Octopus Genome Supplementary Notes

1. GENOME SEQUENCING AND ASSEMBLY 3

1.1. Library preparation 3

1.2. Sequencing and assembly 3

1.3. Assembly statistics 7

1.4. Polymorphism estimation 7

2. TRANSCRIPTOME SEQUENCING AND ANALYSIS 9

2.1. Tissues sampled, RNA preparation and sequencing 9

2.2. Mapping reads to the genome for expression analysis 10

2.3. *de novo* transcriptome assembly using Trinity 10

3. FLUORESCENCE-BASED GENOME SIZE ESTIMATE 12

4. ANNOTATION OF PROTEIN-CODING GENES AND TRANSPOSABLE ELEMENTS 14

4.1. Annotation 14

4.2. Transposable element annotation and expansions 18

4.3. Transposon expression 22

5. RNA EDITING 24

5.1. ADAR genes and phylogeny 24

5.2. Identification of putative edited positions 25

6. ANALYSIS OF SYNTENY 29

6.1. Conservation of synteny 29

6.2. Seeking synteny-based evidence for whole genome duplication 33

7. GENE FAMILY CONSTRUCTION AND PHYLOGENY 35

7.1. Multi-gene cephalopod phylogeny and dating 35

7.2. Construction of orthology groups 38

7.3. Intron and indel analysis 40

7.4. Genome-wide gene family expansions and dating 41

8. DESCRIPTION AND ANALYSES OF SPECIFIC GENE FAMILIES 47

8.1. Annotation methods 47

8.2. Selected developmental control signaling molecules 47

8.2.1. Hedgehog 48

8.2.2. Transforming Growth Factor-β 49

8.2.3. WNT 51

8.2.4. Axon guidance molecules 53

8.3. Protocadherins 54

8.4. C2H2-ZNFs 66

8.5. GPCRs 71

8.6. Chitinases 75

8.7. IL17s and IL17 receptors 76

9. HOX COMPLEMENT AND DISTRIBUTION ON SCAFFOLDS 77

10. NEURONAL GENES 82

10.1. Neurotransmitter-related enzymes 82

10.2. Neurotransmitter vesicular transporters: SLC 17, 18 and 32 82

10.3. SNAREs 86

10.4. Channel and receptor subfamilies 87

10.5. Innexins 90

10.6. Discs, Large (DLG) proteins 93

11. CEPHALOPOD NOVELTIES 95

11.1. Cephalopod-specific gene identification 95

11.2. Reflectins 97

11.3. Octopus-specific gene identification 99

12. REFERENCES 101

1. GENOME SEQUENCING AND ASSEMBLY

1.1 Library preparation

Gonad tissue from an adult male *Octopus bimaculoides* was dissected and frozen in liquid nitrogen. The frozen tissue was ground using a mortar and pestle and 20mg of ground tissue was used per extraction. Twenty extractions were done following the Gentra Invertebrate protocol. Following initial DNA extraction all 20 samples were combined and treated with 8% by volume RNase A at 37°C for one hour. This combined sample was further purified using a standard phenol chloroform isoamyl alcohol extraction. Extraction was followed by EtOH precipitation, and the resulting pellet was resuspended in 50ul Gentra DNA Hydration Solution.

1.2 Sequencing and assembly

Sequencing

Illumina libraries were prepared using Illumina v3 chemistry and sequenced on Illumina HiSeq 2000 instruments at UC Berkeley. Additional Illumina mate pair libraries were prepared by Illumina with 1.5 kb, 4 kb, and 10 kb inserts. Libraries are summarized in Table S1.1.

Sequence reads are deposited in the SRA as BioProject PRJNA270931.

Sequencing lane	read pairs	type	raw bp	genome coverage
OCT210_1	249,309,793	2x150	74,792,937,900	26x
OCT210_2	237,982,553	2x150	71,394,765,900	25x
OCT360_1	199,767,378	2x150	59,930,213,400	21x
OCT360_2	195,122,479	2x150	58,536,743,700	20x
TOTAL			**264,654,660,900**	**92x**
mate pair libraries				**genome coverage (by fragments)**
1.5 kb	176,492,836	2x100	35,298,567,200	92.29
4 kb	158,870,735	2x100	31,774,147,000	221.52
10 kb	166,743,192	2x100	33,348,638,400	581.25

Table S1.1. Illumina genomic libraries and predicted genome coverage based on a 2.8 Gb estimated genome size.

Assembly

Reads were assembled using meraculous (Chapman et al., 2011) with a k-mer size of 51. At k51 the fragment libraries exhibited a single-depth peak at 37x (Figure S1.1). A dmin of 7 was selected to separate the 51-mers likely to contain errors; 51-mers of depth 7 or below were not used in the assembly process.

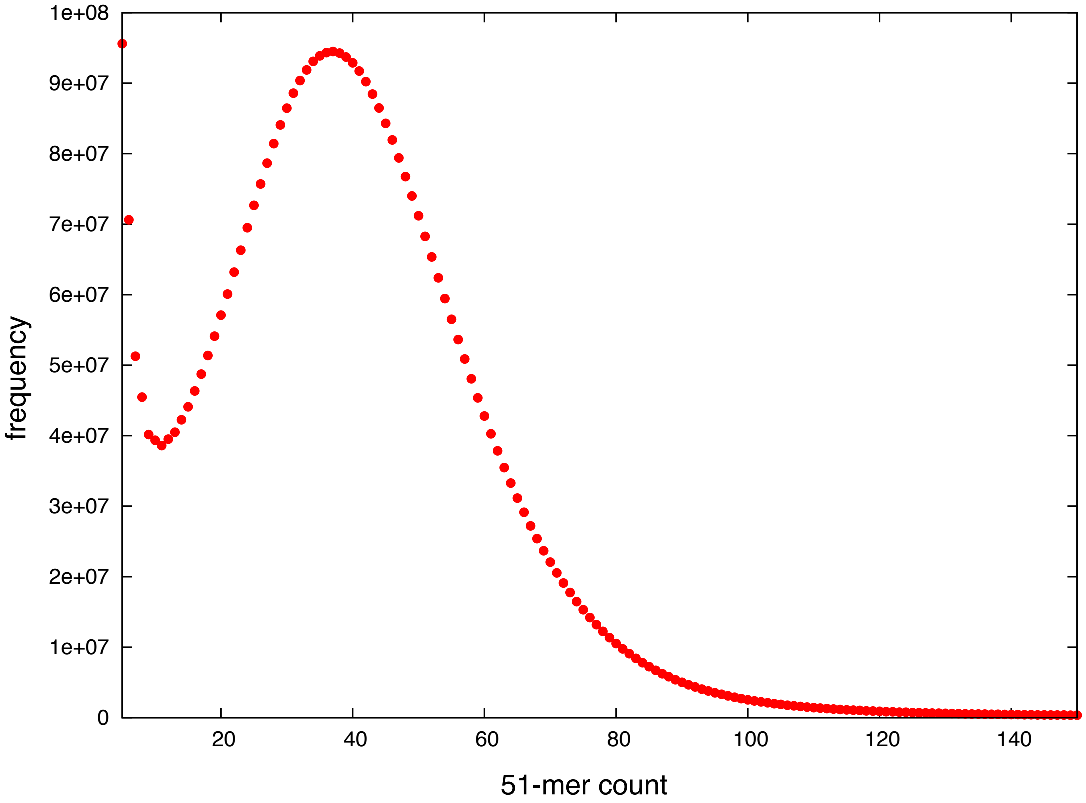


Figure S1.1. 51-mer depth distribution.

Genome size was estimated from these 51-mers. The 51-mer distribution $f(d)$ shows a single peak at depth 37, indicating that the 1x 51mer depth is $d_0 =$ 37. The cumulative fraction of 51-mers at 1-100,000x depth is shown for the genomic fragment libraries (Figure S1.2). The genome size was estimated using the equation

$$\sum_{d=0.25d_o}^{\max Depth} (\frac{d}{d_0}) f(d)$$

where *maxDepth* is the maximum relative depth and $f(d)$ is the number of k-mers with copy number *d* in the dataset. The lower cutoff excludes k-mers that are likely to be due to sequencing errors (Chapman et al., submitted). Note that ~80% of the genome is single copy based on the position of the knee in Figure S1.2. The total estimated genome size based on this method is 2.87 Gb.

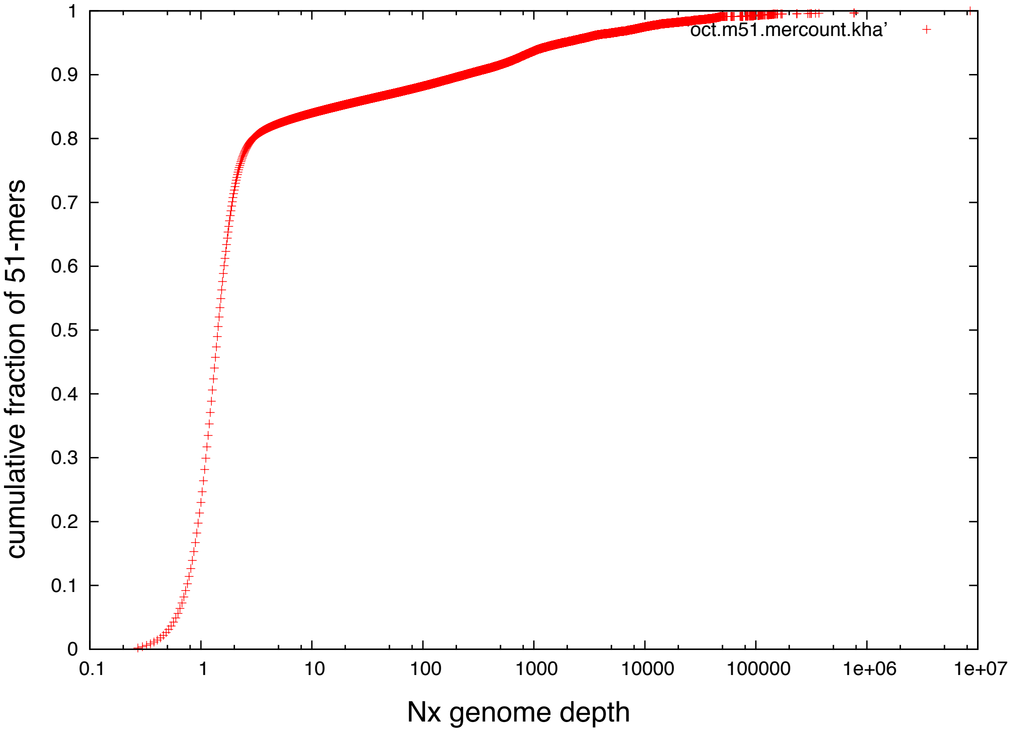
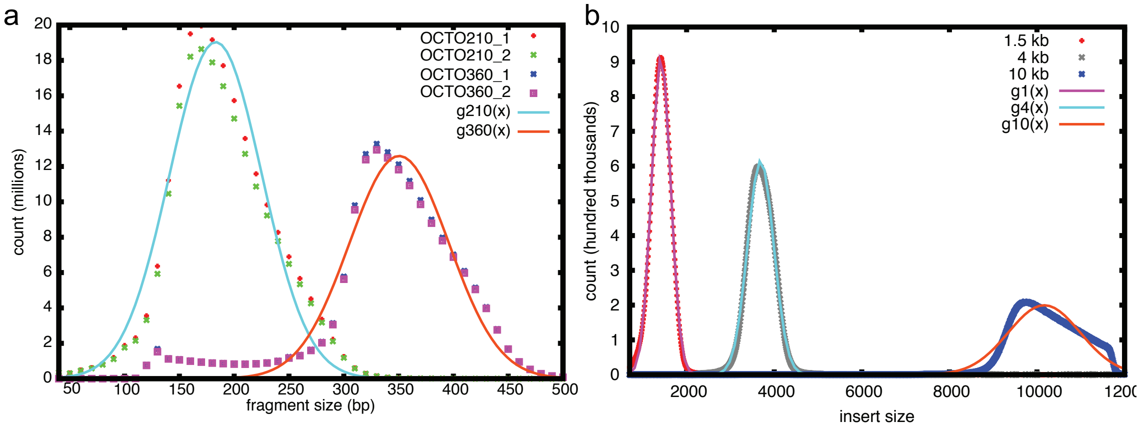


Figure S1.2. Cumulative proportion of sequenced 51-mers present at Nx relative depth.

For assembly, meraculous was used in polymorphic mode. Initial fragment contigs were created by enumerating all unique extensions of the valid 51-mers. These contigs were processed to identify “bubbletigs” (contigs that represent two haplotypes and diverge by a bubble of different sequence) and “isotigs” (contigs that do not have a match with a single bubble difference). The 51-mer depth of these isotigs was graphed and found to exhibit a bimodal depth distribution (data not shown). The lower depth peak represents isotigs for reads realigned at half-depth, indicating only one haplotype is represented by that contig. Isotigs less than 33x 51-mer depth were not used further in the assembly.

Using meraculous, the minimum required number of properly aligned mate-pairs needed to join contigs into a scaffold was 3, 3, and 10 for mate-pair libraries of 1.2k, 4k, and 10k, respectively. Insert size distributions were in accordance with the expected insert range (Figure S1.3). Gaps were then filled with the fragment libraries for the final assembly.



c

Library	Mean insert size, bp	Standard deviation
OCTO210	183	42
OCTO360	351	46
1.5 kb mate pair	1417	218
4 kb mate pair	3692	300
10 kb mate pair	10,165	823

Figure S1.3. a, Insert size distributions and Gaussian fit of fragment libraries. **b**, Insert size distributions and Gaussian fit of mate-pair libraries. **c**, Mean and standard deviation (width) from Gaussian fits**.**

A browser of this genome assembly is available at: http://octopus.metazome.net/.

1.3 Assembly statistics

Sequence total	2,371.5 Mb
Contig sequence total	2,016.2 Mb
Scaffold total	379,696
Contig total	939,190
Scaffold N/L50	1,369/466.1 Kb
Contig N/L50	100,762/5.4 Kb
% genome is scaffolds > 50 Kb	92%

1.4 Polymorphism estimation

In order to estimate heterozygosity, we realigned reads from the fragment libraries to the assembly using BWA-MEM (Li and Durbin, 2009). The realignment shows a single peak depth of coverage (Figure S1.4), indicating the assembly has incorporated both haplotypes evenly across the genome. Picard v1.92 was used to mark duplicates before submission to GATK for SNP calling. GATK v3.2-2 haplotypeCaller was used with options maxAlternateAlleles 2 and maxNumHaplotypesInPopulation 3 defining SNPs as those sites called with genotype quality of 40 or better. The SNP rate was calculated by determining callable loci with GATK's callableLoci walker and limiting both the callable loci and the called SNPs to regions with depth within 2 standard deviations of the peak depth (33-156). A total of 1,424,497 SNPs were found at 1,740,621,467 eligible sites, for a heterozygous SNP rate of 0.0008 per site.

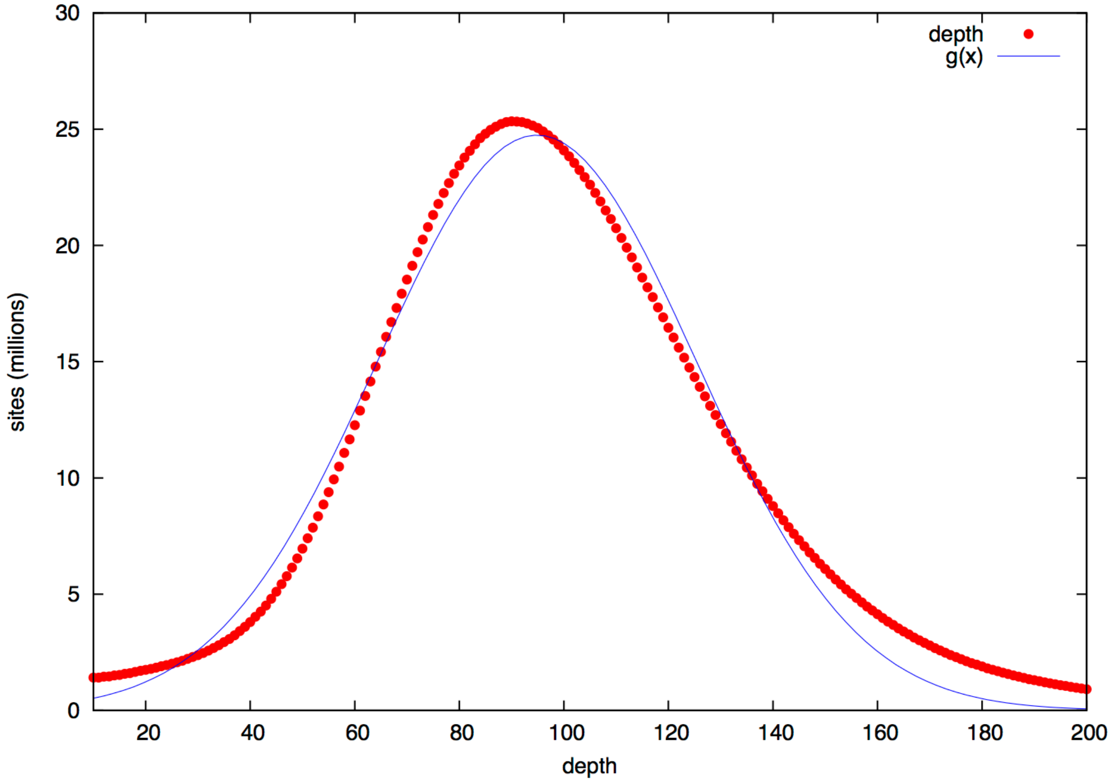


Figure S1.4. Aligned depth of all fragment library reads to genome assembly (red). Blue shows a Gaussian fit.

2. TRANSCRIPTOME SEQUENCING AND ANALYSIS

2.1 Tissues sampled, RNA preparation and sequencing

Transcriptomes from twelve different tissues were generated to aid with gene prediction and for expression analysis. Tissues were chosen to cover a wide range of transcripts, and to identify genes expressed in structures that feature cephalopod innovations. Adult *O. bimaculoides* supplied by Aquatic Research Consultants (Dr. Charles Winkler, San Pedro, CA) were anesthetized in 3% ethanol with 12.5mM $MgCl_2$. The ova sample was isolated from a single, mature egg. The testes sample was isolated from a mature male gonad. The viscera sample combines equal amounts of RNA isolated from the hepatopancreas, kidney, and central heart. Skin was isolated from the eyespots, dorsal and ventral mantle, and the leucophore-rich area between the eyes of month old hatchlings. Suckers were dissected from the distal third of a single adult arm. Stage 15 embryos were staged according to Naef (1928), washed in filtered artificial seawater and dissected from the chorion and off the yolk. All tissues harvested were quickly dissected and either flash frozen on liquid nitrogen or placed in RNAlater (Ambion) at 4°C overnight (Table S2.1). Samples were stored at -70°C. RNA was isolated using Trizol (Invitrogen) following manufacturer's instructions.

RNA integrity was analyzed with a Bioanalyzer 2100; only samples with clean rRNA peaks and little to no degradation were used. Total RNA was polyA selected and directionally sequenced at the University of Chicago Genomics Facility on an Illumina HiSeq2000 per manufacturer's instructions, generating 100bp paired-end reads with an insert size of 300bp.

Tissue	Animals	Initial Storage Conditions	Mb	Reads
Ova	Adult #3, female	Liquid Nitrogen	5,626	55,698,688
Testes	Adult #6, male	Liquid Nitrogen	7,422	73,479,938
Hepatopancreas, Kidney, Heart (Viscera)	Adults #2 (male), #3 (female), #5 (female)	RNAlater	7,176	71,057,530
Posterior Salivary Gland (PSG)	Adult #4, male	Liquid Nitrogen	7,195	71,237,812
Skin	7 hatchlings	Liquid Nitrogen	7,925	78,459,268
Suckers	Adult #4, male	Liquid Nitrogen	7,092	70,218,752
Stage 15	1 clutch, ~15 embryos	Liquid Nitrogen	6,790	67,227,866
Retina	Adult #2, male	Liquid Nitrogen	6,364	63,006,538
Optic Lobe (OL)	Adult #2, male	RNAlater	6,116	60,550,668
Supraesophageal Brain (Supra)	Adult #1, female	Liquid Nitrogen	15,475	154,754,152
Subesophageal Brain (Sub)	Adult #1, female	Liquid Nitrogen	14,103	141,027,874
Axial Nerve Cord (ANC)	Adult #3, female	RNAlater	6,271	62,091,238

Table S2.1. *O. bimaculoides* transcriptome sequencing summary.

2.2 Mapping reads to the genome for expression analysis

Following removal of adapters and low quality sequences, reads were mapped to the genome assembly using TopHat 2.0.11 (Trapnell et al., 2009). A range of 76-90% of reads from different samples could be mapped to the genome. Accepted reads were sorted and indexed with SAMtools (Li et al., 2009). The read counts in each tissue were produced with the bedtools multicov program (Quinlan and Hall, 2010) using the gene model coordinates. The counts were normalized by the total count in each tissue and by the length of the gene. Heatmaps showing expression patterns were generated in R using the heatmap.2 function.

2.3 de novo *transcriptome assembly using Trinity*

Adapters and low quality reads were removed before assembling transcriptomes using the Trinity *de novo* assembly package [version r2013-02-25, (Grabherr et al., 2011; Haas et al., 2013)], both individually and in groups (assembly statistics summarized in Table S2.2). Following assembly, peptide-coding regions were translated using TransDecoder as part of the Trinity package.

Transcriptome	Total Trinity 'genes'	Total Trinity Transcripts	Percent GC	N10 (bp)	N50 (bp)	Median Contig Length (bp)	Average Contig Length (bp)	Total Assembled Bases (bp)
Ova	57,662	68,363	36.87	3,455	1,034	377	667.68	45,644,899
Testes	90,425	118,428	37.66	4,536	1,406	393	780.56	92,440,034
Viscera	95,160	129,052	38.29	4,701	1,446	415	812.59	104,866,269
PSG	68,598	87,232	37.68	5,420	1,643	420	865.29	75,481,024
Suckers	102,168	132,422	37.81	5,947	1,721	388	850.15	112,578,846
Skin	93,008	118,404	37.49	5,262	1,528	393	808.92	95,779,900
St15	87,932	124,062	37.78	5,178	1,544	384	799.54	99,192,901
Retina	97,137	124,585	37.61	5,304	1,508	387	798.77	99,514,962
OL	132,961	167,693	37.52	5,156	1,356	373	748.79	125,567,386
Supra	177,569	235,295	37.2	5,465	1,438	369	765.63	180,150,005
Sub	144,721	192,152	37.4	5,394	1,552	373	793.05	152,385,368
ANC	109,157	142,259	37.43	5,900	1,612	398	837.28	119,110,298
All	305,458	373,396	36.33	4,619	1,152	370	702.83	262,434,425

Table S2.2. *O. bimaculoides* transcriptome assembly summary. Statistics are based on all transcript contigs. Total Trinity ‘genes’ refers to the number of transcript clusters generated by the assembly (isogroups), while Total Trinity Transcripts indicates the number of isoforms (isotigs).

