so we restrict our analysis to regions of highly conserved peptides in the orthologs of all three species. The independent identification of such regions in multiple species provides strong evidence for the accuracy of the gene models in these regions. Hence, we performed multiple alignments of the orthologous clusters and identified gap-free blocks flanked by fully conserved amino acids. We then identified annotated splice sites of all species within these regions, which the additional requirements that 1) none of the peptides must have a gap in the alignment closer than 3 AA from the splice site and 2) no two different peptides must have splice sites at different positions closer than 4 AA. Empirically, these requirements are necessary to avoid spurious detection of "intron losses" due to ambiguities in either the multiple alignment or the gene model's splice sites. While some of these cases may reflect real sliding of donor or acceptor sites, we restrict ourselves to studying gains and losses of introns here. Finally, we required that at least 5 amino acids out of 10 in the flanking regions of the splice sites be either fully conserved or have strong functional similarity among all four species.

9,947 highly reliable intron splice sites were identified by these requirements. The results are summarized as a Venn diagram in figure S6.1, indicating the number of shared introns between the species.

Remarkably, about 81% of the human introns (4,403 of 5,435) are shared with nematostella. Assuming that intron losses have occurred independently in the human and nematostella lineages, and that the probability of independent intron insertion events at the same location is negligible we estimate the loss in *Nematostella* since the last common ancestor (LCA) with human as 158 / (158 + 1258) = 11%. In a similar fashion, we estimate a loss of almost 22% along the human lineage, twice the amount of introns

lost in the *Nematostella* lineage.

The above results also allow us to place upper limits on intron gains within the human and *Nematostella* lineages: 28.6% of all introns shared by human and *Nematostella* (and hence present in their LCA) are also shared by *Arabidopsis*. If additional introns have been independently gained in each lineage we expect a lower fraction of the total introns in each species to be shared with arabidopsis. In fact, we find 26.5% of all *Nematostella* introns and 26.1% of all human introns are shared with *Arabidopsis*, which translate into maximum intron gains of ~9% in human and ~7% in *Nematostella*. These results are strict upper limits, since the lower conservation with *Arabidopsis* can also be explained if the loss rate vary inherently between introns. In this case we will expect introns that are shared between human and *Nematostella* to be less prone to loss, and hence a larger fraction will also have survived in *Arabidopsis*. This scenario is very conceivable since some introns have been shown to contain regulatory elements and the loss of such introns would presumably be selected against.

To the extent that the introns in highly conserved peptide regions studied here are representative of introns in general, the above analysis suggests that the *Nematostella* genome has only lost 11% of its introns since the LCA with human, and gained at most 7%.

We next identified 2,347 clusters of orthologous genes in all bilaterian orthologous clusters with an unambiguous 1:1:1 member relationship in human, *Drosophila melanogaster* (fly), and *C. elegans*. In 1,523 of these clusters, the human gene had a mutual best hit to a *Nematostella* gene, forming clusters of four orthologous genes. 4,951 highly reliable introns were identified by these requirements. The results are summarized in Table S6.1. *Nematostella* has the most introns at these conserved positions, followed by human with a relative intron frequency of about 0.91, whereas nematode and in particular fly have considerably fewer introns (0.37 and 0.21). From these numbers we estimate the intron losses in fly, nematode, and human since their LCA to be 82% , 77%, and 12% respectively. Note that the nematode, although having retained only ~23% of the introns since the LCA with human have ~37% of the number of human introns. This suggests a considerable gain of introns in the nematodes, as also reported by [Logsdon 2004].

This analysis of aligning conserved sequences to identify conservation of introns was further extended to include seven species - *Nematostella vectensis, Homo sapiens, Ciona intestinalis, Drosophila melanogaster, Caenorhabditis elegans, Crytococcus neoformans and Arabodopsis thaliana.* 4342 introns from the seven genomes at 2645 aligned positions which contain an intron in at least one of the seven orthologs.

Methods for Intron Gain/Loss tree

Starting from the binary character matrix compiled as described above of 2,645 intron positions across 7 taxa, we found the most parsimonious solution to the intron gain/loss problem by projecting these characters onto the (known) topology. Weighted parsimony as implemented in PAUP 4.0b10(*40*) was used, with the cost of an intron gain significantly greater (more than 10X) the cost of an intron loss. The parsimony assignment of characters to internal nodes is independent of this gain/loss weight ratio. From the branch lengths produced by PAUP, and the known weights, we solved for the number of losses and gains along each branch as show in the main text figure.

Supplement S7

Local conservation of gene order

To search the human and *Nematostella* genomes for regions of conserved linkage, we performed the following analysis. First, the genes on each genome were assigned unique identifers according to the

order in which they occur on the chromosomes or scaffolds. We then used the sequence alignments described in the clustering section to scan each genome for tandem expanded gene families, defined here as clusters of genes with a maximum of 4 intervening genes, showing similarity at e-values < 1x10-10. All but one member, the longest peptide, were excluded from further analysis at each such region in the genomes.

From the human vs nematostella protein alignments we next excluded all genes with more than 15 hits with e-value < 1x10-10 from consideration. Finally, of the remaining pair-wise hits we included only hits with a score of more than 70% of the value of the highest score of either of the two genes to any of the genes in the opposite genome. This approach enriches the set for orthologous gene pairs while removing weak super-family similarities from the analysis. At this stage we were left with 11,351 pair-wise hits, involving 6,986 nematostella genes and 8,426 human genes. We then recalculated the gene order IDs in the two genomes, featuring only the genes involved in these high-quality alignments, and scanned for regions of conserved synteny or linkage in the following manner:

For the first pair-wise alignment of genes in the proteomes of the two species, the gene locations on the chromosomes were recorded and a one-pair segment of conserved synteny was defined. Subsequent gene pairs either defines new segments, or, if the genes in both species are located within a specified maximum distance, Nmax from a gene pair in an existing segment, the pair is added to that segment. If a pair can be added to two segments, these segments are joined into a larger segment of conserved synteny. Note that this method does not require strict conservation of gene order: inversions on scales smaller than Nmax are tolerated.After traversing all alignments, we have a set of conserved regions, on which we can impose a minimum member limit (typical 3 pairs) to remove potentially spurious regions.

For human-*Nematostella*, we found no strict significant conservation of gene order, but by choosing a large value of Nmax we nonetheless detect regions of conserved linkage in which the local gene order has been scrambled. In order to detect the significance of these regions, we randomly scrambled the order of the genes on each chromosome or scaffold and applied, for the same sequence alignment data, the algorithm to the scrambled data set. This allows us to choose parameters to minimize false positive detection. Note the importance of the filtering out weak hits in this method, as the presence of such hits would significantly increase the false positive rate in the detection of segments of conserved linkage. Using Nmax = 40 and considering only segments of 9 or more participating genes, we find 33 such segments of conserved synteny between human and *Nematostella*, with none expected by chance, as seen by running the algorithm on the scrambled set.

Identification of human genome segments free of recent chromosomal fusions and large-scale rearrangements

To facilitate the search for large-scale conservation of gene linkage in the presence of extensive changes in local gene order between humans and *Nematostella*, we identified 98 segments of the human genome which appear to be uninterrupted by inter-chromosomal translocations or fusions when compared to the genomes of other chordates. To identify likely locations of chromosomal fusions along the human genome which separate such segments, we followed the following procedure:

1. Putatively orthologous gene pairs were identified between the ENSEMBL human gene set and the chordate *Branchiostoma floridae* draft gene set [JGI web page] using the mutual best BLAST hit criterion.
2. Scaffolds of the *B. floridae* assembly were clustered as described below for Nematostella, based on the similarity of the distribution in the human genome of human genes orthologous the genes on the scaffold.
3. A representation of each human chromosome arm was constructed in which each gene along the chromosome was represented by the identifying number of the cluster of scaffolds in which its *B. floridae* ortholog resides.
4. A Hidden Markov Model, constructed and implemented in software for the purpose, was used to segment the human chromosomes into segments with an approximately uniform distribution of hits to a specific subset of the scaffold clusters.

Figure S7.2 illustrates the results of this procedure for human chromosome arms 14q, 15q, 16p and 16q, and Table S7.1 lists the extent of the 98 identified segments in base pair coordinates on the NCBI Human genome build 36.

Construction and Significance Testing of Putative Ancestral Linkage groups (PALs)

To test for conservation of large-scale synteny in the presence of extensive local rearrangement of gene order, we compared 147 of the largest scaffolds of the Nematostella assembly to the segments of the 98 human genome described above. The examined scaffolds were selected because, like the 98 human segments, each contains descendants of 40 or more ancestral eumetazoan genes. For each scaffold-segment pair, we tabulated the number of ancestral gene clusters giving rise to descendants on both members of the pair. This number counts the number of independent orthologs shared by the scaffold and the segment. For each scaffold-segment pair, the number of observed orthologs was compared to a null model in which scaffolds and segments comprise genes descending from genes drawn independently from the set of 7,766 ancestral genes. This method of counting orthologs, and this null model control naturally for independent tandem gene duplicates which could otherwise artifactually inflate the number of observed orthologs in circumstances where there is no remnant of conserved synteny, because tandem duplicates arising independently should be contained in a single reconstructed ancestral gene cluster. The expected number of orthologs under this model is governed by the hypergeometric distribution, allowing us to compute a p-value for consistency for each scaffold-segment comparison with the null model. Since we compared 147 scaffolds with 98 segments, we applied a Bonferroni correction factor of 1/14406. The complete set of these numbers of shared orthologous genes are shown in figure S7.3, for all scaffolds (67/147) and segments (40/98) which participated in a statistically significant shared synteny relationship. Table cell backgrounds are colored yellow when $p < 0.01/14406$, and pink when $p < 0.05 / 14406$. A blue background indicates $p<0.5/14406$.

Table S7.3 has 112 yellow cells, corresponding to 112 cases of statistically significant conservation of synteny between a *Nematostella* scaffold and a segment of the human genome. The rows and columns of this table have been ordered to reveal 13 sets of scaffolds and chromosome segments, defined by the criterion that none can be subdivided without separating into different sets a scaffold-segment pair with significant evidence ($p<0.01$) for conserved synteny. We interpret these collections of modern sequences to be descended from the same chromosomes, or chromosomal segments of the common ancestor of eumetazoa, and refer to them therefore as putative ancestral linkage groups, or PALs.

Table S7.X lists the 255 ancestral gene clusters linked with the HOX clusters in PAL-A.

A clustering method allows more extensive reconstruction of putative ancestral linkage groups.

