Supporting Online Material for

Sea Anemone Genome Reveals Ancestral Eumetazoan Gene Repertoire and Genomic Organization

Nicholas H. Putnam, Mansi Srivastava, Uffe Hellsten, Bill Dirks, Jarrod Chapman, Asaf Salamov, Astrid Terry, Harris Shapiro, Erika Lindquist, Vladimir V. Kapitonov, Jerzy Jurka, Grigory Genikhovich, Igor Grigoriev, Susan M. Lucas, Robert E. Steele, John R. Finnerty, Ulrich Technau, Mark Q. Martindale, Daniel S. Rokhsar*

*To whom correspondence should be addressed. E-mail: dsrokhsar@lbl.gov

Published 6 July, *Science* **317**, 86 (2007)
DOI: 10.1126/science.1139158

This PDF file includes:

Supporting Online Material for "Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization." Putnam et al. 2

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 distribution mainly along the northern Atlantic and Pacific coasts of North America (*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*). Fertilization is followed by cell divisions resulting in a hollow blastula, which gastrulates by invagination and ingression to produce a ciliated, tear-drop-shaped planula larva that swims with an apical tuft of sensory cilia at the front and the blastopore at the rear (Figure S1.1 a-e, h, i). On the seventh day after fertilization, the planula develops into a juvenile polyp, with the blastopore becoming the mouth (*3, 5, 6*) (Figure S1.1f). This metamorphosis results in 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), with sexual maturity reached in 3-4 months. Mature adults are hollow tubes typically 5-10 cm in length, with eight mesenteries and an open (oral) end encircled by 10-20 tentacles a few cm long, and a closed (aboral) end (Figure S1.1l). 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.

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 regenerate when cut. It is not known how tentacle number or body tube length is regulated, either in regeneration or embryogenesis. Voucher specimens of siblings of the original parental strain have been deposited in separate sexes at the Peabody Museum at Yale, New Haven, Connecticut, USA. The sample numbers are as follows: YPM Nos. 39180, 39181, 39182, 39183.

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

Figure S1.1 Methods

Nematocyst staining (Figure S1.1 g): (Methods adapted from (*7*)) Juvenile and small adult Nematostella polyps were relaxed in 7.14% MgCl2 in dH2O for ten minutes and then washed quickly three times in 1X PBS with 10mM EDTA. They were then fixed in 4% paraformaldehyde in 1X PBS with 10mM EDTA for one hour at 4oC. After washing three times for five minutes each in 1X PBS with 10mM EDTA, the animals were stained in a 200uM DAPI solution in 1X PBS for thirty minutes. Animals were mounted in 70% glycerol in Ptw after washing three times for five minutes each in 1X PBS with 10mM EDTA.

In situ hybridization (Figure S1.1 j,k): In situ hybridization was carried out as previously described (*8*).

Supplement S2: Genome sequencing and characterization

Source material for genome sequencing

Genomic DNA was prepared in the laboratory of Ulrich Technau from larval F1 progeny of CH2 males 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, CHORI-219, 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 (http://bacpac.chori.org).

Whole Genome Shotgun (WGS) Sequencing and Assembly.

The genome of *Nematostella vectensis* was sequenced and assembled by whole genome shotgun (WGS) (*9*) as previously described (*10*). 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 (*10, 11*). 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 therefore 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.

Polymorphism

To avoid contamination from commensal microbes common to adult anemones and minimize the impact of allelic variation on assembly quality, we prepared genomic DNA from the larvae of a single mating pair originally isolated from the same lagoon. Our dataset thus nominally contains up to four alleles at each locus. From the shotgun assembly and the analysis of alignments between shotgun reads, we measured a rate of single nucleotide polymorphism among the four parental alleles as 0.8%, or ~1/125 bp, approximately ten times the SNP rate in the human population. Some 16,000 SNPs may be searched at the SNP browser available at StellaBase (http://stellabase.org) (*12*). After correcting for sampling, we estimated that each pair of alleles differs at 0.65% of nucleotide positions (Figure S2.3). Thus the parental anemones whose genomes we sampled have somewhat less allelic variation than broadcast spawning invertebrates such as sea squirts (~2%) (*10*) and sea urchins (5-10%) (*13*), or outbreeding plants like *Populus* (~2%) (*14*), but a comparable amount to the pufferfish (0.5%) (*11*).

Nematostella, however, is not a true broadcast spawner, since while males release sperm into the water, females lay tens to hundreds of eggs encased in a jelly mass that becomes fixed to a benthic substrate. The egg mass may be a derived feature of *Nematostella* that is related to its colonization of the estuarine environment. The relatively low level of intra-specific genetic variation in *Nematostella* vs. marine broadcast spawners might be explained if its estuarine habitat limited gamete dispersal and led to a smaller effective population size. Genetic fingerprinting of wild *Nematostella* populations indicates a high degree of genetic structuring at fine spatial scales, implying extremely low levels of gene flow between neighboring estuaries (*15*). The source population for the genome sequence (Rhode River, Maryland) appears typical in this regard (*16*).

After correcting for ascertainment bias, we estimate that 0.85% of four-fold degenerate sites in predicted proteins are polymorphic in the sampled haplotypes (Figure S2.1). Approximately 0.8% of positions in the assembly contain a polymorphic site (Figure S2.2), and we estimate that the mean pairwise variation between the four alleles represented in the libraries is 0.64 % (Figure S2.3).

We estimate that no more than 5 to 10 percent of the genome is represented redundantly in the assembly due to locally higher rates of heterozygosity between alleles. This estimate is an upper bound because some apparent splits between haplotypes are likely to be recently duplicated regions (*17*). To arrive at this estimate we compared EST contigs to the assembly, the predicted protein sequences to one another, and the assembled scaffolds to one another. 839 (11%) of the EST contigs had alignments of at least 95% identity spanning at least 75% of their length with multiple locations in the

assembly. 4421 predicted genes share at least 99% amino acid identity with another predicted gene. If all these models are assumed to be alternate alleles, this corresponds to from 5% to 11% of loci being represented in multiple alleles. Finally, 898 randomly sampled 2 KB fragments of the assembly that did not include any sequence gaps were aligned to the whole assembly. 132, or 14.7% of the sampled sequences, had alignments of at least 90% identity, covering at least 75% of their length to 1, 2 or 3 (other) locations in the assembly. The vast majority hit a single second location in the assembly, indicating that ~ 8% of genomic loci are represented in multiple alleles in the assembly.

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

Two mixed stage cDNA libraries for Nematostella were prepared in the laboratory of Ulrich Technau and Thomas W. Holstein, cloning polyA RNA from unfertilized eggs through metamorphosis into pSPORT 6.1 (non-normalized library) and pBS-SK(+) (normalized library). The non-normalized 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. The normalized library was prepared from the same source of RNA has an average insert size of 0,9 kb and contains 6.5x105 cfu. In addition, several cDNA libraries were prepared from stage-specific RNA provided by Mark Q. Martindale. Detailed information about the libraries can be obtained upon request.

Two mixed stage cDNA libraries 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 (*18*) (-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 (*19*) (-e 1e-5), indicating that they are likely to represent *bona fide* protein-coding transcripts rather than artifactual sequence. Others may be contaminants of the EST library, or novel genes.

For *Mnemiopsis leidyi*, a cDNA library was created from total RNA prepared from gastrula 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 (*20*) was used to look for tRNAs and BLASTN (*19*) against nr and Repbase (*21*) 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, available from JGI Genome Portal (www.jgi.doe.gov/Nematostella).

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, just a few families represent different classes of autonomous DNA transposons. 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 (*22-24*), 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 (*22*), 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 (*25*). It makes *Nematostella* the first metazoan with Polintons constituting a substantial portion of the genome (*25*).

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 (*22*). Among DNA transposons, the hAT superfamily is well-known for TSDs that are always 8 bp long (*22*). 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. S2.4). 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 (*26*), 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 (*22, 27*). 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. S2.4, 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 (*28*).

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 (*19*)). 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 (*21*).

Chromosome Number

The synchronous cell divisions during early development of Nematostella allowed easy access to large numbers of metaphase plates. Due to loss or overlapping of individual chromosomes, counts tend to underestimate the actual numbers. The analysis shows, however, that the great majority of counts showed 2n = 30 chromosomes (Fig. S2.5). Interestingly, as in Hydra, the chromosomes do not differ significantly in size, which makes it difficult to create a reliable karyotype. No evidence of morphologically distinct sex chromosomes was found. Based on the estimated genome size of 457 Mb, the average size of a chromosome is in the range of 30.5 Mb.

The number of chromosomes found in Nematostella is roughly in the range of what has been reported from other cnidarians: a diploid chromosome set of 30 has been determined in several representatives of the genera *Hydra* and *Pelmatohydra* (*29-31*), which (with the exception of *Hydra viridis*) have a 3-4 fold larger genome size. The chromosome counts from 24 hydrozoan and one scyphozoan species apparently can be anything from 2n = 12 to 32 (reviewed in (*32*)). Among anthozoans the karyotypes of *Aptasiomorpha* sp. (2n = 32), several *Acropora* species (2n = 28) and two species of *Anthopleura* (2n = 18) have been reported (*33-35*).

Supplement S3: Gene prediction

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 (*36*) tools and a custom library of manually curated repeats (deposited in RepBase (*21*)). 146,095 ESTs were clustered into 30,813 consensus sequences and both individual ESTs and consensus sequences were mapped onto genome assembly using BLAT (*18*).

Annotation of Nematostella v1.0 was based on several gene predictors – *ab initio* FGENESH (*37*), homology-based FGENESH+ (*37*), homology-based GENEWISE (*38*), and available ESTs.

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 (*19*). High quality seed proteins were then used to build models using FEGENESH+ and GENEWISE. GENEWISE gene models were then filtered to remove models with frame shifts 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 (*39*) we predicted proteins domains. Using both these sources of information, annotation of each protein was mapped to the terms of Gene Ontology (*40*), KOG clusters of orthologs (*41*), and mapped to KEGG pathways (*42*).

The large set of all predicted models was reduced to a non-redundant set of 27,273 representative models (Filtered Models), where a single best gene model according to the criteria of homology and EST support describes every locus. For this set of representative gene models we assigned GO (*40*) terms to 12,786 proteins, 16,625 (78%) proteins to KOG clusters (*41*), and 695 distinct EC numbers were assigned to 2,822 proteins mapped to KEGG pathways (*42*). Table S3.1 summarizes the set of predicted genes.

The data are available from JGI Genome Portal (http://www.jgi.doe.gov/Nematostella) and has been deposited at DDBJ/EMBL/GenBank under the project accession ABAV00000000. The version described in this paper is the first version, ABAV01000000.

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 choice 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. 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 we 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 proteins 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 (*43*)

- 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 (*44*).

- 16 Pfam domains previously exclusively found only in vertebrates, but not in other bilaterian phyla (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) (*14*). 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.

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 (*45*). 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 trans-membrane 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 created 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 defense proteins. This newly acquired repertoire almost certainly plays a key role in defining vertebrates and mammals. Similarly, 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

A simple way to identify putative orthologs (genes descended from the same gene in the common ancestor) between genomes is through reciprocal best-scoring BLAST hits. The human genome has more such orthologous pairs with *Nematostella* (6,989) than with non-vertebrate bilaterians, including *Drosophila* (5,772), *C. elegans* (4,846), and even the invertebrate chordate *Ciona intestinalis* (6,313).

To understand gene creation and duplication we designed a phylogenetically informed clustering algorithm that 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 that 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, e-value 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. ((*46*) came to the same conclusion using ESTs and BLAST e-value to measure similarity.) Of the 7,766 ancestral metazoan 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 (*47*), and well-aligned regions extracted with GBLOCKS (*48*) 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 (*49, 50*), using a the WAG (*51*) 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 bilaterians, we submitted this dataset to a maximum likelihood analysis using the PHYLIP package's PROML program (*52*), 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 1 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 (*53*) between recent molecular clock estimates (*54*) of the timing of the protostome-deuterostome (95% confidence interval: 640-760 Mya) and choanoflagellate-metazoan (95% CI: 760-960 Mya) divergences. Figure 1 in the main text shows 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.

The relatively slow rate of protein sequence evolution in *Nematostella* compared to fly and nematode can be seen more directly by considering the amino-acid percent identity between reciprocal-best-hits of selected proteomes vs. human. Despite the fact that flies and nematodes share a more recent common ancestor with human than sea anemones do, we find that the anemone peptides are more similar to human than to either of the model protostomes (*46, 55*). Figure S5.1 shows, in 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

Intron conservation can be unambiguously assessed by identifying well-aligned regions of orthologous proteins that are interrupted by introns in one or more species (Figure 3a). Note that this analysis is protected from the effects of gene modeling artifacts, since erroneous predictions in the vicinity of splice sites would disrupt alignment, thereby removing such sequences from consideration.

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 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 varies 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, Cryptococcus neoformans and Arabidopsis thaliana.* 4342

introns from the seven genomes at 2645 aligned positions that 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 (*56*) 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: Analysis of anciently conserved gene linkage

We first searched for regions of approximately conserved gene order between *Nematostella* and human, allowing for local rearrangements as well as independent differential gene loss and/or duplication in each genome (*57*). We found 33 conserved syntenic segments, each containing 9 or more orthologous gene pairs, under conditions for which no such segments are expected when gene order is completely randomized in the two genomes (Figure S7.1). Within each segment, however, local gene order is considerably scrambled. Further relaxing gene order constraints dramatically increases the number of such segments expected by chance, reducing the power of this approach to detect even more ancient conserved genome organization in the face of intra-chromosomal rearrangements. To overcome this limitation we developed a new method to search for statistically significant conserved linkage groups that does not rely on gene order.

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 identifiers 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 that 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 S7.3 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 (*58*). 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 that 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 (*19*), using the NCBI taxonomy database to remove metazoan entries. The Pfam (*45*) 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-animals 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 (*59, 60*) 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 (*60*) 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 (*61*) for use with the Panther ontologies, and computed p-values for enrichment and dearth relative the hypergeometric distribution. For both Panther Pathways and Panther 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 Figures

Figure S1.1 *Nematostella* development and anatomy
a. unfertilized egg (~200 micron diameter) with sperm head; b. early cleavage stage; c. blastula; d. gastrula; e. planula; f. juvenile polyp; g. adult stained with DAPI to show nematocysts with a zoom in on the tentacle in the inset; h, i. confocal images of a tentacle bud stage and a gastrula respectively showing nuclei (red) and actin (green); j. a gastrula showing snail mRNA(purple) in the endoderm and forkhead mRNA (red) in the pharynx and endoderm; k. a gastrula showing Anthox8 mRNA expression; l. an adult *Nematostella*.

Figure S2.1: Observed density of polymorphic 4-fold degenerate codon positions.

The rate of single nucleotide polymorphism observed in four-fold degenerate codon positions of gene

models is 0.85%. Figure S2.1 shows the observed (read) and Poisson ascertainment bias-corrected (green) frequency of polymorphic 4-fold degenerate sites as a function of local depth of assembly for a sampling of 17.3 million 4-fold degenerate codons [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 blue histogram shows the number of positions considered for each depth of coverage [right hand scale].

Figure S2.2: 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 (red) 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 blue curve shows the number of positions considered for each depth of coverage, and the gray curve shows Poisson distributed counts with the same mean.

Figure S2.3: 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).

Figure. S2.4 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 reinhardtii*; 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 (*62*)). Clades and groups of non-LTR retrotransposons are indicated by black and blue rectangles.

Figure S2.5 Number of chromosomes
The number of chromosomes was determined by analyzing 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.

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).

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.

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.

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*.

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 that are recent compared to the time scale of gene order evolution.

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.

Captions for Supplemental Tables

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.

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 (*19*).

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

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

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

The number of proteins with PFAM (*45*) 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*.

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 prediction of a model of the neutral expectation are shown.

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

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 prediction of a model of the neutral expectation are shown.

Table S5.1: Data sources for phylogenetic analysis

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 that are unique to that species. The remaining table rows list the number of introns shared by selected combinations of genomes.

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).

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

Table S7.3 is available as SOM 2 on the Science web site. The table lists the inferred ancestral eumetazoan genes that show ancient conserved linkage with both the vertebrate HOX clusters and Nematostella scaffolds 3 and 61. Each row lists the internal ID number of the cluster, followed by the number of recognized descendant genes in each of the representative genomes used in the analysis, separated by commas: Human, *Xenopus tropicalis*, *Takifugu rubripes*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Nematostella*; the genes in each of these species, separated by a semi-colon, with alternative identifiers for the same gene separated by a forward slash.

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. Y. Kraus, U. Technau, *Dev Genes Evol* **216**, 119 (Mar, 2006).
6. C. A. Byrum, M. Q. Martindale, in *Gastrulation: From Cells to Embryos* C. D. Stern, Ed. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2004) pp. 33-50.
7. S. Szczepanek, M. Cikala, C. N. David, *J Cell Sci* **115**, 745 (Feb 15, 2002).
8. J. R. Finnerty, D. Paulson, P. Burton, K. Pang, M. Q. Martindale, *Evol Dev* **5**, 331 (Jul-Aug, 2003).
9. E. W. Myers *et al.*, *Science* **287**, 2196 (Mar 24, 2000).
10. P. Dehal *et al.*, *Science* **298**, 2157 (Dec 13, 2002).
11. S. Aparicio *et al.*, *Science* **297**, 1301 (Aug 23, 2002).
12. J. C. Sullivan *et al.*, *Nucleic Acids Res* **34**, D495 (Jan 1, 2006).
13. E. Sodergren *et al.*, *Science* **314**, 941 (Nov 10, 2006).
14. G. A. Tuskan *et al.*, *Science* **313**, 1596 (Sep 15, 2006).
15. J. A. Darling, A. M. Reitzel, J. R. Finnerty, *Mol Ecol* **13**, 2969 (Oct, 2004).
16. A. M. Reitzel, *submitted* (2007).
17. J. S. Conery, M. Lynch, *Pac Symp Biocomput*, 167 (2001).
18. W. J. Kent, *Genome Res* **12**, 656 (Apr, 2002).
19. S. F. Altschul *et al.*, *Nucleic Acids Res* **25**, 3389 (Sep 1, 1997).
20. T. M. Lowe, S. R. Eddy, *Nucleic Acids Res* **25**, 955 (Mar 1, 1997).
21. J. Jurka *et al.*, *Cytogenet Genome Res* **110**, 462 (2005).
22. N. L. Craig, *Mobile DNA II* (ASM Press, Washington, D.C., 2002), pp. xviii, 1204 p., [1232] p. of plates.
23. V. V. Kapitonov, J. Jurka, *DNA Cell Biol* **23**, 311 (May, 2004).
24. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* **100**, 6569 (May 27, 2003).
25. V. V. Kapitonov, J. Jurka, *Proc Natl Acad Sci U S A* **103**, 4540 (Mar 21, 2006).
26. K. K. Kojima, H. Fujiwara, *Mol Biol Evol* **21**, 207 (Feb, 2004).
27. H. S. Malik, T. H. Eickbush, *Mol Biol Evol* **15**, 1123 (Sep, 1998).
28. V. V. Kapitonov, J. Jurka, *Mol Biol Evol* **20**, 38 (Jan, 2003).
29. B. a. K. Anokhin, V, *Folia Biol.* **47**, 91 (1999).
30. H. Zacharias, B. Anokhin, K. Khalturin, T. C. Bosch, *Zoology (Jena)* **107**, 219 (2004).
31. B. Anokhin, *Ann. Zool.* **52**, 475 (2002).
32. P. Tardent, in *Morphogenese der Tiere. Handbuch der ontogenetischen Morphologie und Physiologie in Einzeldarstellungen* F. Seidel, Ed. (VEB Gustav Fischer Verlag Jena, 1978) pp. 415.
33. Y. Fukui, *Biological Bulletin* **190**, 6 (2006).
34. J. Kenyon, *Evolution* **51**, 756 (1997).
35. B. a. Q. Choe, H and Song, JI, *Korean J. Biol. Sci* **4**, 103 (2000).
36. A. Smit, P. Green, (2002).
37. A. A. Salamov, V. V. Solovyev, *Genome Res* **10**, 516 (Apr, 2000).
38. E. Birney, R. Durbin, *Genome Res* **10**, 547 (Apr, 2000).
39. E. M. Zdobnov, R. Apweiler, *Bioinformatics* **17**, 847 (Sep, 2001).
40. M. Ashburner *et al.*, *Nat Genet* **25**, 25 (May, 2000).
41. E. V. Koonin *et al.*, *Genome Biol* **5**, R7 (2004).
42. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, *Nucleic Acids Res* **32**, D277 (Jan 1, 2004).
43. J. P. Kastenmayer *et al.*, *Genome Res* **16**, 365 (Mar, 2006).
44. Y. Moran, M. Gurevitz, *Febs J* **273**, 3886 (Sep, 2006).