We compared the *de novo* assembled RNA-Seq output from Trinity to the genome assembly to evaluate the completeness of the assembly. To minimize the number of spuriously assembled transcripts, only transcripts with ORFs predicted by TransDecoder were mapped onto the genome with BLASTN. We found that only 1,130 out of 48,259 transcripts with ORFs (2.34%) did not have a match in the genome with a minimum identity of 95%.

3. FLUORESCENCE-BASED GENOME SIZE ESTIMATE

To estimate the size of the octopus genome experimentally, embryonic cells from *O. bimaculoides* (whole embryos, stage 20, Figure S3.1 bottom panel) and *Danio rerio* (whole embryos, 24 hours post-fertilization, Figure S3.2 top panel) were harvested, stained with DAPI and separated by flow cytometry on a BS FACSCanto analyzer. Relative DAPI content of the cells is indicated in red, with the sharp peaks corresponding to 2C (diploid content) (Vinogradov, 1998). Cell isolation, staining and sorting were run in parallel. *D. rerio* was included for calibration. The ratio of the *O. bimaculoides* peak to the *D. rerio* peak is 1.83-1.85 across multiple runs. Accepting a haploid genome size estimate of 1.454 Gb for *D. rerio* (Freeman et al., 2007; Howe et al., 2013), we estimate the genome size of *O. bimaculoides* to be 2.66-2.68 Gb.

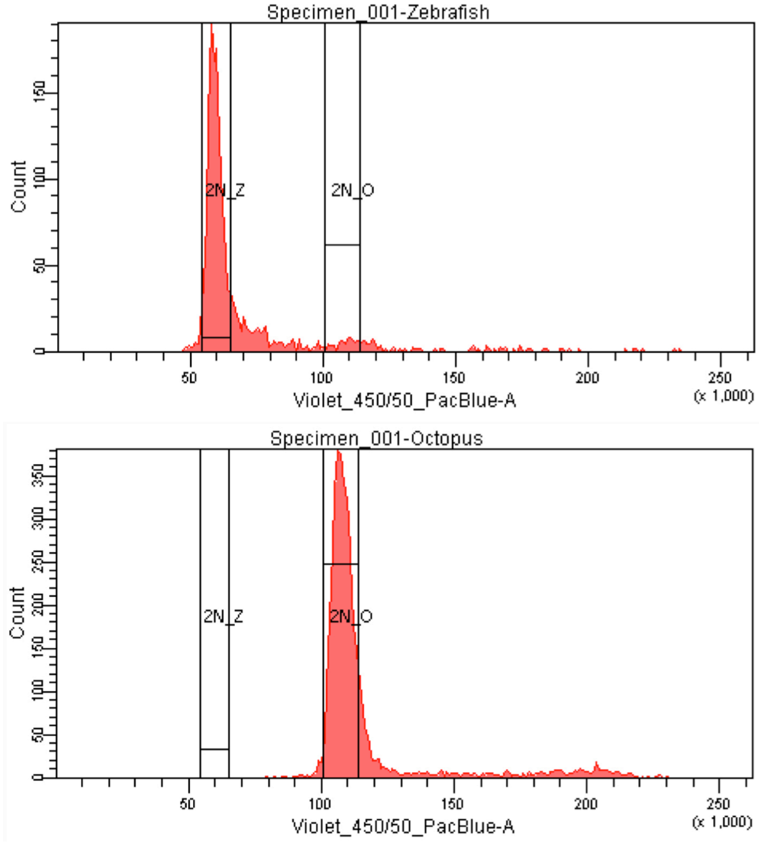


Figure S3.1. Fluorescence peaks from genomic DNA in *D. rerio* (top) and *O. bimaculoides* (bottom) embryonic cells.

This estimate of genome size accords well with the 2.87 Gb *O. bimaculoides* genome estimate based on counting k-mers, given an assumption that genome sequencing reads sample the genome uniformly.

4. ANNOTATION OF PROTEIN-CODING GENES AND TRANSPOSABLE ELEMENTS

4.1 Annotation

A *de novo* repeat library was made by running RepeatModeler (Smit and Hubley, 2008-2010) on the genome to produce a putative library of repeat sequences. Sequences with Pfam domains associated with non-transposable element functions were removed from the library of repeat sequences. A preliminary library was used to mask ~43% of the genome with RepeatMasker prior to annotation (Smit et al., 1996-2010).

RNA-Seq data from ANC, skin, retina, PSG, viscera, testes, ova, St15, suckers, and optic lobe tissues were aligned and assembled on genome using pertran (paired-end transcript assembler; S. Shu, DOE Joint Genome Institute, unpublished). The 107,129 pertran assemblies as well as 156,009 cephalopod EST sequences from NCBI were aligned to the genome using PASA (Haas et al., 2003), which aligns ESTs to the genome assembly sequence via GMAP (Wu and Watanabe, 2005), then filters hits to ensure proper splice boundaries. This process found 88,210 loci with RNA sequence alignment.

The version 2.1 gene annotation set described in the main text was produced using the DOE Joint Genome Institute's gene prediction pipeline which combines homology-based and *ab initio* predictions with transcript-based evidence. Protein sequences from diverse metazoans (*Xenopus tropicalis*, *Lottia gigantea*, *Aplysia californica*, *Crassostrea gigas*, *Nematostella vectensis*, and human) were aligned to the genome by BLASTX and extended by the EXONERATE algorithm (Slater and Birney, 2005). The putative loci from these peptide alignments, as well as the full set of predicted EST assemblies from PASA analysis were submitted to the Fgenesh+ (Salamov and Solovyev, 2000) and GenomeScan (Yeh et al., 2001) gene prediction algorithms, including up to 2 kb additional sequence on each side of peptide alignments. The best of these predictions at each gene locus was selected based on a weighted scoring of metrics, including its predicted peptide to known peptide BLASTP score, EST coverage, protein homolog coverage and degree of support for predicted exon-intron boundaries. This predicted dataset was extended by a second run through the PASA algorithm (Haas et al., 2003) to model UTR and alternative splice variants based on EST support and the final results were filtered to remove genes identified as transposon-related or overlapping the RepeatModeler *de novo* repetitive region predictions by greater than 20%.

This initial annotation dataset was further filtered by requiring annotated genes to have either RNA-Seq coverage, or a “c-score” (*i.e.*, BLASTP score/MBH BLASTP score) of 0.5 or better. EST support was also examined to check that aligned coverage followed the same intron splicing pattern as the gene model.

The final prediction set consists of 33,638 protein-coding loci (Table S4.1.1), 29,844 of which have some degree of transcriptome support. Half of them have support for more than 50% of their length. In addition, 2,819 genes show evidence of alternative splicing as annotated by PASA (Haas et al., 2003). Those genes with ten or more predicted alternative transcripts are reported in Table S4.1.2.

***Octopus bimaculoides* annotation**	
Primary transcripts (loci)	33,368
Alternative transcripts	4,947
Total transcripts	38,585
For primary transcripts:	
Average number of exons	4.3
Median exon length	149 bp
Median intron length	1506 bp
Gene model support:	
Any transcriptome support	29,844
Transcriptome support over 50% of their lengths	19,318
Peptide homology coverage of over 50%	23,138
Pfam annotation	15,319
PANTHER annotation	14,694
KOG annotation	6,083
KEGG orthology annotation	4,151
E.C. number annotation	1,872

Table S4.1.1. Annotation statistics for *O. bimaculoides*.

Gene	**Isoforms**	**SProt ID Best Hit**	**Description**
Ocbimv22017953m.g	128	sp|G5E8K5|ANK3_MOUSE	Ankyrin-3 OS=Mus musculus GN=Ank3 PE=1 SV=1
Ocbimv22030059m.g	64	sp|Q6PD31|TRAK1_MOUSE	Trafficking kinesin-binding protein 1 OS=Mus musculus GN=Trak1 PE=1 SV=1
Ocbimv22019629m.g	42	sp|P41737|LRCH1_FELCA	Leucine-rich repeat and calponin homology domain-containing protein 1 (Fragment) OS=Felis catus PE=2 SV=2
Ocbimv22023057m.g	24	sp|Q8R1S4|MTSS1_MOUSE	Metastasis suppressor protein 1 OS=Mus musculus GN=Mtss1 PE=1 SV=1
Ocbimv22021630m.g	24	sp|Q9IBG7|KCP_XENLA	Kielin/chordin-like protein OS=Xenopus laevis GN=kcp PE=2 SV=1
Ocbimv22017088m.g	24	sp|Q99755|PI51A_HUMAN	Phosphatidylinositol 4-phosphate 5-kinase type-1 alpha OS=Homo sapiens GN=PIP5K1A PE=1 SV=1
Ocbimv22005227m.g	24	sp|O15027|SC16A_HUMAN	Protein transport protein Sec16A OS=Homo sapiens GN=SEC16A PE=1 SV=3
Ocbimv22004770m.g	20	sp|Q70FJ1|AKAP9_MOUSE	A-kinase anchor protein 9 OS=Mus musculus GN=Akap9 PE=2 SV=2
Ocbimv22015048m.g	18	sp|G3MWR8|MICA3_BOVIN	Protein-methionine sulfoxide oxidase MICAL3 OS=Bos taurus GN=MICAL3 PE=3 SV=1
Ocbimv22010506m.g	18	sp|Q8IID4|DYHC2_PLAF7	Dynein heavy chain-like protein PF11_0240 OS=Plasmodium falciparum (isolate 3D7) GN=PF11_0240 PE=3 SV=1
Ocbimv22003188m.g	18	sp|P0C6P5|ARG28_RAT	Rho guanine nucleotide exchange factor 28 OS=Rattus

			norvegicus GN=Arhgef28 PE=1 SV=1
Ocbimv22028814m.g	17	sp|Q6H236|PEG3_BOVIN	Paternally-expressed gene 3 protein OS=Bos taurus GN=PEG3 PE=2 SV=1
Ocbimv22034567m.g	16	sp|Q8C8U0|LIPB1_MOUSE	Liprin-beta-1 OS=Mus musculus GN=Ppfibp1 PE=1 SV=3
Ocbimv22023772m.g	16	sp|P91943|PANG1_DROME	Protein pangolin, isoforms A/H/I/S OS=Drosophila melanogaster GN=pan PE=1 SV=1
Ocbimv22018605m.g	16	sp|O46382|BIG1_BOVIN	Brefeldin A-inhibited guanine nucleotide-exchange protein 1 OS=Bos taurus GN=ARFGEF1 PE=1 SV=1
Ocbimv22013946m.g	16	sp|Q9NZN8|CNOT2_HUMAN	CCR4-NOT transcription complex subunit 2 OS=Homo sapiens GN=CNOT2 PE=1 SV=1
Ocbimv22005183m.g	16	sp|Q17BU3|KIF1A_AEDAE	Kinesin-like protein unc-104 OS=Aedes aegypti GN=unc-104 PE=3 SV=1
Ocbimv22004739m.g	16	sp|Q92609|TBCD5_HUMAN	TBC1 domain family member 5 OS=Homo sapiens GN=TBC1D5 PE=1 SV=1
Ocbimv22031030m.g	15	sp|Q4KKX4|NCOR1_XENTR	Nuclear receptor corepressor 1 OS=Xenopus tropicalis GN=ncor1 PE=2 SV=1
Ocbimv22037811m.g	12	sp|A1A5G0|CLAP1_XENTR	CLIP-associating protein 1 OS=Xenopus tropicalis GN=clasp1 PE=1 SV=1
Ocbimv22037201m.g	12	sp|Q8VCB2|MED25_MOUSE	Mediator of RNA polymerase II transcription subunit 25 OS=Mus musculus GN=Med25 PE=1 SV=1
Ocbimv22036429m.g	12	NR	PREDICTED: LOW QUALITY PROTEIN: plectin, partial [Anas platyrhynchos], putative SMC_N, Pfam02463 domain
Ocbimv22031313m.g	12	sp|O35206|COFA1_MOUSE	Collagen alpha-1(XV) chain OS=Mus musculus GN=Col15a1 PE=1 SV=2
Ocbimv22027614m.g	12	sp|P97603|NEO1_RAT	Neogenin (Fragment) OS=Rattus norvegicus GN=Neo1 PE=2 SV=1
Ocbimv22024792m.g	12	sp|P07909|ROA1_DROME	Heterogeneous nuclear ribonucleoprotein A1 OS=Drosophila melanogaster GN=Hrb98DE PE=2 SV=1
Ocbimv22022931m.g	12	sp|Q9NRA8|4ET_HUMAN	Eukaryotic translation initiation factor 4E transporter OS=Homo sapiens GN=EIF4ENIF1 PE=1 SV=2
Ocbimv22019579m.g	12	sp|Q63532|SPR1A_RAT	Cornifin-A OS=Rattus norvegicus GN=Sprr1a PE=2 SV=1
Ocbimv22018747m.g	12	sp|Q8TD84|DSCL1_HUMAN	Down syndrome cell adhesion molecule-like protein 1 OS=Homo sapiens GN=DSCAML1 PE=1 SV=2
Ocbimv22011363m.g	12	sp|Q0III3|DC1I2_BOVIN	Cytoplasmic dynein 1 intermediate chain 2 OS=Bos taurus GN=DYNC1I2 PE=1 SV=1
Ocbimv22010641m.g	12	sp|Q8VC31|CCDC9_MOUSE	Coiled-coil domain-containing protein 9 OS=Mus musculus GN=Ccdc9 PE=2 SV=1
Ocbimv22006841m.g	12	sp|Q9BVI0|PHF20_HUMAN	PHD finger protein 20 OS=Homo sapiens GN=PHF20 PE=1 SV=2
Ocbimv22000216m.g	12	sp|P98160|PGBM_HUMAN	Basement membrane-specific heparan sulfate proteoglycan core protein OS=Homo sapiens GN=HSPG2 PE=1 SV=4
Ocbimv22018912m.g	11	sp|Q91348|F26L_CHICK	6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase OS=Gallus gallus PE=2 SV=2
Ocbimv22035390m.g	10	sp|Q0KL00|PIEZ1_RAT	Piezo-type mechanosensitive ion channel component 1 OS=Rattus norvegicus GN=Piezo1 PE=2 SV=3
Ocbimv22035154m.g	10	sp|Q9NYQ6|CELR1_HUMAN	Cadherin EGF LAG seven-pass G-type receptor 1 OS=Homo sapiens GN=CELSR1 PE=1 SV=1
Ocbimv22029548m.g	10	sp|A2AWA9|RBGP1_MOUSE	Rab GTPase-activating protein 1 OS=Mus musculus GN=Rabgap1 PE=1 SV=1
Ocbimv22024015m.g	10	NR	hypothetical protein CLF_113084, partial [Clonorchis sinensis]
Ocbimv22018408m.g	10	sp|Q6PJ21|SPSB3_HUMAN	SPRY domain-containing SOCS box protein 3 OS=Homo sapiens GN=SPSB3 PE=1 SV=2

Table S4.1.2. *O. bimaculoides* genes with 10 or more predicted alternative transcripts along with their best SwissProt or NR hit at e-value 1e-3 or better.

4.2 Transposable element annotation and expansions

Repetitive elements were identified *de novo* using RepeatScout 1.0.5 (Price et al., 2005) based on a random sampling of 1/3 of the assembled scaffolds and using lmer size 16. The filtering was done according to the RepeatScout pipeline and the identified repeats were annotated using RepBase version 20140131 (Jurka et al., 2005) with RepeatMasker (Smit et al., 1996-2010), TBLASTX against this database, and a BLASTX against a custom set of transposon-like sequences from NCBI (available upon request). After manual curation, this process allowed for the annotation of 2,851 out of 6,899 sequences. In total, at least 45% of the genome is estimated to be repetitive. Despite the high repeat content, genes involved in transposon silencing (*e.g.*, piwi) are not expanded. The unannotated sequences constitute low-copy repeats, with the highest copy number repeats being SINE retrotransposons (at least 3.6% of the genome) as well as simple repeats (at least 11.01%, Tables S4.2.1 and S4.2.2). Interestingly, all highly abundant SINE repeats belong to the OK_SINE class described in Oshima and Okada (1994); we find no evidence for expansion of other cephalopod SINEs (Ceph-SINE) described by Akasaki et al. (2010).

In addition to the RepeatScout library, we have constructed a RepeatModeler library (Smit et al., 1996-2010, http://www.repeatmasker.com) producing 368 elements (154 unknown). This library masks 43% of the genome, with the same repeat classes showing similar total content as well as age distribution (data not shown). For dating purposes we focused on using a RepeatScout-based library, as it seems to be less aggressive in its consensus reconstruction, thus allowing for more accurate age estimation. Additionally, around 920 (including 508 annotated) elements in the RepeatScout library lack a good (>80% identity) hit in the RepeatModeler library (while only 26 total, including 18 annotated, elements are lacking in the RepeatScout library). Both libraries are available upon request.

Repeat name	Masked base pairs
1. (CA)nSimple_repeat14	63,864,586
2. OK_SINE2/tRNA_Octopus93	31,217,901
3. AmnSINE2SINE/Deu9	23,654,527
4. (TA)n	23,454,084
5. (CA)n	23,211,490
6. AmnSINE2SINE/Deu5	19,383,767
7. (GA)n	14,245,304
8. OR1_SINE2/tRNA_Octopus2	13,003,833
9. (TATG)nSimple_repeat15	12,096,880
10. (TAA)n	11,348,724

Table S4.2.1. Highest copy number repeats in the octopus genome as detected by RepeatMasker.

Repeat class	Count	Bases masked	Percent genome
DNA Transposon (total)	52,645	8,773,457	0.38
Sola	13	5,155	0.00
EnSpm	13,903	895,897	0.04
hAT	11,680	2,642,475	0.11
Kolobok	103	2,514	0.00
Helitron	26,946	5,227,416	0.23
Retrotransposon (total)	935,223	151,588,316	6.59
Tx1	3,218	990,327	0.04
RTE	164,889	42,021,625	1.83
SINE	653,030	82,897,804	3.60
CR1	52,213	11,828,641	0.51
R4	43,817	9,356,147	0.41
L1	18,056	4,493,772	0.20
Endogenous Retrovirus (total)	8,183	697,163	0.03
ERV	8,183	697,163	0.03
Low complexity (total)	2,949,809	253,193,996	11.01
Other low complexity	76,380	9,655,237	0.42
T rich	20,052	2,943,274	0.13
AT rich	545,951	38,076,729	1.66
Simple repeat	2,307,355	202,517,198	8.81
GC rich	71	1,558	0.00
Other	249,115	38,628,072	1.68
Satellite	10,507	1,497,543	0.07
Other transposons	238,608	37,130,529	1.61
rRNA (total)	1,222	121,819	0.01
rRNA	907	92,485	0.00
SSU-rRNA	315	29,334	0.00
tRNA	2,123	162,441	0.01
snRNA	492	39,598	0.00
Other repeats and fragments (unknown)	147,203,282	334,892,743	14.56

Table S4.2.2. Octopus repeat classes and their abundance.

We dated transposable elements with RepeatMasker and adjusted the distances for multiple substitution using the Jukes-Cantor formula JK= -3/4*log(1-4*d/3), where d is the distance estimated by RepeatMasker. This method identified two major expansion waves in the octopus genome at 0.09 and 0.2 (Figure S4.2.1). This corrected distance estimation should be directly comparable to the dS measure (see Supplementary Note 7). Using an estimate for the synonymous substitutions per million years (Supplementary Note 7), we calculated that JK 0.09 corresponds to 0.09 [JK] / 0.0036 [JK/myr] = 25 million years, and 0.2 [JK] / 0.0036 [JK/myr] = 56 million years. This estimated timeframe suggests that those expansions happened specifically in the octopus lineage after its divergence from other cephalopod species included in our transcriptome set (Extended Data Figure 10a).

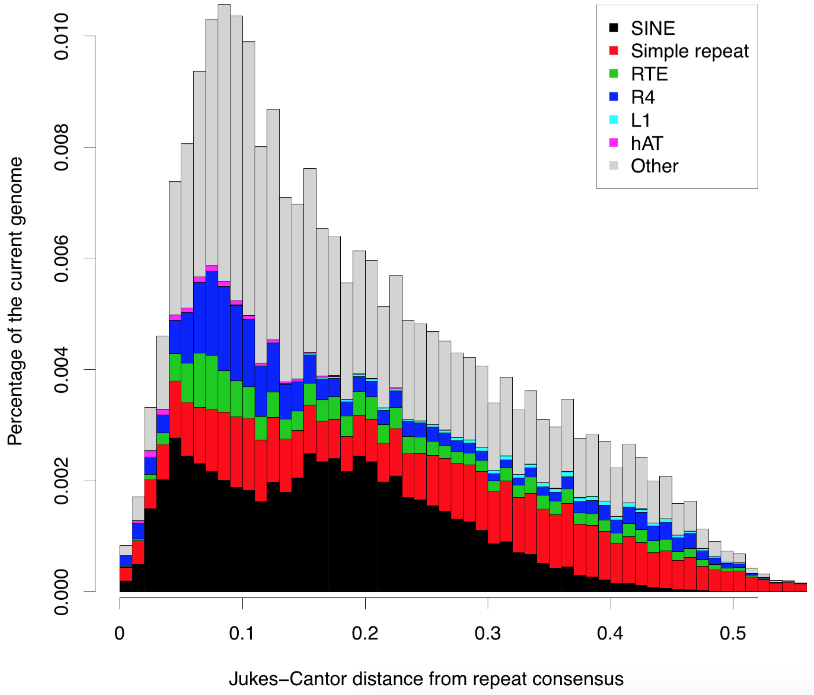


Figure S4.2.1. Transposable element insertion history (Jukes-Cantor distance adjusted). Percentage of transposon occupancy at different age classes (measured as distance from consensus) is shown. The most highly abundant repeat classes are highlighted: SINEs, simple repeats (not showing an expansion), R4, RTE. DNA transposon hAT is shown for comparison.

4.3 Transposon expression

To quantify transposon expression in different tissues, the original transcriptome to genome map was used to obtain counts using the bedtools multicov tool based on RepeatMasker gff output. Counts were normalized by the total expression of transposons in a given tissue. Heatmap representations were computed using heatmap.2 function in the gplots package in R, using row normalization with ‘scale’ function. We found that at least half of a total of 5,496,558 octopus transposable element (TE) loci show expression (2,685,265). Using our gene annotations, which incorporate transcriptomic data, we asked what proportion of the active TE loci overlap with UTR or exonic sequences. We found that 15.2% of active TE loci have such overlaps: 1.6% overlap with UTR sequences, 0.8% overlap with exon regions, and 12.8% lie within intron regions. The overlapping active TE loci likely do not constitute expression related to “single-locus” TE activity; we have removed them from our expression analysis. Notably, looking at the age distribution of repeat copies (of at least 100bp) we found that very young and very old copies show the lowest levels of expression (Figure S4.3.1). The most abundant expression is seen in middle-aged (Jukes-Cantor 0.2-0.3) repeats.