Having demonstrated that there is extensive conservation of linkage relationships among genes using the conservative statistical criteria described above, we developed a more sensitive method to reconstruct ancestral linkage groups based on clustering scaffolds or chromosome segments. In this method, a matrix of ortholog counts similar to that shown in figure S7.3 is constructed. The rows and columns of this table are then clustered hierarchically, using Pearson correlation as a measure of similarity and the average pairwise linkage method with the "cluster" program(*41*). Figure S7.4 shows the result as a "dot plot" as in figure S7.2. Horizontal and vertical lines divide clusters of scaffolds (vertical lines) and human chromosome segments (horizontal lines), defined by a cut of the hierarchical tree at a correlation coefficient of 0.2. This clustering of scaffolds and chromosome segments defines 15 large PALs, each with descendants of more than one hundred ancestral eumetazoan genes. 3055 ancestral genes, or 40% of the ancestral genes are assigned to one of these PALs.

Supplement S8: Eumetazoan Ancestry of Genes

Construction of "Centroid" sequences.

We define the "centroid" of a cluster of orthologous amino acid sequences to be a synthetic amino acid sequence which maximizes the sum of BLAST alignment scores between the centroid and the members of the cluster. This provides a surrogate for the peptide sequence that is ancestral to each cluster.

Classification of eumetazoan genes by ancestry

Centroids (see above) of the ancestral eumetazoan gene clusters were aligned to non-animal entries in SwissProt/TREMBL[Uniprot release 8 from http://www.uniprot.org] with BLAST(*11*), using the NCBI database to remove metazoan entries. The Pfam(*30*) annotation of SwissProt/TREMBL from swisspfam [Version of Sept. 6 2006. Current version available from http://pfam.janelia.org] was parsed to identify Pfam domains found only in animals, as well as pairs of Pfam domains that occur separately in non-aminals but only were found together in animals.

Clusters whose centroid had a BLAST hit to out-group proteins of e-value <1e-6, and also clusters containing a member which is a mutual best hit to an *Arabidopsis*, *Dictyostelium* or *Saccharomyces* were annotated as "ancient," unless one of the following conditions was met:
1) if both the Nematostella peptide and at least one other animal protein had an "animal specific" Pfam domain, the cluster was designated a type II novelty.
2) if both the Nematostella peptide and at least one other animal protein had an "animal specific" Pfam domain combination, the cluster was designated a type III novelty.

Note that type III (animal-specific eukaryotic domain combinations) are based only on pairwise combinations. Thus animal proteins that shuffle the order of domains found within an ancient eukaryotic family are not designated as novel in this analysis.

Functional annotation of ancestral gene clusters

Panther(*42, 43*) family annotations on the sequences of extant species were transferred to the inferred ancestral clusters when both *Nematostella* and bilaterian members of the clusters shared the same Panther annotation. These annotations were mapped to various overlapping functional categories using the Panther Pathways(*43*) and Panther Ontology databases.

To asses whether specific functional categories were over- or underrepresented among the different types of novelties, we adapted the GOstat approach of Beissbarth and Speed (*44*) for use with the Panther ontologies, and computed p-values for enrichment and dearth relative the hypergeometric distribution. For both Panther Pathways and Pather Ontology, we limited our tests to the 100 ontology terms which had the greatest number of inferred ancestral genes assigned to them, and applied a Bonferroni correction for 100 tests, even though this is somewhat conservative, since the categories have significant overlap. Table S8.1 lists the functional categories enriched for novel genes of the three types.

Captions for Supplemental Tables and Figures

Table S1.1 Partial list of the merits of Nematostella as a model organism.

Table S2.1 Summary of WGS libraries

Shotgun libraries are identified by their four-letter name, which is used as a prefix to the identifier of all reads from the library. For each library, the table lists: the mean size of genomic DNA inserts in base pairs; the number of sequencing reads attempted for each library; the number of reads with at least 100 bp of high-quality sequence after removal of vector and low-quality sequence, as described previously[Dehal 2002]; the number of reads which have a detected alignment to other reads in the shotgun data set (see discussion above); the number of reads which are placed in the contigs of the assembly; and the mean read length, after trimming. Column totals are shown in bold for selected columns, and the fraction of reads lost to trimming, lack of alignment, and lack of placement in the assembly is shown as a percentage of the previous total.

Figure S2.1: Observed density of polymorphic sites

The rate of single nucleotide polymorphism observed in the assembled genome sequence is 0.8%. Figure S2.2 shows the observed (orange) and Poisson ascertainment bias-corrected (green) frequency of polymorphic positions as a function of local depth of assembly for a sampling of 14.4 million positions in the assembly [Left hand scale]. Positions are considered polymorphic if two or more WGS reads indicate each of two or more different bases at a given position. The red curve shows the number of positions considered for each depth of coverage, and the dotted curve shows poisson distributed counts with the same mean.

Figure S2.2: Four haplotype polymorphism fit

The number of polymorphic sites (red crosses) as a function of local depth of the assembly is compared with expected values for four independent haplotypes with average pairwise differences of 0.5% (green), 0.64% (blue) and 0.7% (purple).

Table S2.2: Summary of tandem repeat elements from raw WGS reads.

Paired fosmid end reads were screened for highly abundant 16-mer DNA words appearing in both ends of fosmid clones, indicating their presence in the genome in large tandem arrays. Identified 16-mers were assembled with JUGGERNAUT, and their abundance in the whole genome shotgun reads was estimated by alignment to a sample of WGS reads from all libraries using BLAST(*11*).

Table S2.3. Transposable elements in the sea anemone genome.

Figure. S2.3 Neighbor-joining tree of eukaryotic non-LTR retrotransposons constructed for their reverse transcriptase. Black circles mark novel families of non-LTR retrotransposons identified in this study. Unmarked retrotransposons have been described previously and are collected in Repbase Reports. Abbreviations of host species are as follows: NV, *Nematostella vectensis*; XT, frog *Xenopus*

tropicalis; BF, lancelet *Branchiostoma floridae*; AG, mosquito *Anopheles gambiae*; DM, fruit fly *Drosophila melanogaster*; DR, fish *Danio rerio*; CR, green algae *Chlamydomonas reinhardi*; TP, diatom *Thalassiosira pseudonana*; SP, sea urchin *Strongylocentrotus purpuratus*; PS, turtle *Platemys spixii*; SJ, blood fluke *Schistosoma japonica*; Cis, sea squirt *Ciona savignyi*. Only >40% bootstrap values are shown next to corresponding nodes of the tree (based on MEGA3(*45*)). Clades and groups of non-LTR retrotransposons are indicated by black and blue rectangles.

Figure S2.4 Number of chromosomes
The number of chromosomes was determined by analysing over 90 metaphase plates in spreads. The conclusion is that 2N = 30, the same number as in Hydra. A sample metaphase plate is shown, with the histogram of the number of observed chromosomes per plate.

Table S3.1: Summary of gene model statistics For Nematostella Filtered Models 1.0

Figure S3.1: Distribution of C-score

The number of genes with a best C-value (see section S3) greater than Cmin, or Cmin from zero to one, with alignment e-value threshold *Nematostella* (red) and human (blue), with BLAST e-value threshold 1e-10 (solid curves) and 1e-3 (dashed).

Table S3.2: Compared ambundances of PFAM domains for selected domains.

The number of proteins with PFAM(*30*) hits to 10 abundant PFAM domains, along with the abundance rank of that PFAM domain in each genome, is compared among five metazoan genomes, including *Nematostella*.

Figure S3.2: Number of bidirectional BlastP hits (potential 'orthologs') between 22,218 human genes (from Ensembl) and other organisms with known genomes. Despite early divergence, sea anemone shares more hits with human, than other bilaterians, except vertebrates.

Figure S3.3: Fraction of unique multi (Pfam) domain (2 or more domains) gene models from Nematostella (total 983) shared by other metazoans and yeast.

Figure S3.4: 2264 Pfam domains present in all 6 vertebrates with known genomes:human, mouse, dog, chicken, frog and fugu. Below is the histogram of numbers of these domains shared by ciona, fly, nematode and sea anemone.

Table S3.3: Preferentially retained PFAM domains within recent tandem expansions in ***Nematostella***

Tandem gene expansions were identified based on 4DTv as described in the text. PFAM domains with a significantly greater number of observed examples among tandem expansions in the *Nematostella* genome relative to the predication of a model model of the neutral expectation are shown.

Table S3.4: Preferentially retained PFAM domains within recent tandem expansions in ***Homo sapien***

Tandem gene expansions were identified based on 4DTv as described in the text. PFAM domains with a

significantly greater number of observed examples among tandem expansions in the human genome relative to the predication of a model model of the neutral expectation are shown.

Table S5.1: Data sources for phylogenetic analysis

Figure S5.1: Distribution of percent ID Against Human Proteins

The distribution of the percent identity in mutual-best-hit protein alignments between human genes and the genes of the frog, Xenopus tropicalis, pufferfish Takifugu rubripes, Nematostella, fruit fly Drosophila melanogaster, and nematode Caenorhabditis elegans.

Table S6.1: Distribution of 4,951 introns in conserved regions of orthologs in human, fly, nematode, and *Nematostella*. Numbers in parenthesis refer to the number of introns not shared by any other species.

Figure S6.1: Venn diagram for three-way intron conservation comparison
Venn diagram showing the distribution of 9,947 intron splice sites in *Homo sapiens*, *Nematostella vectensis*, and *Arabidopsis thaliana*.

Table S6.1: Four-way intron conservation comparison

The distribution of 4,951 introns in highly conserved, orthologous peptide sequences from human, *Drosophila melanogaster*, and *C. elegans*, and *Nematostella*. The first four lines list the total number of introns in each species, followed in parentheses by the number which are unique to that species. The remaining table rows list the number of introns shared by selected combinations of genomes.

Figure S7.1: Synteny block search
The size distribution of synteny blocks for human vs. *Nematostella* (blue bars) is compared to that for a synthetic data set in which gene positions have been artificially randomized (maroon bars), where synteny blocks are defined as maximal collections of ortholog pairs where pairs of adjacent orthologous pairs have no more than 40 non-participating genes intervening between them.

Figure S7.2: HMM segmentation example
Each graph plots the rank order of human genes along four human chromosome arms (horizontal coordinate) versus the rank position of the *B. floridae* mutual-best-hit ortholog within five clusters of *B. floridae* scaffolds. Vertical red lines indicate the boundaries between human chromosome arms, and horizontal red lines indicate boundaries between scaffold clusters. Discontinuities in the distribution of orthologous gene positions within chromosome arms identified by a hidden markov model are indicated by the addition of vertical black lines on the right. These discontinuities are most easily explained by chromosomal fusions or large-scale re-arrangements in the human lineage which are recent compared to the time scale of gene order evolution.