45. R. D. Finn *et al.*, *Nucleic Acids Res* **34**, D247 (Jan 1, 2006).
46. U. Technau *et al.*, *Trends Genet* **21**, 633 (Dec, 2005).
47. R. C. Edgar, *Nucleic Acids Res* **32**, 1792 (2004).
48. J. Castresana, *Mol Biol Evol* **17**, 540 (Apr, 2000).
49. J. P. Huelsenbeck, F. Ronquist, *Bioinformatics* **17**, 754 (Aug, 2001).
50. F. Ronquist, J. P. Huelsenbeck, *Bioinformatics* **19**, 1572 (Aug 12, 2003).
51. S. Whelan, N. Goldman, *Mol Biol Evol* **18**, 691 (May, 2001).
52. J. Felsenstein. (Distributed by the author., 2004).
53. R. Dawkins, *The ancestor's tale : a pilgrimage to the dawn of evolution* (Houghton Mifflin, Boston, 2004), pp. xii, 673 p.
54. E. J. Douzery, E. A. Snell, E. Bapteste, F. Delsuc, H. Philippe, *Proc Natl Acad Sci U S A* **101**, 15386 (Oct 26, 2004).
55. R. D. Kortschak, G. Samuel, R. Saint, D. J. Miller, *Curr Biol* **13**, 2190 (Dec 16, 2003).
56. D. L. Swofford. (Sinauer Associates, Sinderland, Massachusetts, 2003).
57. A. McLysaght, K. Hokamp, K. H. Wolfe, *Nat Genet* **31**, 200 (Jun, 2002).
58. M. J. de Hoon, S. Imoto, J. Nolan, S. Miyano, *Bioinformatics* **20**, 1453 (Jun 12, 2004).
59. P. D. Thomas *et al.*, *Genome Res* **13**, 2129 (Sep, 2003).
60. H. Mi *et al.*, *Nucleic Acids Res* **33**, D284 (Jan 1, 2005).
61. T. Beissbarth, T. P. Speed, *Bioinformatics* **20**, 1464 (Jun 12, 2004).
62. S. Kumar, K. Tamura, M. Nei, *Brief Bioinform* **5**, 150 (Jun, 2004).

May 17, 2007
Supplemental figures and tables.

Figure S1.1: Nematostella development and anatomy

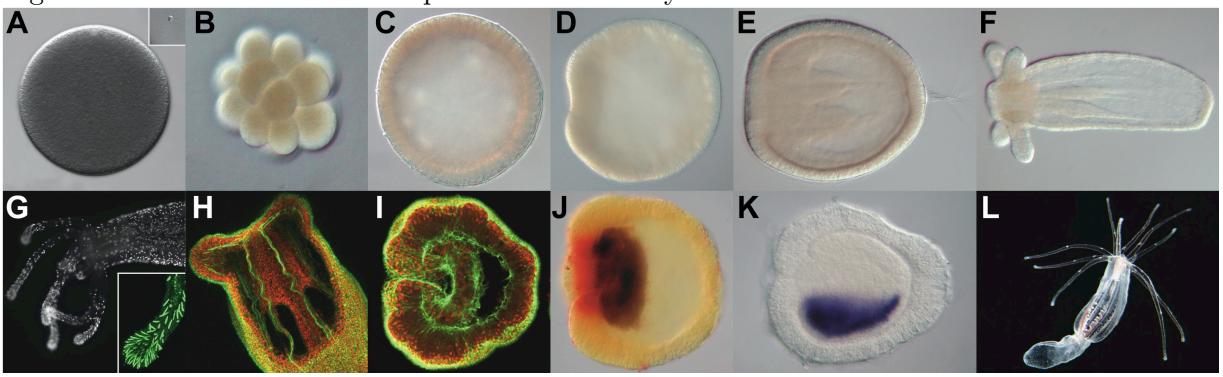

Figure S2.1: Distribution of observed polymorphism rates at 4-fold degenerate codon positions

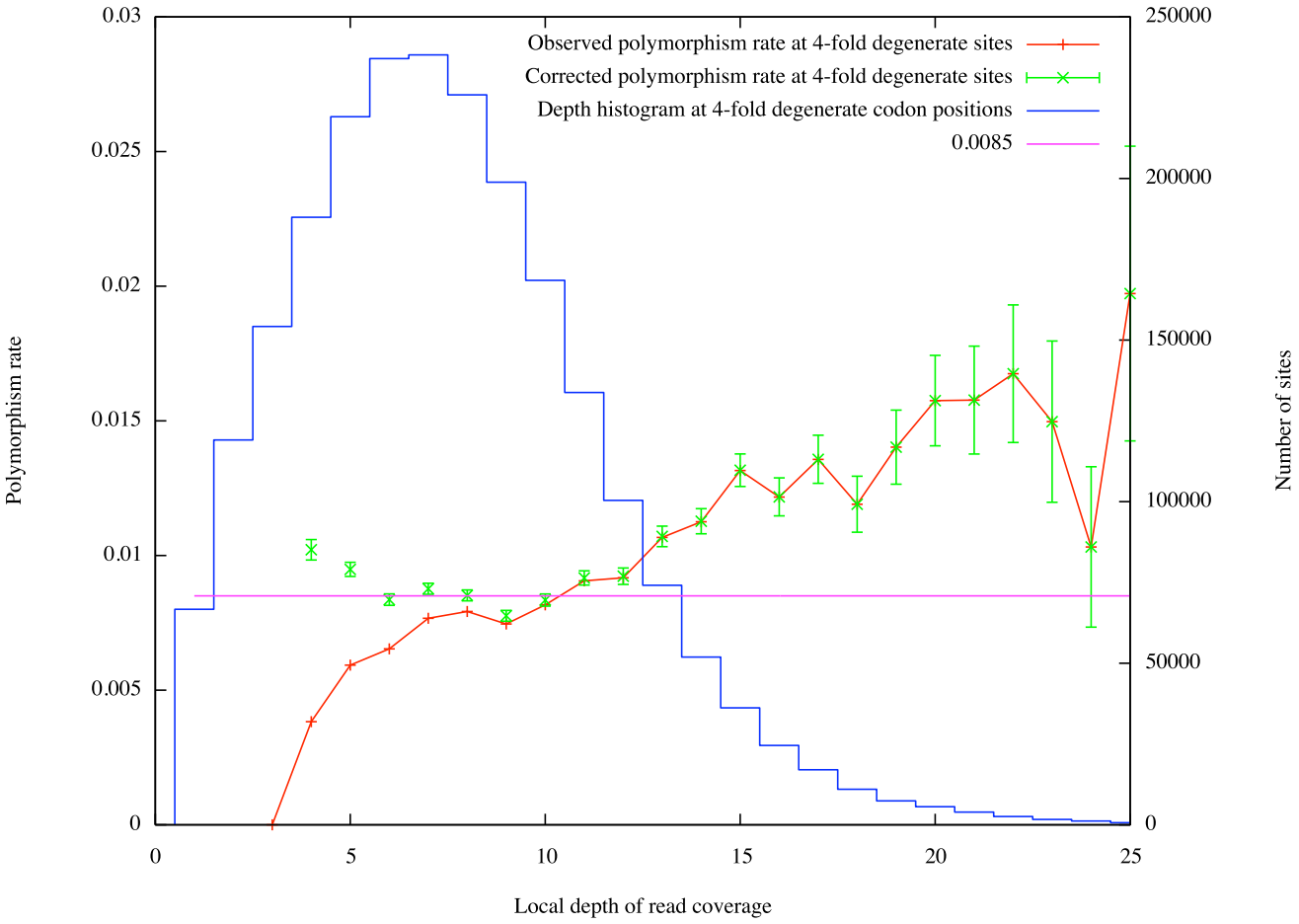

Figure S2.2: Distribution of observed polymorphism rates

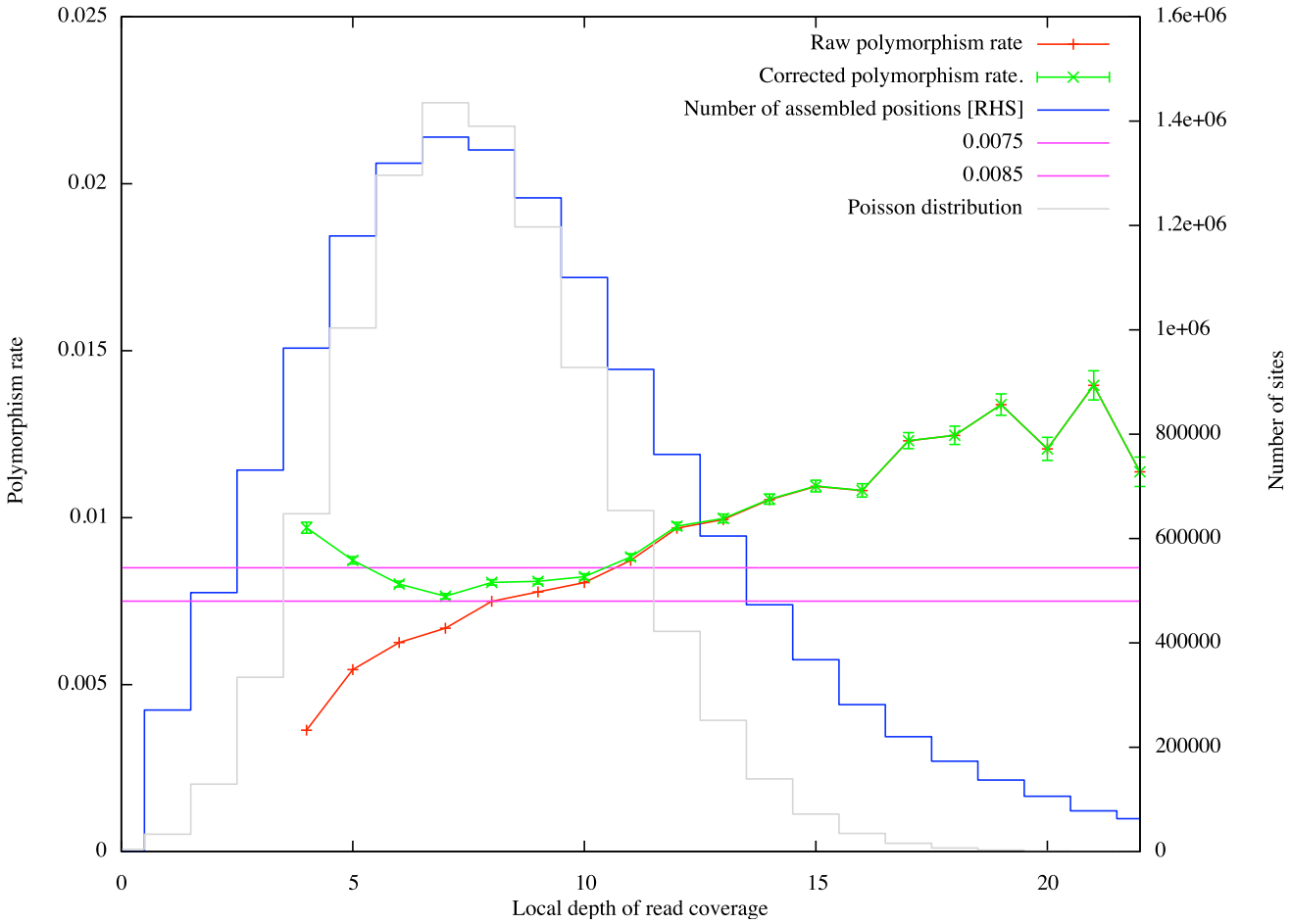

Figure S2.3: Four haplotype polymorphism fit

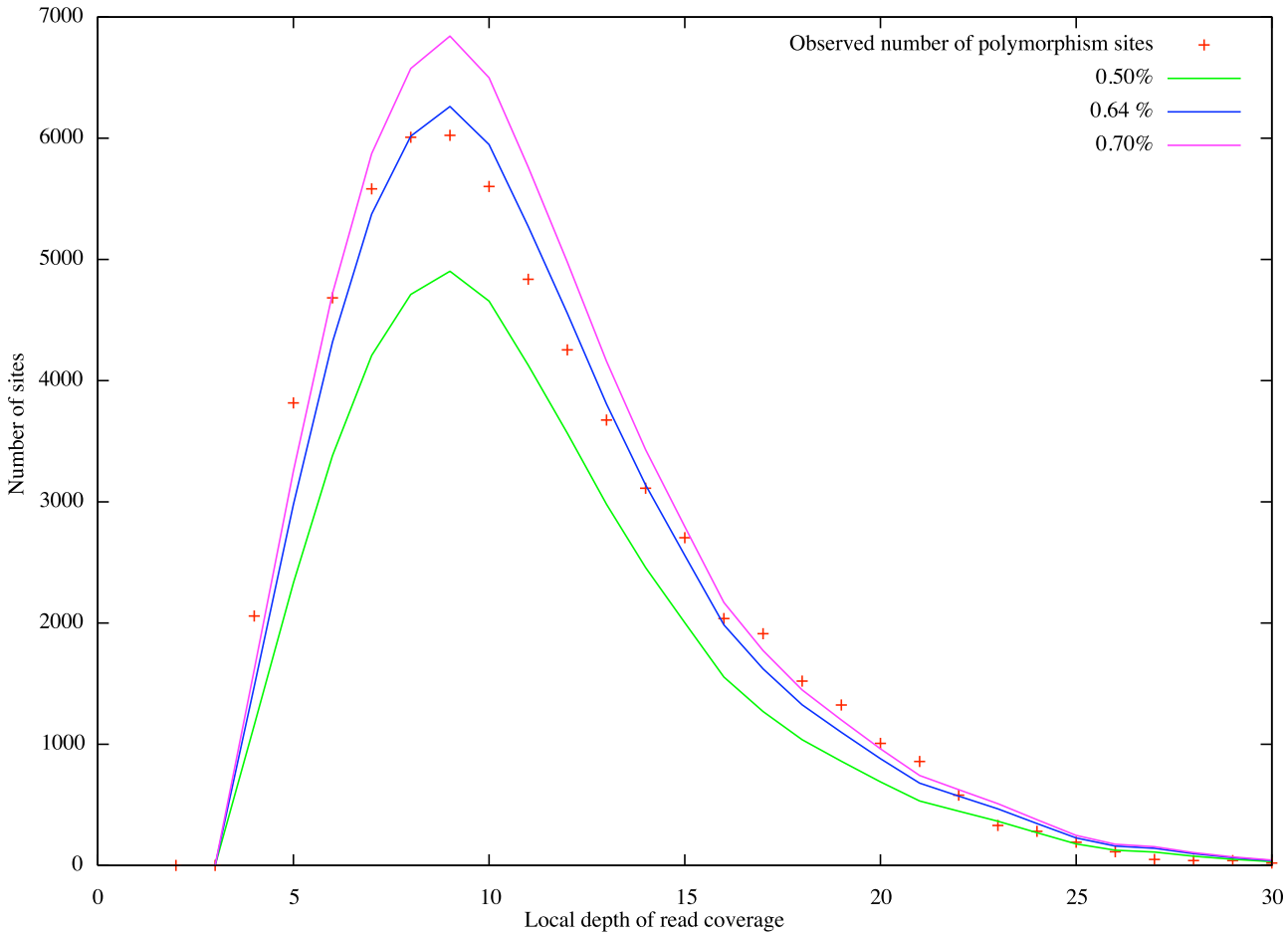

Figure S2.4: Neighbor-joining tree of eukaryotic non-LTR retrotransposons

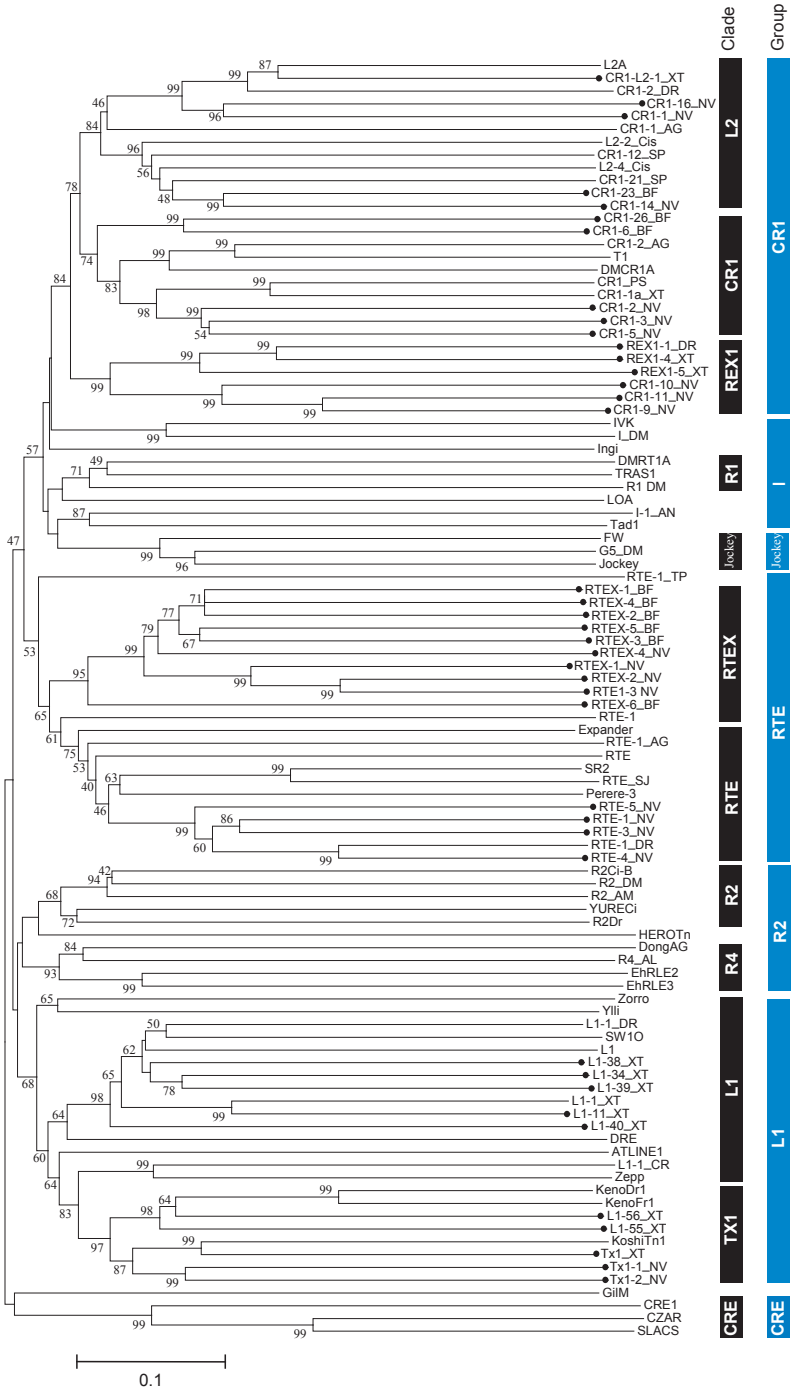

Figure S2.5: Number of chromosomes.