Across tissues, we find that transposable elements are highly expressed in neural tissues, accounting for 8% of total expression in ANC, OL, retina, subesophageal and supraesophageal brain, compared to 5% of total expression in other tissues. The main representative is the most abundant transposon, OK_SINE. We also found transposons that are only active in mature eggs, such as Penelope (Extended Data Figure 8a).

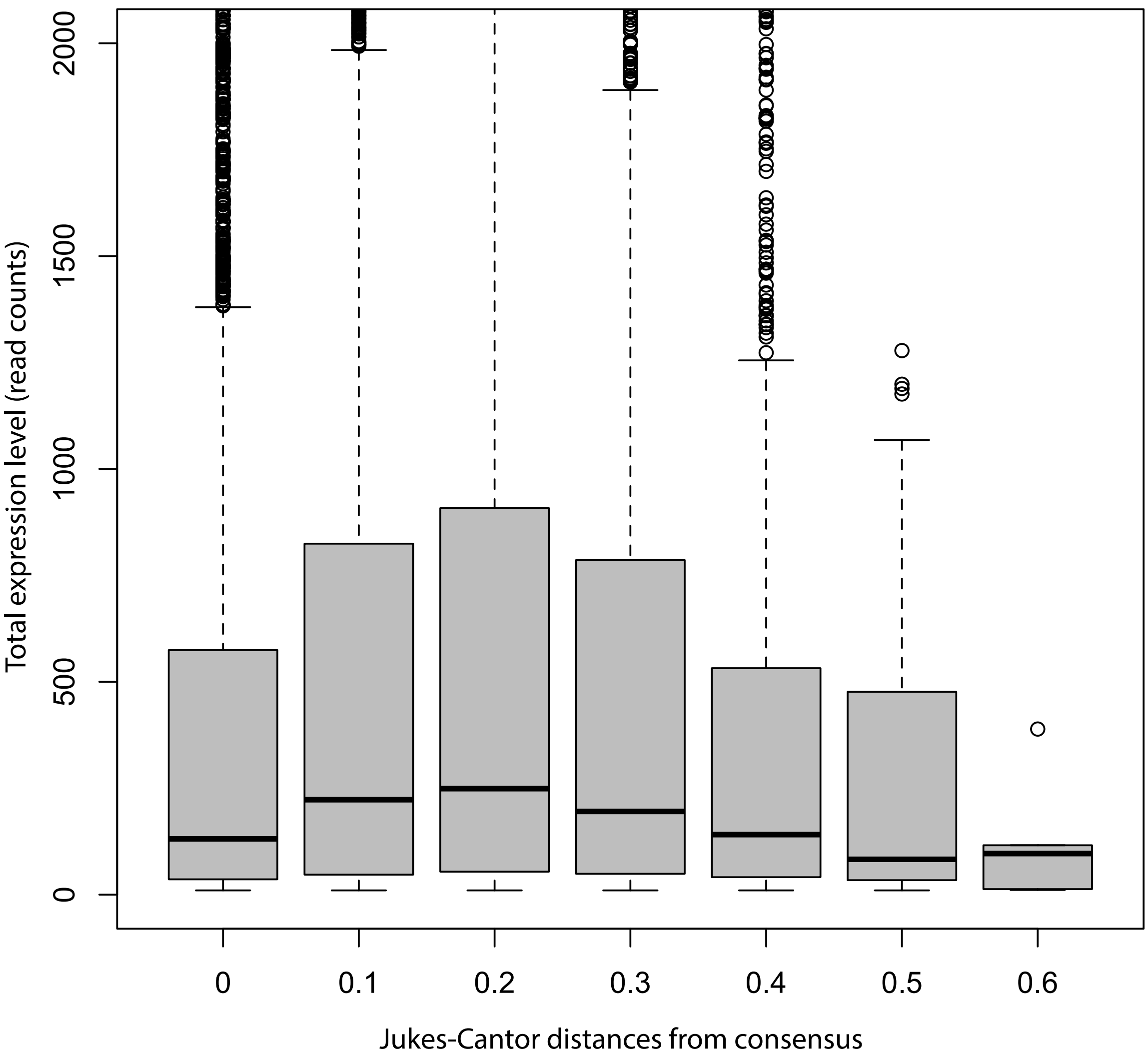


Figure S4.3.1. Total expression level (read counts, y-axis) distribution for repeat loci at different age categories (Jukes-Cantor distances, x-axis).

5. RNA EDITING

RNA editing by adenosine deamination is a process in which adenosines are converted to inosines in mRNA, mediated by adenosine deaminases that act on RNA (ADARs). Editing by ADARs is the most common form of RNA editing described in animals, and it may serve a range of biological functions, including transposon silencing (Savva et al., 2013), targeting virus dsRNA, and regulating gene expression and mRNA stability (Wang et al., 2013). Since the translation machinery interprets inosine as guanosine, A-to-I editing can also alter codon sequences, potentially changing the amino acid sequence of the resulting protein. Many mRNAs are edited at multiple sites: for example, five bases in the human serotonin receptor are edited, yielding 24 protein isoforms (Fitzgerald et al., 1999; Werry et al., 2008). A-to-I editing in coding sequences can therefore increase the diversity of proteins encoded by a single gene.

5.1 ADAR genes and phylogeny

ADARs share a common domain organization, with two to three double-stranded RNA binding domains (dsRBDs) and a C-terminal adenosine deaminase domain. Humans have three ADARs, though only ADAR1 and ADAR2 have been described as having editing capabilities (Chen et al., 2000). ADAR2 has two dsRBDs, while ADAR1 has three dsRBDs and one or two z-DNA binding domains. These domains, called z-alpha domains, have thus far only been described in mammals. *Drosophila* has one *ADAR*, which is most closely related to *ADAR2* (Savva et al., 2012). Two isoforms of *ADAR2* have been described in *D. pealeii*, with two and three dsRBDs, respectively (Palavicini et al., 2009). *O. bimaculoides* has both *ADAR1* and *ADAR2* homologs, as well as an *ADAR-like* gene (Extended Data Figure 1a). Like human *ADAR1*, the octopus *ADAR1* has a z-alpha domain but only one dsRBD (Extended Data Figure 1b). *O. bimaculoides ADAR2* and *ADAR-like* both have two dsRBDs and a single adenosine deaminase domain.

5.2 Identification of putative edited positions

As ADARs have been found in a range of bilaterians and non-bilaterian metazoans, RNA editing is likely an ancient mechanism for increasing the complexity of the proteome. However, there are striking differences in the pattern of editing between humans and invertebrates. Current estimates of A-to-I editing sites in humans range from 1.4 million to 100 million (Ramaswami and Li, 2014; Ulbricht and Emeson, 2014), though relatively few of these edits create non-synonymous changes. Currently, 602 non-synonymous edits are described in the Rigorously Annotated Database of A-to-I RNA editing (www.rnaedit.com). Nearly twice that number has been found in *Drosophila* (Graveley et al., 2011; Ramaswami and Li, 2014), and several hundred editing sites have been described across just a few genes in squid, suggesting that altering protein sequence by A-to-I editing is quite common in invertebrate coding sequences (Rosenthal and Seeburg, 2012). Edited transcripts resulting in recoding have been primarily identified in genes expressed in the nervous system in all three of these animals.

Here, we leveraged our extensive transcriptome sequencing to identify possible sites of RNA editing. We used SAMtools (Li et al., 2009) to call SNPs between the genome and RNA-Seq reads that had been mapped to the genome with TopHat (Trapnell et al., 2012). Genomic SNPs were identified as described in Supplementary Note 1.4. Using bedtools (Quinlan and Hall, 2010), we removed SNPs predicted in both the transcriptome and the genome and discarded SNPs that had a Phred score below 40 or were outside of predicted genes. SNPs were binned according to the type of nucleotide change and taking into account the direction of transcription. Using this method, we found 3,572 unique A-to-G edits in the coding sequences of 2,012 genes, the majority of which are edited in neural tissues, particularly the ANC and optic lobe (Extended Data Figure 1d). This method also identified other DNA-RNA differences (DRDs). It is likely that these DRDs represent polymorphisms rather than edits, as the DNA and RNA were collected from different animals and the non A-to-G calls have the same relative proportions of variants as do the polymorphisms predicted from genomic sequence alone. There were far more A-to-G changes than any of the other types of DRDs, however, which is consistent with RNA editing. These elevated A-to-G changes were also found predominantly in neural tissue, which is where RNA editing by ADARs has been previously described, and which parallels the expression of ADARs in our transcriptomes (Extended Data Figure 1c). We also found that the distribution of A-to-G edits in exonic sequence was distinctive in different types of tissues: neural structures had a higher percentage of edits in the coding sequence than did reproductive or other tissues (Figure S5.1).

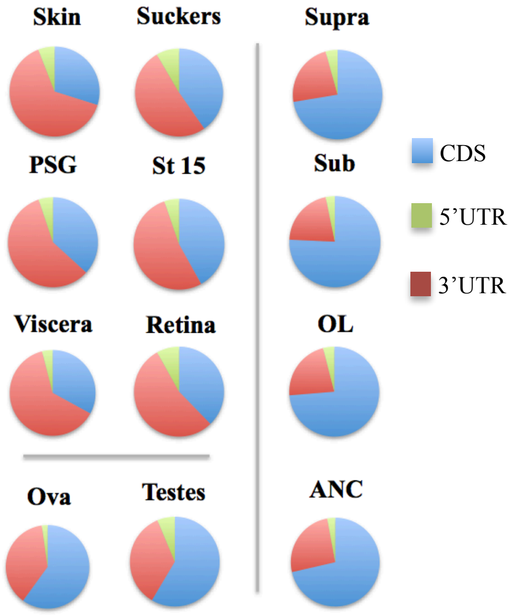


Figure S5.1. Distribution of edits in exonic sequence in *O. bimaculoides* transcriptomes.

We manually examined a number of these edits to identify changes that result in alterations of protein sequence, focusing on the optic lobe and the ANC. These tissues showed extensive editing of delayed rectifying potassium channels, Kv1 and Kv2, which has been described in other cephalopods. In *O. bimaculoides*, we identified two Kv1 genes (*KCNA1A* and *KCNA1B*) and one Kv2 gene (*KCNB1*). In *KNCA1A*, we identified all of the edits described for *KCNA1* in temperate octopus species (Garrett and Rosenthal, 2012) (Table S5.1). A number of edits in *KCNB1* were conserved with squid, two of which are found in octopus, squid and flies (Yang et al., 2008). We also identified extensive editing of *ADAR1*, and a single edit in *ADAR2*, which is highly edited in squid (Palavicini et al., 2009). We were able to identify thousands of candidate RNA editing sites by comparing the genome to the transcriptome and manual searches through these alignments showed many more candidate editing sites, indicating that RNA editing in octopus is likely much more extensive than what we describe here. This finding is consistent with a recent study using next generation sequencing in squid (Alon et al., 2015), which was published while this study was in review.

	# editing sites	**# NS editing sites (nt/aa)**	**NS editing sites**
ADAR1	24	15	K81G, K368R, N335D, T419A, S432G, K744G, K748E, K815E, I911V, H924R, I939M, I940V, N1120D, N1189D, T1208A
ADAR2	1	1	N98G
GRIA2	10	3	N332G, E333G, R632G
Kv1/KCNA1A	10	(10/8)	N12S, S26G, K99E, I107V, I139V, M142V, I293V, I344V
Kv1/KCNA1B	2	2	R27G, I417V
Kv2/KCNB1	13	11	H66R, H70R, H74R, H75R, R154G, S158G, S165G, I407V, I506V, Y591C, I612V

Table S5.1. Non-synonymous (NS) changes in ADARs and ion channel proteins in the optic lobe. Manually identified edits were mapped onto protein sequences to determine if the changes altered the protein sequence. Edits in *KCNA1* are shared with other octopus species (in blue), and edits in *KCNB1* are shared with squid (green) and squid and *Drosophila* (orange).

Our analysis showed that edited transcripts are not limited to those involved in neuronal signaling or in editing enzymes. Rather, we also found evidence of RNA editing in housekeeping genes, such as tubulin, and signaling proteins, such as kinases (Table S5.2). This finding indicates RNA editing that alters coding sequence is frequently employed in a wide variety of genes. Roughly two-thirds of the A-to-G changes in coding sequences we manually examined alter encoded amino acids, which is similar to the ~64% rate described in *Drosophila* (St

Laurent et al., 2013) and squid (Alon et al., 2015). These data support a broad and important role for RNA editing in octopus.

Gene name	Non-synonymous edits	Gene family
Ocbimv22032594m	S126G	tubulin
Ocbimv22010928	I172V, I178V	tubulin
Ocbimv22007862m	N52S	tubulin
Ocbimv22032594	S126G	tubulin
Ocbimv22002831m	C62G	tubulin
Ocbimv22006874m	T33A, I45V, I133V, K135E	tubulin
Ocbimv22010931m	K121R, I177V	tubulin
Ocbimv22035659m	Y49F, T271A, I384V	tubulin
Ocbimv22035660m	I384V	tubulin
Ocbimv22000366m	K111R	tubulin
Ocbimv22013969m	T400A	tubulin
Ocbimv22007878m	T285A	kinase
Ocbimv22024168m	K70R, I123V, E798G, T897A, S900G	kinase
Ocbimv22013896m	K536E, E576G, K807G	kinase
Ocbimv22023316m	I221M, I263V, T288A	kinase
Ocbimv22021615m	T61A, T300A, K649R	-
Ocbimv22001351m	I98V, M237V, S246G	kinase
Ocbimv22011498m	S103G, Y202C, Y322C, E374G, I491V, T497A, I610M, R837G, M870V	-

Table S5.2. Non-synonymous changes in tubulins and kinases in the optic lobe. Manually identified edits were mapped onto protein sequences to determine if the changes altered the protein sequence.

6. ANALYSIS OF SYNTENY

6.1 Conservation of synteny

Conservation of macro-synteny (chromosomal linkage) and micro-synteny (clusters of unrelated genes that retain tight linkage) reflects both the mechanisms and dynamics of chromosome rearrangement as well as possible selection for the retention of linkages between genes and associated regulatory sequences. Micro-synteny was computed based on metazoan node gene families (Supplementary Note 7). We used Nmax 10 (maximum 10 intervening genes) and Nmin 3 (minimum of three genes in a syntenic block) according to the pipeline described in Simakov et al. (2013). To simplify gene family assignments we limited our analyses to gene families shared among human, amphioxus, *Capitella*, *Helobdella*, *Octopus*, *Lottia*, *Crassostrea*, *Drosophila*, and *Nematostella*. This analysis yielded 4,033 gene families. Due to this reduced species sampling, we required ancestral bilaterian syntenic blocks to have a minimum of one species present in both ingroups, or in one ingroup and one outgroup. This analysis identified 198 syntenic blocks (Table S6.1). Octopus retains only 34 blocks, which is fewer than those retained by other lophotrochozoans, such as *Lottia* (96 blocks) and *Capitella* (69 blocks). Even the genome of the Pacific oyster *Crassostrea gigas*, which has a scaffold N50 size comparable to our octopus assembly (401 kb; Zhang et al., 2012), shows greater conservation of synteny (48 blocks).

The genome assemblies used in our synteny calculation varied significantly in the number of genes per scaffold. The average values were as high as 950 genes in human or as low as 4 in *Capitella* or octopus. To account for underestimation of synteny due to fragmented genome assemblies, we simulated shorter assemblies by cutting the scaffolds with more than 5 genes down to sub-scaffolds with a random number of genes drawn from an exponential distribution (expected value of 5). This procedure produced artificial genomes with the median of 2 genes per scaffold (average of 2-3 genes in all genomes). After repeating this procedure 50 times and computing the synteny as described above, we found that octopus still has a significant decrease in synteny as compared to other lophotrochozoans (Figure S6.1). In fact, the median proportion of syntenic blocks is comparable between *Drosophila* (0.19) and the octopus (0.18). The greatest loss in synteny was found in the *Helobdella* genome, which retains only 14% of the syntenic blocks in the simulations.

Species	**Lgi**	**Cte**	**Obi**	**Hro**	**Cgi**	**Dme**	**Nve**	**Bfl**	**Hsa**
Bilaterian block count	96	69	34	22	48	24	67	140	68
Median proportion in simulated data	0.38	0.28	0.18	0.14	0.20	0.19	0.27	0.68	0.32

Table S6.1. Distribution of 198 inferred bilaterian syntenic blocks among individual species.

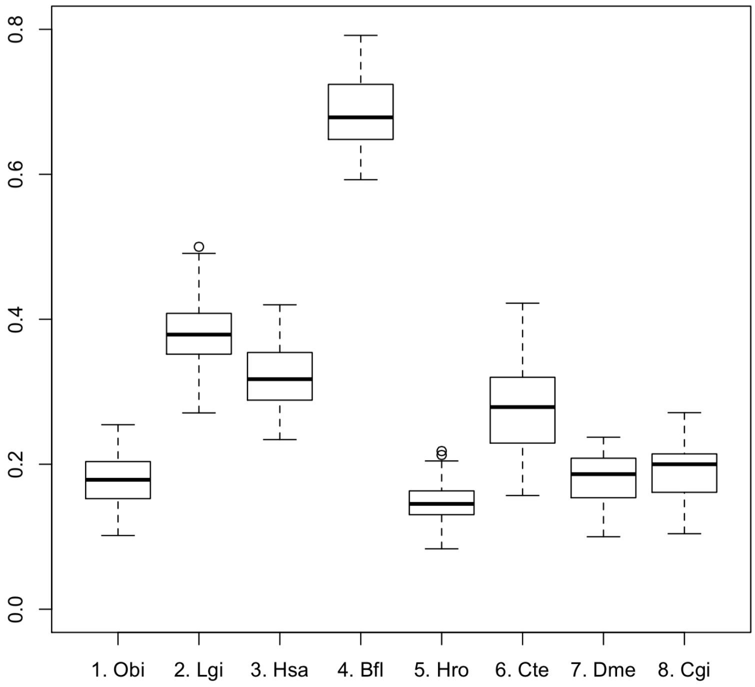


Figure S6.1. Proportion of the identified bilaterian syntenic blocks in each species across 50 simulation runs.

We used the Circos tool (Krzywinski et al., 2009) to plot the shared synteny across 6 different genomes (*Octopus*, *Lottia*, *Helobdella*, *Capitella,* human, amphioxus). Again, octopus has a lower number of outgoing connections (Figure S6.2) as compared to that of *Lottia*, *Capitella*, and amphioxus. This plot affirms that *Helobdella* retained the lowest synteny, with only 114 outgoing connections.

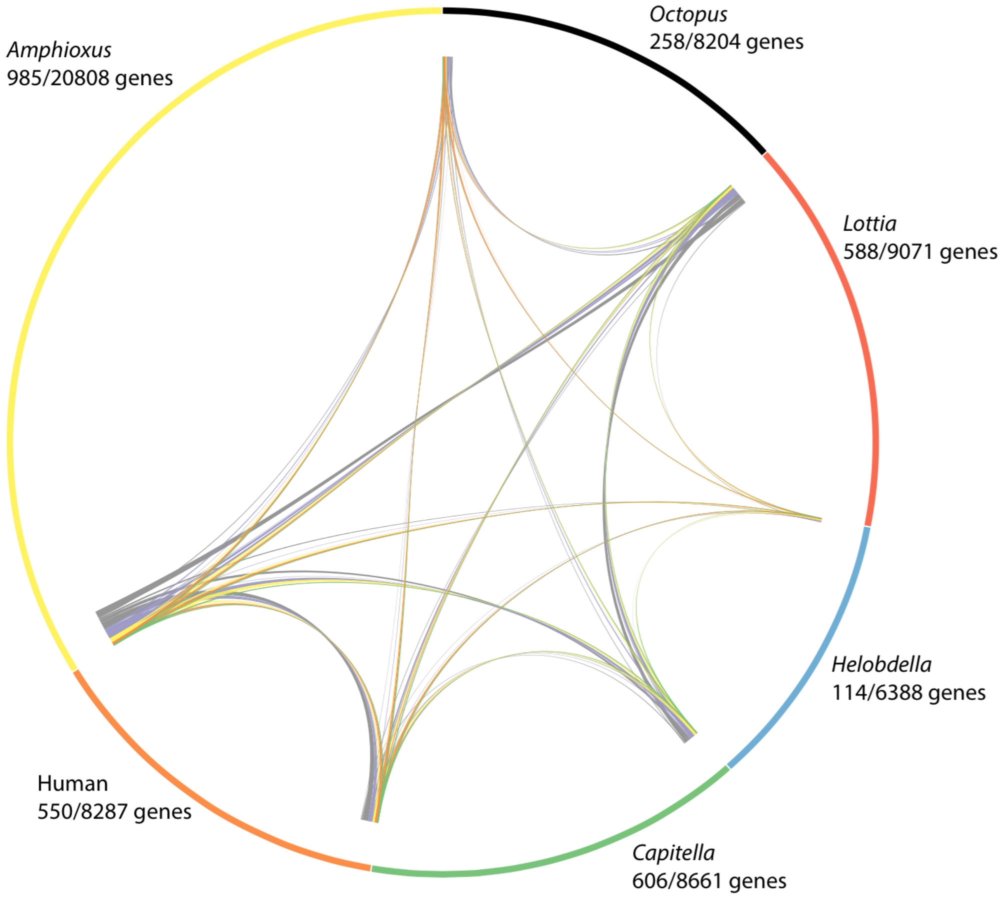


Figure S6.2. Reduced synteny in octopus compared to other species. Each species is represented by a different color around the circumference. The total length of the arc corresponds to the total number of genes in the 4,033 orthologous gene families present in each species. Genes that are in micro-syntenic linkage are connected by lines. Lines are colored according to the number of species that share the syntenic block: green - all 6, orange - 5, yellow - 4, purple - 3, and grey - 2 species.

The most prominent losses of synteny in octopus include the Hox cluster (see main text), as well as Forkhead, *WNT*, and *BMP* linkages (Source Data: ‘microsynteny.xls’). Despite the significant loss of micro-synteny generally, we still found some conserved linkages in octopus. For example, we found conserved linkage of metabolic enzymes around endothelial differentiation factor 1, as well as an ancient linkage of N-acetylgalactosaminyltransferase with two Twist genes (Figure S6.3). Interestingly, the Twist-locus seems to be associated with a NeuroD gene in both *Lottia* and *Capitella*, which are unlinked in octopus, potentially pointing at a partial loss of micro-synteny.