Table S7.1: Table of human chromosome segments used in large-scale synteny search
A list of the human genome segments used in that PAL analysis. For each segment, the segment name, the human chromosome, and the start and end points on the chromosome, in base pair coordinates on the NCBI Human genome build 36.

Table S7.2: Complete Oxford Grid for Human-Nematostella comparison
"Oxford grid" which tabulates the number of ancestral gene clusters shared between the 22 *Nematostella* scaffolds (columns) and 14 segments of the human genome (rows) that are assigned to PALs A, B and C. Cell colors indicate Bonferroni-corrected p-value < 0.01 (yellow), < 0.05 (pink), < 0.5

(blue).

Figure S7.3: Clustering method for constructing putative ancestral linkage groups (PALs) Blue dots mark the position in human chromosome segments (vertical coordinate) and the *Nematostella* scaffolds (horizontal coordinate) of a pair of orthologous genes. Nematostella scaffolds and human chromosome segments have been ordered by a hierarchical clustering procedure, and concatenated together. Gene positions are in rank order rather than base pair coordinate, where only genes descended from the set of 7,766 ancestral gene clusters have been numbered. Descendants of ancestral eumetazoan clusters with more than 25 genes from the six representative animal genomes were excluded from the analysis. Horizontal and vertical lines divide clusters of human chromosome segments and *Nematostella* scaffolds defined by having an average pairwise correlation coefficient of their distribution of hits to the other genome greater than 0.2. The trees along the left and top of the plot are graphical representations of the average pairwise correlation scores among the hierarchically clustered human segments (left) and *Nematostella* scaffolds (top). Terminal branches are centered

Figure S7.4: Detail of Human chromosome 12 showing genes contributing to PAL A. Detail of main text figure 4c, showing the region flanking the HOX C gene cluster on human Chromosome 12. Horizontal tick marks indicate positions of human genes descended from the set of 7,766 inferred ancestral genes. Genes with an ortholog in *Nematostella* on scaffolds 26, 61, 53, 46, 3 and 5 are labeled and connected by a colored line to the position of the *Nematostella* ortholog (See Fig 4c), except where the gene falls into an ancestral metazoan cluster for more than 25 genes from human, frog, fish, fly, nematode and *Nematostella* (Section S4). These large genes families are more likely to have members showing spurious conserved synteny, since they may have members in many regions of the genome. The genes of the HOX C cluster fall into such a large family, but have been labeled to show the position of the HOX cluster.

Table S7.3: The 225 ancestral gene clusters linked with the HOX clusters in PAL-A:

This table is available for download from http://169.229.10.93/~nputnam/palA.clusters.html

Table S8.1: Table of functional categories enriched for novel genes of the three types.

Panther ontology annotations of the inferred ancestral gene set have been tested for enrichment in each of the three categories of novelty (novel sequence, novel domain, and novel combination of domains), as described in section S8, and significant over- and under-representations have been tabulated here for (A) Panther Ontology Terms for Biological Process and Molecular Function, and (B) Panther Pathways. For each term with a significant over or under representation, the table shows: the ontology term ID from the Panther system; the natural log of the p-value for the enrichment; a "+" or "-" to indicate over- and under-representation, respectively; the number of inferred ancestral genes which both have the annotation in question, and belong to the category of novelty being considered [N(ont & cat)]; the number of inferred ancestral genes which have the annotation in question [N(ont)]; the number of inferred ancestral genes belonging to the category of novelty being considered [N(cat)] ; the total number of inferred ancestral genes [N(total) ; the percentage of novelties of the category being considered which are annotated with the ontology term [N(ont & cat)/N(cat)] ; the percentage of all ancestral genes which are annotated with the ontology term [N(ont) / N(cat)]; and a short description of the ontology term.

References:

1. T. A. Stephenson, *London: The Ray Society* II (1935).
2. R. B. Williams, *Journal of Natural History* 9, 51 (1975).
3. C. Hand, K. Uhlinger, *Biological Bulletin* 182, 169 (1992).
4. J. H. Fritzenwanker, U. Technau, *Dev Genes Evol* 212, 99 (Mar, 2002).
5. S. Szczepanek, M. Cikala, C. N. David, *J Cell Sci* 115, 745 (Feb 15, 2002).
6. J. R. Finnerty, D. Paulson, P. Burton, K. Pang, M. Q. Martindale, *Evol Dev* 5, 331 (Jul-Aug, 2003).
7. E. W. Myers *et al.*, *Science* 287, 2196 (Mar 24, 2000).
8. P. Dehal *et al.*, *Science* 298, 2157 (Dec 13, 2002).
9. S. Aparicio *et al.*, *Science* 297, 1301 (Aug 23, 2002).
10. W. J. Kent, *Genome Res* 12, 656 (Apr, 2002).
11. S. F. Altschul *et al.*, *Nucleic Acids Res* 25, 3389 (Sep 1, 1997).
12. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* 25, 955 (Mar 1, 1997).
13. J. Jurka *et al.*, *Cytogenet Genome Res* 110, 462 (2005).
14. N. L. Craig, *Mobile DNA II* (ASM Press, Washington, D.C., 2002), pp. xviii, 1204 p., [1232] p. of plates.
15. V. V. Kapitonov, J. Jurka, *DNA Cell Biol* 23, 311 (May, 2004).
16. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* 100, 6569 (May 27, 2003).
17. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* 103, 4540 (Mar 21, 2006).
18. K. K. Kojima, H. Fujiwara, *Mol Biol Evol* 21, 207 (Feb, 2004).
19. H. S. Malik, T. H. Eickbush, *Mol Biol Evol* 15, 1123 (Sep, 1998).
20. V. V. Kapitonov, J. Jurka, *Mol Biol Evol* 20, 38 (Jan, 2003).
21. A. Smit, P. Green, (2002).
22. A. A. Salamov, V. V. Solovyev, *Genome Res* 10, 516 (Apr, 2000).
23. E. Birney, R. Durbin, *Genome Res* 10, 547 (Apr, 2000).
24. E. M. Zdobnov, R. Apweiler, *Bioinformatics* 17, 847 (Sep, 2001).
25. M. Ashburner *et al.*, *Nat Genet* 25, 25 (May, 2000).
26. E. V. Koonin *et al.*, *Genome Biol* 5, R7 (2004).
27. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, *Nucleic Acids Res* 32, D277 (Jan 1, 2004).
28. J. P. Kastenmayer *et al.*, *Genome Res* 16, 365 (Mar, 2006).
29. G. A. Tuskan *et al.*, *Science* 313, 1596 (Sep 15, 2006).
30. R. D. Finn *et al.*, *Nucleic Acids Res* 34, D247 (Jan 1, 2006).
31. U. Technau *et al.*, *Trends Genet* 21, 633 (Dec, 2005).
32. R. C. Edgar, *Nucleic Acids Res* 32, 1792 (2004).
33. J. Castresana, *Mol Biol Evol* 17, 540 (Apr, 2000).
34. J. P. Huelsenbeck, F. Ronquist, *Bioinformatics* 17, 754 (Aug, 2001).
35. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* 19, 1572 (Aug 12, 2003).
36. S. Whelan, N. Goldman, *Mol Biol Evol* 18, 691 (May, 2001).
37. J. Felsenstein. (Distributed by the author., 2004).
38. R. Dawkins, *The ancestor's tale : a pilgrimage to the dawn of evolution* (Houghton Mifflin, Boston, 2004), pp. xii, 673 p.
39. E. J. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, H. Philippe, *Proc Natl Acad Sci U S A* 101, 15386 (Oct 26, 2004).
40. D. L. Swofford. (Sinauer Associates, Sinderland, Massachusetts, 2003).
41. M. J. de Hoon, S. Imoto, J. Nolan, S. Miyano, *Bioinformatics* 20, 1453 (Jun 12, 2004).
42. P. D. Thomas *et al.*, *Genome Res* 13, 2129 (Sep, 2003).
43. H. Mi *et al.*, *Nucleic Acids Res* 33, D284 (Jan 1, 2005).
44. T. Beissbarth, T. P. Speed, *Bioinformatics* 20, 1464 (Jun 12, 2004).
45. S. Kumar, K. Tamura, M. Nei, *Brief Bioinform* 5, 150 (Jun, 2004).

December 20, 2006
Supplemental figures and tables.

Table S1.1: Partial list of the merits of Nematostella as a model organism

Developmental Biology

Genomic Approaches

Population Genetics and Ecology

Table S2.1: Summary of WGS libraries

ID	Insert (bp)	N Reads	N Trimmed Reads	N reads with alignments	N placed	Mean trimmed read length
AFII	3149	7658	6867	4839	4035	574
AOWB	2840	1764309	1554340	1026838	880357	630
ATSY	2840	993061	881391	573406	494101	624
AFIK	6489	1864687	1549006	1076195	901598	640
ATWA	6489	915891	834861	592875	500265	709
AFIN	35000	163392	111408	66999	58809	525
ASYG	35000	209087	175771	92574	80041	613
AUNF	35000	50688	40845	35617	31468	656
AXOW	35000	19200	16536	14483	12810	666
AZGY	35000	9216	7056	5664	5001	658
		5997189	**5178081**	**3489490**	**2968485**	
			-14%	-33%	-15%	