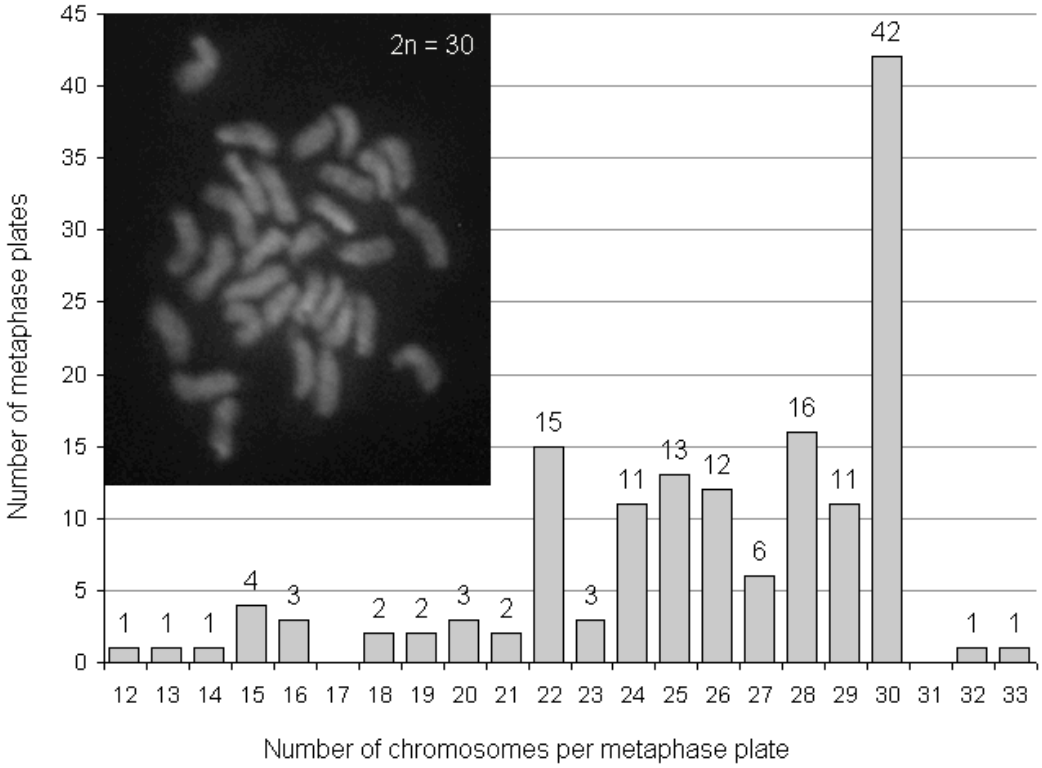

Figure S3.1: Distribution of C-scores

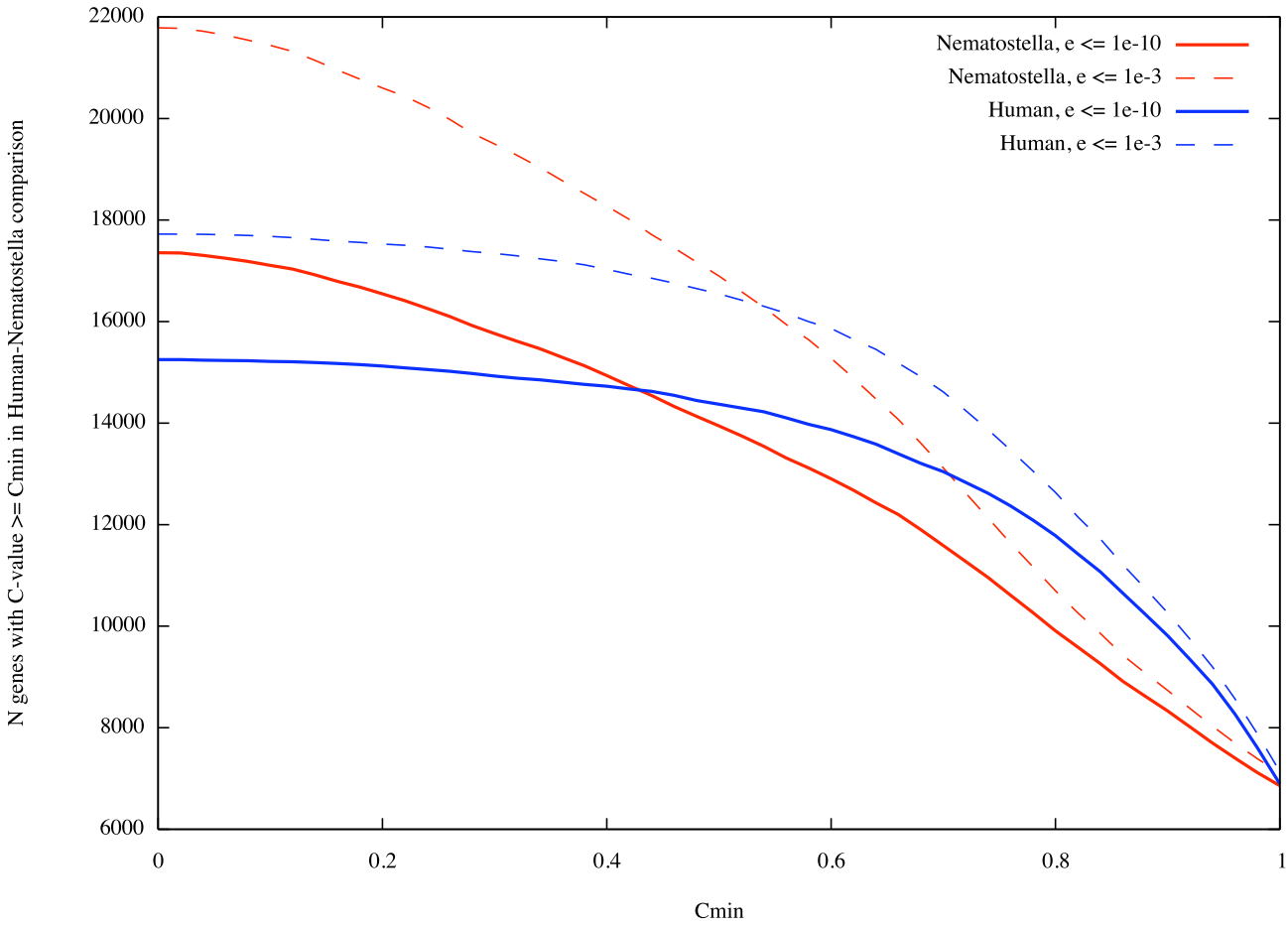

Figure S3.2: Number of bidirectional BlastP hits between 22,218 human genes and other organisms

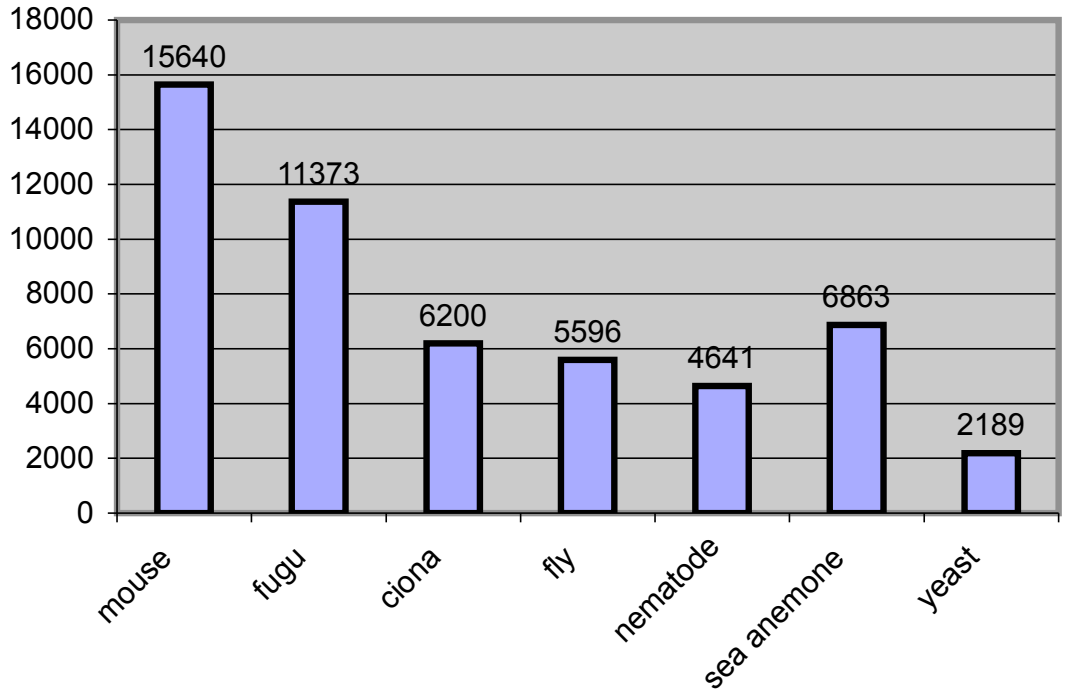

Figure S3.3: Fraction of unique multi (Pfam) domain (2 or more domains) gene models from Nematostella (total 983) shared by other metazoans and yeast.

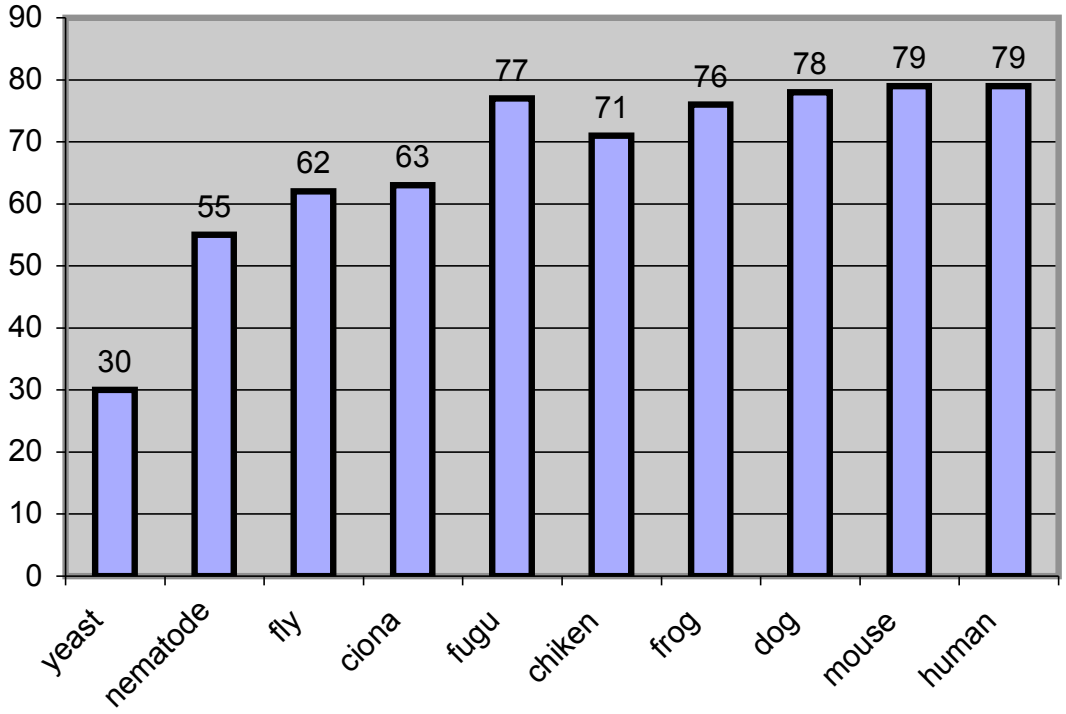

Figure S3.4: 2264 Pfam domains present in all 6 vertebrates

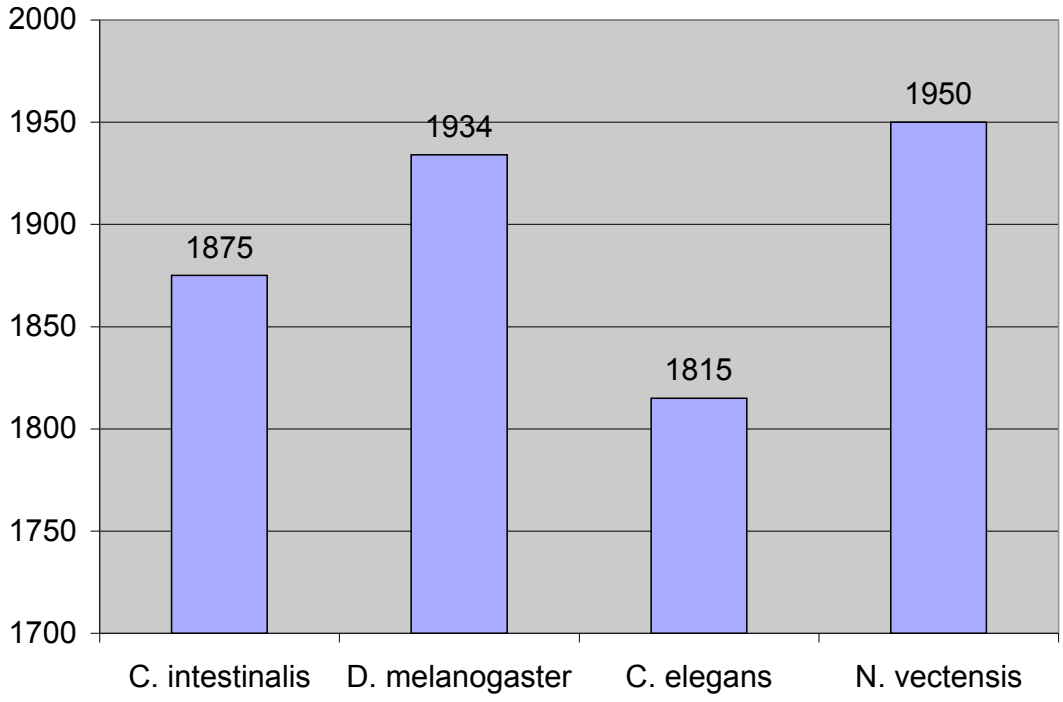

Figure S5.1: Distribution of percent ID Against Human Proteins

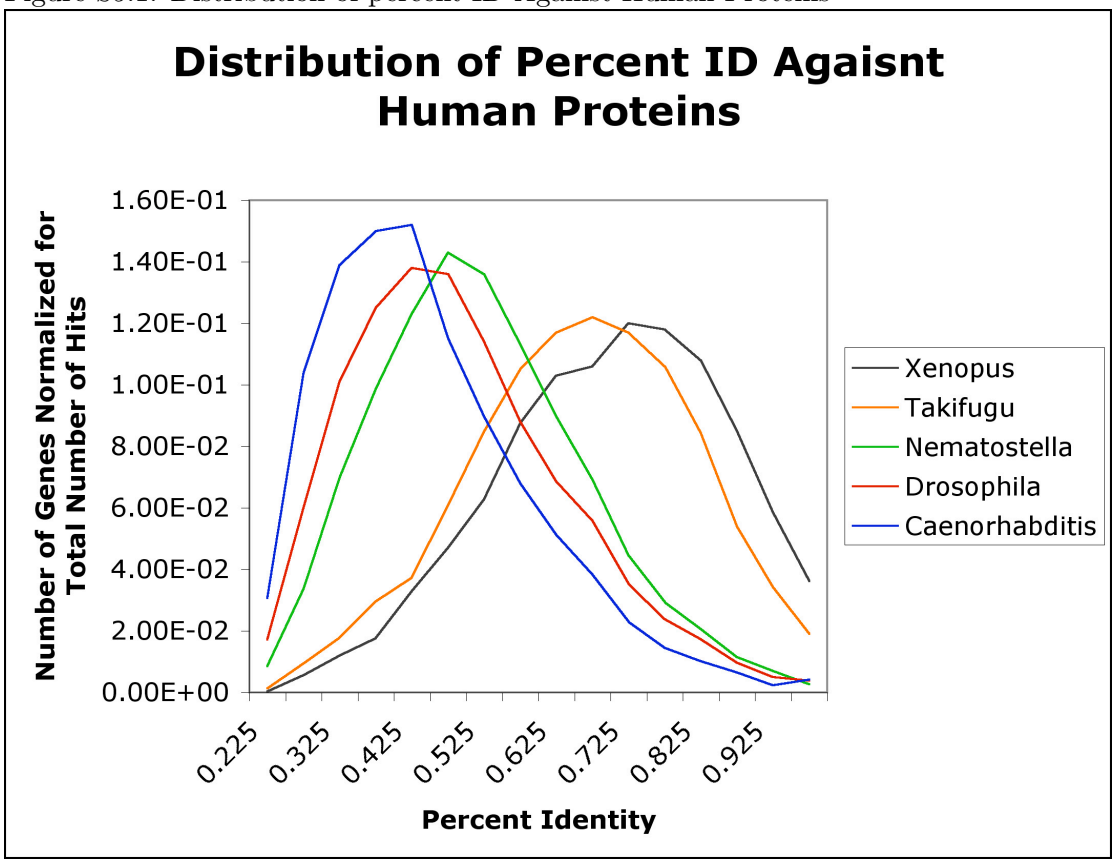

Figure S6.1: Venn diagram for three-way intron conservation comparison

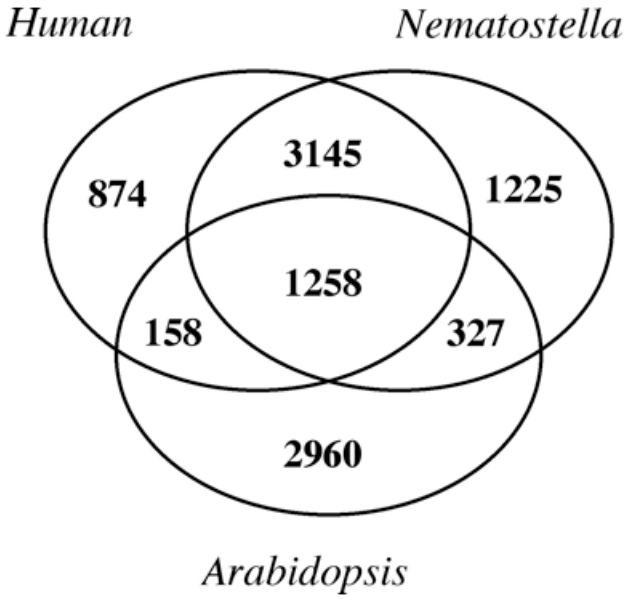

Figure S7.1: Synteny block search

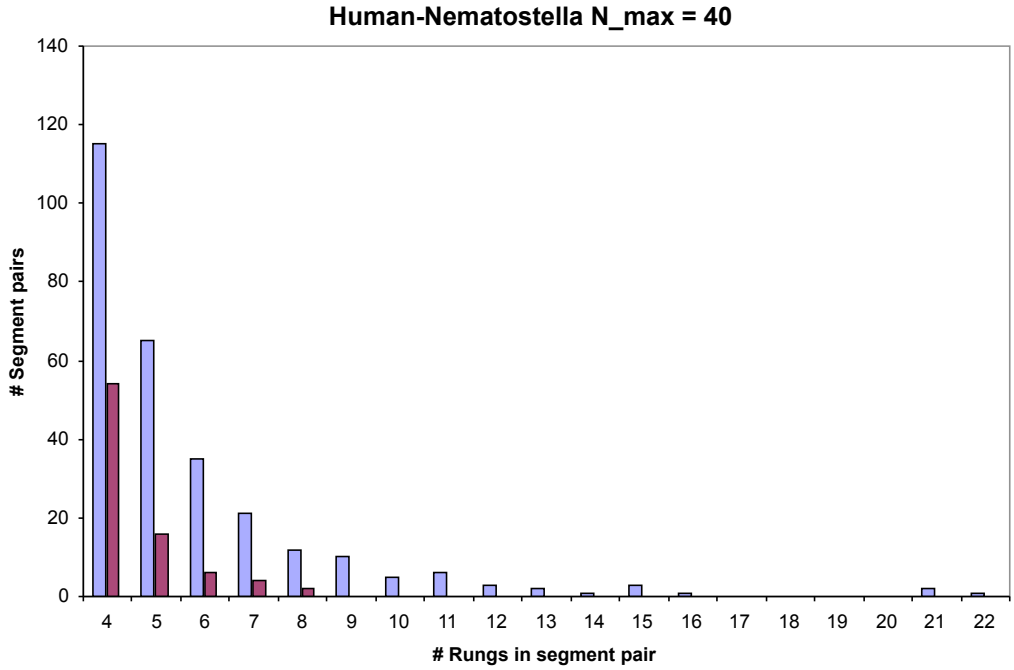

Figure S7.2: HMM segmentation example

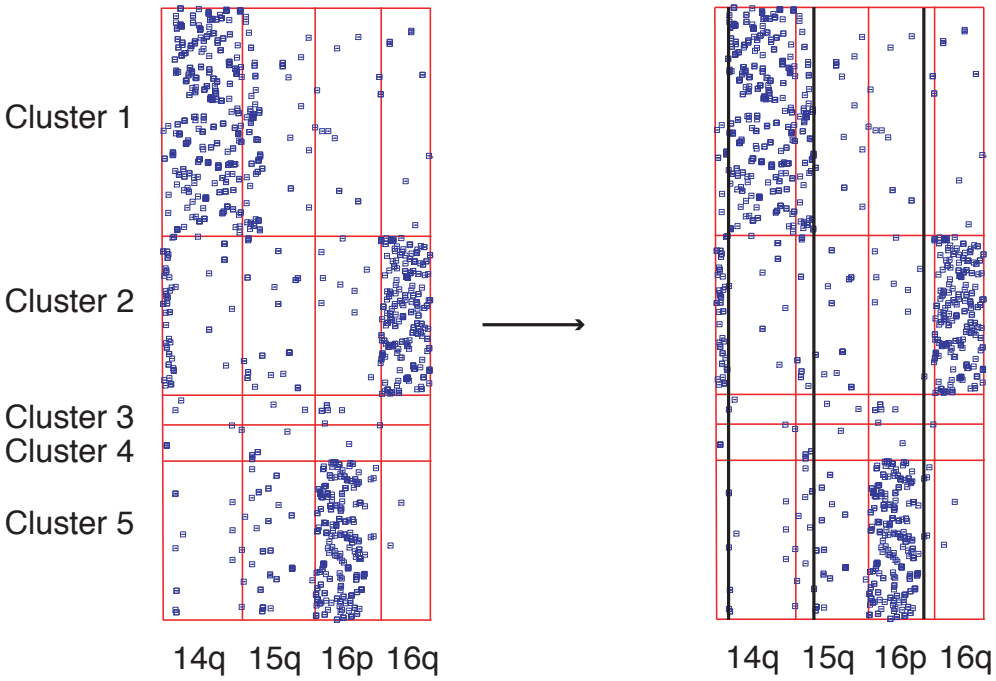

Figure S7.3: Clustering method for constructing putative ancestral linkage groups (PALs)