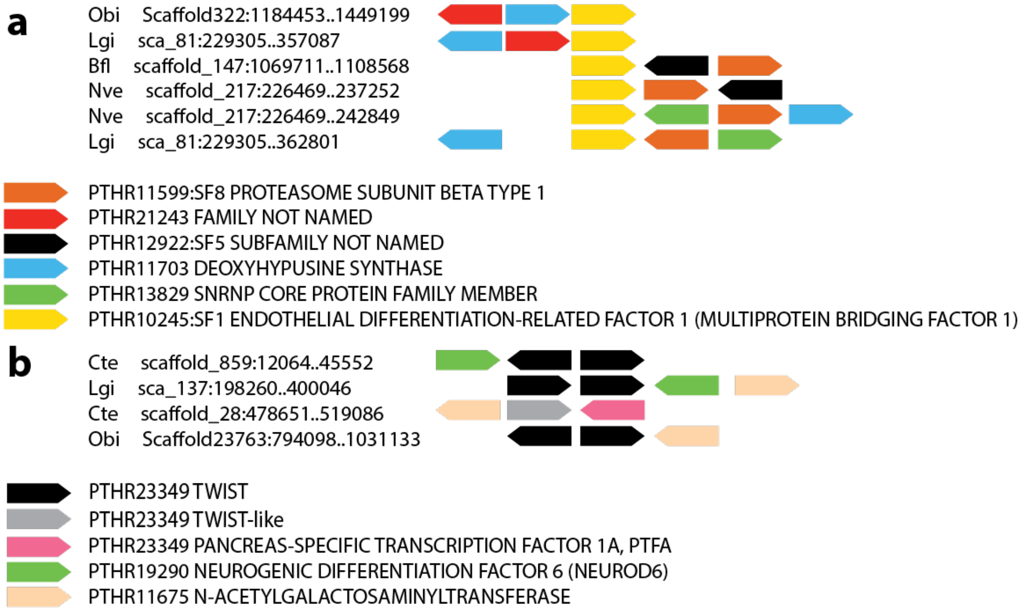

a
Obi Scaffold322:1184453..1449199
Lgi sca_81:229305..357087
Bfl scaffold_147:1069711..1108568
Nve scaffold_217:226469..237252
Nve scaffold_217:226469..242849
Lgi sca_81:229305..362801
PTHR11599:SF8 PROTEASOME SUBUNIT BETA TYPE 1
PTHR21243 FAMILY NOT NAMED
PTHR12922:SF5 SUBFAMILY NOT NAMED
PTHR11703 DEOXYHYPUSINE SYNTHASE
PTHR13829 SNRNP CORE PROTEIN FAMILY MEMBER
PTHR10245:SF1 ENDOTHELIAL DIFFERENTIATION-RELATED FACTOR 1 (MULTIPROTEIN BRIDGING FACTOR 1)
b
Cte scaffold_859:12064..45552
Lgi sca_137:198260..400046
Cte scaffold_28:478651..519086
Obi Scaffold23763:794098..1031133
PTHR23349 TWIST
PTHR23349 TWIST-like
PTHR23349 PANCREAS-SPECIFIC TRANSCRIPTION FACTOR 1A, PTFA
PTHR19290 NEUROGENIC DIFFERENTIATION FACTOR 6 (NEUROD6)
PTHR11675 N-ACETYLGALACTOSAMINYLTRANSFERASE

Figure S6.3. Examples of retained linkage in octopus. **a**, Endothelial differentiation factor linkages in metazoans. **b**, Twist cluster.

The branch length estimation for Extended Data Figure 9b was done based on the binary matrix of presence or absence of syntenic blocks across nine species (only considering gene families with all nine species represented) using MrBayes and a constrained gene family-derived tree topology. In total, the ‘alignment' contained 354 characters, *i.e.*, synteny blocks that are present in at least two species. MrBayes restriction model was used and the transition rates were defined to reflect the less-likely scenario of independently acquiring a syntenic block (prior probability = 0.01) compared to losing linkage (prior probability = 0.99) (see intron analysis, Supplementary Note 7.3). Changing those priors (tested: loss probability from 0.9 to 0.99) does not affect the outcome of the analysis. The scale bar represents the cumulative amount of synteny block loss and gain per block (0.01 corresponds to 354*0.01 or ~3 blocks lost and gained).

6.2 Seeking synteny-based evidence for whole genome duplication

The presence of "doubly conserved synteny" is a strong signature of ancient whole genome duplication (Dietrich et al., 2004; Kellis et al., 2004). In the case of octopus, this would appear as two segments of the octopus genome showing conserved synteny to the one segment of an unduplicated, distantly related mollusc, such as *Lottia*. When we examined the octopus genome, we found only one such short segment. This segment was a paralogous cluster of myosin binding proteins, and not two groups of unrelated genes as expected for doubly conserved synteny.

Another signature of whole gene duplication is the occurrence of blocks of intra-genomic (homeologous) synteny, sometimes called "paralogons" (Leveugle et al., 2003). This signal relies on the presence of linked groups of retained duplicates, and so can be weakened by gene loss after whole genome duplication. Based on our (metazoan node) orthology assignments, we found only 38 regions of intra-genomic conserved microsynteny with three or more genes in octopus, compared with 32 in *Lottia* and 18 in *Crassostrea*. To check whether our gene family clustering artificially biased these results, we repeated the calculation using PANTHER-based clustering. Using this method, octopus showed only 12 such linkages (of at least three genes), similar to *Lottia* (22 blocks) and *Crassostrea* (14 blocks), both of which are not proposed to be ancient polyploids. In contrast, there are 1,370 human paralogons defined using PANTHER containing at least three genes, reflecting the retention of linkages from the two rounds of whole genome duplication at the base of vertebrates (Leveugle et al., 2003).

Both of these signals for ancient whole genome duplication depend on the retention of gene duplicates, and can thus be degraded by genome fragmentation and rearrangement. The octopus genome is highly rearranged relative to other available molluscan genomes, so the absence of documented doubly conserved synteny cannot provide definitive evidence against whole genome duplication. However, we have shown that the octopus genome does not exhibit an unusual number of functionally unrelated gene duplicates (excluding the specific gene family expansions of the C2H2 zinc fingers and protocadherins): what gene duplications are present do not show a detectable signal for whole genome duplication (Supplementary Note 7.4). Overall, we find no strong evidence to support the hypothesis of coleoid genome duplication.

7. GENE FAMILY CONSTRUCTION AND PHYLOGENY

7.1 Multi-gene cephalopod phylogeny and dating

To construct a cephalopod phylogeny, we used available transcriptomic data for *Sepia officinalis* (Bassaglia et al., 2012), *Doryteuthis* (formerly *Loligo*) *pealeii* (Brown et al., 2014; DeGiorgis et al., 2011), *Euprymna scolopes* (Bonaldo et al., 1996), *Idiosepius paradoxus* (Shigeno et al., 2006), *Aplysia californica* (Moroz et al., 2006) as well as available sequences for human, *Capitella teleta*, *Lottia gigantea*, *Nematostella vectensis*, *Branchiostoma floridae*, *Crassostrea gigas*, *Pinctada fucata*, and *Strongylocentrotus purpuratus* (Table S7.2.1). Open reading frame prediction (Parra et al., 2009) for cephalopod species was done using a custom pipeline based on getorf program from EMBOSS (Rice et al., 2000), selecting the longest ORFs for each transcript (minimal ORF length: 300 nucleotides). Using octopus-centered BLASTP, we identified mutual-best-hits in each proteome pair. Proteins from different proteomes matching the same octopus gene were merged to provide a cluster of orthologous genes. This resulted in 116 gene families with no missing data. Individual families were aligned with MUSCLE (Edgar, 2004) and trimmed using default parameters in Gblocks (Talavera and Castresana, 2007), retaining only gapless positions in well-defined alignment blocks. The alignments were concatenated, resulting in 4,009 positions. Phylogenetic analysis was performed using MrBayes (Huelsenbeck and Ronquist, 2001) with 4 chains and 1 million generations producing a highly supported tree (Extended Data Figure 10a). A maximum likelihood tree using TreePuzzle (Schmidt et al., 2002), with one million generations and gamma-distributed sites with 8 categories, is shown in Figure S7.1.1.

Age estimation based on branch length was performed using r8s (Sanderson, 2003), by fixing the molluscan (gastropod-bivalve to cephalopod) radiation to 540 mya (Kroger et al., 2011) and using the Langley-Fitch method for divergence estimation. This calculation resulted in an estimate of 280 mya for the coleoid cephalopod radiation and 670 mya for the diversification of bilaterians. The maximum likelihood tree from TreePuzzle yielded similar results: 260 mya for cephalopod split and 660 mya for bilaterians. These findings are in line with the current estimate for the octopus-squid divergence at 275 mya based on fossil and molecular evidence (Kroger et al., 2011; Strugnell et al., 2005).

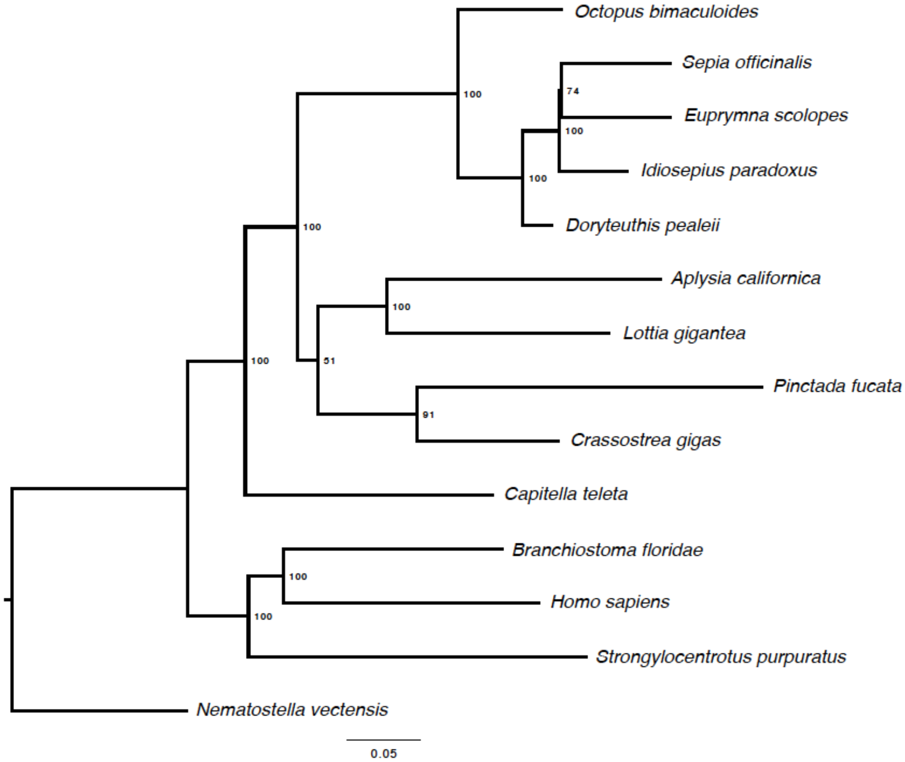


Figure S7.1.1. Maximum likelihood phylogeny of cephalopods using TreePuzzle. Bootstrap supports are shown.

To correlate this separation with dS and Jukes-Cantor distance measures, we considered pairwise mutual best hits between octopus and *Lottia*, octopus and *Sepia*, and octopus and *Idiosepius*. Protein alignments were used to make codon alignments. We disregarded alignments less than 300 bp in length. The yn00 program from the PAML package (Yang, 1997) was then run to estimate dN (amount of non-synonymous substitutions per site) and dS (amount of synonymous substitutions per site) distances between orthologs. We used the Yang and Nielsen (2000) method, instead of Nei-Gojobori, as the latter seems to reach saturation at dS ~2. Figure S7.1.2 shows the distance distributions and identifies a major peak at dS ~2 for the cephalopods and a single peak at dS ~4 for the octopus-*Lottia* split. This profile would fit with the protein phylogeny estimate, since it suggests that the bivalve-cephalopod split is twice as old as the octopus-squid (decapodiform) split. However, these estimates should be taken with caution, as dS >1 values are prone to saturation artifacts. For example, the older peak at dS ~4 (resulting in a second peak in the cephalopod divergence plots) may arise from older duplicated sequences merged together due to saturation effects. If one assumes that dS ~4 is the molluscan divergence peak and dS 2 is the cephalopod divergence peak, then the dS ~1 range covers approximately 130 million years. This number suggests dS ~ 2 / (2 * 275 million years) = 0.0036 neutral substitutions per million years (a factor of 2 in the denominator corrects for the total elapsed time for substitutions, which is twice the divergence time).

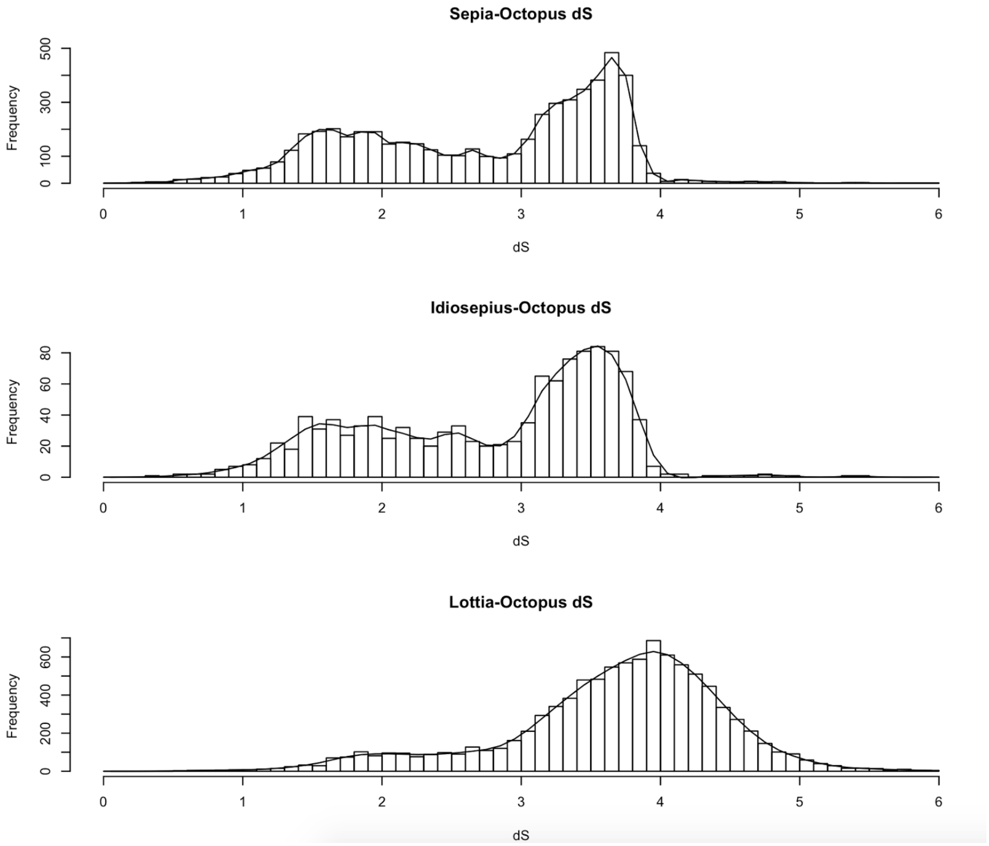


Figure S7.1.2. Divergence dating with dS between cephalopod and mollusc sequences.

7.2 Construction of orthology groups

Gene family assignments were done using a pipeline described in Simakov et al. (2013). The genomes used in the analysis are listed in Table S7.2.1. We used *Monosiga brevicollis*, *Salpingoeca rosetta*, *Pleurobrachia bachei* and *Mnemiopsis leidyi* as outgroups, in accordance with recent publications (Fairclough et al., 2013; King et al., 2008; Moroz et al., 2014; Ryan et al., 2013). We used gene families from the metazoan node of the tree, aligned them with MUSCLE, and analyzed intron/indel characters and synteny.

Species name	Code	Reference/Gene models
Mus musculus	Mmu	NCBI m36 from ensembl 41
Homo sapiens	Hsa	NCBI 36 from Ensembl 41
Gallus gallus	Gga	Ensembl vs 55 (Aug 2009) on WASHUC2 assembly
Xenopus tropicalis	Xtr	Annotation v4.2 on assembly v4.1
Latimeria chalumnae	Lch	Amemiya et al. (2013)
Danio rerio	Dre	Zv 6 from ensembl 41
Ciona intestinalis	Cin	JGI finalized models Dec 2005 REPLACES proteome 16
Branchiostoma floridae	Bfl	Brafl1 JGI annotation, Released April 11 2006
Strongylocentrotus purpuratus	Spu	Spur_v2.1 from NCBI build 2.1
Capitella teleta	Cte	Simakov et al. (2013)
Helobdella robusta	Hro	Simakov et al. (2013)
Lottia gigantea	Lgi	Simakov et al. (2013)
Crassostrea gigas	Cgi	Zhang et al. (2012)
Pinctada fucata	Pfu	Takeuchi et al. (2012)
Adineta vaga	Ava	Flot et al. (2013)
Schmidtea mediterranea	Sme	Mk4 models from http://smedgd.neuro.utah.edu/
Schistosoma mansoni	Sma	Version 080508 from ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/
Drosophila melanogaster	Dme	BDGP 4 from ensembl 41
Tribolium castaneum	Tca	NCBI gene models build 1 version 1 based on the assembly Tcas_2.0, September 2005
Daphnia pulex	Dpu	FilteredModels8 from JGI 09.04.2007
Ixodes scapularis	Isc	Version 1.1 from ftp://ftp.vectorbase.org/public_data/organism_data/iscapularis/
Strigamia maritima	Stm	http://metazoa.ensembl.org/Strigamia_maritima/Info/Index
Caenorhabditis elegans	Cel	Wormbase release WS164
Nematostella vectensis	Nve	JGI Annotation of Nematostella genome version 1
Acropora digitifera	Adi	Shinzato et al. (2011)
Hydra magnipapillata	Hma	Chapman et al. (2010)
Trichoplax adhaerens	Tad	FilteredModels2 from JGI 08.14.2007
Mnemiopsis leidyi	Mle	Ryan et al. (2013)
Amphimedon queenslandica	Aqu	Aqu1 models 8.04.08
Pleurobrachia bachei	Pba	Moroz et al. (2014)
Monosiga brevicolis	Mbr	JGI Annotation of Monosiga genome version 1
Salpingoeca rosetta	Sro	Origins of Multicellularity Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)

Table S7.2.1. Genomes and gene model versions used.

We defined ancestral bilaterian gene families as those having at least 2 species represented in both ingroups (protostomes and deuterostomes), or in one ingroup and the non-bilaterians. This definition resulted in 8,822 bilaterian gene families. The octopus genome includes representatives in 6,348 of these families (72%), similar to other lophotrochozoans such as *Lottia* (7,077 or 80%) and *Capitella* (6,996 or 79%), and deuterostomes such as human (6,579 or 75%) and amphioxus (7,236 or 82%). For comparison, *Drosophila* and *Schistosoma* have only 4,573 (52%) and 3,082 (35%) representatives, respectively.

7.3 Intron and indel analysis

In order to analyze the dynamics of introns and coding indels, we considered 2,816 metazoan gene families with at least 20 species and screened for conserved sites, as described in Simakov et al. (2013). Briefly, we required a conserved splice site to have at least 3 out of 8 flanking amino acid residues with the same biochemical properties, and no additional splice site within 4 amino acid residues (to exclude possibly misaligned regions). Similarly, based on amino acid alignments, we defined indels as regions that have gaps in at least one sequence in the alignment with fully conserved flanking amino acids (and otherwise 3 out of 8 conserved amino acids). This analysis yielded 5,779 high-confidence intron and 3,323 indel sites in the 2,816 families considered. Using the same criteria to define ancestral bilaterian gene families (*i.e.*, by considering intron presence in outgroups, see above), we identified 2,077 ancestral bilaterian introns within these genes. Introns are likely to represent ancestral characters and their absence in some extant bilaterian species most likely reflects losses. Indels appear to be more dynamic. Out of the 2,077 ancestral bilaterian introns identified here, *O. bimaculoides* retains 1,839 (or 85%) introns, comparable to *L. gigantea* (1,860, 90%), *C. teleta* (1,676, 81%), human (1,794, 86%), and amphioxus (1,764, 85%). For comparison, *D. melanogaster* and *C. elegans*, which have undergone extensive intron loss, retain only 316 (15%) and 442 (21%) introns, respectively (Putnam et al., 2007).

We did not find a significant increase in the number of novel introns in octopus when compared to other species. For example, in the 2,816 genes examined, octopus has around the same number of novel introns (8) as do *Lottia* and *Capitella*.

For a more accurate assessment of the total turnover rate, we generated trees using the presence of indels and introns as binary characters. To obtain the strongest phylogenetic signal, we focused only on octopus, *Lottia*, *Capitella*, *Helobdella*, *Drosophila*, human, amphioxus, *Nematostella*, and *Crassostrea*, requiring all 9 species to be present in the alignment. This condition yielded 9,108 intron characters and 8,995 indels characters. We ran MrBayes (Huelsenbeck and Ronquist, 2001) with a constrained tree topology (1 million generations). We used the default transition frequency for indels (both gain and loss being equally likely) and set prior probabilities of 0.01 for 0→1 (gain of intron) and 0.99 for 1→0 (loss of intron) transitions, to reflect the assumption that the independent gain of introns is much less likely than the loss of existing introns. The final trees are displayed in Extended Data Figure 3. Octopus shows a similar rate of loss and gain in both indels and introns compared to other "slow-evolving" spiralians such as *Lottia* and *Capitella*.

7.4 Genome-wide gene family expansions and dating

The proposal of one or two rounds of whole genome duplication in the coleoid lineage derives from the haploid chromosome number in octopus species (N = 28 or 30) (Adachi et al. (2014), and references therein), which is approximately double the putative chromosome number of the molluscan ancestor (N=15-20) (Hallinan and Lindberg, 2011; Simakov et al., 2013). Although some of the gene families expanded in octopus seem to be expanded in vertebrates as well (C2H2, Interleukin 17), we do not find any evidence for a genome-wide convergence signal in the gene family sizes in octopus. We conducted principal component analysis (PCA) using the prcomp function in R on the Pfam domain counts (Extended Data Figure 3a). The PC1 component separates vertebrates from invertebrates and explains 9% of variance, while PC2 (7% of the variance) correlates with the ecdysozoan group. Other genomes, including octopus, tend to form a central group with no particular affiliation.

Our analysis of specific gene families involved in development and neuronal function in octopus revealed a general trend of single copy orthologs and absence of any duplicates (*e.g.*, Wnt, Hh, Hox, Supplementary Notes 8.2, 9). This finding is confirmed by general Pfam composition analysis that shows that domains with a single representative in the octopus dominate the genome (similar to *Lottia*, *Crassostrea*, *Capitella*), while human and *Xenopus tropicalis* show the opposite trend, with the majority of Pfams having at least 2 gene copies (Table S7.4.1). These data are consistent with the absence of whole genome duplication in octopus relative to other non-cephalopod molluscs.

	Single ortholog	**2-5 orthologs**	**Ratio**
Octopus	2394	1345	1.779925651
Lottia	2396	1306	1.834609495
Crassostrea	2255	1444	1.561634349
Capitella	2457	1568	1.566964286
Human	521	1914	0.272204807
Xenopus	1662	2125	0.782117647

Table S7.4.1. Counts of Pfam families with single and low-copy multiple (2-5) genes. The signal for duplication is present in the vertebrates and absent in octopus and other lophotrochozoans.