Figure S2.2: Four haplotype polymorphism fit

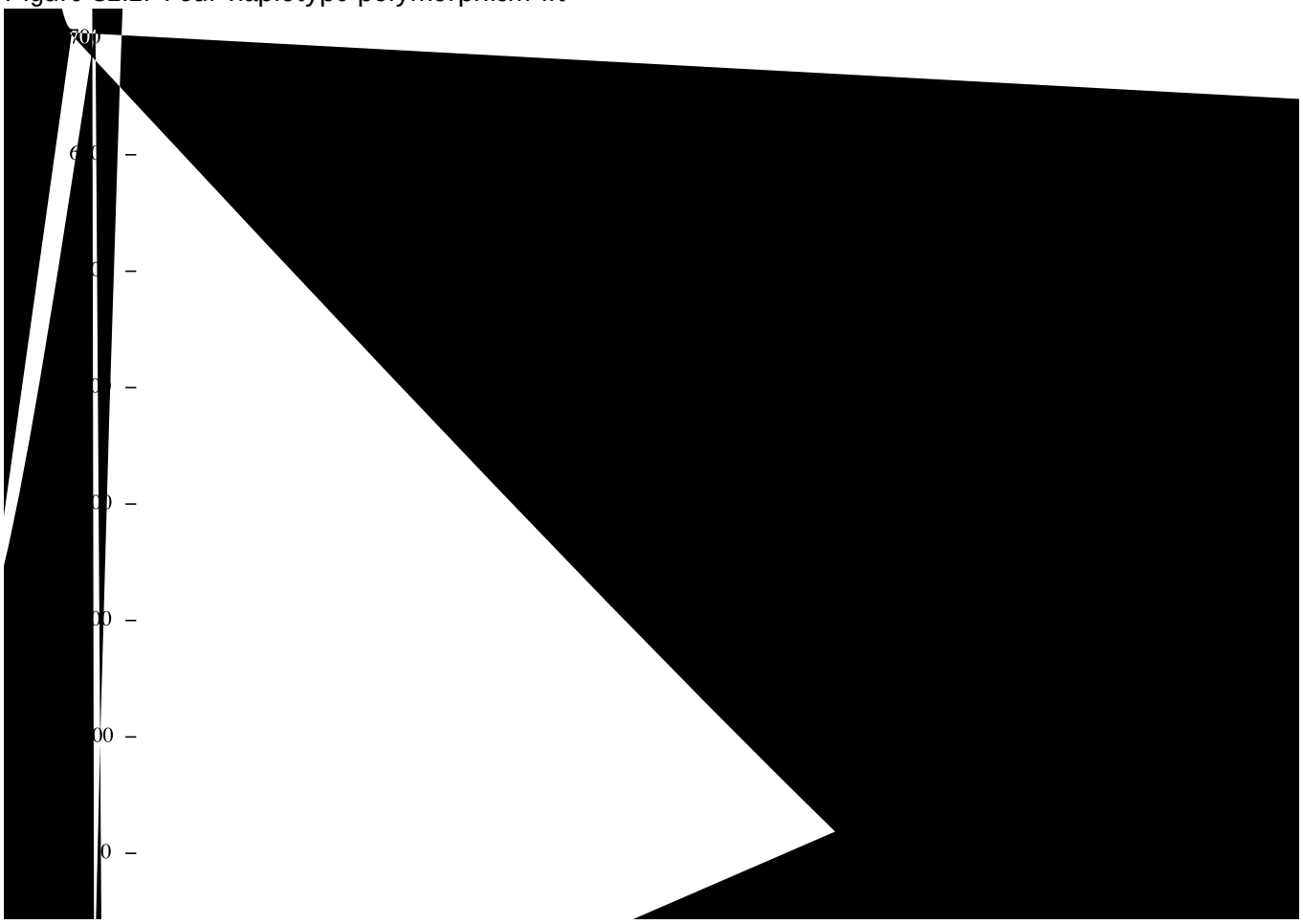


Table S2.2: Summary of tandem repeat elements

Element name	len(bp)	%WGS	Est. Tandem Array size (kb)	Notes
TCTTTGATGTGCTCATjuggernaut	522	10.3%	300	Unclassified cut & paste DNA transposon
AAAAAAAAATCGAACAjuggernaut	7,146	8.8%	2,250	18S, 28S rRNA operon
TTCACGGGTTAATGAAjuggernaut	2,001	7.6%	130	Mariner-3_NVDNA transposon
AAACAAAAGACGCTTTjuggernaut	930	2.3%	360	
GTGTTTGTGGTGTTTTjuggernaut	175	0.8%	2,130	Met-tRNA
GTGATCGGACGAGAACjuggernaut	186	0.8%	1,040	5S rRNA
CCAATCTTAACGTGCAjuggernaut	622	0.6%	350	
CAAAGTCGGCTTCACGjuggernaut	200	0.4%	710	
TTTTTGATCAAAAAAAjuggernaut	770	0.2%	470	U6 snRNA
GTAGACGAAAGATCTCjuggernaut	1,702	0.1%	230	U2 snRNA, 5S rRNA
	Total:	31.9%		

Table S2.3: Transposable elements in the sea anemone genome

Classes of TEs	Percent of the genome %
Total DNA transposons	18.5
"cut and paste":	
Mariner (*Tc1, Pogo* groups)	2.3
hAT	2.1
Kolobok	1.6
PiggyBac	1.0
Harbinger	1.0
P	0.5
MuDR	0.3
En/Spm	0.05
Merlin	0.01
IS4EU	<0.01
Unclassified	5.2
"self-synthesizing" *Polintons*	3.0
"rolling circle" *Helitrons*	1.4
Total retrotransposons	4.6
LTR retrotransposons:	
Gypsy	1.5
BEL	0.2
Copia	0.05
Unclassified	0.2
DIRS	0.4
Non-LTR retrotransposons:	
CR1 (CR1, L2, and REX1 groups)	1.0
RTE (RTE, RTEX)	0.4
L1 (L1, Tx1)	0.1
R2	<0.01
Penelope	0.7
Unclassified TEs	3.1
Total TEs	26.2

Figure. S2.3: Neighbor-joining tree of eukaryotic non-LTR retrotransposons

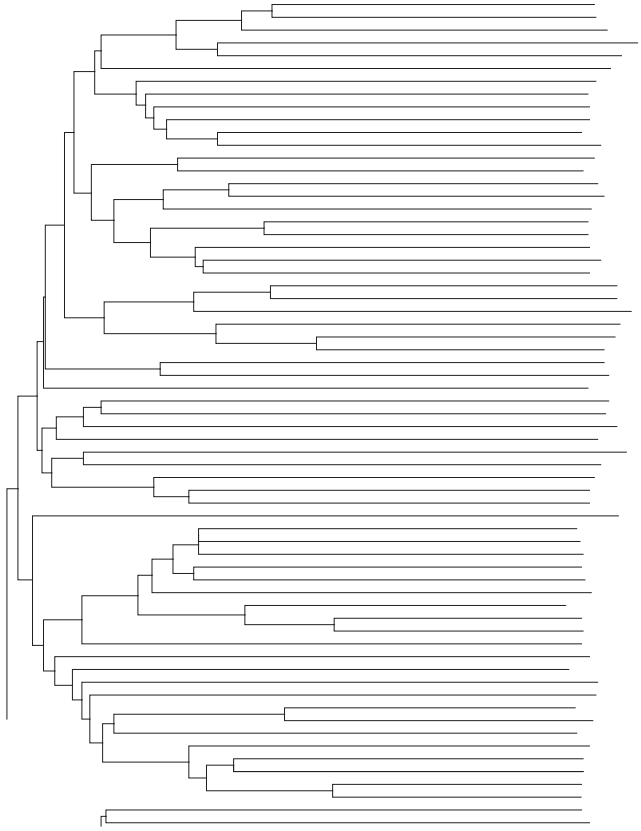


Figure S2.4: Number of chromosomes.

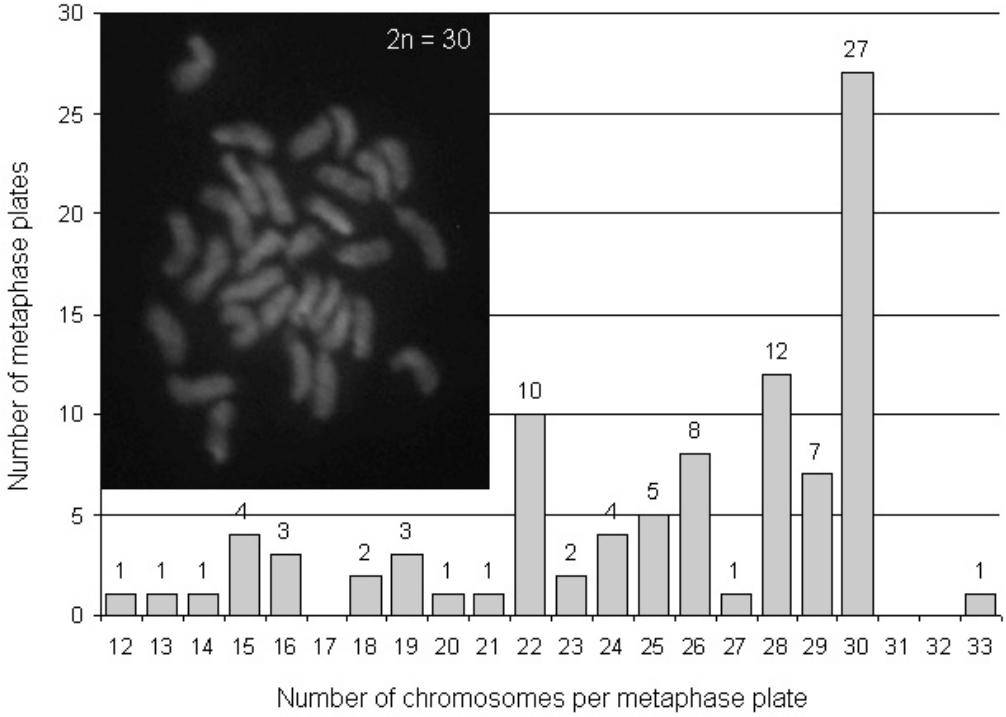


Table S3.1: Summary of gene models

Total number of filtered models	27,273
Models without homology to know proteins from NR	896 (3.3%)
Complete models (ATG and Stop codons)	13,343
Half-complete models	6,975
Incomplete models	6,955
Models exactly predicted by fgeneh and genewise	2,182 (8%)
Models extened to UTRs by ESTs	6,144
Number of single-exon genes (some fraction may be psuedo-genes)	8,460 (31%)
Average number of exons per gene	5.3
Average number of exons per gene (excluding single-exon genes)	7.2
Average transcript length	1,092 bp
Average gene length	

Figure S3.1: Distribution of C-scores

Table S3.2: Compared ambundances of PFAM domains for selected domains

	N – number R - rank	*N. vectensis*		*H. sapiens*		*C. intestinalis*		*D. melanogaster*		*C. elegans*	
		N	**R**	**N**	**R**	**N**	**R**	**N**	**R**	**N**	**R**
PF00001	**7tm_1**	617	1	546	2	59	32	53	27	63	31
PF00008	**EGF domain**	356	2	152	20	162	3	40	39	53	42
PF00069	**protein kinase**	278	3/4	448	3	251	1	201	3	326	2
PF00754	**F5/8 type C**	278	3/4	20	179	14	150	5	418	3	687
PF00400	**WD domain**	262	5	244	7	201	2	156	4	118	11
PF00096	**Zinc finger**	213	6	711	1	160	4	296	1	117	12
PF00023	**Ankyrin repeat**	181	7	236	8	117	5	84	13	84	22
PF00097	**RING finger**	175	8	204	12	71	19	64	19	86	21
PF00036	**EF hand**	162	9	166	18	110	8	83	15	63	33
PF00046	**Homeobox**	152	10	221	10	83	14	99	9	19	89