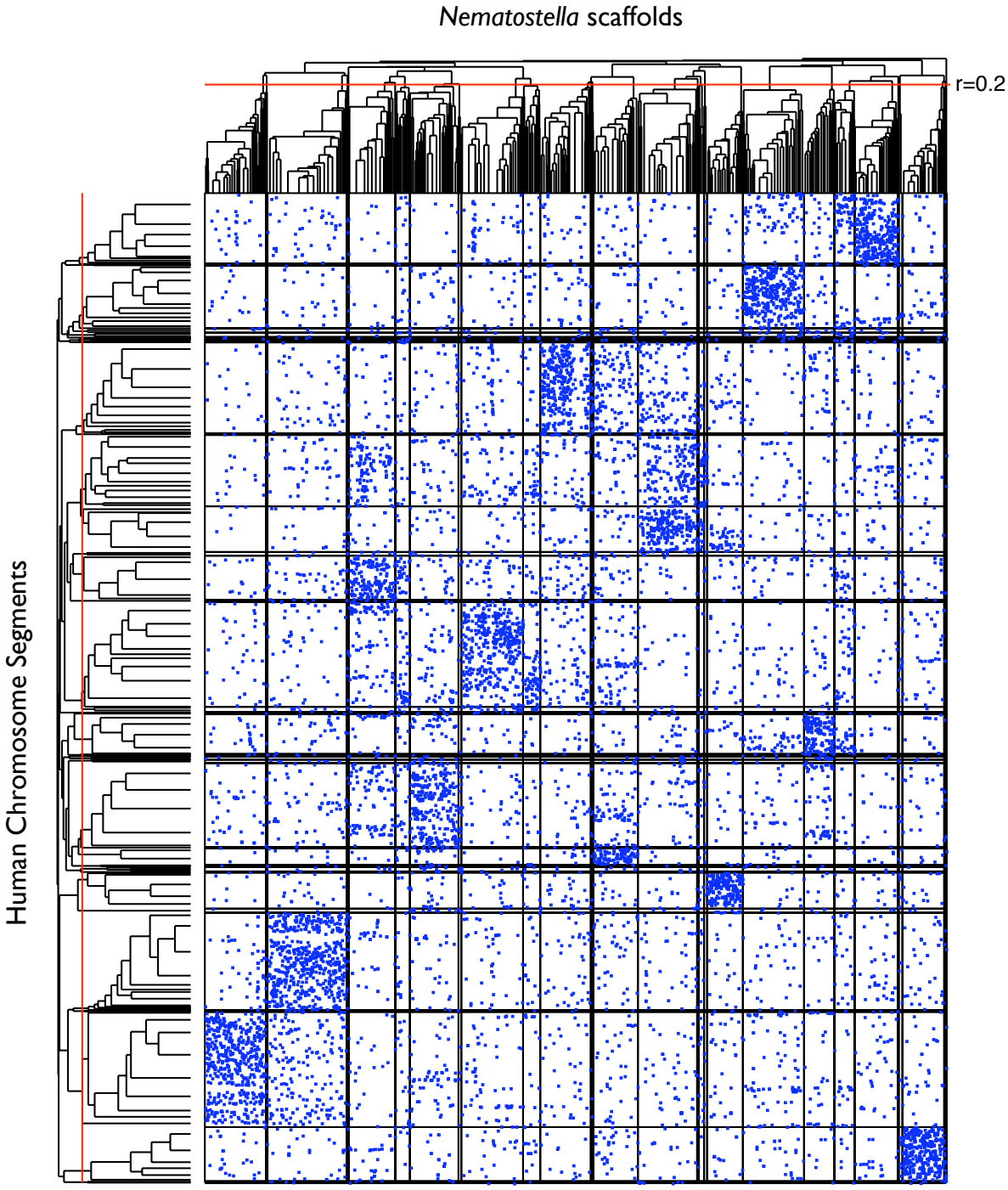

Figure S7.4: Detail of Human chromosome 12 showing genes contributing to PAL A.

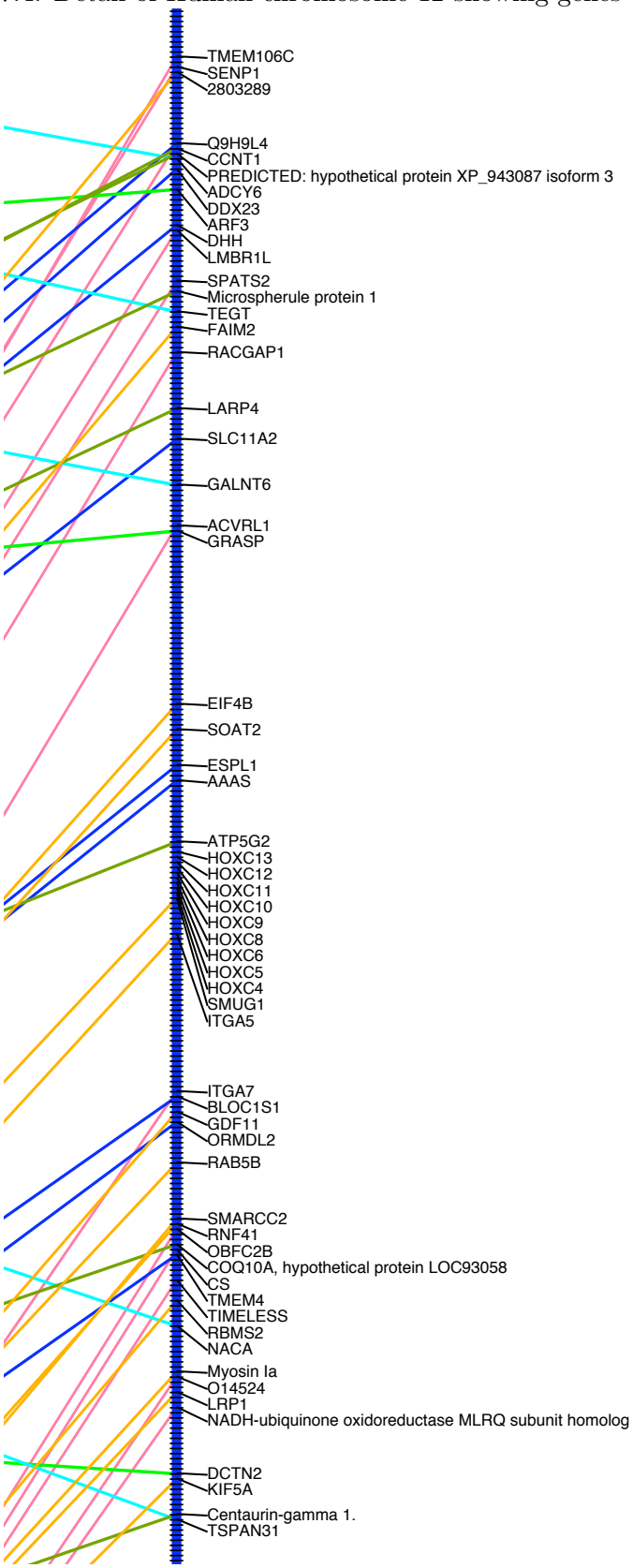

Table S1.1: Partial list of the merits of Nematostella as a model organism

Developmental Biology

- short generation time (8-10 weeks from fertilization to spawning)
- sexes are separate; sex determination is stable
- prolific sexual and asexual reproduction in the lab
- rapid regeneration and ease of generating clonal populations
- in situ hybridization protocols have been optimized

Genomic Approaches

- relatively small genome (450 million base pairs haploid size)
- most primitive living eumetazoan
- outgroup to other animal genomic models (fly, mouse, nematode)
- most developmental gene families known from other animal systems have been found
- gene families generally appear simpler (fewer members) than in bilaterians
- cDNA and BAC libraries available
- EST projects under way, along with those from other Cnidarians

Population Genetics and Ecology

- easily collected over a wide geographic range well documented since the 1950s
- representative of benthic marine invertebrates with sessile adults and planktonic larvae
- collected in both pristine and polluted sites
- native versus invasive populations may be compared
- asexual reproduction & gravid state are easily visualized in transparent animals

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%	

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**

Table S3.1: Summary of gene models

Filtered 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	4.5 kb
Average protein length	331 aa
Average exon length	208 bp
Average intron length	800 bp

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

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
PF00001	7 transmembrane receptor (rhodopsin family)	112	15.8
PF01352	KRAB box	62	14.4
PF00143	Interferon alpha/beta domain	12	13.2
PF00201	UDP-glucoronosyl and UDP-glucosyl transferase	9	10.4
PF00038	Intermediate filament protein	21	9.6
PF01500	Keratin high sulfur B2 protein	8	9.5
PF00047	Immunoglobulin domain	60	9.2
PF00048	Small cytokines (intecrine/chemokine), interleukin-8 like	13	8.9
PF00028	Cadherin domain	20	8.5
PF02841	Guanylate-binding protein, C-terminal domain	5	8.5
PF00067	Cytochrome P450	16	8.5
PF00248	Aldo/keto reductase family	8	8.2
PF00808	Histone-like transcription factor (CBF/NF-Y) and archaeal histone	16	8.2
PF02806	Alpha amylase, C-terminal all-beta domain	4	8.1
PF00125	Core histone H2A/H2B/H3/H4	19	7.8
PF07686	Immunoglobulin V-set domain	47	7.7
PF00129	Class I Histocompatibility antigen, domains alpha 1 and 2	7	7.6
PF00096	Zinc finger, C2H2 type	70	7.5
PF02798	Glutathione S-transferase, N-terminal domain	9	7.4
PF06623	MHC_I C-terminus	4	7.4
PF00128	Alpha amylase, catalytic domain	4	7.4
PF00043	Glutathione S-transferase, C-terminal domain	9	7.3
PF01454	MAGE family	9	6.8
PF02263	Guanylate-binding protein, N-terminal domain	5	6.6
PF04722	Ssu72-like protein	4	6.2
PF05831	GAGE protein	5	5.9
PF07654	Immunoglobulin C1-set domain	11	5.9
PF05296	Mammalian taste receptor protein (TAS2R)	6	5
PF02736	Myosin N-terminal SH3-like domain	4	4.6
PF06409	Nuclear pore complex interacting protein (NPIP)	4	4.6
PF00007	Cystine-knot domain	4	4.6
PF01576	Myosin tail	4	4.4
PF00622	SPRY domain	10	3.5
PF00059	Lectin C-type domain	8	3.1

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
Hydra magnipapillata
Monosiga brevicollis
Renieria spp.
Saccharomyces cerevisiae From genome-ftp.stanford.edu, version released on July 7, 2004

ESTs:

Mnemiopsis leidyi

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

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	23731462
14q12-32.33	14	23754046	105032519
15p13-q13.3	15	1	30805828
15q13.3-15.2	15	30805828	41269660
15q15.3-26.3	15	41529412	100004844
16p11.2-13.3	16	72004	27846219
16p11.2	16	28333242	31029424
16q11.2-24.3	16	45265844	88626264
17p13.2-13.3	17	621332	6608690
17p13.1-13.2	17	6608691	8299690
17p11.2-13.1	17	8299690	20458232
17q11.2-12	17	23674170	32434314
17q12-21.32	17	33964971	44369806
17q21.33-22	17	45136933	52026908
17q22-23.2	17	53308380	57331627
17q23.3-25.3	17	59268418	78471871
18q12.2-21.31	18	31310368	53429455
18q21.33-23	18	57933823	75983560
19p13.2-13.3	19	966940	9814179
19p13.11-13.2	19	10080470	18166924
19p13.11-q13.11	19	19470851	37834416
19q13.11-13.33	19	37834416	53822936
19q13.33-13.42	19	54106710	60560743
19q13.42-13.43	19	60560743	63811651
20p11.21-12.3	20	5873116	25355542
20q11.21-13.33	20	29693425	61045004
20q13.33	20	61045004	62435964
21p13-q21.3	21	1	29291902
21q21.3-22.3	21	29291902	44280308
21q22.3	21	44280308	46944323
22q11.1-12.3	22	16056470	30556650
22q12.3-13.2	22	32322586	41325288
22q13.2-13.33	22	41887065	49313184