To conduct a genome-wide screen for gene family expansions, we used Pfam domain assignments (Finn et al., 2014) to define gene categories. Our BLASTP-based gene family clustering method is too specific for this purpose, often splitting several gene subfamilies (*e.g.*, *WNT1*, *WNT5*, etc.) into different clusters, while global enrichment analysis necessitates that different subfamilies be pooled together. In addition, since Pfam uses Hidden Markov Models (HMM), it is more accurate in identifying highly divergent copies of genes than our BLASTP-based clustering. We constructed a table with 8,624 different Pfam domains and the counts for genes with those domains in 23 species. If a domain occurred multiple times in a protein sequence, it was counted only once. To exclude transposon derived domains, mispredictions, or unknown domains, we removed Pfams that were categorized as "unknown," "not named," "uncharacterized," "transposase," "helitron," "helicase," "DUF," or "DDE_Tnp." We then iteratively conducted a Fisher's exact test in R (Team, 2014), comparing the number of counts in Pfam families found in the following groups of species, to the background, defined as the average of the counts in the remaining species:

- Group 1 - Spiralians. *Helobdella*, *Capitella*, *Pinctada*, *Crassostrea*, *Lottia*, and octopus vs. the average of the remaining species
- Group 2 - Molluscs. *Pinctada*, *Crassostrea*, *Lottia*, and octopus vs. the average of the remaining species
- Group 3 - Octopus vs. the average of the remaining species

Each species was also tested individually against the background. Multiple testing correction was done with the Bonferroni method in R. A Pfam domain was considered expanded for a phylogenetic group only if at least half of the species in the group showed a significant corrected p-value (0.01). These results are shown in Figure 1b. To represent and account for different gene numbers across species in the heatmap, the individual counts were normalized by the total gene count in each species, then normalized within each gene family (across species) by the scale function in R.

To confirm independence of annotation biases, we conducted the same analysis using PANTHER (Thomas et al., 2003). The resulting heatmap of identified expansions is shown in Figure S7.4.1. This analysis identified enriched categories similar to those identified by our BLASTP-based Pfam analysis: for example, Pfams including G-protein coupled receptors, Interleukin-17s, Zinc-fingers, sialic acid metabolic proteins, as well as extracellular matrix glycoproteins (which contain the protocadherin expansion) showed enrichment in the octopus.

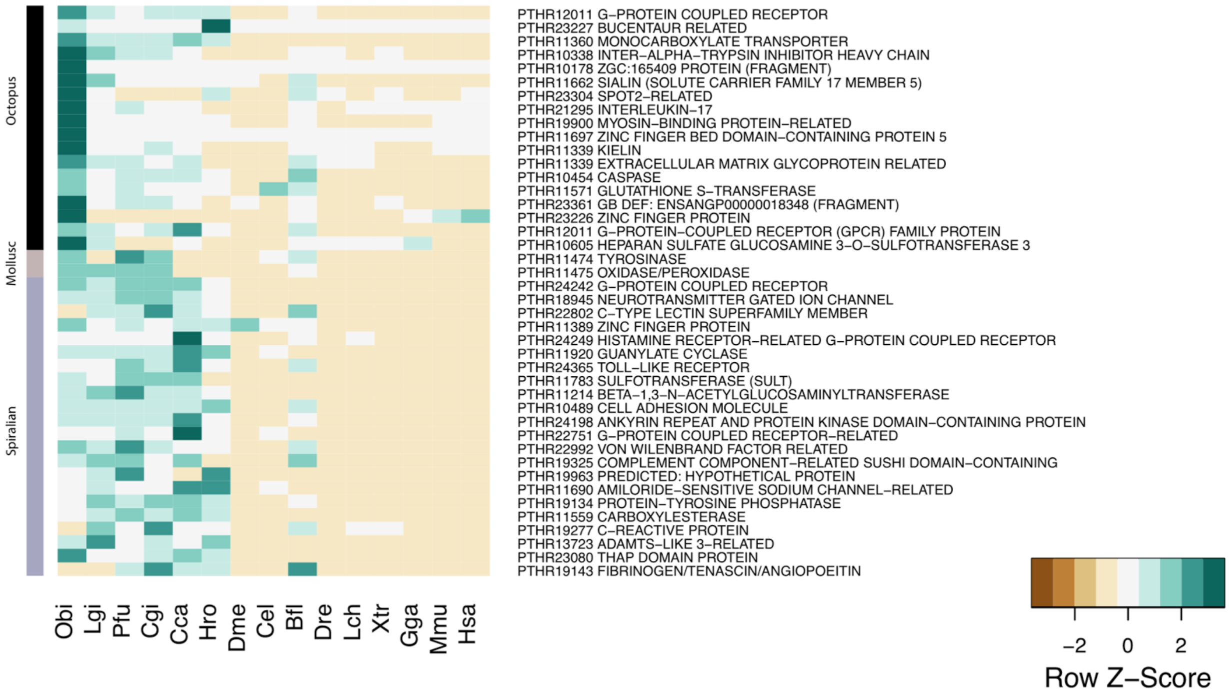


Figure S7.4.1. Heatmap of expansions in the three phylogenetic groups using PANTHER annotations.

We applied dS measures to assign approximate timings to these Pfam expansions. Based on Pfam classification, we conducted pairwise protein alignments between all full-length members (with a start and a stop codon) of a given Pfam group, and estimated the dS for each of them, after converting to CDS-based alignment (using the protein alignment as an anchor). We used yn00 from PAML (Yang, 1997) to estimate divergence. To avoid biases towards ancient separation of paralogs due to counting all N(N-1)/2 combinations, we constructed a neighbor-joining tree based on pairwise dS distances, and counted only the dS values for the different nodes. The total distribution of dS values across all Pfams is shown in Figure S7.4.2. The distribution shows that there is no significant recent expansion, with only 218 out of 3,940 Pfams (~5%) having a dS <1. The peak at dS ~2 presumably corresponds in time to the cephalopod radiation. However, manual curation revealed that many of the gene families within that peak are old duplicates dating back to the bilaterian radiation, and likely appear at dS ~2 due to saturation effects. We also observe a peak at dS 2 in other species that have no indication of whole genome duplication (such as *Lottia*, *Capitella*, and *Crassostrea*, Figure S7.4.2). Additionally, the peak at dS 2 becomes less prominent when considering unlinked genes from low-copy Pfams (not more than 5 members), which removes genes found to be recently tandemly expanded in octopus, such as the C2H2-ZNFs and PCDHs (Supplementary Note 8.2). The remaining peak at dS 4 is also clearly in the saturation range. Gene families expanded past dS ~3 constitute old expansions, the exact date of which is beyond the scope of this method.

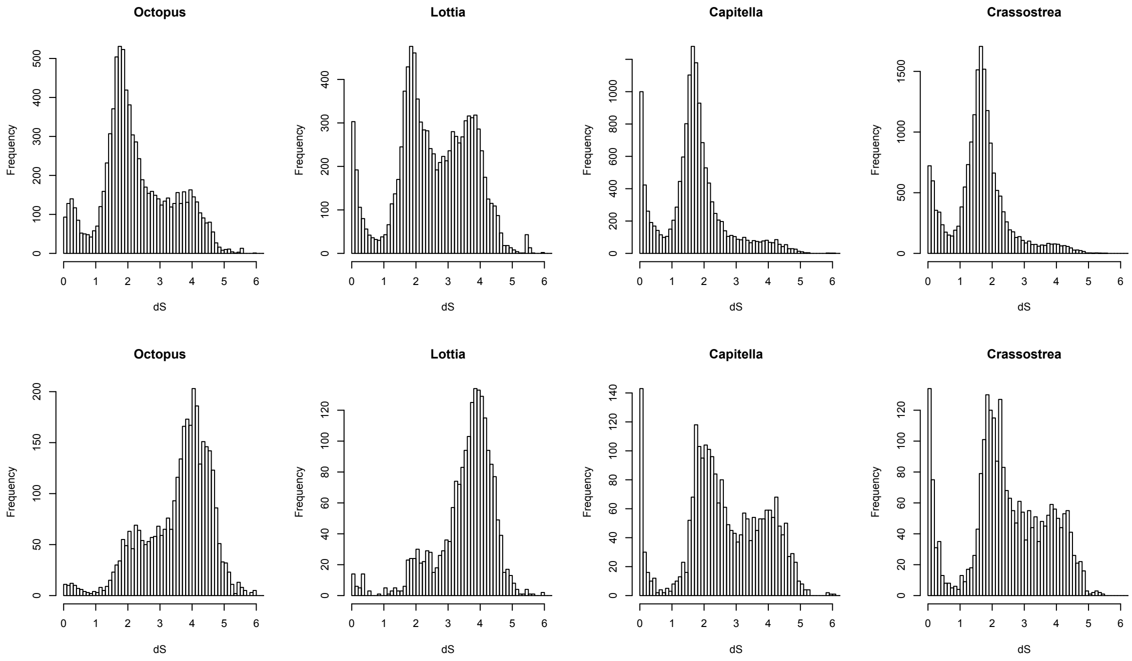


Figure S7.4.2. Histogram of dS distance distribution among paralogs belonging to each Pfam category. Upper row: all Pfam categories and distances, lower row: Pfam categories with less than 5 genes in a given species and with genes on different scaffolds reveal a less pronounced peak at dS 2.

Using our coarse dating, we looked to see if there was tissue specific expansion in different gene families. A gene was defined as tissue specific if at least 75% of its expression was localized to a single tissue. For tissue specific genes, we plotted the timing the gene's duplication against the tissue it is specific to (Figure S7.4.3). Testes and retina contain the highest levels of duplication prior to or during the cephalopod radiation. More recent duplicates have higher representation in the axial nerve cord, skin, and suckers, with the proportion in testes diminishing.

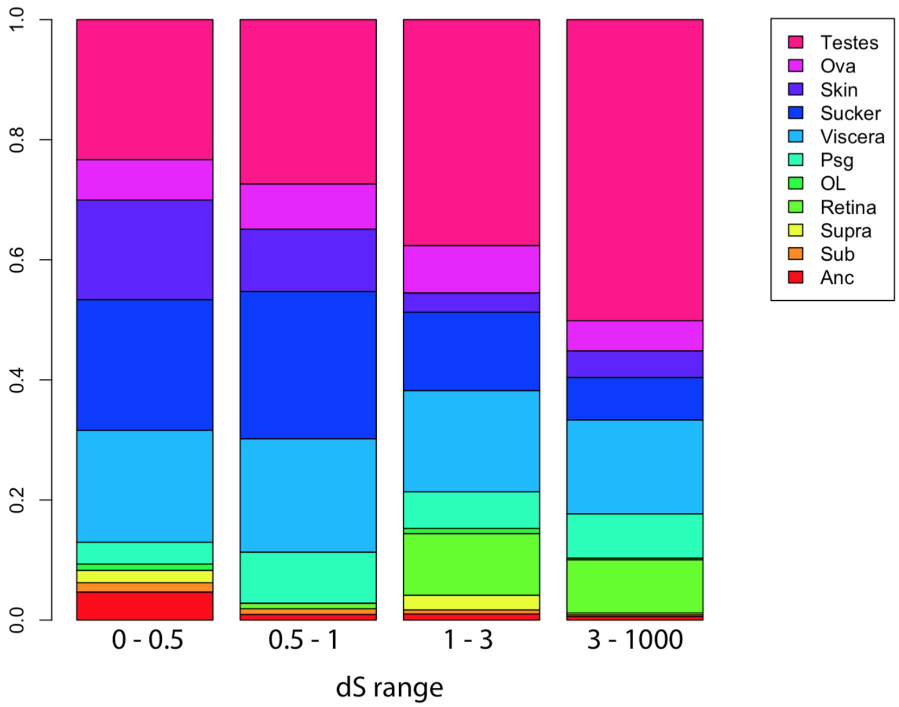


Figure S7.4.3. Relative distribution of tissue-specific genes based on duplication timing.

8. DESCRIPTION AND ANALYSES OF SPECIFIC GENE FAMILIES

Gene families of particular interest, including developmental regulatory genes, neural-related genes, and gene families that appear to be expanded in *O. bimaculoides*, were manually curated and analyzed.

8.1 Annotation methods

We searched the octopus genome and transcriptome assemblies for candidate genes using BLASTP and TBLASTN searches with sequences from human, mouse, and *D. melanogaster*. We also searched protein sets deposited for *L. gigantea, C. gigas, A. californica, C. teleta, T. castaneum, D. melanogaster, C. elegans, B. floridae, C. intestinalis, D. rerio, M. musculus,* and *H. sapiens* for members of these gene families. Candidate genes were verified using BLAST and Pfam. Genes identified in the octopus genome were confirmed and extended using the transcriptomes, and multiple gene models that matched the same transcript were combined. The identified sequences from octopus and other bilaterians were aligned using either MUSCLE (Edgar, 2004) or CLUSTALO (Sievers et al., 2011). Phylogenetic trees were constructed with FastTree (Price et al., 2010) using full-length sequences, and members of each family were counted.

8.2 Selected developmental control signaling molecules

Developmental control genes are often highly conserved across phylogenetically disparate groups of animals, and may serve as indicators of genomic events, such as duplications, or dramatic losses. Gene families known to play a role in metazoan development, including ligands (FGF, WNT, TGFβ, NOTCH ligands, HEDGEHOG, axon guidance molecules) and transcription factors (C2H2-ZNFs, homeodomain, high mobility group, basic helix-loop-helix, nuclear hormone receptors, Fox, Tbox) were manually catalogued using the methods described in SN8.1. The results are summarized in Table 1.

8.2.1 Hedgehog

Hedgehog (Hh) signaling is an ancient part of the bilaterian developmental toolkit: *Hh*s have been found in the sea anemone *N. vectensis*, as well as most bilaterian animals. *Hhs* participate in a range of developmental processes, including segment polarity regulation and patterning of the gut, wing disc and eyes in *D. melanogaster* (Lee et al., 1992; Tabata and Kornberg, 1994), patterning of the gut, neural tube and limbs in vertebrates (Hooper and Scott, 2005; McGlinn and Tabin, 2006), and gut formation in *H. robusta* (Kang, 2003). While mammals have three *Hh*s (*Shh, Ihh,* and *Dhh*), a single *Hh* is present in the invertebrate bilaterian genomes examined, with the exception of *C. elegans*, which lacks a clear *Hh* gene (Burglin and Kuwabara, 2006) (Table 1). We found a single *Hh* in the *O. bimaculoides* genome (Figure S8.2.1), which codes for both the N-terminal Hh signaling domain and the C-terminal Hint domain characteristic of authentic *Hh* proteins.

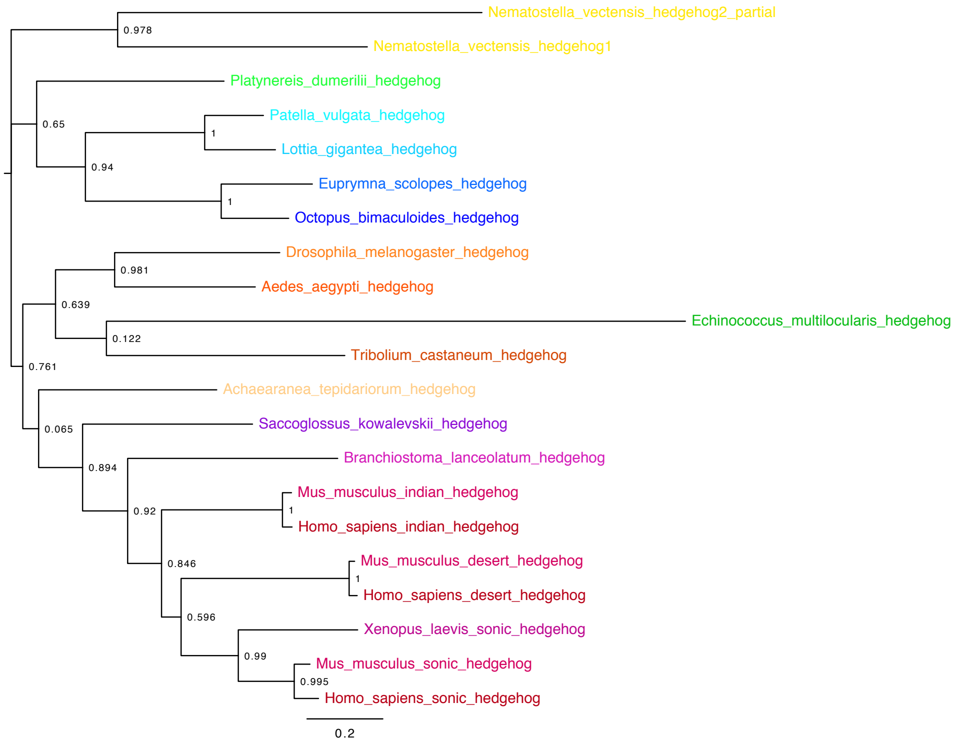

Nematostella_vectensis_hedgehog2_partial
0.978
Nematostella_vectensis_hedgehog1
Platynereis_dumerilii_hedgehog
0.65
Patella_vulgata_hedgehog
1
Lottia_gigantea_hedgehog
0.94
Euprymna_scolopes_hedgehog
1
Octopus_bimaculoides_hedgehog
Drosophila_melanogaster_hedgehog
0.981
Aedes_aegypti_hedgehog
0.639
Echinococcus_multilocularis_hedgehog
0.122
Tribolium_castaneum_hedgehog
0.761
Achaearanea_tepidariorum_hedgehog
0.065
Saccoglossus_kowalevskii_hedgehog
Branchiostoma_lanceolatum_hedgehog
0.894
Mus_musculus_indian_hedgehog
0.92
1
Homo_sapiens_indian_hedgehog
Mus_musculus_desert_hedgehog
0.846
1
Homo_sapiens_desert_hedgehog
0.596
Xenopus_laevis_sonic_hedgehog
0.99
Mus_musculus_sonic_hedgehog
0.995
Homo_sapiens_sonic_hedgehog
0.2

Figure S8.2.1. Phylogenetic tree of eumetazoan hedgehogs.

8.2.2 Transforming Growth Factor-β

The Transforming Growth Factor-β (TGF-β) family of ligands include the TGF-βs *sensu stricto*, Activins, Leftys, Growth Differentiation Factors (GDFs), and Bone Morphogenetic Proteins (BMPs). The TGF-βs play important roles during animal development. Mammalian genomes encode 33 TGF-β family members (Wharton and Derynck, 2009) while only handfuls are reported in ecdysozoans (Gesualdi and Haerry, 2007; Gumienny and Savage-Dunn, 2013). We found 12 *TGF-β* genes in *O. bimaculoides*, including homologs of *BMP2/4*, *BMP3*, *BMP5-8*, *GDF2*, *GDF8/11*, *Nodal*, and *ADMP* (Figure S8.2.2). While other lophotrochozoans appear to have *BMP10/GDF2* homologs, we did not retrieve an octopus homolog in this analysis. Overall, lophotrochozoan genomes have roughly twice as many TGF-βs than do model ecdysozoans, and have homologs of genes absent in ecdysozoan lineages, including *Nodal, BMP3,* and *ADMP*.

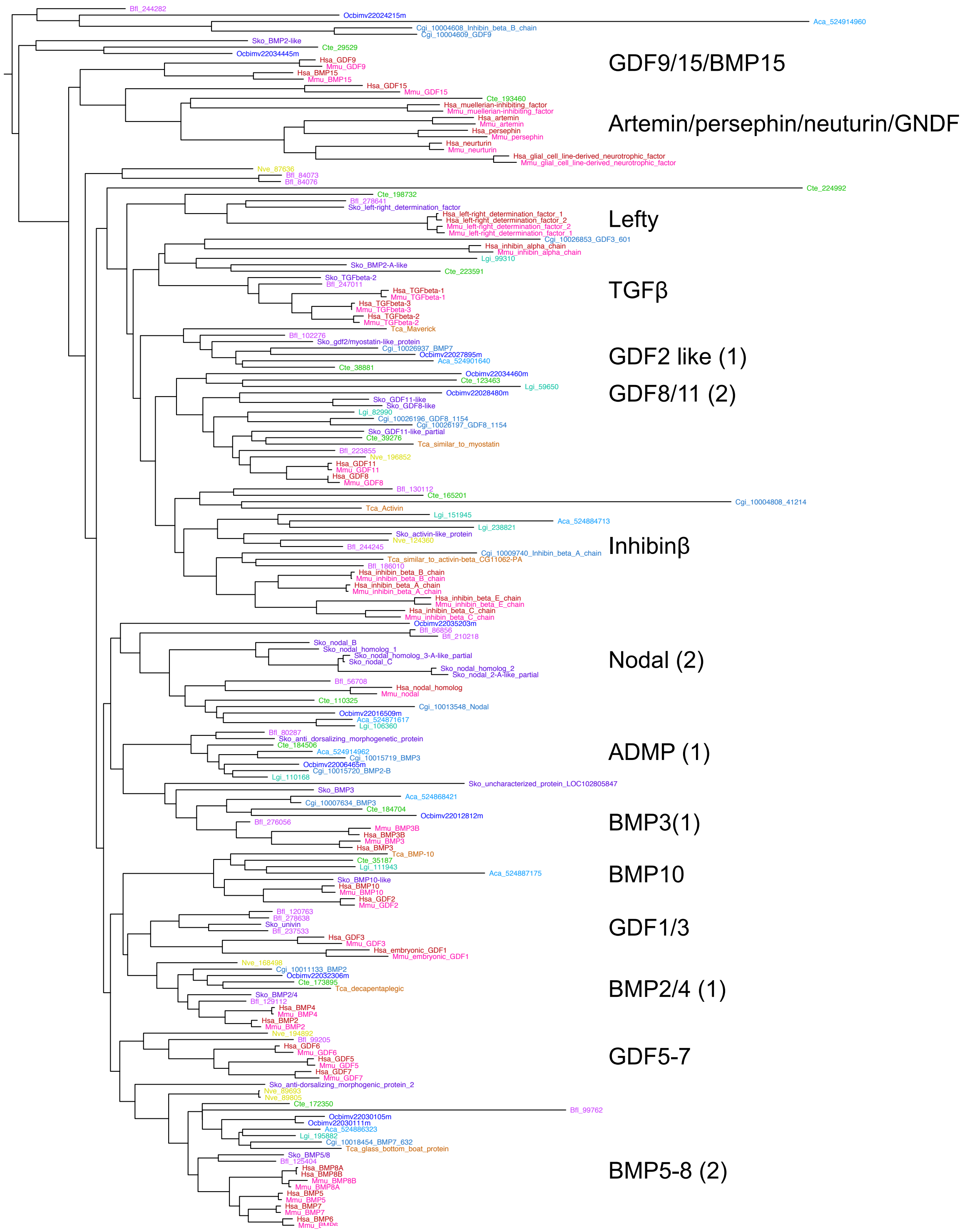


Figure S8.2.2. Phylogenetic tree showing distribution of eumetazoan TGFβ superfamily proteins. Branch labels are colored according to species: Hsa (red), Mmu (pink), Nve (yellow), Tca (orange), Cgi (cornflower blue), Aca (sky blue), Lgi (teal), Obi (dark blue), Cte (green). Gene families are listed to the right; parentheses indicate number of homologs found in the *O. bimaculoides* genome.