Figure S3.2: Number of bidirectional BlastP hits between 22,218 human genes and other organisms

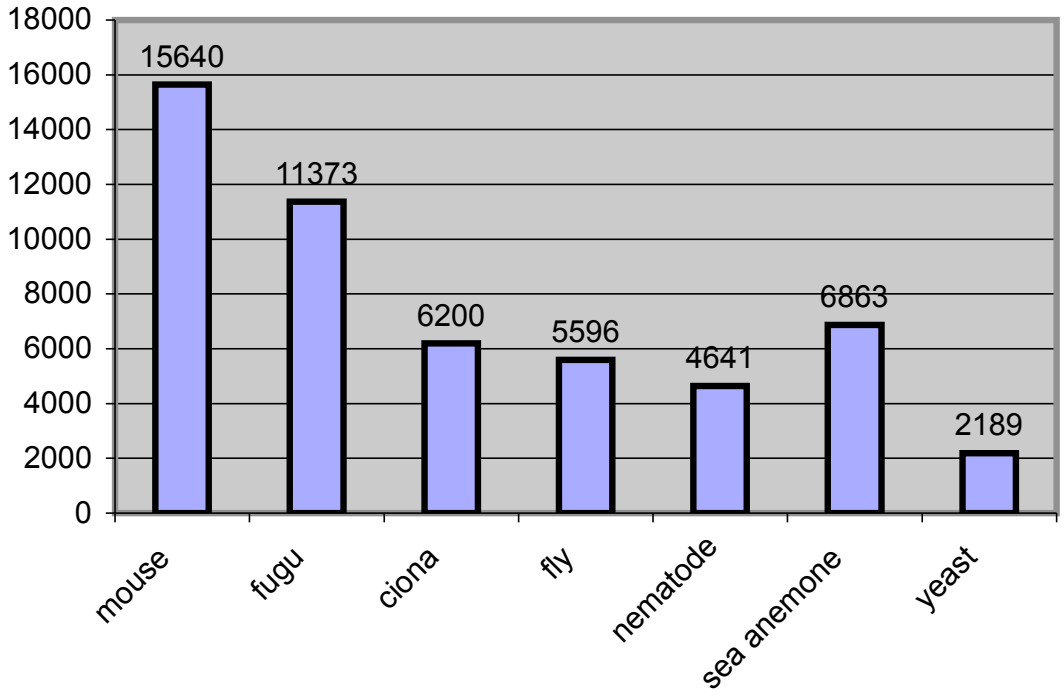


Figure S3.3: Fraction of unique multi (Pfam) domain (2 or more domains) gene models from Nematostella (total 983) shared by other metazoans and yeast.

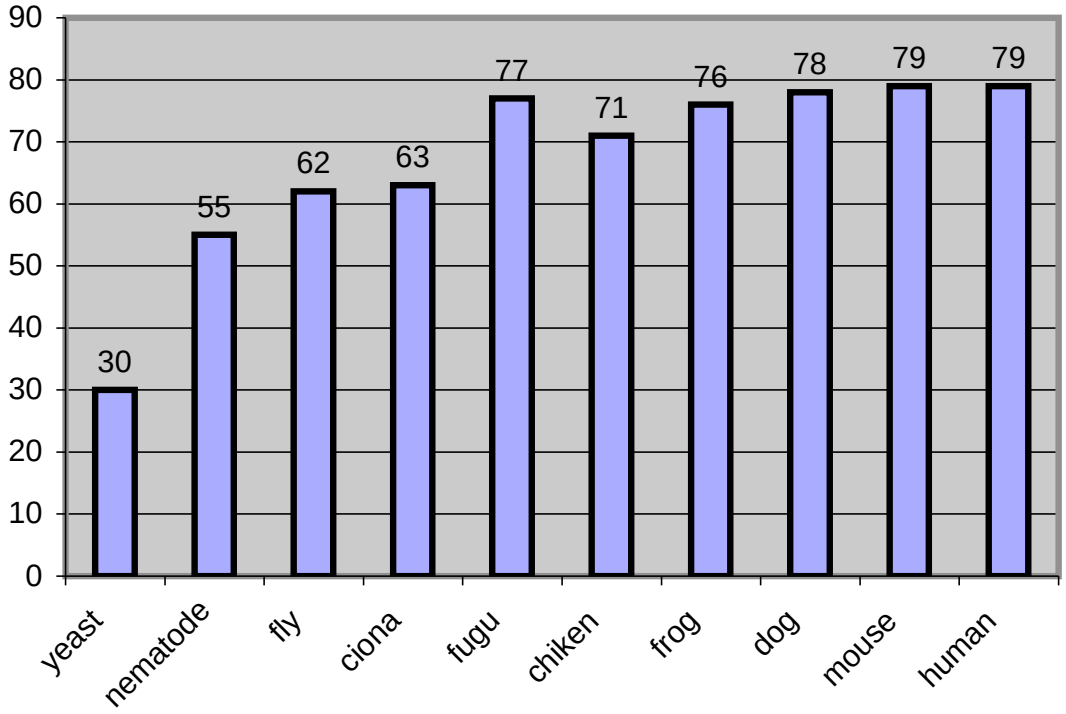


Figure S3.4: 2264 Pfam domains present in all 6 vertebrates

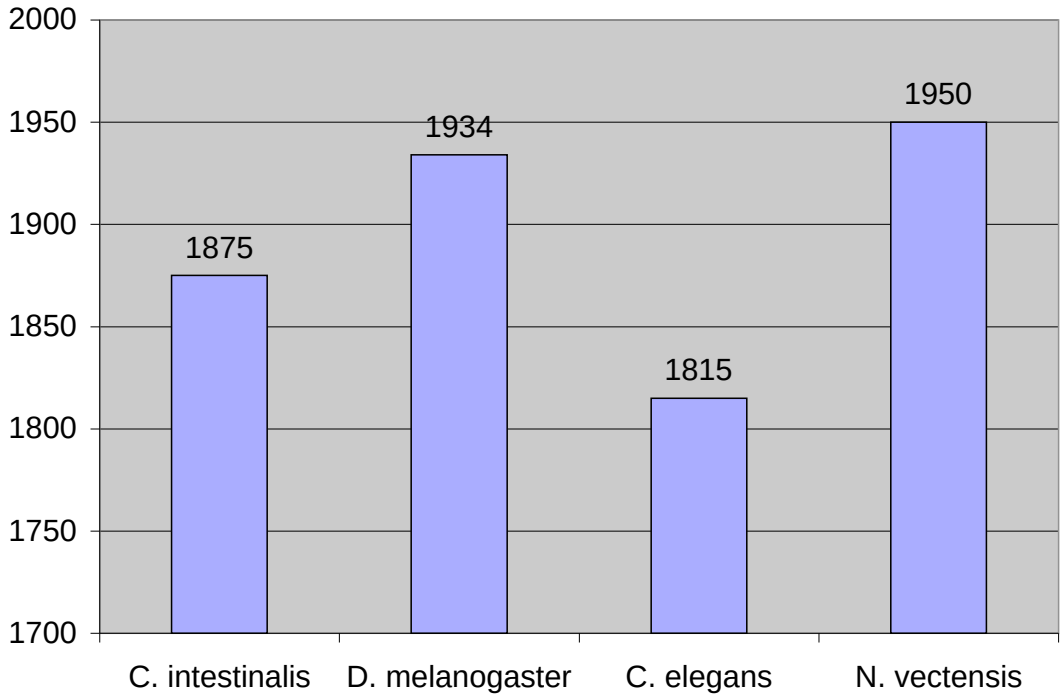


Table S3.3: Preferentially retained PFAM domains within recent tandem expansions in Nematostella

PFAM ID	PFAM Description	#recent	sigma
PF00147	Fibrinogen beta and gamma chains, C-terminal globular domain	18	9.4
PF00112	Papain family cysteine protease	11	7.9
PF00067	Cytochrome P450	18	7.8
PF03953	Tubulin/FtsZ family, C-terminal domain	10	7.6
PF00643	B-box zinc finger	16	6.9
PF02140	Galactose binding lectin domain	12	6.8
PF00091	Tubulin/FtsZ family, GTPase domain	9	6.6
PF00515	TPR Domain	22	6.5
PF07719	Tetratricopeptide repeat	22	5.5
PF00110	wnt family	5	4.4
PF00125	Core histone H2A/H2B/H3/H4	17	4.3
PF03160	Calx-beta domain	5	4.1
PF00754	F5/8 type C domain	27	3.9
PF00106	short chain dehydrogenase	10	3.7
PF00102	Protein-tyrosine phosphatase	6	3.2

Table S3.4: Preferentially retained PFAM domains within recent tandem expansions in Homo sapien

PFAM ID	PFAM Description	#recent	sigma

Table S5.1: Table of data sources for phylogenetic analysis

Data sources for phylogenetic analysis

Whole or partial genome sequences

Xenopus tropicalis JGI v4.1
Takifugu rubripes JGI v4.0
Homo sapiens Ensembl build 38

Drosophila melanogaster Ensembl build 38
Caenorhabditis elegans Ensembl build 38

Nematostella vectensis JGI V1.0

Ciona intestinalis JGI v2.0

Lottia gigantea [J. Chapman, unpublished]
Hydra magnipapillata [Steele et al, unpublished]
Monosiga brevicollis [JGI unpublished]
Renieria spp. [JGI unpublished]
Saccharomyces cerevisiae From genome-ftp.stanford.edu, version released on July 7, 2004

ESTs:

Mnemiopsis leidyi

Figure S5.1: Distribution of percent ID Against Human Proteins

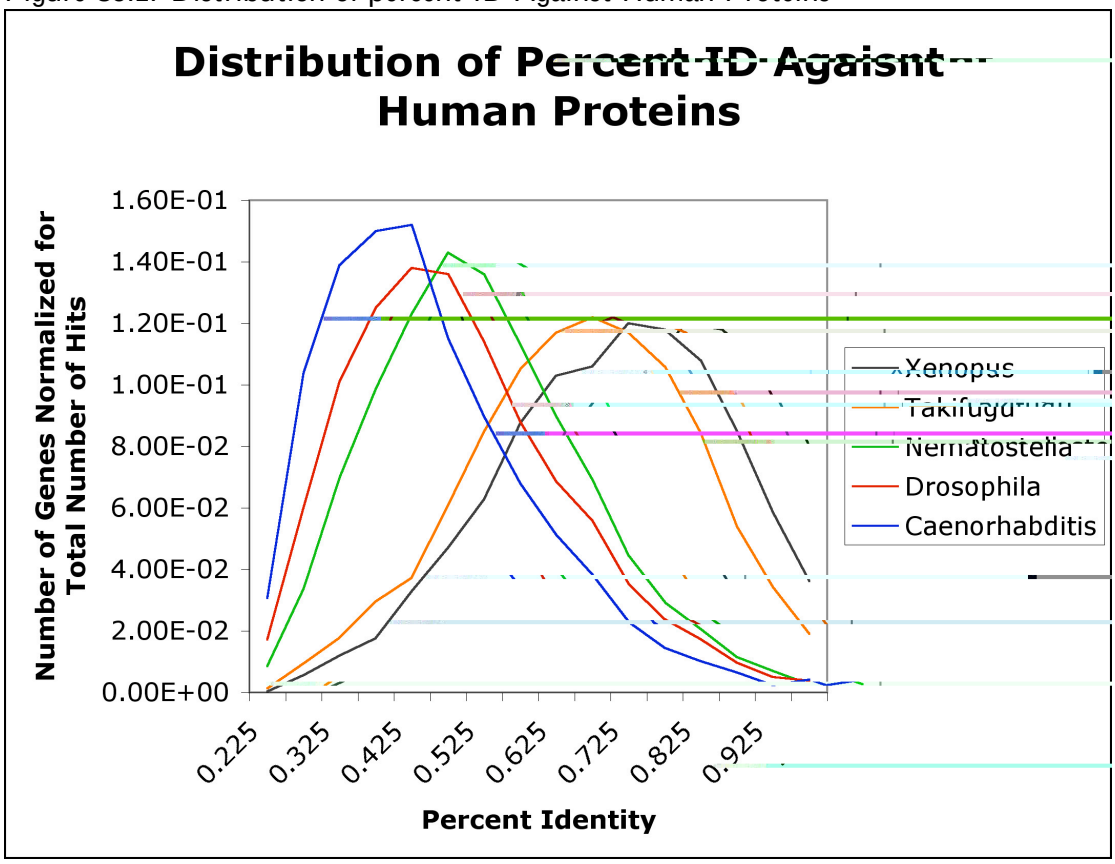


Figure S6.1: Venn diagram for three-way intron conservation comparison

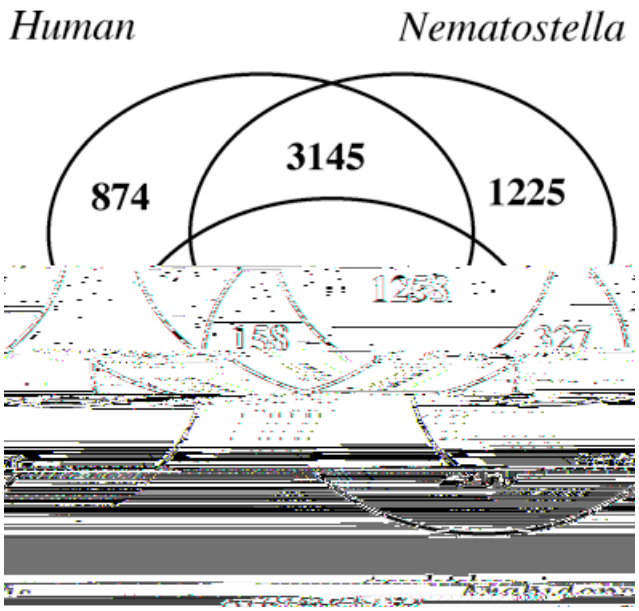


Table S6.1: Four-way intron conservation comparison

Species	**Total Introns**
H. sapiens	3326 (476)
N. vectensis	3647 (771)
D. melanogaster	761 (171)
C. elegans	1363 (551)
H.sapiens + N. vectensis	2751
H. sapiens + C.elegans	714
H. sapiens + D.melanogaster	536
C.elegans + D.melanogaster	232
H.sapiens + N.vectensis + D. melanogaster	495
H.sapiens + N.vectensis + C.elegans	640
shared by all four species	196

Figure S7.1: Synteny block search

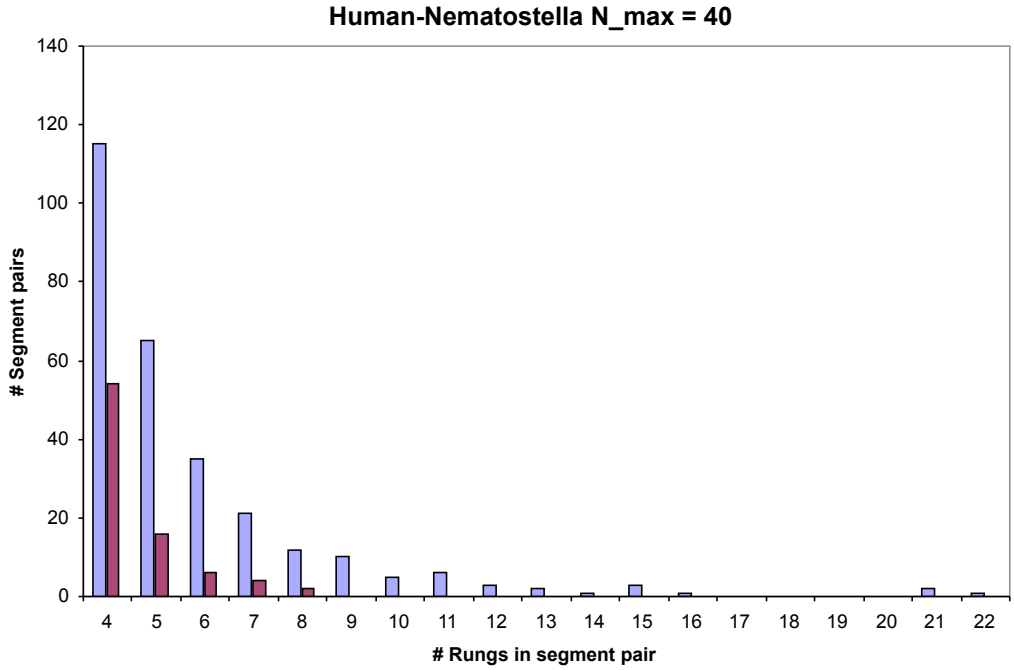


Figure S7.2: HMM segmentation example

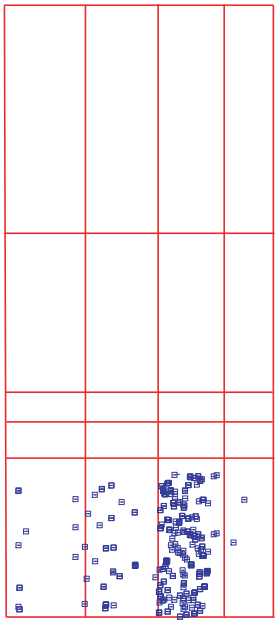


Table S7.1: Table of human chromosome segments used in large-scale synteny search

Name	Chromosome	Start	End
Xp11.4-22.2	X	9673696	37588240
Xp11.21-11.3	X	46887841	55047087
Xp11.21-q13.1	X	55047088	68655440
Xq13.1-28	X	68655440	153978722
Yp11.32-q12	Y	1	57657766
1p36.12-36.33	1	877210	20855970
1p36.11-36.12	1	20855971	25549674
1p34.3-36.11	1	25549674	39269870
1p31.1-34.2	1	40008738	74859196
1p13.3-31.1	1	78330448	110388420
1p12-13.3	1	110388421	118243306
1p12-q21.2	1	119430925	148345068
1q21.2-23.1	1	148345068	155020532
1q23.1-24.2	1	155163302	166097526
1q24.2-31.2	1	168062712	191336756
1q31.2-32.2	1	191336757	208079724
1q32.2-44	1	208079724	244976017
2p24.3-25.3	2	1	15421694
2p13.2-24.3	2	15421694	73578474
2p11.2-13.1	2	74513568	86693774
2p11.2-q11.2	2	86693775	96287750
2q11.2-35	2	96287750	220120257
2q37.1-37.3	2	233900764	242339685
3p24.3-26.3	3	3181960	14740598
3p22.1-24.3	3	15310713	42757316
3p13-22.1	3	43109152	73163221
3p13-q12.2	3	73163222	101930675
3q12.2-27.3	3	101930675	187872602
3q28-29	3	191514553	199135808
4p15.2-16.3	4	929333	25008576
4p12-15.2	4	25278016	48189124
4q12-35.2	4	52592031	190392426
5p12-15.31	5	6704566	43577691
5p12-q12.1	5	43577692	62108653
5q12.1-23.3	5	62108653	128467978
5q31.1-35.3	5	132114396	179586409
6p22.1-25.3	6	1	27327284
6p21.2-22.1	6	27327284	37533628
6p21.2-q14.1	6	37533629	76036806
6q14.1-25.3	6	76036806	158925275
6q27	6	165628122	170899992
7p22.1-22.3	7	762350	6605590
7p11.2-21.3	7	7683932	55720376
7q11.21-11.23	7	65073872	75458076
7q21.3-35	7	96616718	143142128
7q35-36.3	7	143896277	156273990
8p22-23.3	8	1	16976821
8p11.21-22	8	16976821	43145466
8q11.22-24.3	8	51668647	145706329
9p13.3-22.3	9	15431371	35804014
9p13.3-q13	9	35804015	70248716
9q13-31.3	9	70248716	113718168
9q32-34.3	9	114961559	139558315
10p11.22-13	10	15220868	32652190
10q11.21-24.1	10	42623200	98406664
10q24.1-26.3	10	99128910	134856173
11p11.2-15.5	11	188669	47791440
11q12.1-13.1	11	57183558	66019773
11q13.1-25	11	66045396	133689416
12p11.21-13.33	12	2832566	30786824
12q12-14.3	12	42480106	64833745
12q15-23.3	12	67504578	105913314
12q23.3-24.33	12	107435346	131912602
13q12.11-14.11	13	21020596	40815702
13q14.11-34	13	41236742	114076856
14q11.2-12	14	19835780	