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

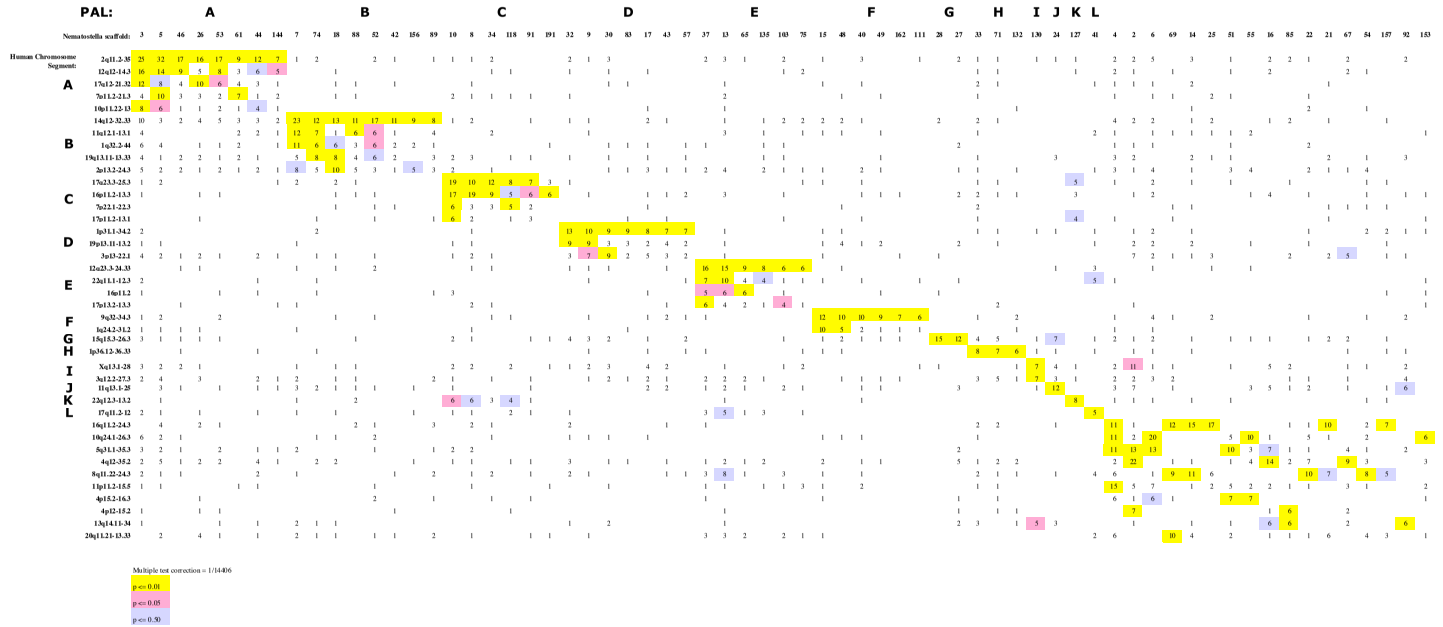

Table S7.3

Cluster ID, counts per genome	Human	*Xenopus tropicalis*	Fugu	*Drosophila*	*C. elegans*	*Nematostella*
4938402 3,3,4,1,1,1,	2771030 / ENSG00000115252; ENSG00000154678 / 2789868; 2803530 / ENSG00000123360 / PDE1B / phosphodiesterase 1B, calmodulin-dependent	1305475; 1316596; 1317690	1356856; 1357681; 1360294; 1382788	CG14940 / 2948801 / Phosphodiesterase 1c	PhosphoDiEsterase / T04D3.3 / 2969903	1731980
4938670 2,1,2,1,1,1,	ENSG00000144118 / 2770588; v-ral simian leukemia viral oncogene homolog A (ras related) / ENSG00000006451 / 2789950 / RALA	1317927	1363990; 1372881	CG2849 / Ras-related protein / 2963312	2975544 / Y53G8AR.3 / RAL (Ras-related GTPase) homolog	1744929
4938678 2,1,1,1,1,1,	ENSG00000119013 / NDUFB3 / NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 3, 12kDa / 2771148; 2806876	1322331	1366061	2953345 / CG10320	C18E9.4 / 2973219	1744378
4938691 2,1,3,2,1,1,	adenylate cyclase 5 / 2788843 / ENSG00000173175 / ADCY5; 2803310 / ADCY6 / adenylate cyclase 6 / ENSG00000174233	1308452	1363744; 1373689; 1379629	2956821 / CG32158; 2964804 / CG9533 / rutabaga	2984440 / T01C2.1 / Adenylyl CYclase	1730971
4938740 3,3,5,1,1,1,	2771369 / protein kinase, AMP-activated, gamma 3 non-catalytic subunit / ENSG00000115592 / PRKAG3; 2791141 / protein kinase, AMP-activated, gamma 2 non-catalytic subunit / ENSG00000106617 / PRKAG2; protein kinase, AMP-activated, gamma 1 non-catalytic subunit / PRKAG1 / 2803324 / ENSG00000181929	1297175; 1307663; 1321430	1359026; 1363145; 1368767; 1370004; 1374205	SNF4/AMP-activated protein kinase gamma subunit / 2960709 / CG17299	2978320 / Y111B2A.8	1740260
4938743 6,3,5,2,1,1,	2770773; 2789792 / ENSG00000122565 / CBX3 / chromobox homolog 3 (HP1 gamma homolog, Drosophila) / similar to chromobox homolog 3; 2792406; ENSG00000094916 / chromobox homolog 5 (HP1 alpha homolog, Drosophila) / 2803519 / CBX5; 2805764 / ENSG00000108468 / CBX1 / chromobox homolog 1 (HP1 beta homolog Drosophila); 2809947	1302387; 1316333; 1319398	1361004; 1361074; 1377184; 1378429; 1378758	2948138 / CG8409 / Suppressor of variegation 205; 2964023 / CG7041 / HP1b	K08H2.6 / 2991575 / HP1 Like (heterochromatin protein)	1749173
4938764 1,0,2,0,0,1,	similar to hypothetical protein / 2770581		1357902; 1376288			1740523
4938799 2,2,2,1,1,2,	ENSG00000148450 / methionine sulfoxide reductase B2 / MSRB2 / 2758437; ENSG00000174099 / 2803751 / methionine sulfoxide reductase B3 / MSRB3	1317622; 1318677	1360982; 1366407	2959004 / SelR / CG6584	2977412 / F44E2.6	1740500; 1741185
4938973 2,1,2,2,2,2,	DEAH (Asp-Glu-Ala-His) box polypeptide 16 / ENSG00000204560 / 2773981 / DHX16; DHX8 / 2805610 / ENSG00000067596 / DEAH (Asp-Glu-Ala-His) box polypeptide 8	1298697	1374669; 1376327	2949728 / CG10689; CG8241 / 2951956	EEED8.5 / Masculinisation Of Germline / 2971944; Masculinisation Of Germline / 2974673 / C04H5.6	1734659; 1744640
4938977 1,1,1,1,1,1,	ribosomal protein L27 / 2805586 / RPL27 / ENSG00000131469	1305236	1363093	CG4759 / 2961584 / Ribosomal protein L27	2966676 / Ribosomal Protein, Large subunit / C53H9.1	1734402
4939034 2,1,2,0,1,1,	ENSG00000164897 / 2791120 / transmembrane and ubiquitin-like domain containing 1; transmembrane and ubiquitin-like domain containing 2 / 2805640	1313923	1374845; 1377555		B0303.4 / 2977362	1729929
4939049 1,1,2,1,1,1,	2789817 / 3-hydroxyisobutyrate dehydrogenase / ENSG00000106049 / HIBADH	1318569	1359830; 1379797	CG15093 / 2952874	B0250.5 / 2988625	1733573
4939089 1,1,1,1,1,1,	ESPL1 / 2803481 / extra spindle poles like 1 (S. cerevisiae) / ENSG00000135476	1298527	1383019	CG10583 / 2955081 / Separase	Y47G6A.12 / SEParase / 2967022	1730767
4939127 1,1,1,0,1,2,	ANKZF1 / 2771389	1309731	1380349		2977172 / K06H7.3	1734510; 1744938
4939190 2,2,2,1,0,2,	hypothetical protein FLJ20309 / 2771243; chromosome 12 open reading frame 41 / 2803307	1313617; 1316615	1357662; 1377913	CG18041 / 2962194		1742027; 1752366
4939278 13,1,1,1,1,1,	2757876 / ENSG00000166593 / similar to 40S ribosomal protein S26; 2760603; 648199 / ENSG00000196656 / 649899 / 649912 / 2777706 / ribosomal protein S26 / 648739 / similar to 40S ribosomal protein S26 / RPS26; ENSG00000196089 / 2782195; ENSG00000137021 / 2782416 / similar to 40S ribosomal protein S26; ENSG00000170847 / 643516 / 2787460; 2789754 / ENSG00000173534; 2791672 / ENSG00000196121 / similar to 40S ribosomal protein S26; 2792573 / ENSG00000196933; 2796861 / ENSG00000142832 / similar to 40S ribosomal protein S26; 648199 / 649899 / ENSG00000197728 / 649912 / ribosomal protein S26 / 2803590 / 648739 / similar to 40S ribosomal protein S26 / RPS26; 2805701 / ENSG00000204652 / similar to 40S ribosomal protein S26; ENSG00000183462 / 2808418 / similar to 40S ribosomal protein S26 / ribosomal protein S26 pseudogene 10	1294754	1363182	2949609 / Ribosomal protein S26 / CG10305	2970203 / F39B2.6 / Ribosomal Protein, Small subunit	1733059
4939282 2,1,3,1,1,1,	ENSG00000138411 / 2771100 / HECT, C2 and WW domain containing E3 ubiquitin protein ligase 2 / HECW2; 2789964 / HECT, C2 and WW domain containing E3 ubiquitin protein ligase 1 / HECW1 / ENSG00000002746	1311473	1357911; 1365047; 1371990	2964109 / CG3099	2975554 / F45H7.6	1756526
4939408 2,2,2,1,0,1,	similar to basic leucine zipper and W2 domains 1 / 649561 / 2771137 / ENSG00000082153 / BZW1; 2789721 / ENSG00000136261 / basic leucine zipper and W2 domains 2 / BZW2	1310365; 1313284	1366966; 1381472	CG2922 / eukaryotic initiation factor 5C / 2958087		1734348
4939410 1,1,1,1,1,1,	2805569 / vacuolar protein sorting 25 homolog (S. cerevisiae) / VPS25 / ENSG00000131475	1306249	1372548	2950813 / lethal (2) 44Db / CG14750	2969759 / W02A11.2	1734600
4939448 1,0,1,1,1,1,	solute carrier family 35, member F5 / SLC35F5 / 2770558 / ENSG00000115084		1377060	CG8195 / 2952201	2967364 / B0041.5	1742307
4939478 1,1,1,0,0,1,	hypothetical protein MGC72075 / 2789775	1311162	1373294			1742055
4939559 3,2,1,0,0,1,	2789705 / transmembrane protein 106B / TMEM106B; transmembrane protein 106C / 2803282 / TMEM106C; transmembrane protein 106A / 2805602 / TMEM106A / similar to hypothetical protein MGC20235	1306746; 1321196	1368472			1745116
4939637 1,1,1,0,1,1,	ENSG00000123607 / tetratricopeptide repeat domain 21B / 2770855 / TTC21B	1299651	1380282		ZK328.7a / 2976464	1734404

Cluster ID, counts per genome	Human	*Xenopus tropicalis*	Fugu	*Drosophila*	*C. elegans*	*Nematostella*
4939682 4,5,7,1,1,2,	ENSG00000078114 / 2758389; ENSG00000197893 / NRAP / 2760326; ENSG00000183091 / nebulin / NEB / 2770758; 2805366 / ENSG00000002834	1303359; 1303623; 1312588; 1314457; 1315214	1358432; 1368571; 1371274; 1372010; 1373014; 1376961; 1380966	2956857 / CG3849 / Lasp	F42H10.3 / 2977315	1733759; 1740918
4939696 1,1,1,1,0,1,	ENSG00000123415 / SMUG1 / 2803517 / single-strand-selective monofunctional uracil-DNA glycosylase 1	1316938	1373983	2960104 / CG5285		1742638
4939823 1,1,5,1,0,2,	2771426 / similar to RIKEN cDNA A230078I05 gene	1303281	1358971; 1364842; 1365540; 1375078; 1377059	CG14234 / 2965658		1731311; 1739330
4939862 1,1,1,0,1,1,	2803683 / ENSG00000135506 / amplified in osteosarcoma	1299600	1371568		F48E8.4 / 2976228	1731977
4939876 3,1,1,1,1,1,	similar to nascent polypeptide-associated complex alpha polypeptide / ENSG00000121089 / 2777953; NACA / 2803628 / ENSG00000196531 / nascent-polypeptide-associated complex alpha polypeptide; 2805966 / NACAL / ENSG00000196861 / nascent-polypeptide-associated complex alpha polypeptide-like	1310333	1358674	2951715 / CG8759 / Nascent polypeptide associated complex protein alpha subunit	Y65B4BR.5a / Y65B4BR.5 / 2966478	1751203
4939909 1,1,2,1,0,1,	GPR155 / ENSG00000163328 / 2770949 / G protein-coupled receptor 155	1303861	1366877; 1375972	2957008 / CG7510		1729901
4940295 1,1,2,1,1,1,	ENSG00000136758 / 2758515 / YME1L1 / YME1-like 1 (S. cerevisiae)	1308114	1375674; 1378256	CG3499 / 2953592	2977886 / M03C11.5	1742693
4940326 1,1,1,1,1,1,	2803617 / timeless homolog (Drosophila) / TIMELESS / ENSG00000111602	1302169	1374167	timeout / CG7855 / 2959415	2978265 / Y75B8A.22 / TIMeless (Drosophila/mammal) related	1741013
4940369 1,1,1,1,0,1,	2758467 / chromosome 10 open reading frame 63 / ENSG00000151023	1313655	1372407	2954585 / CG16984		1742835
4940377 1,3,2,0,2,1,	2771066 / SLC40A1 / ENSG00000138449 / solute carrier family 40 (iron-regulated transporter), member 1	1297970; 1317283; 1321612	1359225; 1363964		2966726 / Y37E3.16; 2983223 / R09B5.4	1741095
4940402 4,3,5,0,0,1,	ENSG00000144583 / 2771330 / MARCH4 / membrane-associated ring finger (C3HC4) 4; 2778213; 2778214; 2803690 / MARCH9 / membrane-associated ring finger (C3HC4) 9 / ENSG00000139266	1311588; 1321222; 1321345	1359054; 1360167; 1374831; 1379417; 1379510			1749052
4940403 3,3,3,0,0,1,	2771383 / chromosome 2 open reading frame 17; 2778217 / hypothetical protein FLJ20152; hypothetical protein LOC162427 / 2805558	1299409; 1309676; 1317778	1360161; 1370794; 1377485			1745266
4940429 1,1,1,1,0,1,	2803342 / hypothetical protein FLJ13236	1307024	1379445	wurst / 2965243 / CG9089		1733507
4940441 1,1,0,0,0,1,	2771176 / amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 12 / ENSG00000155749 / ALS2CR12	1296660				1751145
4940444 5,6,6,1,2,2,	2770716 / ENSG00000115850 / lactase / LCT; 2772381 / ENSG00000188501 / lactase-like / LCTL; 2777008 / ENSG00000176201 / glucosidase, beta, acid 3 (cytosolic) / GBA3; KLB / ENSG00000134962 / klotho beta / 2777084; ENSG00000133116 / klotho / 2786842 / KL	1299827; 1305138; 1311543; 1311768; 1311878; 1312039	1358515; 1362228; 1365496; 1367713; 1372605; 1378485	CG9701 / 2956866	2975366 / E02H9.5; 2980526 / C50F7.10	1746559; 1751543
4940507 2,2,3,1,1,1,	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase 1 / 2771357 / ENSG00000144579 / CTDSP1; CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) small phosphatase-like / CTDSPL / 2788032 / ENSG00000144677	1305749; 1315790	1357485; 1361761; 1364142	CG5830 / 2956688	Temporarily Assigned Gene name / 2969084 / B0379.4	1750922
4940572 6,4,7,1,2,2,	MYL1 / 2771293 / ENSG00000168530 / myosin, light polypeptide 1, alkali; skeletal, fast; ENSG00000160808 / myosin, light polypeptide 3, alkali; ventricular, skeletal, slow / MYL3 / 2788180; 2803595 / ENSG00000196465 / myosin, light polypeptide 6B, alkali, smooth muscle and non-muscle; ENSG00000092841 / 2803598; 2805738 / myosin, light polypeptide 4, alkali; atrial, embryonic / ENSG00000198336 / MYL4; 2806096 / ENSG00000187141 / similar to Myosin light polypeptide 6 (Myosin light chain alkali 3) (Myosin light chain 3) (MLC-3) (LC17)	1308206; 1309772; 1312717; 1316858	1357946; 1362578; 1362676; 1367874; 1378261; 1378822; 1383313	Myosin light chain cytoplasmic / CG3201 / 2963520	T12D8.6 / 2978553; 2991785 / K04C1.4	1734630; 1748253
4940652 1,1,1,0,1,1,	DNPEP / aspartyl aminopeptidase / 2771401 / ENSG00000123992	1321395	1357286		2976380 / F01F1.9	1744225
4940740 1,1,2,1,1,1,	2770940 / ENSG00000138430 / GTP-binding protein 9 (putative)	1317341	1366791; 1374309	CG1354 / 2964122	Temporarily Assigned Gene name / W08E3.3 / 2969910	1742723
4940788 1,1,1,1,1,1,	2771331 / SMARCAL1 / ENSG00000138375 / SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a-like 1	1321150	1376936	CG3753 / 2947548 / Marcal1	2976614 / C16A3.1 / C16A3.1a	1735999
4940804 1,1,1,1,1,1,	2805359 / proteasome (prosome, macropain) subunit, beta type, 3 / PSMB3 / ENSG00000108294	1314222	1370606	2958646 / CG11981 / Pros3	2971778 / Y38A8.2	1746548
4940839 1,1,1,1,1,1,	ENSG00000144366 / GULP, engulfment adaptor PTB domain containing 1 / 2771059 / GULP1	1309437	1364668	CG11804 / ced-6 / 2950997	2976291 / F56D2.7 / CEll Death abnormality	1737757
4941003 2,2,4,1,1,1,	ENSG00000198586 / 2770905; 2805974 / TLK2 / ENSG00000146872	1314455; 1317722	1364396; 1371257; 1373038; 1381555	CG32782 / Tousled-like kinase / 2963317	C07A9.3 / 2977675 / Tousled-Like Kinase	1731975
4941037 1,1,1,1,1,1,	ENSG00000115839 / RAB3GAP1 / 2770708 / RAB3 GTPase activating protein subunit 1 (catalytic)	1302292	1366490	2947012 / CG31935	2991930 / RaB GAP related / F20D1.6	1732709
4941122 2,2,3,1,1,2,	ADP-ribosylation factor-like 5B / ENSG00000165997 / 2758377 / ARL5B; ENSG00000162980 / 2770759 / ARL5A / ADP-ribosylation factor-like 5A	1298595; 1321228	1360746; 1367974; 1374117	CG7197 / 2955506	ZK632.8 / 2977718 / ARF-Like	1735111; 1745542

Cluster ID, counts per genome	**Human**	***Xenopus tropicalis***	**Fugu**	***Drosophila***	***C. elegans***	***Nematostella***
4941199 4,5,4,1,2,3,	LON peptidase N-terminal domain and ring finger 2 / 2770393 / LONRF2; ENSG00000175556 / 2793209 / LON peptidase N-terminal domain and ring finger 3 / LONRF3; BRCA1 / ENSG00000012048 / breast cancer 1, early onset / 2805591; LON peptidase N-terminal domain and ring finger 1 / LONRF1 / 2807782	1304004; 1305020; 1313046; 1315369; 1318207	1367210; 1369428; 1370201; 1382750	CG32369 / 2955419	2975677 / C36A4.8 / BRCa homolog (tumor suppressor gene Brca1); 2977538 / T02C1.1	1741472; 1754072; 1756579
4941231 4,2,4,1,0,1,	2788671; ENSG00000189043 / NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa / NDUFA4 / 2789702; 2797659; ENSG00000185633 / NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4-like 2 / 2803653	1304270; 1304912	1360316; 1361534; 1361683; 1374755	2957822 / CG32230		1747744
4941343 2,2,3,1,1,1,	2770738 / activin A receptor, type IIA / ENSG00000121989 / ACVR2A; ENSG00000114739 / activin A receptor, type IIB / 2788050 / ACVR2B	1297169; 1299490	1360785; 1378209; 1379131	CG7904 / 2959655 / punt	2976300 / C05D2.1 / abnormal DAuer Formation	1735192
4941370 5,4,6,2,1,2,	ENSG00000107779 / bone morphogenetic protein receptor, type IA / BMPR1A / 2759505; activin A receptor, type I / ACVR1 / ENSG00000115170 / 2770784; 2777582 / ENSG00000138696 / bone morphogenetic protein receptor, type IB / BMPR1B; 2803415 / ACVRL1 / activin A receptor type II-like 1 / ENSG00000139567; 2811414	1298144; 1301514; 1311375; 1320416	1359360; 1360825; 1364182; 1366372; 1370510; 1380211	CG14026 / 2947619 / thickveins; saxophone / 2950642 / CG1891	C32D5.2 / 2972303 / SMAll	1733787; 1747335
4941399 1,1,1,1,1,1,	pyridoxamine 5'-phosphate oxidase / PNPO / 2805757 / ENSG00000108439	1320817	1359177	CG31472 / 2958447	2976796 / F57B9.1	1737704
4941437 2,2,3,1,1,1,	cyclin T2 / CCNT2 / ENSG00000082258 / 2770703; ENSG00000129315 / CCNT1 / 2803308 / cyclin T1	1299785; 1321220	1357903; 1377932; 1378335	CG6292 / Cyclin T / 2957002	2977141 / CyclIn T / F44B9.4	1749344
4941503 1,1,1,1,1,1,	ENSG00000115524 / SF3B1 / 2771110 / splicing factor 3b, subunit 1, 155kDa	1314372	1358684	2946872 / CG2807	2975795 / T08A11.2	1734339
4941576 6,3,3,2,2,3,	2770908; 2787898 / ENSG00000206562 / methyltransferase like 6 / METTL6; 2790177 / NSUN5C / NOL1/NOP2/Sun domain family, member 5C; NOL1/NOP2/Sun domain family, member 5 / NSUN5 / 2790192 / ENSG00000130305; ENSG00000165055 / 2790837 / METTL2B; 2805972 / ENSG00000087995 / methyltransferase like 2A	1317752; 1317753; 1322519	1377994; 1378609; 1379844	methyltransferase-like / CG13929 / 2954419; 2960407 / CG5558	2974753 / Y53F4B.4b / Y53F4B.4; 2975702 / ZK1058.5	1731004; 1749873; 1753358
4941595 2,3,2,3,2,3,	SSB / 2770885 / Sjogren syndrome antigen B (autoantigen La) / ENSG00000138385; ENSG00000144357 / ZNF650 / zinc finger protein 650 / 2770888	1297371; 1304956; 1320605	1357666; 1367700	CG10922 / La autoantigen-like / 2949895; CG1530 / 2963852; 2963855 / CG1531	2967351 / C44E4.4; 2971684 / F10G7.10 / F10G7.10a	1739272; 1748317; 1749587
4941686 3,2,1,1,0,1,	ENSG00000114784 / eukaryotic translation initiation factor 1B / EIF1B / 2788075; similar to suppressor of initiator codon mutations, related sequence 1 / 2800229 / ENSG00000198747; 2805513 / eukaryotic translation initiation factor 1 / ENSG00000173812	1298887; 1319819	1369271	2954711 / CG17737		1729898
4941802 3,3,4,0,1,1,	GRB14 / 2770836 / ENSG00000115290 / growth factor receptor-bound protein 14; ENSG00000106070 / 2790066 / GRB10 / growth factor receptor-bound protein 10; 2805393 / ENSG00000141738 / GRB7 / growth factor receptor-bound protein 7	1297266; 1307488; 1319070	1365576; 1369665; 1377574; 1379357		2978362 / Y37D8A.4	1737654
4941806 2,2,2,1,1,1,	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 1 / 2788204 / ENSG00000173473 / SMARCC1; SMARCC2 / ENSG00000139613 / SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily c, member 2 / 2803599	1296443; 1297764	1365121; 1378262	CG18740 / moira / 2959908	Y113G7B.23 / Phasmid Socket Absent / 2988579	1743259
4941889 1,1,0,1,1,1,	cell division cycle 27 / CDC27 / ENSG00000004897 / similar to Cell division cycle protein 27 homolog (CDC27Hs) (H-NUC) / 2805736	1318609		2955312 / CG8610 / Cdc27	Metaphase-to-Anaphase Transition defect / Y110A7A.17 / 2967497	1748109
4941910 3,3,5,1,1,1,	2758602 / kinesin family member 5B / KIF5B / ENSG00000170759; kinesin family member 5C / KIF5C / ENSG00000168280 / 2770744; ENSG00000155980 / kinesin family member 5A / KIF5A / 2803671	1301325; 1307942; 1311514	1363233; 1368358; 1371716; 1373855; 1378379	CG7765 / 2952350 / Kinesin heavy chain	2977277 / UNCoordinated / R05D3.7	1737753
4941984 4,4,6,1,2,1,	2770866 / LASS6 / LAG1 longevity assurance homolog 6 (S. cerevisiae) / ENSG00000172292; ENSG00000090661 / LAG1 longevity assurance homolog 4 (S. cerevisiae) / LASS4 / 2780181; ENSG00000143418 / LAG1 longevity assurance homolog 2 (S. cerevisiae) / 2798542 / LASS2; LAG1 longevity assurance homolog 5 (S. cerevisiae) / 2803373 / ENSG00000139624 / LASS5	1304431; 1306318; 1307411; 1311041	1360285; 1364425; 1367203; 1367313; 1372742; 1377007	Longevity assurance gene 1 / 2963619 / CG3576	2980802 / C09G4.1 / Homolog of Yeast Longevity gene; K02G10.6 / 2989688 / Homolog of Yeast Longevity gene	1740387
4942029 2,2,2,1,0,1,	ALS2 / 2771186 / amyotrophic lateral sclerosis 2 (juvenile) / ENSG00000003393; ENSG00000178038 / ALS2CL / 2788172 / ALS2 C-terminal like	1294945; 1314398	1373942; 1378173	CG7158 / 2957660		1737250
4942105 4,2,4,3,1,1,	2762436 / ENSG00000124216 / snail homolog 1 (Drosophila) / SNAI1; 2771284; snail homolog 3 (Drosophila) / SNAI3 / ENSG00000185669 / 2786480; SNAI2 / snail homolog 2 (Drosophila) / ENSG00000019549 / 2808199	1300942; 1314715	1360874; 1378803; 1382121; 1382239	2949283 / CG3758 / escargot; worniu / 2949287 / CG4158; snail / 2949289 / CG3956	2978673 / K02D7.2	1737486
4942339 2,1,2,0,1,1,	DNA polymerase-transactivated protein 6 / 2771129; 2803343 / spermatogenesis associated, serine-rich 2 / ENSG00000123352 / SPATS2	1307691	1359904; 1381477		2966910 / Y71F9AL.9	1744268
4942393 1,2,2,1,2,2,	2770755 / ENSG00000080345	1321752; 1321794	1363923; 1364937	2952239 / CG30085	2982867 / F11E6.7; 2991430 / F19H6.3	1731155; 1752061
4942500 1,1,1,1,1,1,	2771105 / ENSG00000197121 / GPI deacylase	1311374	1372104	CG3160 / 2963537	2986198 / T19B10.8	1734671
4942730 1,1,1,1,1,1,	2803366 / RACGAP1 / Rac GTPase activating protein 1 / ENSG00000161800	1306316	1375409	CG13345 / 2951913 / RacGAP50C	K08E3.6 / 2978601 / CYtoKinesis defect	1744521
4942845 2,2,1,0,0,1,	LanC lantibiotic synthetase component C-like 1 (bacterial) / LANCL1 / ENSG00000115365 / 2771294; LANCL2 / LanC lantibiotic synthetase component C-like 2 (bacterial) / 2790079 / ENSG00000132434	1308943; 1318967	1362499			1733785