8.2.3 WNT

WNTs are secreted intercellular signaling glycoproteins that play important roles in developmental processes in metazoans. Thirteen *Wnt* subfamilies have been identified in eumetazoans. Apart from chordates, each subfamily, when present, is represented by one gene. Twelve of the 13 subfamilies are present in the cnidarian *N. vectensis* (Kusserow et al., 2005), and 11 and 12 of the subfamilies are found in the lophotrochozoans *L. gigantea* and *C. teleta*, respectively (Cho et al., 2010). This bilaterian *Wnt* gene set is expanded in vertebrates (humans have 19), and reduced in ecdysozoans (*D. melanogaster* and *C. elegans* have 7 and 5 *Wnts*, respectively) (Eisenmann, 2005; Rubin et al., 2000). *O. bimaculoides*, like *C. teleta*, has 12 *WNTs* (Figure S8.2.3). All protostomes examined thus far lack *WNT3* family members (Bolognesi et al., 2008; Cho et al., 2010). While *WNT8* is absent in *L. gigantea*, it is present in *O. bimaculoides* and in *C. gigas*, indicating that it is part of the ancestral molluscan *Wnt* complement.

Unlike the Hox cluster and other transcription factor families, three of the *Wnts* have maintained genomic linkage in the *O. bimaculoides* genome. *WNT1*, *WNT6*, *WNT9* and *WNT10* are found on the same chromosome in a number of animals, including *B. floridae, D. melanogaster, C. teleta*, *L. gigantea,* and *N. vectensis* (only *WNT10* and *WNT6*) (Cho et al., 2010). In *O. bimaculoides*, *WNT1*, *WNT6*, and *WNT10* are all located on the same scaffold. While *WNT5* and *WNT7* are also linked in *N. vectensis, B. floridae* and *L. gigantea*, we do not find evidence that this linkage is retained in octopus.

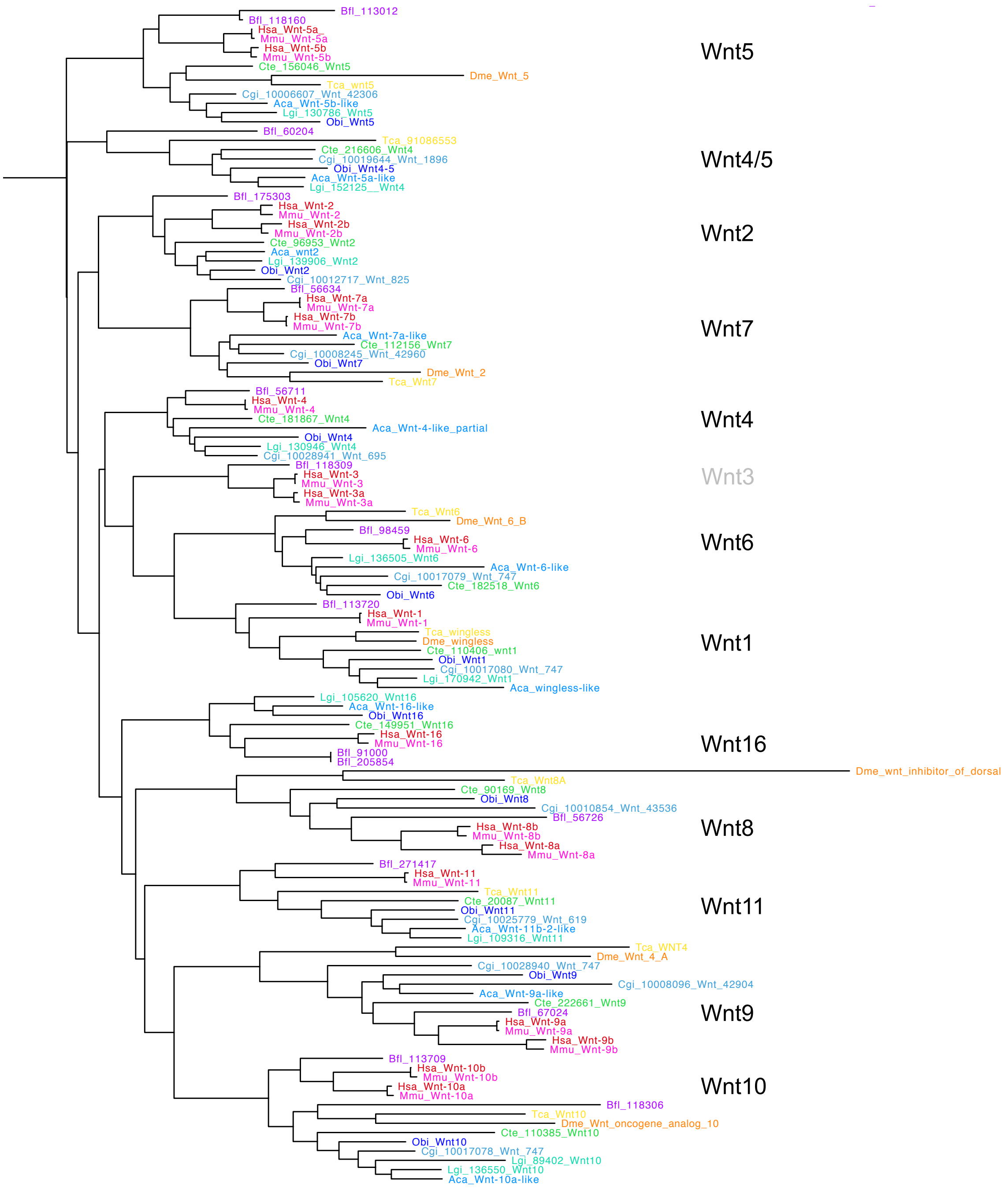


Figure S8.2.3. Phylogenetic tree showing distribution of eumetazoan Wnts. Branch labels are colored according to species: Hsa (red), Mmu (pink), Tca (yellow), Dme (orange), Cgi (cornflower blue), Aca (sky blue), Lgi (teal), Obi (dark blue), Cte (green). Gene families are listed to the right.

8.2.4 Axon guidance molecules

Axon guidance cues and receptors are necessary for the normal development of complex neural circuitry present in the adult nervous system. Combinations of chemo-attractive and chemo-repulsive cues actively direct neuronal processes to their correct targets. These cues, which include the Netrins, Slits, Semaphorins, Ephrins, and their respective receptors, signal over both long and short distances in the developing nervous system (Kolodkin and Tessier-Lavigne, 2011). We identified 25 axon guidance cues and their receptors in the octopus genome (Table S8.2.4). Consistent with their critical roles in neuronal development in genetic model systems, these genes show strongest expression in the St15 transcriptome and elevated expression in adult nervous tissues (Figure S8.2.4).

Gene	***O. bimaculoides***	***L. gigantea***	***H. sapiens***
Netrin	1	1	3
UNC-5	3	1	4
DCC family	1	-	2
Slit	4	1	3
ROBO	3	2	3
SrGAP	1	1	3
Ephrin	1	1	8
Ephrin receptor	2	1	14
Semaphorin	3	3	20
Plexin	6	3	9
Total	25	14	71

Table S8.2.4. Summary of axon guidance molecules.

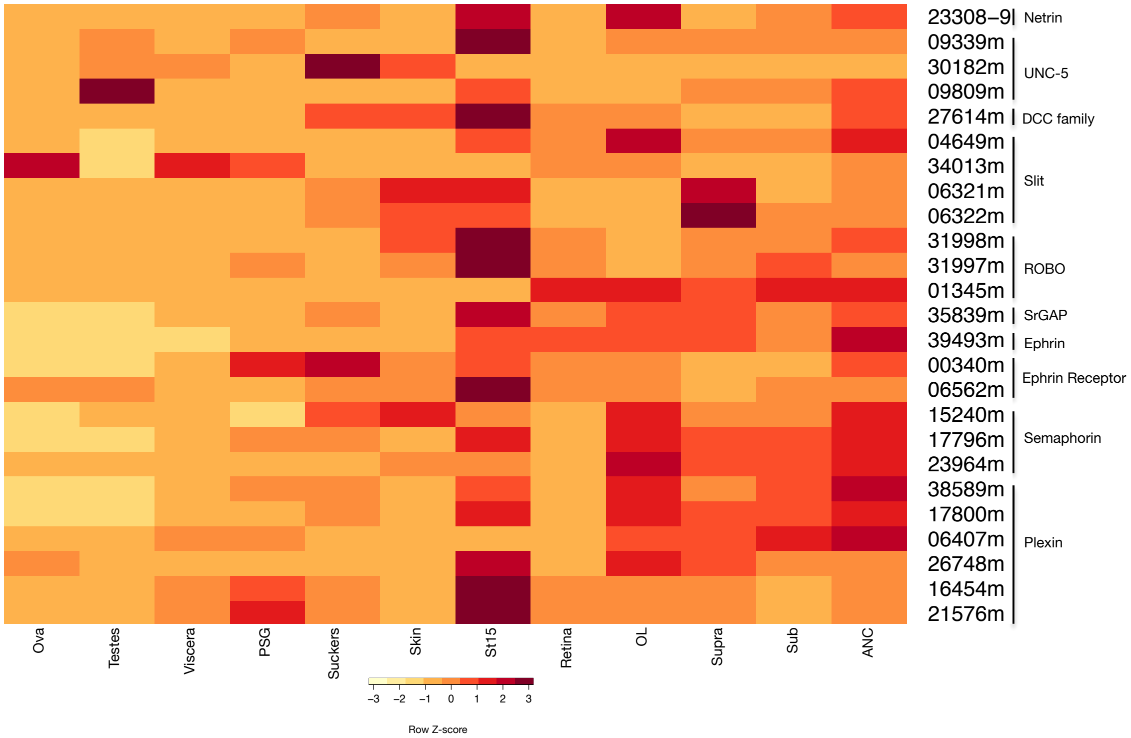


Figure S8.2.4. Expression profiles of axon guidance molecules.

8.3 Protocadherins

In vertebrates, protocadherins are the largest group within the cadherin superfamily; humans and mice have more than 60 isoforms (Frank and Kemler, 2002). Protocadherins are cell adhesion molecules with 6 or 7 extracellular cadherin repeats (EC), a single transmembrane region (TM), and a cytoplasmic domain (CD) (Hulpiau and van Roy, 2009; Morishita and Yagi, 2007) (Figure S8.3.1). Protocadherins interact promiscuously to form tetramers in the same cell (in *cis*) while the EC domain tetramers have been shown to mediate strictly homophilic intercellular binding (in *trans*) (Schreiner and Weiner, 2010). Once assembled, the combined EC domains of the tetramer potentially allow for hundreds of thousands of specific interactions. As protocadherins generate combinatorial complexity at the cell surface and are predominantly expressed in the nervous system, this gene family may provide a molecular basis for the circuit diversity and specificity of the vertebrate nervous system, analogous to the role proposed for DSCAM in flies (Chen and Maniatis, 2013). Indeed, studies in mice have suggested that protocadherins may be important in dendritic patterning in cortical neurons and neurite self-avoidance in Purkinje and starburst amacrine cells (Lefebvre et al., 2012).

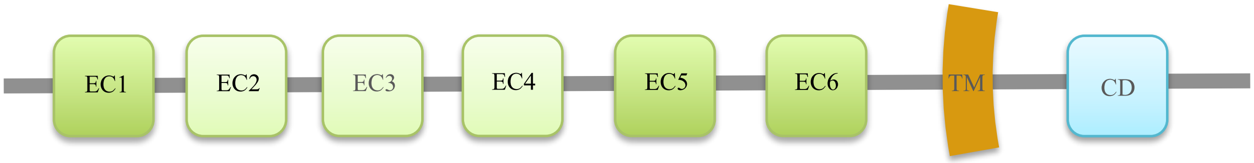


Figure S8.3.1. Structure of a typical protocadherin. Protocadherin proteins have 6-7 extracellular cadherin domains (EC) and a transmembrane domain (TM). Some vertebrate protocadherins also have a highly conserved cytoplasmic domain (CD).

Of the roughly 60 protocadherins present in mammalian genomes, more than 50 are found on a single chromosome in three clusters: α, β, and γ. α and γ protocadherins are generated by alternative splicing. Each gene is made up of one variable exon (containing the extracellular ECs, TM, and part of the intracellular domain) and three constant exons that are shared by all genes in the cluster (Chen and Maniatis, 2013). This intron-exon structure, with the ECs and TM being encoded by one or two large exons, is found in many of the non-clustered protocadherins as well (Vanhalst et al., 2005).

Until recently, protocadherins were thought to be a vertebrate innovation. While they are absent from the genomes of the ecdysozoan model organisms *D. melanogaster* and *C. elegans*, solitary protocadherins have previously been identified in *A. californica* and *N. vectensis*, indicating that this family predates the Bilateria (Hulpiau and van Roy, 2011). While this gene was apparently lost in the lineage leading to the Ecdysozoa, here we find a substantial number in lophotrochozoan genomes, including *O. bimaculoides*, *L. gigantea*, *C. gigas,* and *C. teleta* (Figure S8.3.2).

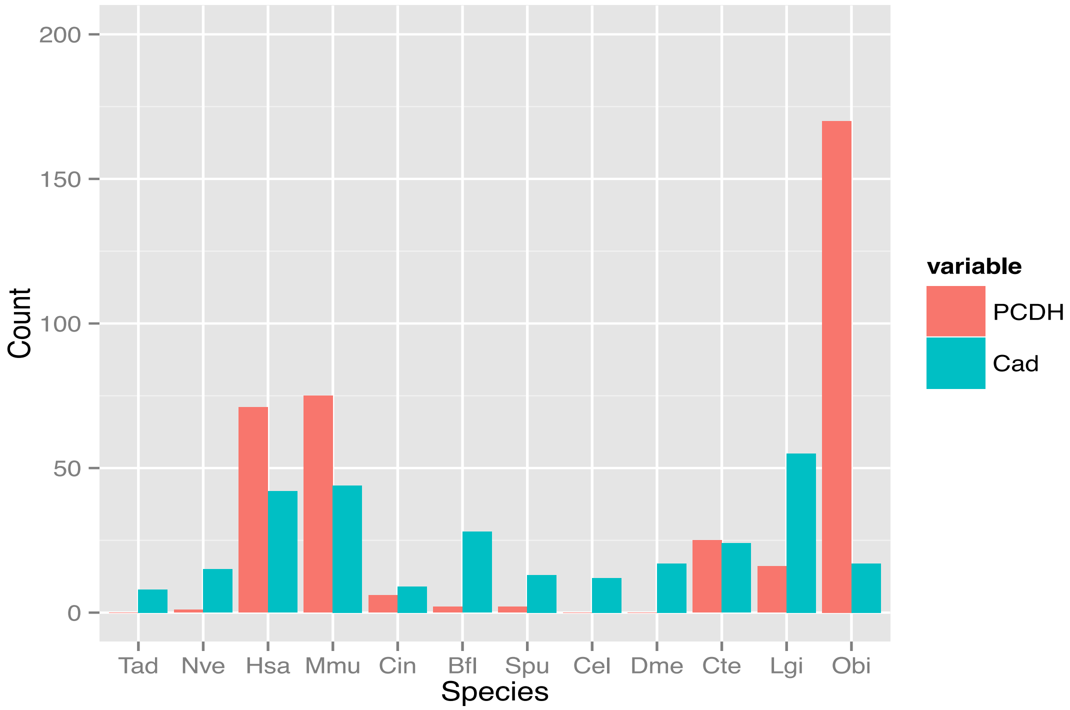


Figure S8.3.2. Distribution of cadherins and protocadherins across metazoans. Numbers protocadherins and other cadherins identified in the genomes of *H. sapiens, M. musculus, C. intestinalis, B. floridae, S. purpuratus, C. elegans, D. melanogaster, C. teleta, L. gigantea, O. bimaculoides, N. vectensis,* and *T. adhaerens.*

These lophotrochozoan protocadherins group with vertebrate protocadherins on a cadherin superfamily tree with separate, lineage-specific expansions in vertebrates, annelids, molluscs, squid, and octopuses (Figure S8.3.3). In annelids, protocadherins from different species interdigitate on the tree, suggesting the protocadherins expanded prior to the divergence of the polychaetes. Snail and oyster protocadherins also interdigitate, which indicates that the gene family expanded before the divergence of the bivalves and the gastropods, but independently from cephalopods and annelids. The topology of this tree also supports the affinity of the bivalves and the gastropods, as proposed by Smith et al. (2011) and Kocott et al. (2011).

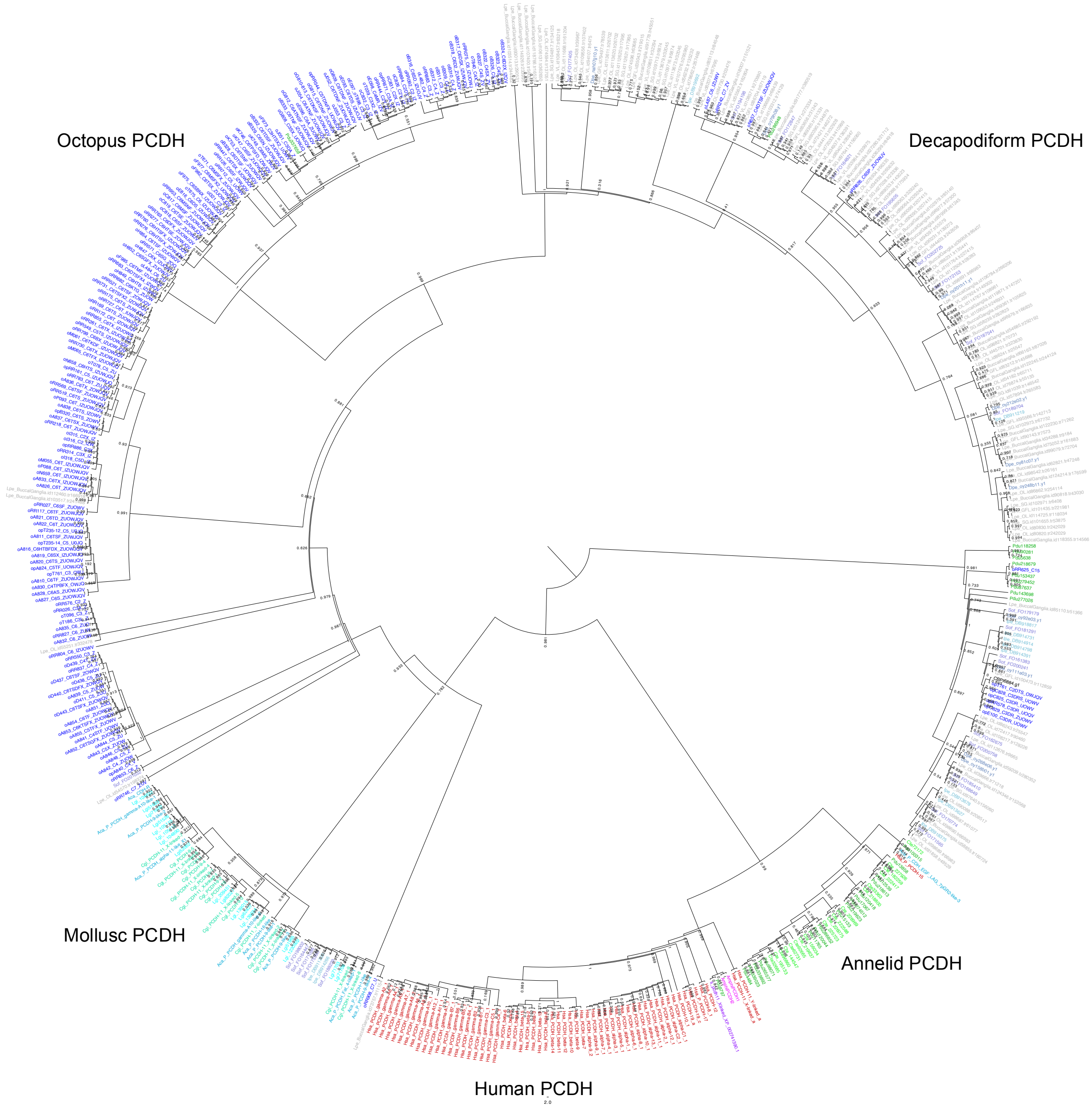


Figure S8.3.3. Metazoan protocadherin expansions. Phylogenetic tree of protocadherin genes in *H. sapiens* (red), *C. teleta* (dark green), *P. dumerilii* (light green), *L. gigantea* (teal), *A. californica* (sky blue), *C. gigas* (turquoise) *O. bimaculoides* (blue), *D. pealeii* (grey), *S. officinalis* (slate), and *I. paradoxus* (light blue-grey). We assigned the 15 Scaffolds with 3+ PCDHs an alphabetic name from A-P (skipping O for clarity). Protocadherins found in the transcriptome but not identified in the genome are designated T, and all other scaffolds (<2 PCDHs) are called RR.

The expansion of the protocadherins in cephalopods is particularly striking: we identified 168 protocadherins in the *O. bimaculoides* genome, more than twice the number found in mammals. The vast majority of these genes are contained in a single lineage-specific expansion, with many of them arranged in clusters in the genome (Figure 2, Extended Figure 4). The three largest scaffolds have 31, 17, and 10 protocadherins clustered together, respectively, and at least 25 other scaffolds have two or more protocadherins. Given the fragmented nature of the genome assembly, the number of clustered cadherins is likely much larger than what we describe here. Protocadherins that are clustered on the same scaffold are transcribed in the same direction (Figure 2), and many have a similar intron-exon structure to that found in vertebrates: the first exon contains the ECs, TM, and part of the intracellular domain, and precedes three smaller exons. We were unable to find evidence of a splicing mechanism similar to the one found in vertebrates in our *de novo* assembled transcriptomes.

The human γ protocadherin cluster contains 22 genes. If all of these genes can assemble into tetramers in *cis*, they could produce more than 230,000 specific molecular interfaces (Schreiner and Weiner, 2010), which is far more than the roughly 19,008 possible extracellular domains for *Drosophila* DSCAM (Zipursky and Sanes, 2010). If the octopus protocadherins assemble into tetramers, then the largest cluster (Scaffold 30672) alone could generate nearly one million different tetramers. Together, all octopus protocadherins could generate hundreds of millions of different tetramers, an order of magnitude greater than the 25 million tetramers that could be produced by 71 human protocadherins. As the *O. bimaculoides* protocadherins are predominantly expressed in nervous tissues (Figure 2), this gene family could play a role in providing a diverse molecular substrate for the development of large, complex nervous systems.

We also identified 155 protocadherins in the transcriptomes of the longfin inshore squid, *Doryteuthis pealeii* (Brown et al., 2014), as well as a number of other decapodiform protocadherins from cephalopod ESTs deposited in GenBank.