Table S7.2: Complete Oxford Grid for Human-Nematostella comparison

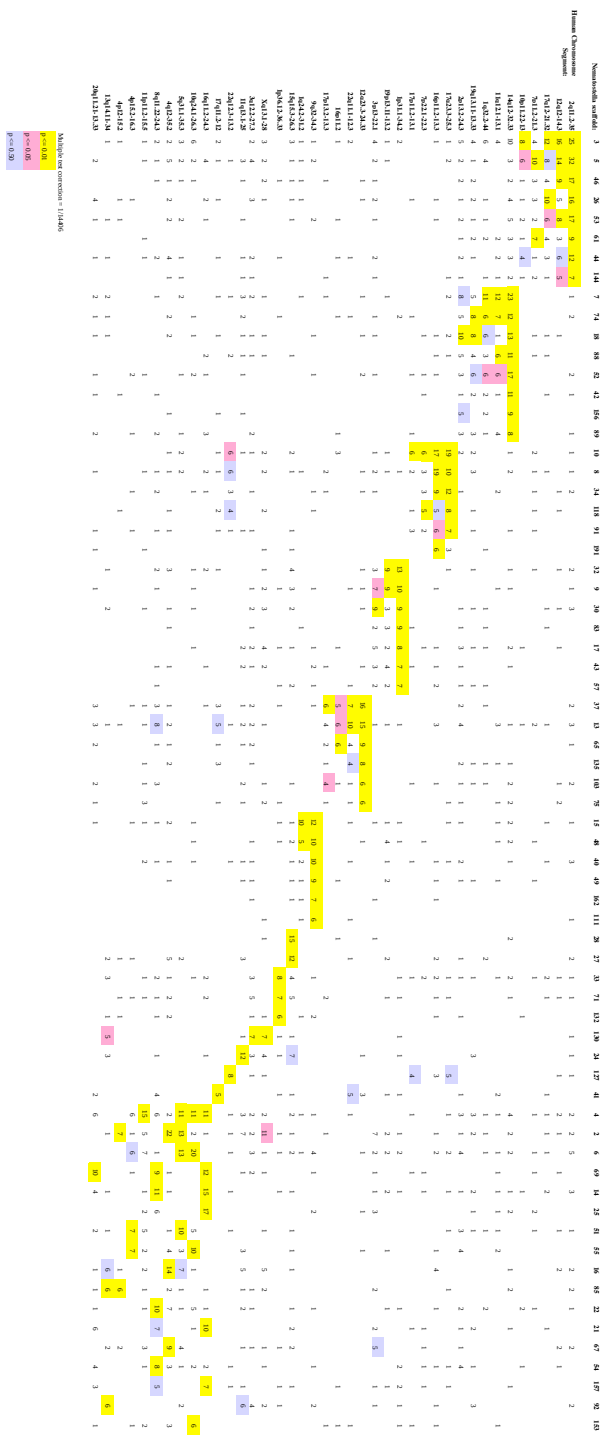


Figure S7.3: Clustering method for constructing putative ancestral linkage groups (PALs)

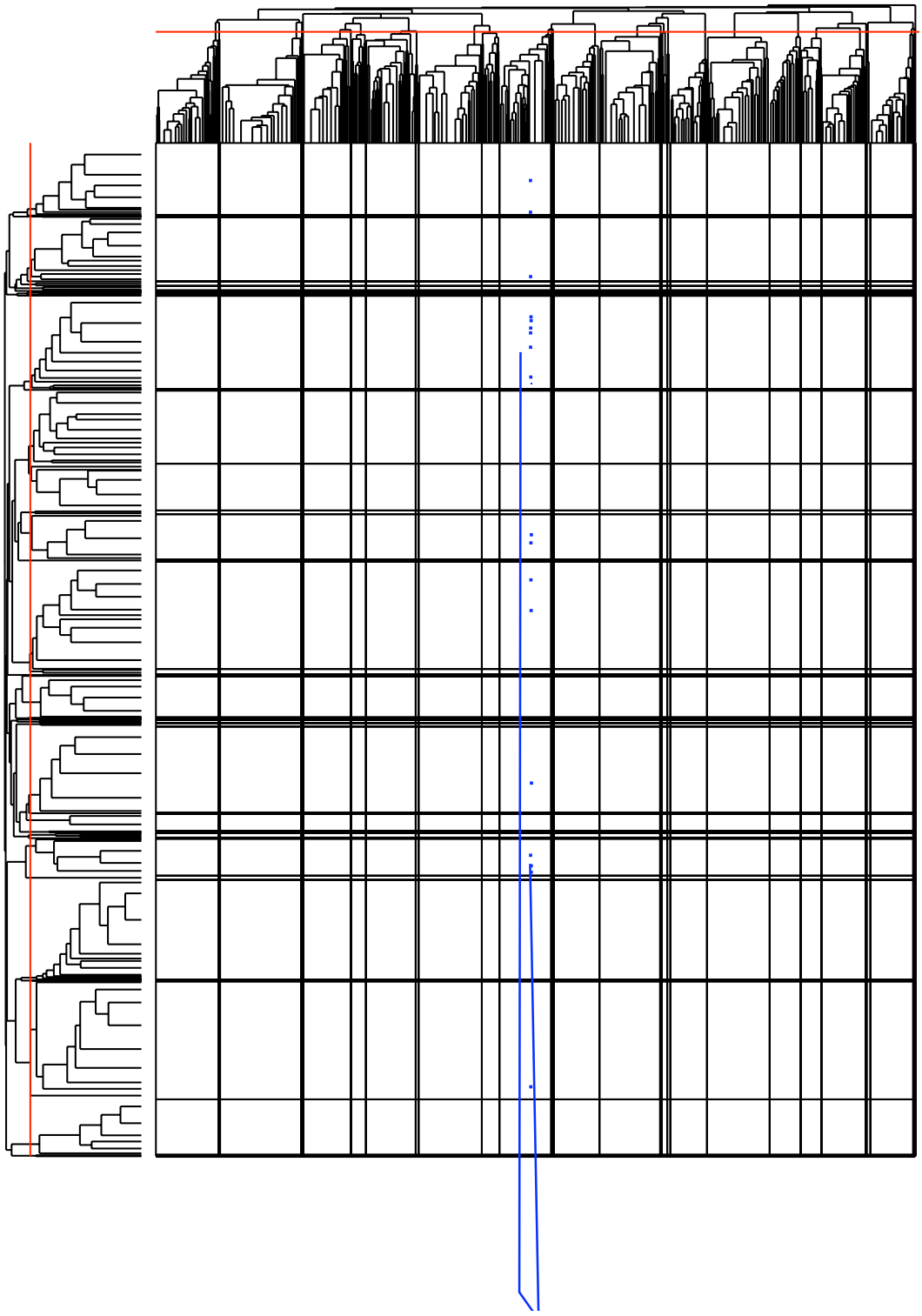


Figure S7.4: Detail of Human chromosome 12 showing genes contributing to PAL A.

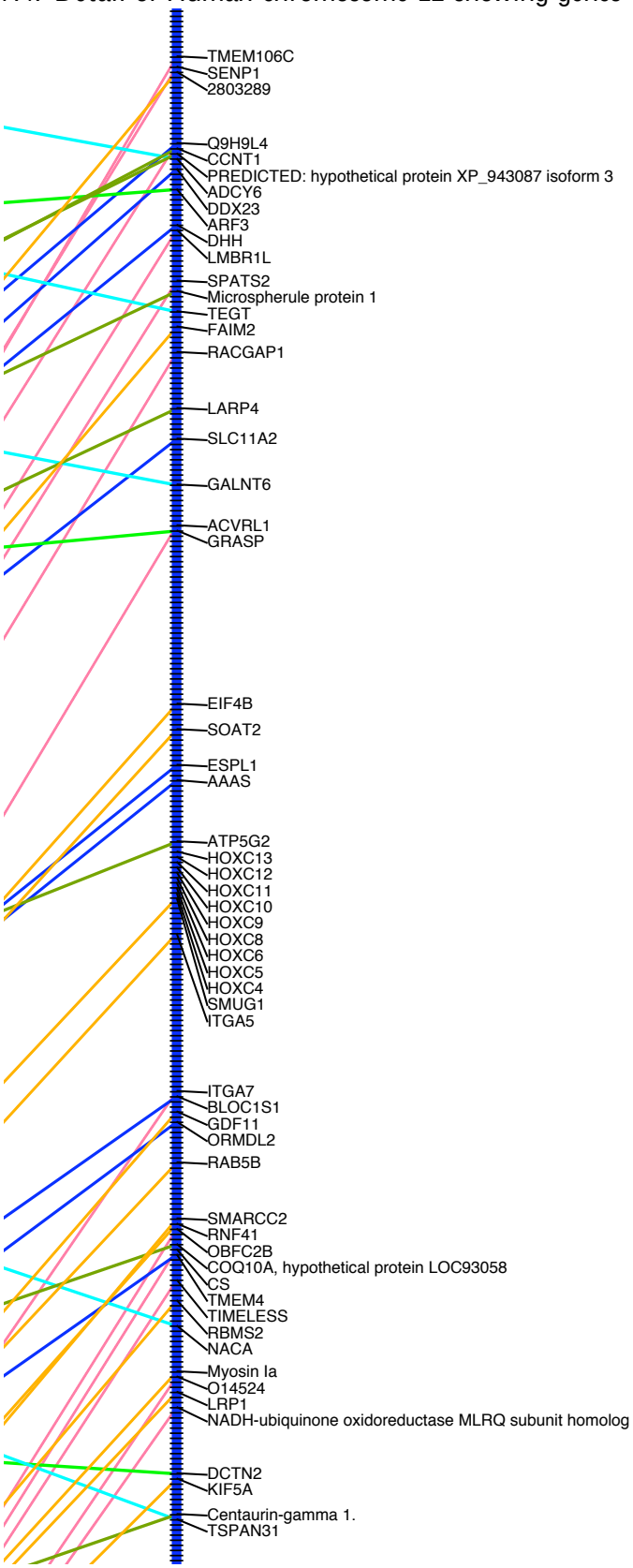


Table S81.a: Panther Ontology Terms for Biological Process and Molecular Function:

Ontology ID	ln(p-value) enrichment /depeletion	+/-	N(ont & cat)	N(ont)	N(cat)	N(total)	N(ont& cat) / N(cat)	N(ont)/ N(total)	Ontology Term Desc.
Type III novelty, p<0.05/100 enriched ontology terms:									
BP00102	-52.3	+	68	575	240	7766	28%	7%	Signal transduction
MF00100	-26.5	+	23	125	240	7766	10%	2%	G-protein modulator
BP00285	-21.7	+	29	246	240	7766	12%	3%	Cell structure and motility
BP00111	-20.7	+	29	257	240	7766	12%	3%	Intracellular signaling cascade
MF00093	-20.6	+	36	379	240	7766	15%	5%	Select regulatory molecule
BP00103	-19.3	+	24	192	240	7766	10%	2%	Cell surface receptor mediated signal transduction
MF00212	-18.7	+	14	65	240	7766	6%	1%	Other G-protein modulator
BP00124	-16.7	+	13	64	240	7766	5%	1%	Cell adhesion
MF00261	-16.6	+	16	101	240	7766	7%	1%	Actin binding cytoskeletal protein
BP00166	-16.1	+	16	104	240	7766	7%	1%	Neuronal activities
BP00104	-15.6	+	14	82	240	7766	6%	1%	G-protein mediated signaling
BP00274	-12.5	+	16	135	240	7766	7%	2%	Cell communication
BP00199	-12.3	+	14	107	240	7766	6%	1%	Neurogenesis
BP00064	-11.7	+	21	231	240	7766	9%	3%	Protein phosphorylation
BP00286	-11.6	+	16	145	240	7766	7%	2%	Cell structure
BP00246	-11.3	+	14	116	240	7766	6%	1%	Ectoderm development
MF00107	-11.1	+	22	259	240	7766	9%	3%	Kinase
MF00091	-11.1	+	20	222	240	7766	8%	3%	Cytoskeletal protein
BP00119	-10.0	+	10	69	240	7766	4%	1%	Other intracellular signaling cascade
BP00193	-9.4	+	27	396	240	7766	11%	5%	Developmental processes
Type II novelty, p<0.05/100 enriched ontology terms:									
BP00193	-39.7	+	40	396	158	7766	25%	5%	Developmental processes
BP00102	-38.4	+	47	575	158	7766	30%	7%	Signal transduction
MF00001	-25.2	+	18	115	158	7766	11%	1%	Receptor
BP00274	-24.6	+	19	135	158	7766	12%	2%	Cell communication
BP00246	-20.5	+	16	116	158	7766	10%	1%	Ectoderm development
BP00199	-19.5	+	15	107	158	7766	9%	1%	Neurogenesis
BP00103	-13.4	+	16	192	158	7766	10%	2%	Cell surface receptor mediated signal transduction
BP00287	-11.7	+	9	68	158	7766	6%	1%	Cell motility
BP00044	-11.5	+	18	273	158	7766	11%	4%	mRNA transcription regulation
MF00016	-10.3	+	10	100	158	7766	6%	1%	Signaling molecule
BP00166	-10.0	+	10	104	158	7766	6%	1%	Neuronal activities
BP00111	-9.7	+	16	257	158	7766	10%	3%	Intracellular signaling cascade
MF00036	-9.3	+	19	352	158	7766	12%	5%	Transcription factor
BP00285	-8.9	+	15	246	158	7766	9%	3%	Cell structure and motility
BP00248	-7.8	+	8	89	158	7766	5%	1%	Mesoderm development
BP00040	-7.7	+	19	398	158	7766	12%	5%	mRNA transcription
Type I novelty, p<0.05/100 enriched ontology terms:									
MF00016	-8.0	+	29	100	1186	7766	2%	1%	Signaling molecule
All types of novelty, p<0.05/100 enriched ontology terms:									
BP00102	-24.4	+	182	575	1584	7766	11%	7%	Signal transduction
BP00103	-24.4	+	79	192	1584	7766	5%	2%	Cell surface receptor mediated signal transduction
BP00193	-23.1	+	134	396	1584	7766	8%	5%	Developmental processes
MF00016	-22.8	+	49	100	1584	7766	3%	1%	Signaling molecule
BP00274	-22.5	+	60	135	1584	7766	4%	2%	Cell communication
BP00166	-16.2	+	45	104	1584	7766	3%	1%	Neuronal activities
BP00246	-12.5	+	45	116	1584	7766	3%	1%	Ectoderm development
BP00248	-12.4	+	37	89	1584	7766	2%	1%	Mesoderm development
BP00124	-12.1	+	29	64	1584	7766	2%	1%	Cell adhesion
BP00104	-11.5	+	34	82	1584	7766	2%	1%	G-protein mediated signaling
MF00001	-11.0	+	43	115	1584	7766	3%	1%	Receptor
BP00199	-10.3	+	40	107	1584	7766	3%	1%	Neurogenesis
BP00281	-7.9	+	35	99	1584	7766	2%	1%	Oncogenesis
BP00111	-7.8	+	75	257	1584	7766	5%	3%	Intracellular signaling cascade
Type III novelty, p<0.05/100 depleted ontology terms:									
MF00131	-10.2	-	1	398	240	7766	0%	5%	Transferase
Type II novelty, p<0.05/100 depleted ontology terms:									
Type I novelty, p<0.05/100 depleted ontology terms:									
BP00060	-114.4	-	24	1056	1186	7766	2%	14%	Protein metabolism and modification
MF00042	-55.4	-	46	915	1186	7766	4%	12%	Nucleic acid binding
BP00031	-50.7	-	62	1034	1186	7766	5%	13%	Nucleoside, nucleotide and nucleic acid metabolism

BP00063	-41.2	-	13	447	1186	7766	1%	6% Protein modification
MF00141	-40.1	-	5	330	1186	7766	0%	4% Hydrolase
MF00107	-33.8	-	3	259	1186	7766	0%	3% Kinase
MF00123	-31.4	-	6	289	1186	7766	1%	4% Oxidoreductase
MF00131	-30.9	-	15	398	1186	7766	1%	5% Transferase
BP00019	-30.5	-	4	254	1186	7766	0%	3% Lipid, fatty acid and steroid metabolism
BP00125	-30.3	-	16	405	1186	7766	1%	5% Intracellular protein traffic
BP00141	-29.6	-	14	377	1186	7766	1%	5% Transport
BP00001	-29.4	-	3	231	1186	7766	0%	3% Carbohydrate metabolism
BP00064	-29.4	-	3	231	1186	7766	0%	3% Protein phosphorylation
MF00170	-28.0	-	1	188	1186	7766	0%	2% Ligase
BP00071	-27.0	-	7	273	1186	7766	1%	4% Proteolysis
BP00203	-27.0	-	13	346	1186	7766	1%	4% Cell cycle
MF00082	-23.0	-	5	219	1186	7766	0%	3% Transporter
MF00108	-21.9	-	3	183	1186	7766	0%	2% Protein kinase
MF00126	-21.7	-	0	130	1186	7766	0%	2% Dehydrogenase
BP00282	-21.6	-	0	129	1186	7766	0%	2% Mitosis
BP00013	-21.4	-	0	128	1186	7766	0%	2% Amino acid metabolism
BP00061	-20.8	-	4	190	1186	7766	0%	2% Protein biosynthesis
BP00289	-19.1	-	9	241	1186	7766	1%	

BP00034	-13.3	-	12	167	1584	7766	1%	2%	DNA metabolism
BP00019	-12.5	-	25	254	1584	7766	2%	3%	Lipid, fatty acid and steroid metabolism
MF00097	-12.5	-	5	106	1584	7766	0%	1%	G-protein
MF00044	-12.1	-	3	86	1584	7766	0%	1%	Nuclease
MF00284	-12.1	-	3	86	1584	7766	0%	1%	Other ligase
MF00170	-12.0	-	16	188	1584	7766	1%	2%	Ligase
BP00141	-11.9	-	45	377	1584	7766	3%	5%	Transport
BP00076	-11.8	-	6	111	1584	7766	0%	1%	Electron transport
BP00020	-11.7	-	2	74	1584	7766	0%	1%	Fatty acid metabolism
BP00276	-11.0	-	9	130	1584	7766	1%	2%	General vesicle transport
BP00048	-10.8	-	5	97	1584	7766	0%	1%	mRNA splicing
MF00264	-10.8	-	5	97	1584	7766	0%	1%	Microtubule family cytoskeletal protein
MF00065	-10.7	-	3	79	1584	7766	0%	1%	mRNA processing factor
MF00051	-10.1	-	6	101	1584	7766	0%	1%	Helicase
MF00099	-9.8	-	4	83	1584	7766	0%	1%	Small GTPase
MF00127	-9.8	-	5	91	1584	7766	0%	1%	Reductase
BP00062	-9.5	-	4	81	1584	7766	0%	1%	Protein folding
MF00086	-9.1	-	9	118	1584	7766	1%	2%	Other transporter
MF00153	-8.7	-	15	158	1584	7766	1%	2%	Protease
BP00071	-8.7	-	33	273	1584	7766	2%	4%	Proteolysis
BP00081	-8.5	-	5	84	1584	7766	0%	1%	Coenzyme and prosthetic group metabolism
MF00077	-8.4	-	5	83	1584	7766	0%	1%	Chaperone
MF00157	-7.9	-	6	88	1584	7766	0%	1%	Lyase
BP00129	-7.6	-	7	94	1584	7766	0%	1%	Endocytosis

Table S8.1b: Panther Pathways

Ontology ID	ln(p-value) enrichment/ depeletion	+/-	N(ont& cat)	N(ont)	N(cat)	N(total)	N(ont& cat) / N(cat)	N(ont)/ N(total)	Ontology Term Desc.
Type III novelty, p<0.05/100 enriched ontology categories:									
P00031	-12.91	+	11	62	240	7766	5%	1%	Inflammation mediated by chemokine and cytokine signaling pathway
P00019	-9.85	+	7	33	240	7766	3%	0%	Endothelin signaling pathway
P04385	-7.92	+	4	12	240	7766	2%	0%	Histamine H1 receptor mediated signaling pathway
P00027	-7.91	+	5	21	240	7766	2%	0%	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway
Type II novelty, p<0.05/100 enriched ontology categories:									
P00004	-20.96	+	10	34	158	7766	6%	0%	Alzheimer disease-presenilin pathway
P00057	-17.06	+	12	77	158	7766	8%	1%	Wnt signaling pathway
P00005	-14.74	+	10	62	158	7766	6%	1%	Angiogenesis
P00031	-12.47	+	9	62	158	7766	6%	1%	Inflammation mediated by chemokine and cytokine signaling pathway
P00034	-8.03	+	7	65	158	7766	4%	1%	Integrin signalling pathway
P00045	-7.81	+	4	18	158	7766	3%	0%	Notch signaling pathway
Type I novelty, p<0.05/100 enriched ontology categories:									
All types of novelty, p<0.05/100 enriched ontology categories:									
P00031	-10.44	+	27	62	1584	7766	2%	1%	Inflammation mediated by chemokine and cytokine signaling pathway
P00005	-8.28	+	25	62	1584	7766	2%	1%	Angiogenesis
P00057	-7.98	+	29	77	1584	7766	2%	1%	Wnt signaling pathway
Type III novelty, p<0.05/100 depleted ontology categories:									
Type II novelty, p<0.05/100 depleted ontology categories:									
Type I novelty, p<0.05/100 depleted ontology categories:									
P00049	-7.81	-	0	47	1186	7766	0%	1%	Parkinson disease
All types of novelty, p<0.05/100 depleted ontology categories:									