Cluster ID, counts per genome	**Human**	***Xenopus tropicalis***	**Fugu**	***Drosophila***	***C. elegans***	***Nematostella***
4943219 1,1,1,1,1,1,	dephospho-CoA kinase domain containing / 2805677 / DCAKD	1320072	1371007	2958345 / CG1939	2977688 / T05G5.5	1738260
4943222 5,4,6,1,1,2,	ALS2CR7 / ENSG00000138395 / amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 7 / 2771187; ENSG00000058091 / 2790393 / PFTK1; ENSG00000102225 / PCTAIRE protein kinase 1 / PCTK1 / 2791920; ENSG00000117266 / PCTAIRE protein kinase 3 / 2800747; ENSG00000059758 / 2803958 / PCTAIRE protein kinase 2 / PCTK2	1299715; 1300079; 1303196; 1304621	1358714; 1363047; 1367296; 1367945; 1373759; 1378162	Ecdysone-induced protein 63E / CG10579 / 2954774	PCTAIRE class cell cycle kinase / 2980672 / C07G1.3	1729730; 1747956
4943372 5,5,6,2,2,2,	ENSG00000177283 / frizzled homolog 8 (Drosophila) / FZD8 / 2758689; ENSG00000155760 / frizzled homolog 7 (Drosophila) / FZD7 / 2771189; 2771267 / FZD5 / frizzled homolog 5 (Drosophila) / ENSG00000163251; 2790396 / frizzled homolog 1 (Drosophila) / ENSG00000157240 / FZD1; 2805657 / FZD2 / frizzled homolog 2 (Drosophila) / ENSG00000180340	1300695; 1306024; 1311370; 1315221; 1315282	1357320; 1362213; 1363087; 1369501; 1373946; 1378167	2956461 / frizzled / CG17697; CG9739 / 2957176 / frizzled 2	T23D8.1 / More Of MS / 2969054; Caenorhabditis FriZzled homolog / F27E11.3 / 2983810	1732254; 1739871
4943423 2,2,2,1,1,1,	p300/CBP-associated factor / PCAF / 2787933 / ENSG00000114166; ENSG00000108773 / 2805537 / GCN5L2 / GCN5 general control of amino-acid synthesis 5-like 2 (yeast)	1296829; 1314947	1372380; 1376754	CG4107 / Pcaf / 2956241	2967039 / P300/CBP Associated Factor homolog / Y47G6A.6	1741240
4943589 3,2,3,1,0,1,	2771072 / ENSG00000128699 / ORMDL1 / ORM1-like 1 (S. cerevisiae); 2803570 / ORMDL2 / ORM1-like 2 (S. cerevisiae) / ENSG00000123353; 2805411 / ORMDL3 / ORM1-like 3 (S. cerevisiae) / ENSG00000172057	1297320; 1312687	1372043; 1376537; 1381782	ORMDL / 2957652 / CG14577		1730731
4943613 1,1,0,1,1,1,	2803390 / LETM1 domain containing 1 / ENSG00000050426 / LETMD1	1312580		2955577 / CG5989	2968412 / F30F8.9 / F30F8.9a	1756286
4943644 4,5,9,4,2,1,	DPP10 / 2770560 / ENSG00000175497 / dipeptidyl-peptidase 10; ENSG00000197635 / 2770827 / dipeptidyl-peptidase 4 (CD26, adenosine deaminase complexing protein 2) / DPP4; ENSG00000078098 / FAP / 2770830 / fibroblast activation protein, alpha; 2791157 / ENSG00000130226 / dipeptidyl-peptidase 6 / DPP6	1303698; 1304226; 1305147; 1312429; 1319817	1357929; 1359926; 1368146; 1371157; 1372099; 1376256; 1379753; 1379802; 1380993	2947698 / CG11034; CG11319 / 2947840; CG32145 / 2956496 / omega; 2965249 / CG9059	2987441 / T23F1.7 / Dipeptidyl Peptidase Four (IV) family; 2991897 / C27C12.7 / Dipeptidyl Peptidase Four (IV) family	1748696
4943754 4,3,6,1,1,1,	region containing chromosome 2 open reading frame 12; RNA binding motif, single stranded interacting protein 1 / 2770820 / RNA binding motif, single stranded interacting protein 1 / RBMS1 / ENSG00000153250; RNA binding motif, single stranded interacting protein / ENSG00000144642 / 2787975 / RBMS3; ENSG00000076067 / RNA binding motif, single stranded interacting protein 2 / RBMS2 / 2803622; region containing chromosome 2 open reading frame 12; RNA binding motif, single stranded interacting protein 1 / 2803759 / ENSG00000174082	1304076; 1307597; 1308735	1365571; 1370738; 1378299; 1380168; 1380501; 1382160	2955037 / CG32423	Temporarily Assigned Gene name / 2975803 / R10E4.2	1734374
4943812 1,1,1,2,1,1,	ENSG00000144381 / HSPD1 / 2771112 / heat shock 60kDa protein 1 (chaperonin)	1305699	1365895	CG7235 / 2947676; 2964305 / CG12101 / Heat shock protein 60	Heat Shock Protein / 2975208 / Y22D7AL.5	1732806
4943831 1,2,2,2,1,1,	2758440 / ENSG00000168267 / pancreas specific transcription factor, 1a / PTF1A	1298383; 1303389	1365900; 1378302	2958360 / CG33323 / 48 related 1; CG5952 / 2959936 / 48 related 2	2989576 / F48D6.3 / Helix Loop Helix	1736617
4943867 3,3,4,1,1,1,	2787931 / RAB5A / RAB5A, member RAS oncogene family / ENSG00000144566; 2803585 / ENSG00000111540 / RAB5B / RAB5B, member RAS oncogene family; ENSG00000108774 / 2805539 / RAB5C / RAB5C, member RAS oncogene family	1303799; 1306075; 1307855	1372343; 1374743; 1375562; 1376117	CG3664 / 2947128 / Rab-protein 5	2968851 / RAB family / F26H9.6	1743390
4943966 2,2,2,1,1,1,	PSCDBP / 2770780 / pleckstrin homology, Sec7 and coiled-coil domains, binding protein / ENSG00000115165; 2803416 / GRP1 (general receptor for phosphoinositides 1)-associated scaffold protein / ENSG00000161835 / GRASP	1307637; 1309978	1380200; 1382624	2955159 / CG6619	2968411 / F30F8.3	1732502
4944093 1,1,1,1,0,1,	2803487 / ENSG00000094914 / AAAS / achalasia, adrenocortical insufficiency, alacrimia (Allgrove, triple-A)	1313329	1363773	CG16892 / 2964082		1753609
4944134 2,3,3,1,0,1,	potassium voltage-gated channel, subfamily H (eag-related), member 8 / ENSG00000183960 / 2787928 / KCNH8; ENSG00000089558 / KCNH4 / 2805540 / potassium voltage-gated channel, subfamily H (eag related), member 4	1306757; 1313282; 1314982	1359784; 1370137; 1378479	eag-like K+ channel / 2952700 / CG5076		1742192
4944156 2,1,1,2,1,1,	LSM12 homolog pseudogene / 654166 / 2805630 / LSM12 homolog (S. cerevisiae) / LSM12; LSM12 homolog pseudogene / 654166 / 2808075 / LSM12 homolog (S. cerevisiae) / LSM12	1318571	1362200	CG14164 / 2955866; CG15735 / 2964494	2978022 / M142.5	1737454
4944180 1,2,6,1,1,1,	ENSG00000122691 / 2789731 / twist homolog 1 (acrocephalosyndactyly 3; Saethre-Chotzen syndrome) (Drosophila) / TWIST1	1304257; 1315836	1357959; 1361096; 1369849; 1369900; 1371199; 1372408	2953620 / CG2956 / twist	C02B8.4 / 2990494 / Helix Loop Helix	1733735
4944355 1,1,1,1,1,1,	2770612 / ENSG00000163161	1319094	1371254	2955919 / CG8019 / haywire	Y66D12A.15 / 2978131	1740431
4944455 1,1,1,0,1,1,	DnaJ (Hsp40) homolog, subfamily C, member 10 / ENSG00000077232 / 2771032 / DNAJC10	1296920	1380971		2969516 / Y47H9C.5 / DNaJ domain (prokaryotic heat shock protein)	1740091
4944629 5,3,5,1,1,1,	2762520; SMT3 suppressor of mif two 3 homolog 1 (S. cerevisiae) / 2771193 / SUMO1 / ENSG00000116030; similar to SMT3 suppressor of mif two 3 homolog 2 / 2790084 / ENSG00000184763; ENSG00000188612 / similar to SMT3 suppressor of mif two 3 homolog 2 / SMT3 suppressor of mif two 3 homolog 2 (S. cerevisiae) / SUMO2 / 2793179; 2809270 / SMT3 suppressor of mif two 3 homolog 3 (S. cerevisiae) / ENSG00000184900 / SUMO3	1296696; 1314395; 1315915	1362709; 1363202; 1366400; 1373981; 1380388	CG4494 / smt3 / 2947928	2966592 / SUMO (ubiquitin-related) homolog / K12C11.2	1744154

Cluster ID, counts per genome	**Human**	***Xenopus tropicalis***	**Fugu**	***Drosophila***	***C. elegans***	***Nematostella***
4944687 3,8,5,1,6,2,	ABCB11 / ENSG00000073734 / 2770871 / ATP-binding cassette, sub-family B (MDR/TAP), member 11; ATP-binding cassette, sub-family B (MDR/TAP), member 4 / ENSG00000005471 / 2790367; ABCB1 / ENSG00000085563 / 2790370 / ATP-binding cassette, sub-family B (MDR/TAP), member 1	1295505; 1296282; 1297329; 1298178; 1301408; 1313124; 1313346; 1315474	1359809; 1361295; 1364424; 1381412; 1382471	CG8523 / 2951988 / Multi drug resistance 50	2967754 / P-GlycoProtein related / C34G6.4; P-GlycoProtein related / 2982038 / K08E7.9; C47A10.1 / 2988117 / P-GlycoProtein related; 2991205 / P-GlycoProtein related / ZK455.7; F42E11.1 / 2991206; F22E10.3 / P-GlycoProtein related / 2991490	1738223; 1745530
4944795 3,3,7,1,1,1,	ENSG00000091409 / 2770927 / ITGA6; ENSG00000135424 / 2803562 / ITGA7; ITGA3 / 2805813 / ENSG00000005884 / integrin, alpha 3 (antigen CD49C, alpha 3 subunit of VLA-3 receptor)	1302132; 1306250; 1321933	1357667; 1358724; 1359567; 1363177; 1368263; 1371677; 1374094	2964617 / multiple edematous wings / CG1771	F54G8.3 / 2977523 / INtegrin Alpha	1744152
4944858 3,3,4,3,1,1,	ENSG00000119231 / SUMO1/sentrin specific peptidase 5 / 2789530 / SENP5; SENP1 / 2803287 / ENSG00000079387 / SUMO1/sentrin specific peptidase 1; 2804785 / SENP3 / SUMO1/sentrin/SMT3 specific peptidase 3 / ENSG00000161956	1302532; 1310253; 1319454	1360941; 1374118; 1376591; 1378529	2946751 / CG11023; 2956280 / CG32110; 2965549 / CG12359 / Ulp1	2976129 / T10F2.3 / Ubiquitin-Like Protease	1748345
4945009 2,2,4,2,1,1,	2770874 / LRP2 / ENSG00000081479 / low density lipoprotein-related protein 2; 2803650 / low density lipoprotein-related protein 1 (alpha-2-macroglobulin receptor) / ENSG00000123384 / LRP1	1297302; 1298380	1379661; 1380708; 1380937; 1381586	2950753 / CG33087; CG12139 / 2964090	Low-density lipoprotein Receptor Related / F29D11.1 / 2968360	1749310
4945034 4,2,3,1,1,1,	2770959 / ENSG00000154518 / ATP5G3 / ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C3 (subunit 9); 2787529; 2803504 / ENSG00000135390 / ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C2 (subunit 9) / ATP5G2; ATP5G1 / ENSG00000159199 / ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C1 (subunit 9) / 2805784	1305779; 1309962	1364351; 1364964; 1377299	2962453 / CG1746	Y82E9BR.3 / 2975176	1734540
4945044 1,1,1,1,1,1,	ENSG00000106443 / PHF14 / PHD finger protein 14 / 2789703	1295628	1368471	2947476 / CG15439	2988141 / Y59A8A.2	1741325
4945321 2,2,1,1,1,1,	2771388 / ENSG00000198925 / ATG9 autophagy related 9 homolog A (S. cerevisiae) / ATG9A; ATG9B / ENSG00000181652 / ATG9 autophagy related 9 homolog B (S. cerevisiae) / 2791099	1308694; 1308748	1383156	CG3615 / 2952362 / Autophagy-specific gene 9	2982954 / T22H9.2a / T22H9.2	1749273
4945396 2,1,2,1,1,1,	2770584 / erythrocyte membrane protein band 4.1 like 5 / EPB41L5 / ENSG00000115109; 2783777 / EPB41L4B / ENSG00000095203 / erythrocyte membrane protein band 4.1 like 4B	1302604	1363426; 1378271	CG9764 / yurt / 2959474	T04C9.6 / 2976448 / FERM domain (protein4.1-ezrin-radixin-moesin) family	1746764
4945416 2,3,3,1,1,1,	2801289 / ENSG00000143761 / ADP-ribosylation factor 1 / ARF1; 2803318 / ADP-ribosylation factor 3 / ARF3 / ENSG00000134287	1307431; 1310645; 1312630	1360625; 1368671; 1370778	2957788 / CG8385 / ADP ribosylation factor 79F	2976338 / B0336.2	1731764
4945473 1,1,2,1,1,1,	NDUFS1 / ENSG00000023228 / 2771245 / NADH dehydrogenase (ubiquinone) Fe-S protein 1, 75kDa (NADH-coenzyme Q reductase)	1312888	1364550; 1366994	NADH:ubiquinone reductase 75kD subunit precursor / CG2286 / 2963917	NADH Ubiquinone Oxidoreductase / Y45G12B.1 / 2983577	1755366
4945573 1,1,1,1,1,1,	2770875 / ENSG00000163093 / BBS5 / Bardet-Biedl syndrome 5	1297679	1364420	CG1126 / 2957959	R01H10.6 / 2977848	1737229
4945694 2,2,2,1,1,2,	NR4A2 / 2770774 / nuclear receptor subfamily 4, group A, member 2 / ENSG00000153234; NR4A3 / ENSG00000119508 / 2783637 / nuclear receptor subfamily 4, group A, member 3	1304109; 1317626	1358913; 1362207	CG1864 / 2949936 / Hormone receptor-like in 38	2975610 / C48D5.1 / Nuclear Hormone Receptor family	1733675; 1750437
4945714 1,1,1,1,1,7,	2771081 / FLJ20160 protein	1301103	1362104	CG12858 / 2952067	2977017 / R13A5.9	1729890; 1732703; 1733805; 1734412; 1735386; 1735388; 1739380
4945728 5,4,6,1,1,2,	ARHGAP12 / ENSG00000165322 / 2758598 / Rho GTPase activating protein 12; Rho GTPase activating protein 15 / 2770731 / ARHGAP15 / ENSG00000075884; BCR / breakpoint cluster region / 2802054 / ENSG00000186716 / similar to breakpoint cluster region isoform 1; 2804483 / active BCR-related gene / ENSG00000159842 / ABR; ENSG00000185602 / 2805693	1308671; 1309724; 1321827; 1322202	1363186; 1368454; 1369746; 1370340; 1378376; 1380280	CG40494-PA.3 / CG40494 / 2966008	C38D4.5 / 2975981	1736778; 1747171
4945733 2,3,4,2,1,3,	PDE11A / phosphodiesterase 11A / ENSG00000128655 / 2770985; ENSG00000138735 / 2777735 / PDE5A / phosphodiesterase 5A, cGMP specific	1304820; 1318410; 1320724	1360376; 1366693; 1366847; 1379145	Phosphodiesterase 11 / CG10231 / 2949621; Phosphodiesterase 6 / 2959639 / CG8279	PhosphoDiEsterase / 2967523 / C32E12.2	1732991; 1734836; 1753536