Maximum likelihood trees show segregated octopus and decapodiform protocadherin expansions (Figure S8.3.3). This finding may be a signature of concerted evolution, or it may indicate that this gene family underwent massive expansions in parallel after the two major clades of coleoid cephalopods diverged.

Paralog distance calculations for octopus protocadherins support recent parallel expansions in cephalopod lineages (Figure S8.3.4). Pairwise alignments for paralog sequences were constructed with MUSCLE (Edgar, 2004) and CDS alignments were computed using protein alignments as anchors. dS estimation was done with yn00 in PAML (Yang, 1997) as described in Nielsen and Yang (2003). All pairwise distances were then combined into a matrix to generate a neighbor-joining tree. Only ages for the most recent duplicates and the nodes on the neighbor-joining tree were retained. This analysis showed separate peaks for clustered and non-clustered octopus protocadherins: the clustered protocadherins have a mean pairwise dS ~0.4 while the non-clustered protocadherins have a mean pairwise dS ~1. Our analysis dates these divergences to ~135 mya and ~55 mya. Thus, these divergences happened well after the decapodiform–octopod split, which we estimate to have occurred ~270 mya (Supplementary Note 7.4).

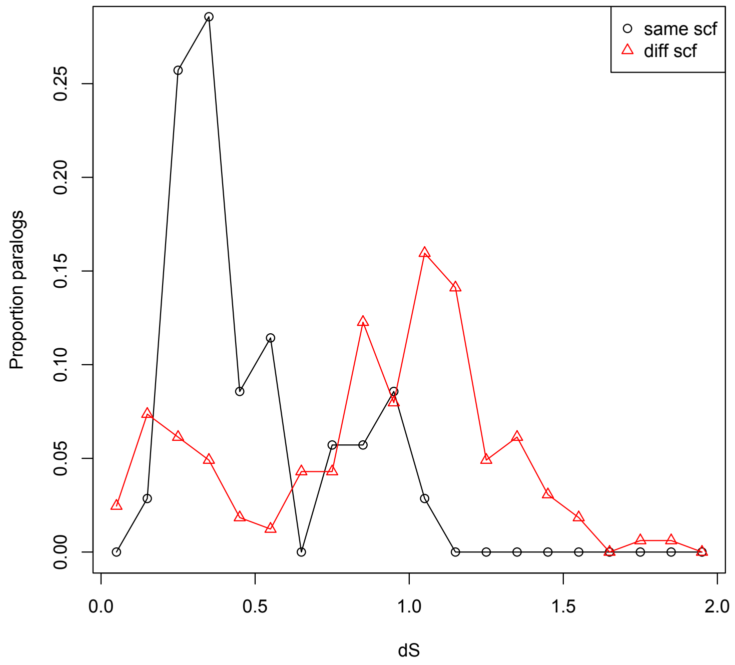


Figure S8.3.4. Protocadherin paralog distance calculation for genes located on the same or different scaffolds (scf).

Gene conversion has been described in vertebrate protocadherins (Noonan et al., 2004), largely at the 3' end of the sequence: in EC5, EC6, and in the CD. Concerted evolution of these sequences can obscure the evolutionary history of these genes. In contrast to the arrangement in vertebrates, trees of each of the individual EC domains show that the octopus and decapodiform expansions remain largely distinct (Figures S8.3.4-6). We also see that a number of protocadherins found on the same scaffold cluster together in these trees (asterisks, Figures S8.3.6-8), as they do in trees based on full-length sequence (Figure 2). Sequence alignments indicate a very low degree of nucleotide and amino acid differences across certain regions of the sequence. For example, only ~4% of the nucleotides differ over a 560nt stretch across EC4 and EC5 of 14 genes on Scaffold 30672 and Scaffold 9600 (Figure S8.3.5). Other regions of sequence, particularly in EC2 and EC3, show much greater sequence variation. The protocadherins with highly similar sequences are located in clusters on scaffolds, which could be a signature of gene conversion. However, as each of these genes have different variations at the nucleotide level, it could also indicate that these are the result of recent duplication.

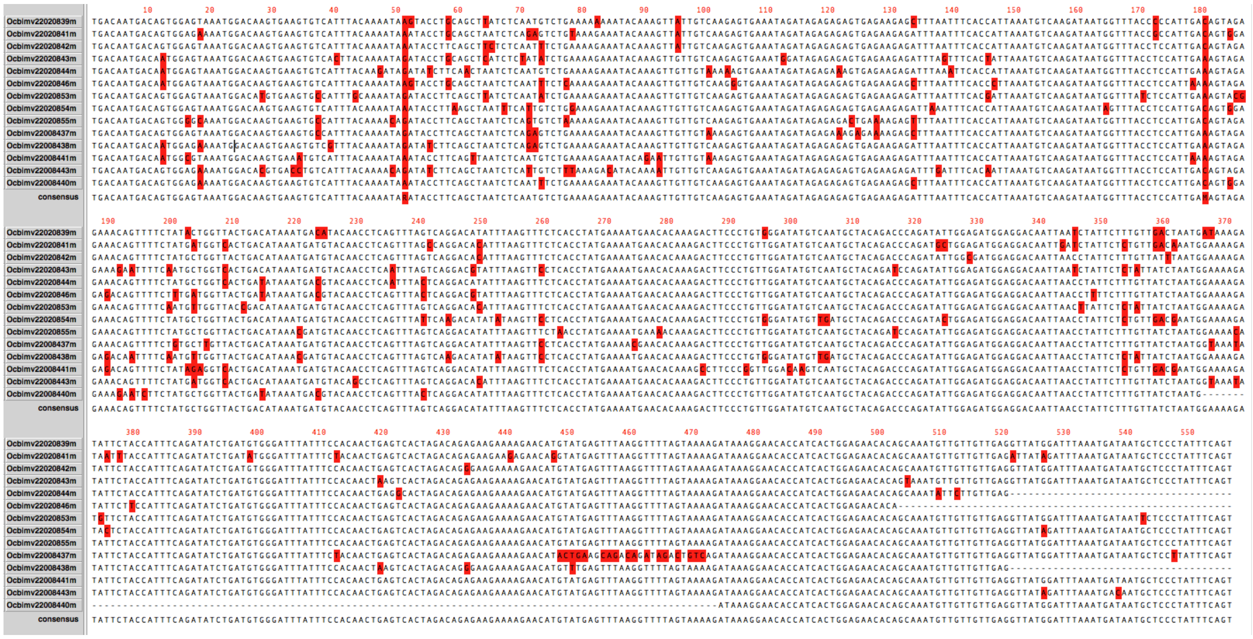


Figure S8.3.5. Multiple sequence alignment of EC4 and EC5 of 14 protocadherins located on Scaffold 30672. Nucleotides marked in red do not match the consensus sequence.

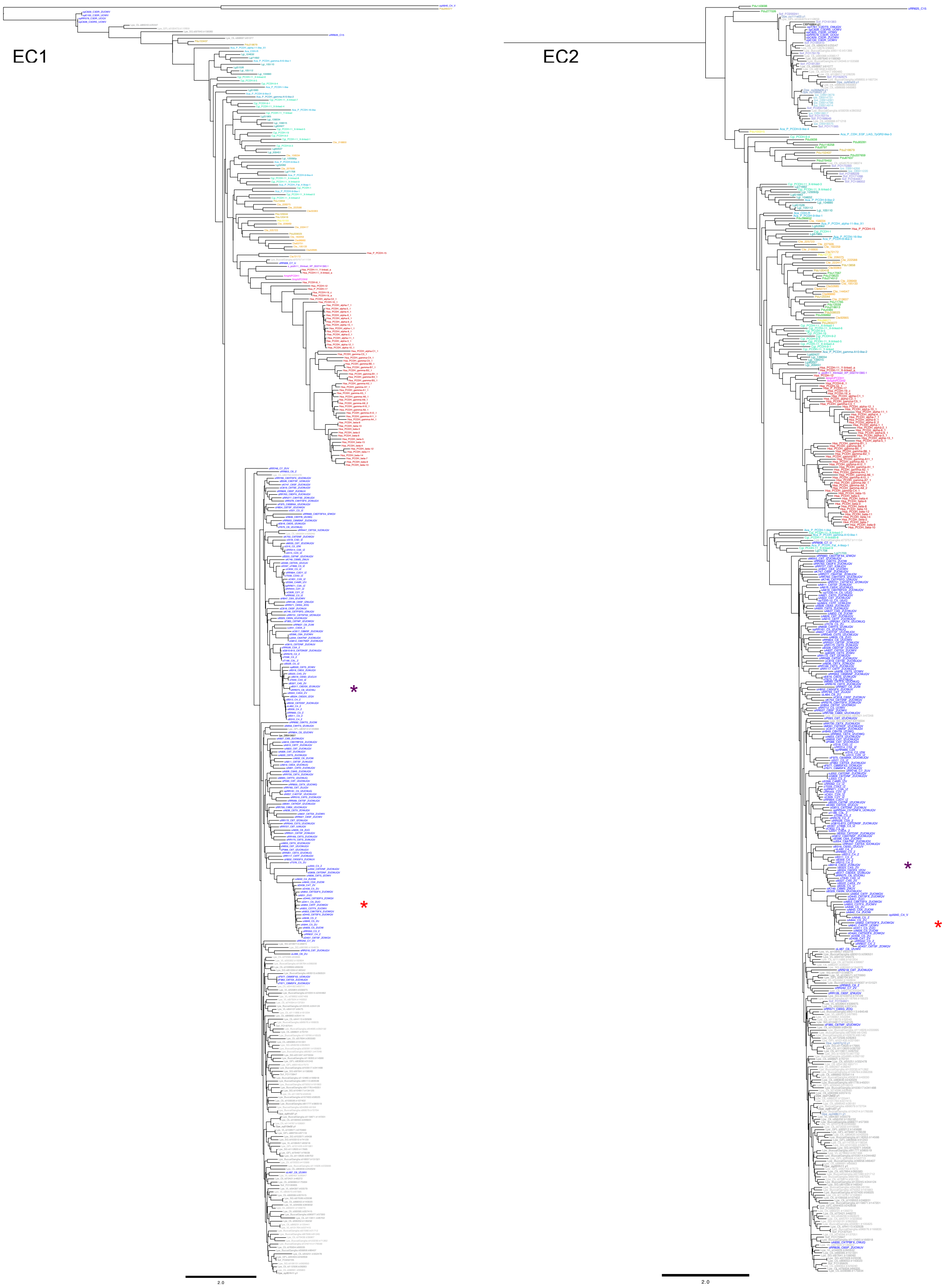


Figure S8.3.6. ML trees of EC1 and EC2 domains of metazoan protocadherins. Phylogenetic trees of protocadherin EC1 and EC2 in *H. sapiens* (red), *C. teleta* (orange), *P. dumerilii* (yellow), *L. gigantea* (teal), *A. californica* (sky blue), *C. gigas* (turquoise) *O. bimaculoides* (blue), *D. pealeii* (grey), *S. officinalis* (slate), and *I. paradoxus* (light blue-grey). Red asterisk: clustering of sequences on Scaffolds 30672 and 159035. Purple asterisk: clustering of sequences on Scaffold 9600.

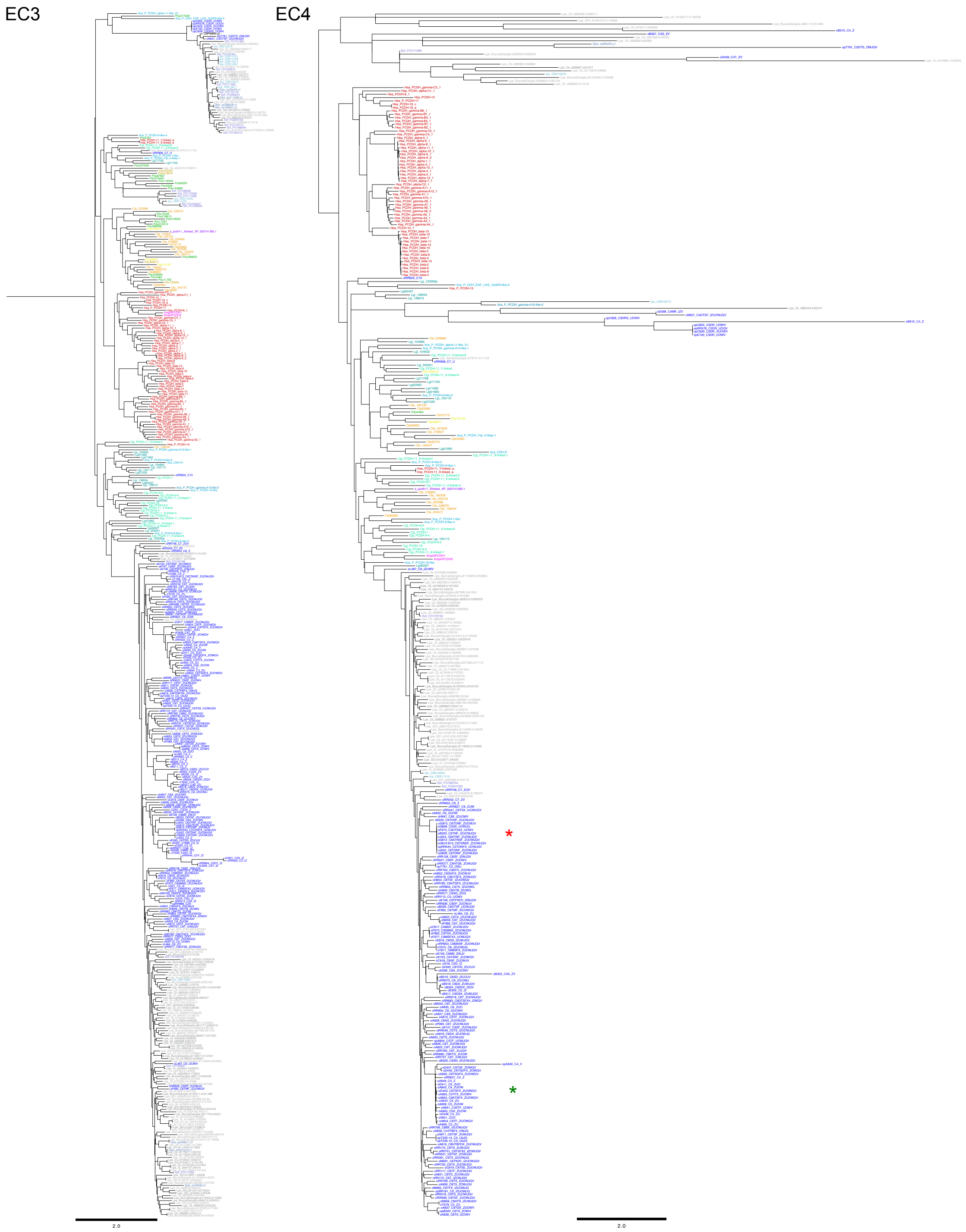


Figure S8.3.7. ML trees of EC3 and EC4 domains of metazoan protocadherins. Phylogenetic trees of protocadherin EC3 and EC4 in *H. sapiens* (red), *C. teleta* (orange), *P. dumerilii* (yellow), *L. gigantea* (teal), *A. californica* (sky blue), *C. gigas* (turquoise) *O. bimaculoides* (blue), *D. pealeii* (grey), *S. officinalis* (slate), and *I. paradoxus* (light blue-grey). Red asterisk: clustering of sequences on Scaffolds 30672 and 159035. Green asterisk: clustering of sequences on Scaffolds 93179 and 309453.

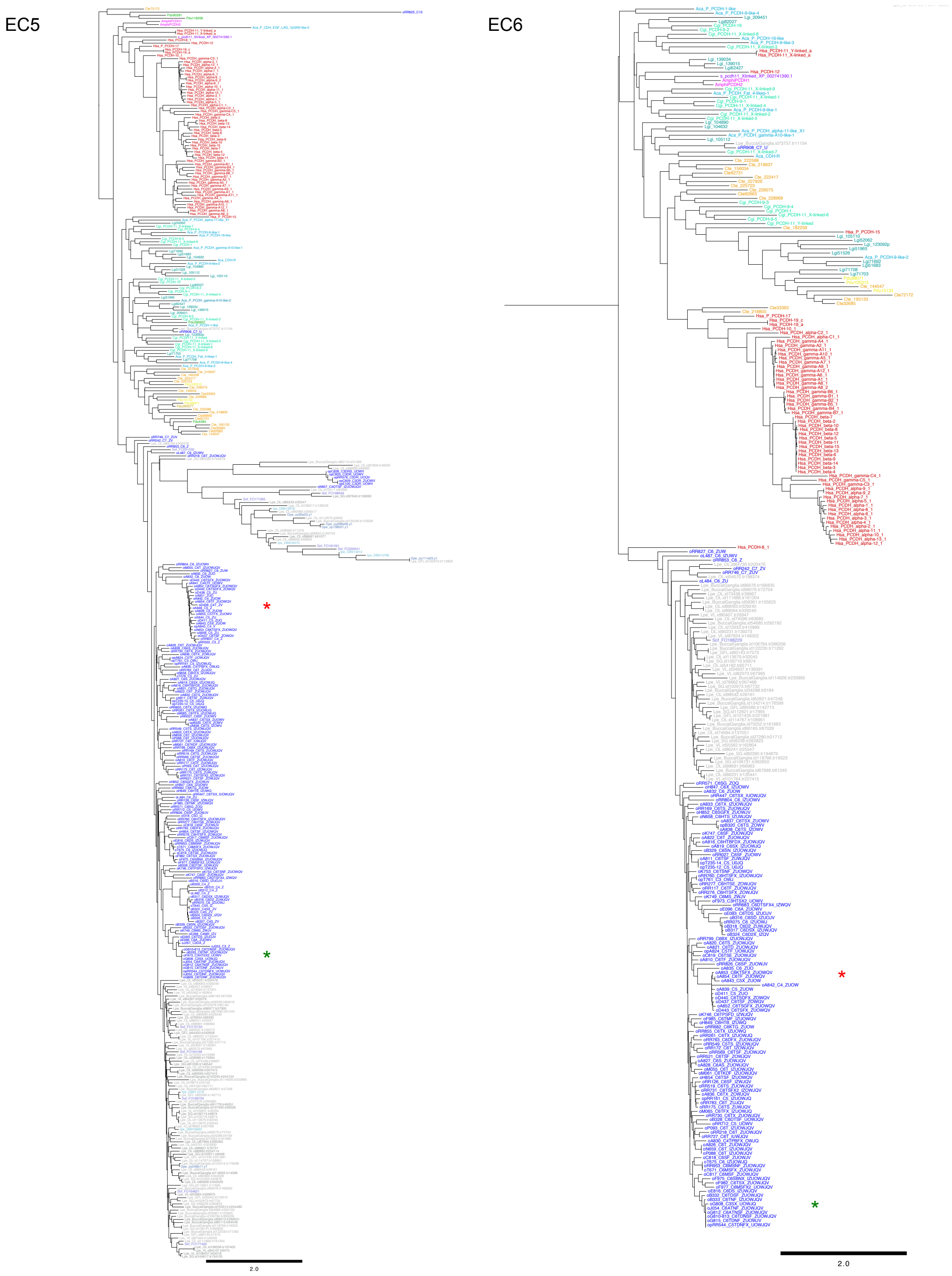


Figure S8.3.8. ML trees of EC5 and EC6 domains of metazoan protocadherins. Phylogenetic trees of protocadherin EC5 and EC6 in *H. sapiens* (red), *C. teleta* (orange), *P. dumerilii* (yellow), *L. gigantea* (teal), *A. californica* (sky blue), *C. gigas* (turquoise) *O. bimaculoides* (blue), *D. pealeii* (grey), *S. officinalis* (slate), and *I. paradoxus* (light blue-grey). Red asterisk: clustering of sequences on

Scaffolds 30672 and 159035. Green asterisk: clustering of sequences on Scaffolds 93179 and 309453.

We searched for novel amino acid motifs in octopus cadherins and protocadherins. We identified 7 new motifs (Table S8.3.1): 6 in EC domains and 1 in the TM (domains predicted by Pfam). Variations of these motifs also appear in *D. pealeii* (marked in red, Table S8.3.2). Four octopus motifs were found unchanged in *Doryteuthis*. These motifs were not found in the genomes of other species that we examined, suggesting that cephalopod protocadherins may have unique structures.

Domain	Motif Structure	# of Cadherins with motif
EC1	X_2[Y/L/F][I/V/L/A][G/A][D/N][I/V]XA[D/N]	55
EC1	D[A/T]EX_2C	139
EC5	[I/V][L/F/S][I/V/A][K/T/S/I/R]D[N/C/S/K]GXPXL	107
EC5	XDXNDN[A/P/V/T/S]PY	113
EC6	L[R/K][A/V/S][S/L/V/A]D[R/K/I/N]DX[H/R/G]XN	74
EC6	QNDAG	80
Trans-membrane	[I/V][I/V][I/V/A]X_3[A/V][V/I]X_2[S/A]X_2	94

Table S8.3.1. Amino acid motifs in *O. bimaculoides* cadherins. Conserved sequences in octopus range from 5-12 amino acids.

Domain	Motif Structure	# of Cadherins with motif
EC1	X_2[Y/L/F][I/V/L/A][G/A][D/N][I/V]X_2D	78
EC1	D[A/T]EX_2C	87
EC5	[I/V][L/F/S][I/V/A][K/T/S/I/R]D[N/C/S/K]GXPXL	51
EC5	XDXNDN[APVTS]PY	55
EC6	L[N/R/K][A/V/E/S/R][S/L/V/A/T]D[R/K/C/S/I/G]DX[H/R/G]XN	43
EC6	QXDAG	25
Trans-membrane	[I/V][I/V][I/V/A]X_3[A/V][V/I]X_2[S/A]X_2	35

Table S8.3.2. Amino acid motifs in *D. pealeii* cadherins. Conserved sequences in squid range from 5-12 amino acids. Residues marked in red show divergence from *O. bimaculoides* motifs.

We also identified protocadherins in non-cephalopod lophotrochozoans, including the polychaete annelids *C. teleta* and *P. dumerilii*, and the molluscs *L. gigantea* and *C. gigas. L. gigantea* and *C. gigas* each have 17-25 protocadherins; in *L. gigantea* 14 are clustered on one scaffold, while *C. gigas* has two clusters of four and three doublets. Unlike the clusters in *O. bimaculoides*, these clustered protocadherins in other molluscs are transcribed in alternating directions. The other molluscan protocadherins group together in phylogenetic analyses of full-length sequence, but show a very different pattern from the cephalopod genes. *L. gigantea* and *C. gigas* protocadherins interdigitate, which could indicate that the protocadherins started to expand before bivalves and snails diverged. We see a similar pattern in the annelids, in which the polychaete protocadherins group together on the tree, with the *C. teleta* and *P. dumerilii* genes interdigitating. Unlike the molluscan protocadherins, the largest grouping of the 25 *C. teleta* protocadherins is two genes on one scaffold. Similar lineage-specific patterns of protocadherin expansions has also been described in vertebrates (Hulpiau and van Roy, 2009; Morishita and Yagi, 2007; Yu et al., 2008).