Cluster ID, counts per genome	Human	*Xenopus tropicalis*	Fugu	*Drosophila*	*C. elegans*	*Nematostella*
4945737 4,2,2,0,4,1,	ENSG00000165312 / OTU domain containing 1 / 2758446 / OTUD1; ENSG00000135913 / USP37 / 2771359 / ubiquitin specific peptidase 37; 2781908 / ENSG00000131864 / USP29 / ubiquitin specific peptidase 29; ENSG00000134588 / USP26 / ubiquitin specific peptidase 26 / 2793470	1310653; 1317113	1358509; 1375906		2969184 / Y106G6H.12 / Deubiquitylating with USP/UBP and OTU domains; F29C4.5 / 2978623 / Deubiquitylating with USP/UBP and OTU domains; C04E6.5 / 2984508; 2986293 / Deubiquitylating with USP/UBP and OTU domains / F38B7.5	1741507
4945854 1,1,2,0,0,3,	2770699 / transmembrane protein 163	1320627	1357821; 1379152			1731075; 1731861; 1753980
4945951 3,4,4,1,1,1,	ENSG00000077943 / integrin, alpha 8 / ITGA8 / 2758302; 2771051 / ENSG00000138448; ITGA5 / ENSG00000161638 / integrin, alpha 5 (fibronectin receptor, alpha polypeptide) / 2803527	1295131; 1295516; 1305632; 1321184	1361259; 1362309; 1363713; 1381785	inflated / CG9623 / 2965214	2977407 / F54F2.1	1749001
4946163 1,1,1,1,0,1,	TEGT / ENSG00000139644 / testis enhanced gene transcript (BAX inhibitor 1) / 2803354	1310576	1377988	2955511 / CG7188		1751172
4946212 3,1,1,2,1,1,	2758283 / acyl-Coenzyme A binding domain containing 7; 2770578 / ENSG00000155368 / diazepam binding inhibitor (GABA receptor modulator, acyl-Coenzyme A binding protein); ENSG00000140238 / similar to Acyl-CoA-binding protein (ACBP) (Diazepam binding inhibitor) (DBI) (Endozepine) (EP) / 2772243	1309179	1372798	2948127 / CG8498; 2955294 / Diazepam-binding inhibitor / CG8627	Acyl-Coenzyme A Binding Protein / 2967353 / C44E4.6	1734030
4946841 2,1,1,1,1,1,	2805544 / signal transducer and activator of transcription 5B / ENSG00000173757 / STAT5B; ENSG00000126561 / 2805545 / signal transducer and activator of transcription 5A / STAT5A	1303993	1382824	2960570 / Signal-transducer and activator of transcription protein at 92E / CG4257	Y51H4A.17 / 2982710	1734414
4947135 1,1,1,2,1,1,	2790075 / ENSG00000132432 / Sec61 gamma subunit / SEC61G	1298932	1372206	CG8860 / 2951544; CG14214 / 2965636	2986083 / EndoMitotic Oocytes / F32D8.6	1733779
4947258 3,2,2,1,1,1,	ENSG00000115935 / Wiskott-Aldrich syndrome protein interacting protein / WASPIP / 2770951; similar to SH3 domain binding protein CR16 / 641823 / 2789830 / ENSG00000122574 / 648464; ENSG00000171475 / WIRE protein / 2805424	1297316; 1314368	1375185; 1377449	2953429 / CG13503	R144.4a / 2976077 / R144.4	1748734
4947328 1,1,1,1,1,1,	ENSG00000108306 / FBXL20 / F-box and leucine-rich repeat protein 20 / 2805378	1305109	1374154	2951461 / CG9003	2977226 / C02F5.7a	1751229
4947407 2,1,1,1,1,1,	oligonucleotide/oligosaccharide-binding fold containing 2A / 2771092 / OBFC2A; 2803603 / oligonucleotide/oligosaccharide-binding fold containing 2B / OBFC2B	1301294	1368651	2947989 / CG5181	C06G3.8 / 2980305	1750418
4947708 1,1,1,1,1,1,	ENSG00000187778 / microspherule protein 1 / 2803346	1317564	1379028	CG1135 / 2954843	H28O16.2 / 2969722	1748250
4948425 2,2,2,1,1,1,	2771147 / hypothetical protein MGC39518; down-regulated by Ctnnb1, a / 2789756 / ENSG00000122591	1305231; 1313463	1358943; 1368662	CG6406 / 2952661	D1069.3b / 2970409 / D1069.3	1752589
4948560 1,1,1,1,0,1,	2803457 / eukaryotic translation initiation factor 4B / EIF4B / ENSG00000063046	1298701	1359145	2966296 / CG10837 / CG10837-PB.3		1741443
4948566 2,3,4,1,1,1,	2771372 / CDK5R2 / cyclin-dependent kinase 5, regulatory subunit 2 (p39) / ENSG00000171450; CDK5R1 / 2805224 / ENSG00000176749 / cyclin-dependent kinase 5, regulatory subunit 1 (p35)	1306731; 1309767; 1321364	1359646; 1365556; 1376491; 1377544	CG5387 / Cdk5 activator-like protein / 2948536	2975934 / Cyclin-Dependent Kinase 5 Activating protein homolog	1734407
4948683 1,1,1,1,1,1,	2790040 / ENSG00000136273 / HUS1 checkpoint homolog (S. pombe) / HUS1	1319025	1382683	Hus1-like / 2957956 / CG2525	2967080 / H26D21.1 / human HUS1 related	1744579
4949345 3,2,6,1,2,1,	transmembrane BAX inhibitor motif containing 1 / ENSG00000135926 / TMBIM1 / 2771351; 2803359 / ENSG00000135472 / Fas apoptotic inhibitory molecule 2 / FAIM2; 2808747 / ENSG00000178719 / GRINA / glutamate receptor, ionotropic, N-methyl D-asparate-associated protein 1 (glutamate binding)	1305660; 1319826	1360373; 1360624; 1362803; 1363780; 1373095; 1379210	N-methyl-D-aspartate receptor-associated protein / 2951733 / CG3798	2985700 / X-BoX promoter element regulated / F40F9.1; Temporarily Assigned Gene name / 2985702 / F40F9.2	1741441
4949376 8,2,3,2,0,2,	polycystic kidney disease 1 (autosomal dominant) / PKD1 / 2785294 / ENSG00000008710 / hypothetical protein LOC339047; ENSG00000183793 / 2785527; 2785535 / hypothetical protein LOC339047 / similar to nuclear pore complex interacting protein; ENSG00000196908 / hypothetical protein LOC339047 / 2785538 / similar to nuclear pore complex interacting protein; 2785559 / hypothetical protein LOC339047 / ENSG00000183458; 2785566; ENSG00000158683 / 2790038 / polycystic kidney disease 1 like 1 / PKD1L1; 2802634 / polycystic kidney disease (polycystin) and REJ (sperm receptor for egg jelly homolog, sea urchin)-like / ENSG00000130943 / PKDREJ	1307191; 1311841	1364678; 1370628; 1382176	2949206 / CG12636; CG30048 / 2951607		1750001; 1755043
4949482 2,2,1,0,0,1,	ENSG00000115271 / GCA / grancalcin, EF-hand calcium binding protein / 2770832; ENSG00000075142 / sorcin / SRI / 2790377	1305322; 1312222	1358769			1741952
4950094 1,1,0,1,0,1,	DNAH7 / 2771097 / dynein, axonemal, heavy polypeptide 7 / ENSG00000118997	1319768		Dynein heavy chain at 36C / 2949546 / CG5526		1734689

Cluster ID, counts per genome	Human	*Xenopus tropicalis*	Fugu	*Drosophila*	*C. elegans*	*Nematostella*
4950119 3,2,4,1,3,5,	2771378 / IHH / ENSG00000163501 / Indian hedgehog homolog (Drosophila); 2791173 / SHH / ENSG00000164690 / sonic hedgehog homolog (Drosophila); 2803328 / DHH / desert hedgehog homolog (Drosophila) / ENSG00000139549	1306563; 1320292	1363027; 1369628; 1378454; 1379648	hedgehog / 2961034 / CG4637	ZK1290.12 / WaRThog (hedgehog-like family) / 2972720; GRounDhog (hedgehog-like family) / 2988661 / F46B3.5; WaRThog (hedgehog-like family) / ZK377.1 / 2989412	1729900; 1732171; 1732351; 1733652; 1734185
4950365 1,1,1,1,0,1,	hypothetical protein LOC339287 / 2805421	1319788	1379389	male-specific lethal 1 / CG10385 / 2949638		1751326
4950409 3,1,4,1,0,1,	ENSG00000114923 / 2771438 / solute carrier family 4, anion exchanger, member 3; ENSG00000164889 / solute carrier family 4, anion exchanger, member 2 (erythrocyte membrane protein band 3-like 1) / 2791115 / SLC4A2; 2805644 / solute carrier family 4, anion exchanger, member 1 (erythrocyte membrane protein band 3, Diego blood group) / SLC4A1 / ENSG00000004939	1321127	1363550; 1370096; 1371707; 1377681	CG8177 / 2955830		1742162
4950416 2,2,2,1,0,1,	2758577 / ENSG00000107951 / PAPD1; 2810438 / RNA binding motif protein 21 / RBM21 / ENSG00000149016	1300870; 1321730	1367345; 1370846	CG11418 / 2962953		1741422
4950573 1,0,0,1,0,1,	2770898 / similar to CG14853-PB / ENSG00000204334			2959653 / CG14853		1730102
4950740 1,1,1,1,1,1,	MKI67IP / ENSG00000155438 / MKI67 (FHA domain) interacting nucleolar phosphoprotein / 2770598	1322278	1371041	CG6937 / 2960945	T04A8.6 / 2975943	1745713
4951031 2,1,1,1,1,1,	2760173; ribulose-5-phosphate-3-epimerase / RPE / ENSG00000197713 / similar to Ribulose-phosphate 3-epimerase (Ribulose-5-phosphate-3-epimerase) (HUSSY-17) / rcRPE / 2771287	1314964	1374271	2950589 / CG30499	2976952 / F08F8.7	1744610
4951119 1,1,1,0,1,1,	ENSG00000152127 / 2770698 / mannosyl (alpha-1,6-)-glycoprotein beta-1,6-N-acetyl-glucosaminyltransferase / MGAT5	1302197	1356965		C55B7.2 / 2968000 / GLYcosylation related	1743571
4951204 2,1,1,1,0,1,	2770950; 2789695 / replication protein A3, 14kDa / RPA3 / ENSG00000106399	1315936	1381953	2964405 / CG15220		1737482
4951224 1,1,1,1,1,1,	2805425 / CDC6 cell division cycle 6 homolog (S. cerevisiae) / CDC6 / ENSG00000094804	1296908	1361810	2955591 / CG5971	C43E11.10 / 2967268 / Cell Division Cycle related	1738348
4951588 4,3,5,1,1,1,	ENSG00000152256 / pyruvate dehydrogenase kinase, isozyme 1 / PDK1 / 2770929; ENSG00000004799 / pyruvate dehydrogenase kinase, isozyme 4 / PDK4 / 2790451; ENSG00000067992 / pyruvate dehydrogenase kinase, isozyme 3 / PDK3 / 2791673; ENSG00000005882 / 2805814 / pyruvate dehydrogenase kinase, isozyme 2 / PDK2	1298709; 1303933; 1318269	1364736; 1376326; 1377580; 1379468; 1380889	CG8808 / Pyruvate dehydrogenase kinase / 2950996	2977374 / ZK370.5	1752557
4951645 2,2,2,1,1,1,	enhancer of polycomb homolog 1 (Drosophila) / EPC1 / ENSG00000120616 / 2758609; enhancer of polycomb homolog 2 (Drosophila) / ENSG00000135999 / EPC2 / 2770742	1317581; 1321137	1362623; 1367305	Enhancer of Polycomb / CG7776 / 2951446	Y111B2A.11 / 2978327 / Enhancer of PolyComb-like	1745156
4951710 2,2,2,1,1,1,	2791184 / ENSG00000105983; LMBR1L / ENSG00000139636 / 2803330	1307018; 1317790	1367984; 1379487	2961380 / CG5807	R05D3.2 / 2977283	1747938
4951737 2,2,4,1,1,1,	2758000 / ENSG00000107929 / LARP5; ENSG00000161813 / 2803379 / La ribonucleoprotein domain family, member 4 / LARP4	1304450; 1322094	1368971; 1378127; 1379252; 1380499	CG11505 / 2954714	T12F5.5 / Prion-like-(Q/N-rich)-domain-bearing protein / 2967086	1747431
4951811 2,2,2,5,1,1,	dynein, cytoplasmic 1, intermediate chain 2 / 2770916 / DYNC1I2 / ENSG00000077380; DYNC1I1 / 2790453 / ENSG00000158560 / dynein, cytoplasmic 1, intermediate chain 1	1297360; 1298411	1357575; 1376263	CG9580-PA / CG9580 / 2965707; CG33497 / CG33499-PA / CG33497-PA / 2965709; 2965711 / CG32823 / CG32823-PB; 2965713 / CG33499 / CG33499-PA / CG33497-PA; CG18000 / 2965724 / short wing	DYnein Chain, light Intermediate / C17H12.1 / 2980234	1750320
4951862 2,1,1,0,0,1,	MYO1B / 2771091 / ENSG00000128641; 2803641 / ENSG00000166866	1301178	1363563			1733813
4951877 1,2,1,1,0,1,	2789957 / ENSG00000175600	1310980; 1319564	1357763	CG10877 / 2960582		1739093
4952077 1,0,0,0,0,1,	2771054 / ENSG00000163012 / ZSWIM2 / zinc finger, SWIM-type containing 2					1741985
4952248 1,0,2,0,0,2,	amyotrophic lateral sclerosis 2 (juvenile) chromosome region, candidate 11 / ENSG00000155754 / 2771179 / ALS2CR11		1363089; 1366323			1749231; 1754251
4952264 1,1,1,0,0,1,	2758397 / hypothetical protein LOC387640	1303393	1360988			1739135
4952559 1,1,1,1,1,1,	phosphatidylinositol-3-phosphate/phosphatidylinositol 5-kinase, type III / PIP5K3 / 2771276 / ENSG00000115020	1312243	1358715	2952676 / CG6355	PIP Kinase / 2991521 / VF11C1L.1	1746523

Cluster ID, counts per genome	Human	*Xenopus tropicalis*	Fugu	*Drosophila*	*C. elegans*	*Nematostella*
4952569 7,4,7,1,0,1,	2770840 / hypothetical protein FLJ39822; 2788313 / ENSG00000188338 / solute carrier family 38, member 3 / SLC38A3; ENSG00000017483 / 2791994 / solute carrier family 38, member 5 / SLC38A5; SLC38A1 / solute carrier family 38, member 1 / ENSG00000111371 / 2803260; 2803261 / solute carrier family 38, member 2 / SLC38A2 / ENSG00000134294; 2803264 / ENSG00000139209 / solute carrier family 38, member 4 / SLC38A4; ENSG00000139974 / 2806957 / SLC38A6	1295621; 1299547; 1307908; 1321071	1359126; 1367608; 1368789; 1377691; 1378548; 1379916; 1381970	2950905 / CG13743		1734542
4952722 1,1,1,1,1,1,	2770943 / ENSG00000138433	1304796	1365184	CG6843 / 2957144	Temporarily Assigned Gene name / 2967680 / F55F8.4	1731156
4952932 1,1,1,1,1,1,	2758490 / PDSS1 / prenyl (decaprenyl) diphosphate synthase, subunit 1 / ENSG00000148459	1311711	1375719	CG31005 / 2962459	COenzyme Q (ubiquinone) biosynthesis / C24A11.9 / 2967587	1744266
4953118 2,2,2,1,1,1,	2789926 / STARD3 N-terminal like / ENSG00000010270 / STARD3NL; STARD3 / 2805384 / ENSG00000131748 / START domain containing 3	1297285; 1306686	1358270; 1376100	CG3522 / Start1 / 2953993	Temporarily Assigned Gene name / F26F4.4 / 2976028	1729925
4953406 2,2,3,1,1,1,	protein tyrosine phosphatase, receptor type, N / 2771399 / ENSG00000054356 / PTPRN; 2791199 / ENSG00000155093	1308623; 1321259	1363076; 1378010; 1380196	CG31795 / 2946929 / ia2	related to Islet cell Diabetes Autoantigen / 2976349 / B0244.2	1739987
4953424 1,1,1,0,0,1,	UBE2Z / ubiquitin-conjugating enzyme E2Z (putative) / 2805785	1316585	1377990			1745404
4953428 1,1,1,1,1,1,	2805748 / ENSG00000159111 / mitochondrial ribosomal protein L10	1314807	1379436	CG11488 / 2946803 / mitochondrial ribosomal protein L10	K01C8.6 / 2972980	1737658
4953495 1,1,2,2,1,1,	2758320 / cubilin (intrinsic factor-cobalamin receptor) / ENSG00000107611 / CUBN	1305731	1368719; 1381183	CG32094 / 2956123; 2964100 / CG32702	2986582 / ZC116.3	1745800
4953538 2,1,1,1,1,1,	ATP6V1F / 2790850 / ENSG00000128524 / ATPase, H+ transporting, lysosomal 14kDa, V1 subunit F; 2803289	1302622	1356888	Vacuolar H+ ATPase 14kD subunit / 2952215 / CG8210	2973618 / Vacuolar H ATPase / ZK970.4	1754505
4953553 2,2,2,1,1,1,	ENSG00000057252 / 2800164; ENSG00000167780 / 2803465 / sterol O-acyltransferase 2 / SOAT2	1307847; 1315000	1371387; 1372895	2958586 / CG8112	2992109 / B0395.2	1737728
4953581 1,1,1,1,1,1,	PERLD1 / ENSG00000161395 / 2805388 / per1-like domain containing 1	1315842	1358427	CG3271 / 2950394	R01B10.4 / 2984549	1736860
4953651 2,2,2,2,1,1,	2770785 / ENSG00000007001 / uridine phosphorylase 2 / UPP2; ENSG00000183696 / 2790045 / uridine phosphorylase 1 / UPP1	1311867; 1314865	1364119; 1380127	CG3788 / 2953608; 2961770 / CG6330	2977029 / ZK783.2	1729985
4953748 2,1,2,1,0,1,	ENSG00000095777 / myosin IIIA / 2758476 / MYO3A; 2770891 / MYO3B / ENSG00000071909	1303929	1366026; 1371258	2947981 / neither inactivation nor afterpotential C / CG5125		1736536
4953957 1,1,2,1,0,1,	2805641 / ATXN7L3 / ENSG00000087152	1313961	1376694; 1380751	2957109 / CG13379		1739292
4954335 1,1,1,1,1,1,	ENSG00000198130 / 3-hydroxyisobutyryl-Coenzyme A hydrolase / 2771079 / HIBCH	1320762	1372861	2959804 / CG5044	F09F7.4 / F09F7.4a / 2976271	1749379
4954464 2,2,2,1,0,1,	TSPAN13 / 2789722 / tetraspanin 13 / ENSG00000106537; TSPAN31 / tetraspanin 31 / ENSG00000135452 / 2803687	1303934; 1304140	1375438; 1376584	Tetraspanin 97E / 2961773 / CG6323		1734650
4954552 1,1,2,1,1,1,	2758569 / ENSG00000197321	1310187	1380109; 1381860	CG33232 / 2954563	C10H11.1 / 2967389	1744791
4955343 1,1,1,1,0,1,	ARMC4 / ENSG00000169126 / armadillo repeat containing 4 / 2758542	1308495	1367322	CG5155 / 2947984		1741297
4955719 1,1,1,0,0,1,	2770787 / hypothetical protein BC015395	1312147	1380411			1743920
4956110 1,1,1,1,1,1,	2789963 / MRPL32 / ENSG00000106591 / mitochondrial ribosomal protein L32	1318529	1363095	CG12220 / mitochondrial ribosomal protein L32 / 2962460	2977311 / C30C11.1	1744228
4956308 3,2,4,1,1,2,	ENSG00000178662 / TGF-beta induced apoptosis protein 2 / 2770849; AXUD1 / 2788062 / AXIN1 up-regulated 1 / ENSG00000144655; chromosome 12 open reading frame 22 / ENSG00000110925 / 2803394	1296227; 1300418	1358856; 1360186; 1368829; 1381671	2947134 / CG4272	C41D11.3 / 2967296	1731078; 1756571
4956739 2,2,4,1,1,1,	trafficking protein, kinesin binding 2 / 2771177 / TRAK2 / ENSG00000115993; TRAK1 / ENSG00000182606 / 2788089	1305188; 1314061	1365474; 1371817; 1377896; 1379127	2947950 / CG13777 / milton	T27A3.1 / T27A3.1a / 2967846	1730136
4957314 1,1,1,1,1,1,	2803610 / ENSG00000062485 / citrate synthase / CS	1303827	1367545	CG3861 / lethal (1) G0030 / 2963644	2977834 / CiTrate Synthase / T20G5.2	1738130
4957613 1,1,1,0,0,1,	2758424 / ENSG00000077327	1301059	1369522			1733688
4957817 1,1,2,1,1,1,	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily e, member 1 / SMARCE1 / ENSG00000073584 / 2805438	1299439	1377930; 1379239	2964043 / CG7055 / dalao	2975432 / Y71H2AM.17	1732994
4957880 1,1,1,1,1,1,	LSM5 / ENSG00000106355 / 2789872 / LSM5 homolog, U6 small nuclear RNA associated (S. cerevisiae)	1308192	1379150	2955158 / CG6610	2987465 / F28F8.3 / LSM Sm-like protein	1755359
4958062 1,1,1,0,0,2,	2805739 / ENSG00000056345 / integrin, beta 3 (platelet glycoprotein IIIa, antigen CD61) / ITGB3	1321109	1381808			1743220; 1746183
4958105 1,1,1,0,0,1,	ENSG00000138400 / 2771254 / MDH1B	1312564	1375508			1747237
4958193 2,2,3,1,1,1,	2758456 / ARHGAP21 / ENSG00000107863 / Rho GTPase activating protein 21; 2805350 / Rho GTPase activating protein 23 / ARHGAP23	1303124; 1314238	1358979; 1371219; 1372140	RhoGAP19D / CG1412 / 2965763	C04D8.1 / 2977320	1737227

Cluster ID, counts per genome	**Human**	***Xenopus tropicalis***	**Fugu**	***Drosophila***	***C. elegans***	***Nematostella***
4958300 1,1,1,1,1,1,	nucleolar protein NOP5/NOP58 / 2771194 / ENSG00000055044	1308438	1358798	2947909 / CG10206 / nop5	2966998 / W01B11.3	1745270
4958365 1,1,1,1,1,1,	WD repeat domain 12 / 2771202 / ENSG00000138442 / WDR12	1312474	1359778	CG6724 / 2948657	Temporarily Assigned Gene name / 2967683	1737202
4958478 1,1,1,1,1,1,	ENSG00000163466 / 2771344 / actin related protein 2/3 complex, subunit 2, 34kDa / ARPC2	1298929	1358964	Arc-p34 / CG10954 / 2949905	Temporarily Assigned Gene name / Y6D11A.2 / 2975030	1729980
4958626 1,1,1,1,1,1,	2803643	1295823	1382421	2965181 / CG9723	F10C5.2 / 2974909	1744344
4958637 1,1,1,1,1,1,	2758531 / ACBD5 / ENSG00000107897	1311752	1375574	2947240 / CG8814	Membrane Associated Acyl-CoA binding protein / 2978187 / C18D11.2	1734350
4958996 2,1,2,1,1,1,	2758404 / ENSG00000078403; myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 6 / ENSG00000108292 / 2805355 / MLLT6	1297575	1360340; 1379229	Alhambra / CG1070 / 2958353	2977399 / Zinc Finger Protein	1739576
4959042 1,1,1,0,0,1,	PMS1 postmeiotic segregation increased 1 (S. cerevisiae) / 2771073 / ENSG00000064933 / PMS1	1314431	1371108			1739986
4959088 3,2,4,1,1,1,	ENSG00000157985 / CENTG2 / centaurin, gamma 2 / 2771610; ENSG00000133612 / 2791121 / CENTG3 / centaurin, gamma 3; ENSG00000135439 / 2803686	1297691; 1301162	1362766; 1374180; 1377925; 1383197	2949107 / CG31811 / centaurin gamma 1A	Y39A1A.15 / 2977958 / CeNTaurin	1754386
4959098 1,1,1,0,0,1,	BRCA1 associated RING domain 1 / 2771310 / ENSG00000138376 / BARD1	1294964	1378944			1747129
4959221 2,2,1,1,0,1,	hypothetical protein FLJ23861 / 2771289; 2805718 / KIAA1267	1308627; 1311667	1379509	2959863 / CG4699		1751170
4959318 2,2,2,1,3,1,	solute carrier family 11 (proton-coupled divalent metal ion transporters), member 1 / 2771355 / SLC11A1 / ENSG00000018280; SLC11A2 / ENSG00000110911 / 2803388	1302282; 1307468	1358344; 1368355	Malvolio / CG3671 / 2960678	yeast SMF (divalent cation transporter) homolog / 2979117 / Y69A2AR.4; yeast SMF (divalent cation transporter) homolog / 2990158 / K11G12.3; yeast SMF (divalent cation transporter) homolog / 2990159 / K11G12.4	1736476
4959887 1,1,1,1,0,1,	2805528 / TTC25 / tetratricopeptide repeat domain 25	1297694	1375759	2953425 / CG13502		1748333
4960081 1,2,1,1,1,2,	ENSG00000144785 / TMEM4 / 2803611 / transmembrane protein 4	1303830; 1315760	1367511	2951133 / CG12918	2976385 / F01F1.15	1739287; 1739326
4960159 3,1,1,2,1,1,	2771113 / HSPE1 / ENSG00000115541 / heat shock 10kDa protein 1 (chaperonin 10); 2786333; 2800665	1297791	1365921	2956311 / CG11267; CG9920 / 2959575	2975206 / Y22D7AL.10	1732823
4960255 1,1,1,1,1,1,	RPL37A / ENSG00000197756 / 2771333 / ribosomal protein L37a	1307336	1379001	2947602 / Ribosomal protein L37A / CG5827	2974581 / Y48B6A.2 / Ribosomal Protein, Large subunit	1755385
4960409 1,1,1,1,0,1,	ENSG00000155636 / developmentally regulated RNA-binding protein 1 / 2770986	1312773	1376109	2954915 / CG1316		1740228
4960925 1,0,1,0,0,1,	hypothetical LOC129881 / 2770881		1358177			1740978
4960949 2,2,3,1,1,1,	ENSG00000115806 / GORASP2 / 2770904 / golgi reassembly stacking protein 2, 55kDa; ENSG00000114745 / GORASP1 / golgi reassembly stacking protein 1, 65kDa / 2788060	1304670; 1306224	1359194; 1367409; 1378786	2957311 / Grasp65 / CG7809	2980639 / Y42H9AR.1	1736938
4961015 1,1,1,0,0,1,	2771127 / hypothetical protein FLJ37953	1322003	1371260			1749207
4961127 1,1,1,1,1,1,	2770867 / nitric oxide synthase trafficker / NOSTRIN / ENSG00000163072	1297385	1375370	2964017 / CG4040	C36E8.4 / 2975720	1740919
4961318 1,2,2,1,1,1,	2770804 / similar to death-associated protein	1295552; 1318372	1367861; 1378218	2951421 / CG12384	2968310 / T28F4.5	1739027
4961341 5,6,8,1,1,1,	ENSG00000138379 / GDF8 / 2771077 / growth differentiation factor 8; TGFB1 / transforming growth factor, beta 1 (Camurati-Engelmann disease) / 2781109 / ENSG00000105329; 2801068 / ENSG00000092969; 2803568 / growth differentiation factor 11 / GDF11 / ENSG00000135414; 2807172 / ENSG00000119699 / TGFB3 / transforming growth factor, beta 3	1303308; 1303311; 1306556; 1308547; 1309598; 1319957	1359325; 1362615; 1364948; 1369567; 1373262; 1374088; 1376829; 1380962	myoglianin / CG1838 / 2962695	2975011 / B0412.2 / abnormal DAuer Formation	1750198
4961380 2,2,2,2,2,1,	ENSG00000141429 / UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 1 (GalNAc-T1) / 2757725 / GALNT1; 2770770 / UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 13 (GalNAc-T13) / ENSG00000144278 / GALNT13	1315993; 1318224	1358915; 1361221	CG31651 / 2947604 / polypeptide GalNAc transferase 5; 2952444 / CG30463	2977102 / ZK688.8; Y39E4B.12 / 2978429 / GLYcosylation related	1742190
4961686 1,1,1,1,1,1,	2805362 / CCDC49 / coiled-coil domain containing 49	1312169	1374121	CG2843 / 2947193	F52C9.7 / 2976184	1729877
4961757 1,1,1,0,0,1,	PSMC3IP / 2805557 / ENSG00000131470 / PSMC3 interacting protein	1321835	1376146			1749964
4962139 2,1,4,1,1,1,	plexin domain containing 2 / ENSG00000120594 / PLXDC2 / 2758381; PLXDC1 / 2805371 / ENSG00000161381 / plexin domain containing 1	1311732	1358031; 1367911; 1376852; 1377982	2964197 / CG2221 / lethal (1) G0289	C36E8.3 / 2975719	1741508

Cluster ID, counts per genome	**Human**	***Xenopus tropicalis***	**Fugu**	***Drosophila***	***C. elegans***	***Nematostella***
4962259 1,1,0,0,0,1,	ENSG00000151687 / 2771069	1321608				1742030
4962493 1,1,1,1,1,1,	2805786 / SNF8 / SNF8, ESCRT-II complex subunit, homolog (S. cerevisiae) / ENSG00000159210	1321386	1377698	2960811 / CG6637	2976062 / C27F2.5	1732804
4962524 2,2,2,1,1,1,	ENSG00000071967 / CYBRD1 / 2770913 / cytochrome b reductase 1; cytochrome b-561 / 2805983 / ENSG00000008283 / CYB561	1303884; 1320653	1357687; 1369968	CG1275 / 2954531	2977611 / F55H2.5	1733569
4963080 1,1,1,1,0,1,	2770739 / ORC4L / ENSG00000115947 / origin recognition complex, subunit 4-like (yeast)	1310935	1378070	2954050 / CG2917 / Origin recognition complex subunit 4		1747427
4963127 1,1,2,1,0,1,	AARSD1 / alanyl-tRNA synthetase domain containing 1 / 2805583	1306274	1365970; 1380296	CG10802 / 2963293		1749376
4963196 1,1,1,1,1,1,	ENSG00000135441 / 2803565 / BLOC1S1 / biogenesis of lysosome-related organelles complex-1, subunit 1	1304121	1374219	2952039 / CG30077	2977832 / T20G5.10	1736503
4963296 1,1,0,0,0,1,	ENSG00000205522 / hypothetical LOC255411 / 2803309	1310581				1733651
4963650 1,1,1,1,1,1,	ENSG00000175203 / DCTN2 / 2803670 / dynactin 2 (p50)	1308039	1366680	Dynamitin / CG8269 / 2950861	DyNactin Complex component / 2976416	1732438
4964174 1,1,1,1,0,1,	2803601 / ENSG00000181852 / ring finger protein 41 / RNF41	1303811	1381687	2956673 / CG17033		1744409
4965012 1,1,1,1,1,1,	2789962 / ENSG00000106588 / PSMA2 / proteasome (prosome, macropain) subunit, alpha type, 2	1314434	1379667	Proteasome 25kD subunit / CG5266 / 2959307	D1054.2 / Proteasome Alpha Subunit / 2986040	1743624
4965109 1,1,1,1,1,1,	glycyl-tRNA synthetase / ENSG00000106105 / 2789845 / GARS	1297025	1377883	2956549 / CG6778 / Glycyl-tRNA synthetase	2976127 / T10F2.1 / Glycyl tRNA Synthetase	1729756
4966532 2,2,4,1,1,1,	par-3 partitioning defective 3 homolog (C. elegans) / 2758645 / ENSG00000148498 / PARD3; 2771232 / ENSG00000116117 / ALS2CR19	1300651; 1312433	1362255; 1369944; 1373256; 1378297	CG5055 / 2965268 / bazooka	F54E7.3 / 2976307 / abnormal embryonic PARtitioning of cytoplasm	1751253
4966800 2,2,2,1,1,1,	2770956 / CHN1 / chimerin (chimaerin) 1 / ENSG00000128656; 2789827 / CHN2 / chimerin (chimaerin) 2 / ENSG00000106069	1317895; 1318508	1359702; 1362045	CG3208 / 2963519 / RhoGAP5A	BE0003N10.2 / 2975081	1734441
4967503 2,4,2,0,0,1,	GALNT3 / UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 3 (GalNAc-T3) / ENSG00000115339 / 2770854; 2803403 / ENSG00000139629 / GALNT6 / UDP-N-acetyl-alpha-D-galactosamine:polypeptide N-acetylgalactosaminyltransferase 6 (GalNAc-T6)	1300460; 1301506; 1306228; 1308116	1380376; 1380406			1747814
4968299 1,1,1,1,0,1,	COMMD3 / 2758414 / ENSG00000148444 / COMM domain containing 3	1301081	1364669	CG12106 / 2964056		1741053
4968689 1,1,1,1,1,1,	2803313 / DEAD (Asp-Glu-Ala-Asp) box polypeptide 23 / DDX23 / ENSG00000174243	1319330	1366832	CG10333 / 2949627	2976369 / F01F1.7	1741267
4968705 1,1,1,1,1,1,	2771386 / ABCB6 / ATP-binding cassette, sub-family B (MDR/TAP), member 6 / ENSG00000115657	1311823	1370829	2959878 / CG4225	Heavy Metal Tolerance factor / 2978058 / W09D6.6	1748272
4969887 1,1,1,1,1,1,	2771047 / likely ortholog of mouse immediate early response, erythropoietin 4 / ENSG00000065548	1295515	1377138	2950816 / CG8635	2968390 / F27D4.4	1747432
4971083 2,1,2,1,1,1,	2771111 / COQ10B / coenzyme Q10 homolog B (S. cerevisiae); 2803609 / ENSG00000135469	1303825	1364856; 1382563	CG9410 / 2950390	2976080 / R144.3	1732225
4971538 1,1,1,0,0,1,	2789923 / ENSG00000086288 / TXNDC3 / thioredoxin domain containing 3 (spermatozoa)	1304941	1380111			1731911
4972775 2,1,2,1,1,1,	2771200 / ALS2CR15 / ENSG00000163596 / islet cell autoantigen 1,69kDa-like; islet cell autoantigen 1, 69kDa / ENSG00000003147 / 2789697 / ICA1	1297094	1358935; 1375405	2957529 / CG10566	2967096 / C32E8.7 / Resistance to Inhibitors of Cholinesterase	1733481
4973650 1,2,1,0,1,1,	2805649 / ENSG00000030582 / granulin / GRN	1305607; 1316049	1357377		T22H2.6a / 2969484 / T22H2.6	1752880
4974430 1,0,0,0,0,1,	2805749 / leucine rich repeat containing 46 / LRRC46 / ENSG00000141294					1745663
4975333 2,1,1,0,0,1,	ENSG00000106086 / 2789836 / PLEKHA8; 2803253 / ENSG00000134297 / pleckstrin homology domain containing, family A (phosphoinositide binding specific) member 9 / PLEKHA9	1309973	1361155			1732037
4976561 1,1,1,0,0,2,	SPBC25 / 2770868 / spindle pole body component 25 homolog (S. cerevisiae) / ENSG00000152253	1319130	1365181			1740890; 1746707
4976639 1,1,1,1,1,1,	aspartyl-tRNA synthetase / DARS / 2770718 / ENSG00000115866	1295977	1366237	2951739 / CG3821 / Aspartyl-tRNA synthetase	2977605 / B0464.1 / aspartyl(D) tRNA Synthetase	1737500

Table S8.1

Ontology ID	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%	3%	Other metabolism
MF00153	-18.1	-	3	158	1186	7766	0%	2%	Protease
MF00213	-17.6	-	1	124	1186	7766	0%	2%	Non-receptor serine/threonine protein kinase
BP00034	-17.4	-	4	167	1186	7766	0%	2%	DNA metabolism
BP00036	-17.0	-	0	102	1186	7766	0%	1%	DNA repair
MF00051	-16.9	-	0	101	1186	7766	0%	1%	Helicase
BP00047	-16.5	-	2	133	1186	7766	0%	2%	Pre-mRNA processing
MF00156	-16.4	-	0	98	1186	7766	0%	1%	Other hydrolase
MF00264	-16.2	-	0	97	1186	7766	0%	1%	Microtubule family cytoskeletal protein
MF00093	-16.1	-	25	379	1186	7766	2%	5%	Select regulatory molecule
BP00276	-16.0	-	2	130	1186	7766	0%	2%	General vesicle transport
MF00113	-15.2	-	1	109	1186	7766	0%	1%	Phosphatase
MF00097	-14.7	-	1	106	1186	7766	0%	1%	G-protein
MF00118	-14.6	-	3	135	1186	7766	0%	2%	Synthase and synthetase
MF00284	-14.3	-	0	86	1186	7766	0%	1%	Other ligase
MF00166	-14.0	-	0	84	1186	7766	0%	1%	Isomerase
MF00077	-13.8	-	0	83	1186	7766	0%	1%	Chaperone
MF00099	-13.8	-	0	83	1186	7766	0%	1%	Small GTPase
MF00075	-13.7	-	2	115	1186	7766	0%	1%	Ribosomal protein
BP00062	-13.5	-	0	81	1186	7766	0%	1%	Protein folding
BP00048	-13.3	-	1	97	1186	7766	0%	1%	mRNA splicing
BP00076	-13.1	-	2	111	1186	7766	0%	1%	Electron transport
BP00020	-12.3	-	0	74	1186	7766	0%	1%	Fatty acid metabolism
MF00127	-12.3	-	1	91	1186	7766	0%	1%	Reductase
MF00086	-12.2	-	3	118	1186	7766	0%	2%	Other transporter
BP00285	-12.1	-	15	246	1186	7766	1%	3%	Cell structure and motility
MF00157	-11.8	-	1	88	1186	7766	0%	1%	Lyase
MF00091	-11.6	-	13	222	1186	7766	1%	3%	Cytoskeletal protein
BP00081	-11.2	-	1	84	1186	7766	0%	1%	Coenzyme and prosthetic group metabolism
MF00133	-9.5	-	1	73	1186	7766	0%	1%	Methyltransferase
BP00273	-8.9	-	1	69	1186	7766	0%	1%	Chromatin packaging and remodeling
BP00129	-8.8	-	3	94	1186	7766	0%	1%	Endocytosis
MF00100	-8.5	-	6	125	1186	7766	1%	2%	G-protein modulator
BP00286	-8.5	-	8	145	1186	7766	1%	2%	Cell structure
MF00065	-8.4	-	2	79	1186	7766	0%	1%	mRNA processing factor
MF00119	-8.3	-	2	78	1186	7766	0%	1%	Synthase
BP00142	-8.2	-	8	143	1186	7766	1%	2%	Ion transport
BP00207	-8.0	-	8	141	1186	7766	1%	2%	Cell cycle control
MF00087	-7.9	-	4	99	1186	7766	0%	1%	Transfer/carrier protein
MF00044	-7.7	-	3	86	1186	7766	0%	1%	Nuclease

All types of novelty, p<0.05/100 depleted ontology terms:

BP00060	-64.4	-	91	1056	1584	7766	6%	14%	Protein metabolism and modification
MF00042	-42.9	-	91	915	1584	7766	6%	12%	Nucleic acid binding
BP00001	-37.8	-	5	231	1584	7766	0%	3%	Carbohydrate metabolism
MF00131	-31.0	-	28	398	1584	7766	2%	5%	Transferase
BP00031	-28.7	-	128	1034	1584	7766	8%	13%	Nucleoside, nucleotide and nucleic acid metabolism
MF00141	-27.4	-	22	330	1584	7766	1%	4%	Hydrolase
MF00123	-23.1	-	20	289	1584	7766	1%	4%	Oxidoreductase
BP00125	-23.0	-	36	405	1584	7766	2%	5%	Intracellular protein traffic

BP00061	-21.5	-	9	190	1584	7766	1%	2%	Protein biosynthesis
MF00126	-21.1	-	3	130	1584	7766	0%	2%	Dehydrogenase
MF00075	-20.3	-	2	115	1584	7766	0%	1%	Ribosomal protein
BP00063	-19.5	-	46	447	1584	7766	3%	6%	Protein modification
MF00156	-19.2	-	1	98	1584	7766	0%	1%	Other hydrolase
MF00082	-16.7	-	16	219	1584	7766	1%	3%	Transporter
BP00289	-16.7	-	19	241	1584	7766	1%	3%	Other metabolism
BP00013	-16.6	-	5	128	1584	7766	0%	2%	Amino acid metabolism
MF00166	-16.2	-	1	84	1584	7766	0%	1%	Isomerase
BP00047	-15.9	-	6	133	1584	7766	0%	2%	Pre-mRNA processing
BP00203	-15.9	-	35	346	1584	7766	2%	4%	Cell cycle
BP00282	-15.1	-	6	129	1584	7766	0%	2%	Mitosis
MF00118	-14.6	-	7	135	1584	7766	0%	2%	Synthase and synthetase
BP00036	-13.4	-	4	102	1584	7766	0%	1%	DNA repair
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