8.4 C2H2-ZNFs

The zinc finger (ZNF) domains are small peptide domains that bind zinc ions at Cys and His residues to form “finger-like” protrusions (Iuchi, 2001). The best characterized and most prevalent of the ZNFs, the Cys2His2 type (C2H2), coordinates a single zinc ion to stabilize its unique secondary structure. Two or more fingers, connected by short linker sequences, are needed to bind to unique DNA or RNA sequences. Binding specificity, or the protein’s “fingerprint,” depends on both the identity of the ZNFs and that of the linker sequences (Liu et al., 2014). C2H2-ZNFs have been implicated in many cellular processes, including cell fate determination and early development (Liu et al., 2014). However, much remains unclear about the function and evolution of these proteins.

ZNFs are prevalent among eukaryotes, forming several large superfamilies throughout the animal kingdom. Humans have over 700 C2H2-ZNF proteins, making up 3% of the protein-coding genes in the genome (Klug 2010, Figure S8.4.1). Lineage-specific ZNF gene expansions have been described in a number of vertebrates, raising the interesting possibility that they function in species-specific innovations (Liu et al., 2014; Shannon et al., 2003). For example, the KRAB ZNF gene family descended from a single ancestral sequence and has expanded dramatically in many tetrapod lineages to become the largest family of transcription factors in humans, with roughly 400 members. A subset of these KRAB ZNFs is the result of independent expansion in primates (Nowick et al., 2010).

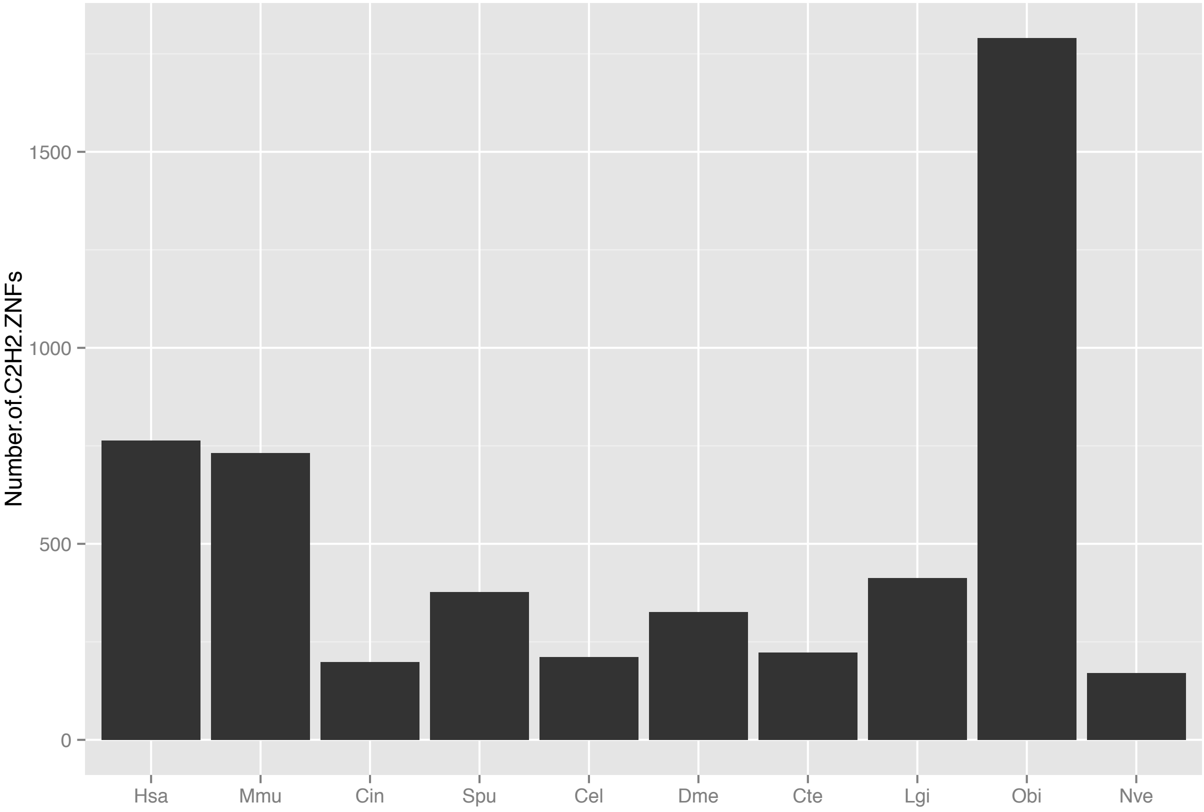


Figure S8.4.1. C2H2 genes in different species. While most invertebrate species investigated have far fewer C2H2-ZNFs than do mammals, *O. bimaculoides* presents a dramatic expansion with almost 1,800 individual C2H2 genes.

We found 1,790 C2H2-ZNF proteins in the genome of *O. bimaculoides*. A phylogenetic tree of *O. bimaculoides* C2H2-ZNFs is available as Supplementary Information (C2H2_tree.pdf). The tree was built using 1,452 sequences from octopus, 539 from human, 388 from mouse, 159 from zebrafish, 7 from *A. queenslandica*, 88 from *A. californica*, 24 from *C. intestinalis*, 65 from *S. kowalevskii*, 44 from *T. castaneum*, 39 from *N. vectensis*, 87 from *L. gigantea*, and 72 from *C. teleta*. The alignments for this tree were made with CLUSTALO and the tree was built with FastTree.

Octopus ZNFs are clustered along scaffolds but transcribed in different directions (Figure 3), which has also been found in vertebrate ZNF clusters (Shannon et al., 2003). Whereas the human ZNF family is dominated by an expansion of the KRAB-type ZNFs, the octopus ZNF complement appears to be characterized by the elaboration of multiple C2H2 types. All invertebrates we investigated, from well-characterized genetic models like *D. melanogaster* and *C. elegans* to other lophotrochozoans, only have a complement of, at most, a few hundred C2H2-ZNF genes, indicating that the expansion in octopuses is unusually dramatic. Despite the differences in numbers, it is interesting to note that the mechanism of genomic expansion may be similar across different species. Our transposable element analyses suggest that the amphioxus, octopus, and human C2H2-ZNF expansions are all linked to beta-satellite repeat activity.

ZNF proteins require at least two or three finger domains to bind DNA selectively; ZNFs with multiple C2H2 arrays may have multiple binding sites mediated by different finger sets (Iuchi, 2001). We identified 1,790 octopus C2H2s with three or more C2H2 domains, as well as other known ZNF domains such as Elf1 and FYVE (Figure S8.4.2). 52% (931/1,790) of these genes contain over 20 C2H2 domains. We found the helix capping linker sequence, TGEKP, to be present in octopus C2H2-ZNF sequences. An analysis of the C2H2 gene sequences did not reveal any octopus-specific linker sequences or effector domains.

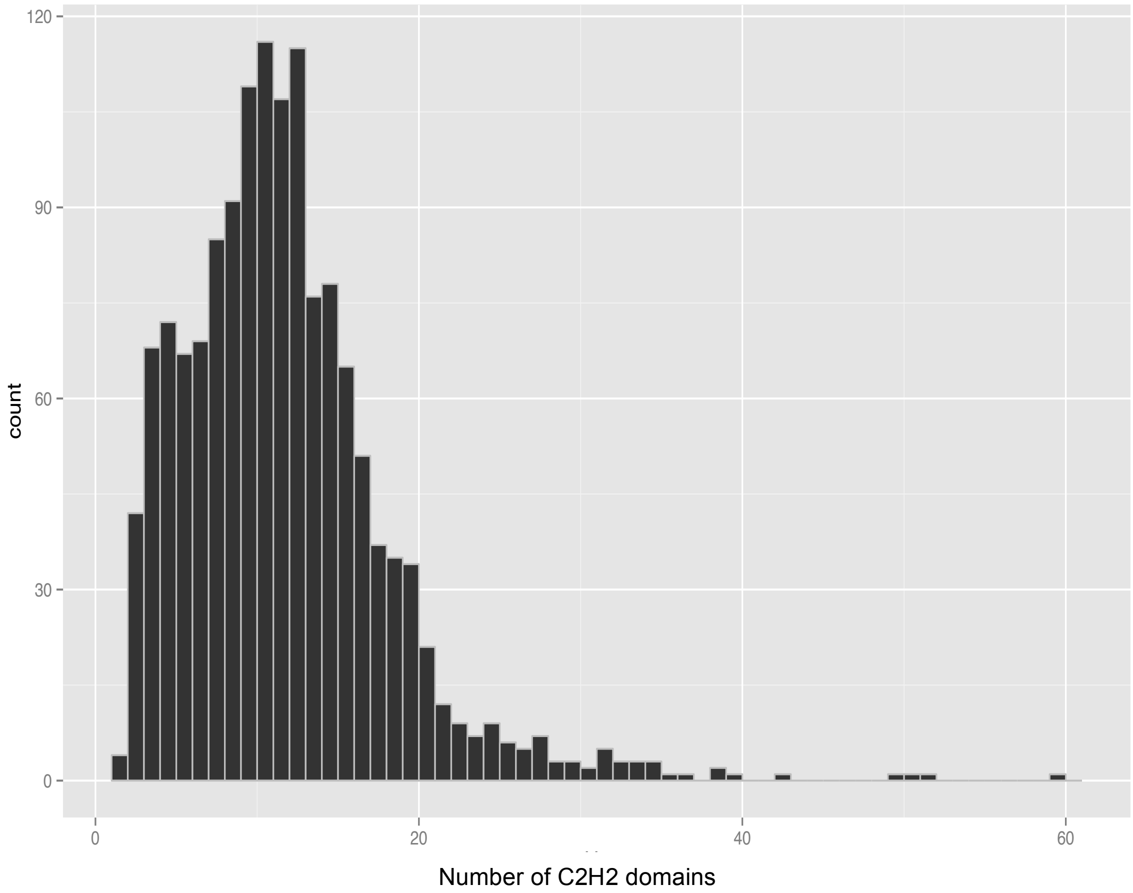


Figure S8.4.2. Distribution of C2H2 Domains in octopus ZNFs. Octopus ZNF genes have between 2 and 60 C2H2 domains.

Of the 1,790 octopus C2H2 genes, 1,429 were confirmed as expressed by our RNA-Seq data. Transcriptome analyses revealed strong enrichment of C2H2 gene expression in neural tissues, especially the OL, the supraesophageal brain, and ANC (Figure S8.4.3). Our data support the long-standing idea that ZNFs function in neural complexity and development (Layden et al., 2010; Liu et al., 2014). The axial nerve cords together contain nearly 350 million cells and represent a unique octopus neural innovation (Young, 1971). The optic lobes and supraesophageal brain are major centers of sensory processing, learning and memory. It is interesting to note that the subesophageal brain shows a less dramatic enrichment of C2H2 genes. Overall, this expression profile is consistent with the observation that a greater number of C2H2-ZNFs is associated with greater neural complexity in vertebrate species (Vinogradov, 2013). We also found C2H2 expression in St15 embryos. High activity of the unusually large C2H2 repertoire during embryogenesis may indicate a role in the development of cephalopod-specific innovations. The octopus C2H2-ZNF complement is one of the largest gene family expansions in the animal kingdom, and our evidence lends support to the importance of these genes in conserved transcriptional pathways.

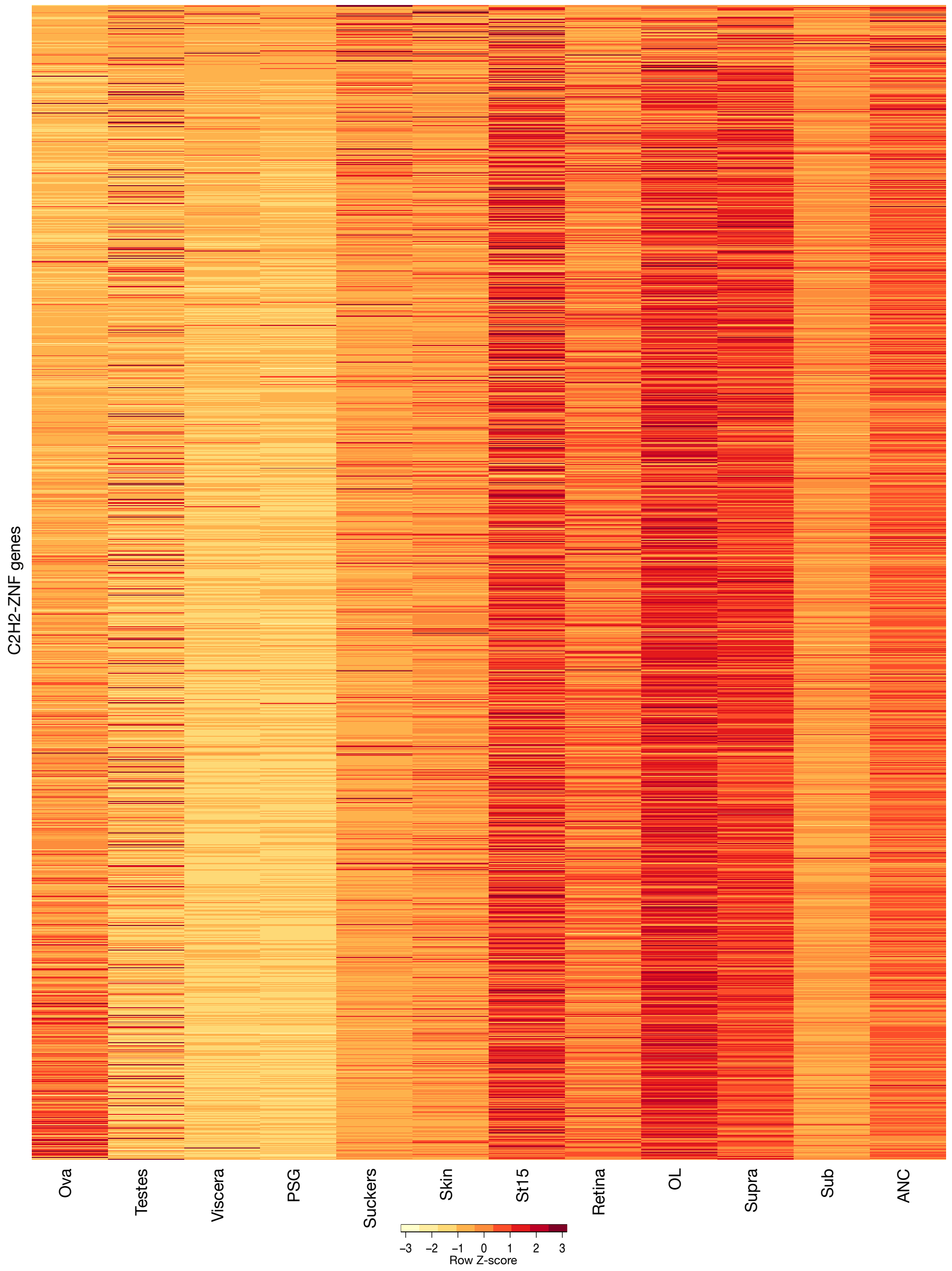


Figure S8.4.3. Expression of 1,429 C2H2-containing genes in 12 octopus transcriptomes.

8.5 GPCRs

The diverse G protein-coupled receptor (GPCR) family is divided into 4 classes. Class A (rhodopsin-type) receptors represent the largest class of GPCRs and include opsins, chemokine receptors, and the vertebrate olfactory receptors. Class B (secretin-type) receptors include the adhesion GPCRs, calcitonin receptors, as well as several hormone receptors. Class C (glutamate-type) receptors consist of the $GABA_B$ and metabotropic glutamate receptors. Class F, the smallest class, comprises the Frizzled and Smoothened genes.

A recent study found GPCR family expansions in 3 lophotrochozoan species (Simakov et al., 2013). We identified 329 GPCRs in *O. bimaculoides*. A phylogenetic tree of *O. bimaculoides* GPCRs is available as Supplementary Information (GPCR_tree.pdf). The tree was built using the 328 GPCR sequences from octopus, 185 from human, 160 from mouse, 226 from zebrafish, 204 from *S. kowalevskii*, 125 from *L. gigantea*, 258 from *C. teleta*, and 87 from *A. californica.* For clarity, a subset of vertebrate olfactory receptors was used. Members of each class and their distribution by species are given in Table S8.5.1. The octopus Class A rhodopsins are centered around 7 o'clock in the tree. Notable features of the octopus GPCR repertoire include the high number of secretin, latrophilin, and metabotropic glutamate receptors relative to *Lottia*, *Capitella*, and humans.

Class	Subtype	Obi	Lgi	Cte	Hsa
A (rhodopsin)	Amine receptors	30	37	122	80
	Lipid-like receptors	6	12	22	37
	Nucleotide-like receptors	12	19	85	36
	Short peptide receptors	119	129	415	57
	Orphan/Other	51	10	12	77
	Total	218	207	656	287[1]
B (secretin)	Calcitonin	7	4	7	2
	Secretin	14	5	1	1
	Latrophilin	7	-	-	3
	Corticotropin releasing factor receptor	4	-	-	2
	Parathyroid hormone receptor	6	-	-	2
	CELSR	1	-	-	3
	Orphan/Other	27	32	23	36
	Total	66	41	31	49
C (glutamate)	GABA	5	6	25	2
	Metabotropic glutamate receptor	15	8	14	8
	Calcium-sensing receptors	1	-	-	2
	Orphan/Other	0	1	3	7
	Total	21	15	42	19
F (frizzled)	Frizzled	5	3	2	10
	Smoothened	1	1	1	1
	Total	6	4	3	11
U (unclassified)	***Total***	17	23	30	52
TOTAL		328	267	732	418

Table S8.5.1. Distribution of GPCRs by class and species. Lgi, Cta and Hsa numbers derived from NCBI protein databases and Simakov et al. (2013).
[1]excludes olfactory receptors.

Tissue-wide expression of all GPCRs, organized by class, is shown in Figure S8.5.1. GPCRs of all classes are enriched in the ANC, OL, and supra- and subesophageal brains. The suckers, skin, testes, and PSG also contain small groups of strongly expressed GPCRs. Other than the Frizzled receptors, which are expressed predominantly in the developing embryo, the St15 transcriptome is not enriched with a large number of highly expressed GPCRs. Taken together, the data suggest that GPCRs are important for signal transduction in mature animals.

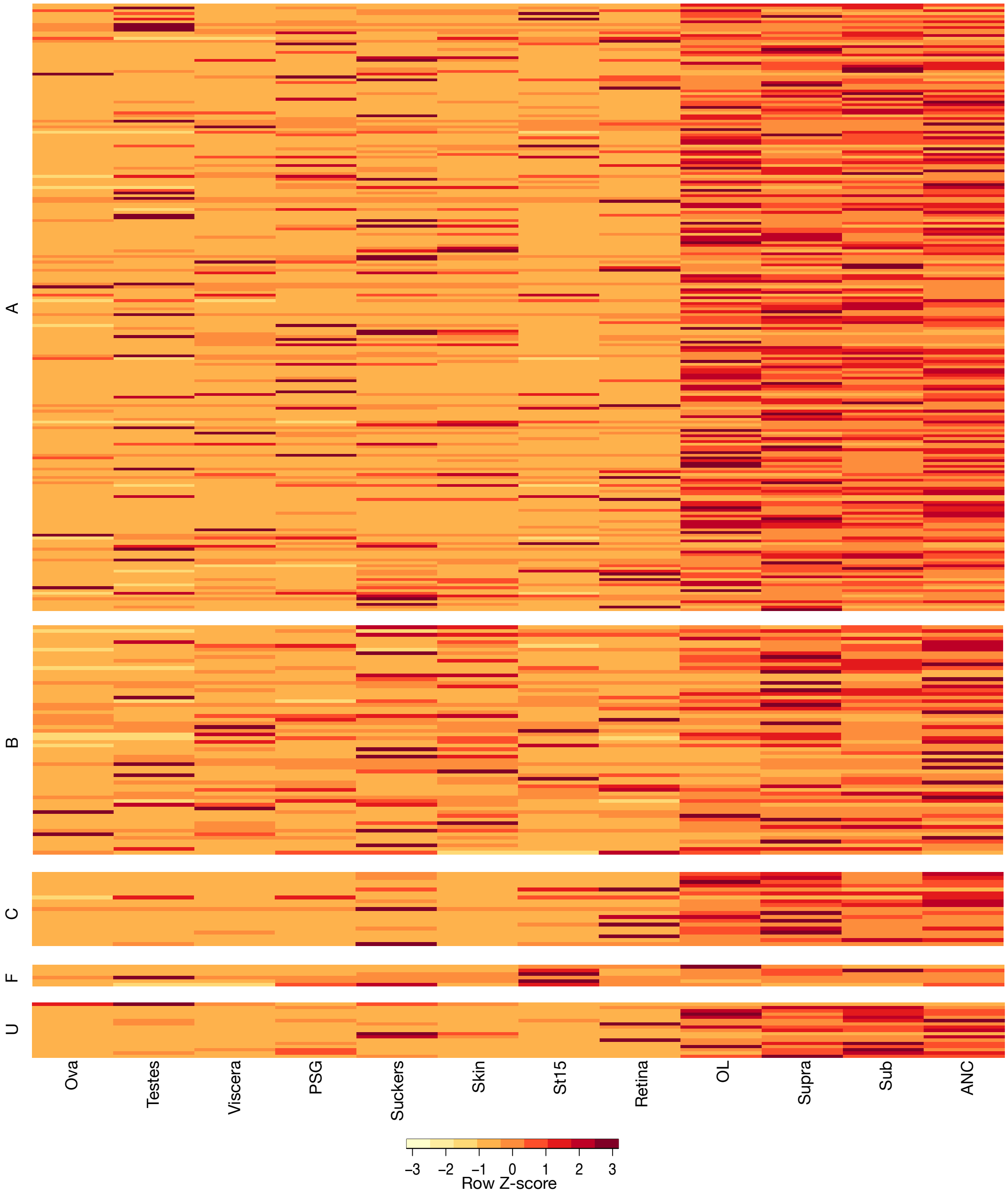


Figure S8.5.1. Expression profiles of 328 GPCR proteins in 12 octopus transcriptomes. A, rhodopsin-type: 220 genes; B, secretin-type: 62 genes; C, glutamate-type: 19 genes; F, Frizzled-type: 6 genes; U, unclassified: 23 genes.

8.6 Chitinases

Chitinases are involved in chitin degradation and are found in bacteria, fungi, plants, insects, and mammals (Adrangi and Faramarzi, 2013). Based on Fisher's exact tests for enrichment in PANTHER categories (similar to Pfams, Supplementary Note 7.4) we identified chitinase-related genes (PTHR11177) as a significantly expanded category in octopus (p-value 5E-03). The main chitinase class has 18 proteins in octopus, similar to the complement in *L. gigantea* (19 chitinases). The di-N-acetylchitobiase (or chitobiase, PTHR11177:SF8) class is expanded in octopus, with 15 members, compared to the 9 members found in *L. gigantea*, 8 in *C. gigas*, 1 in mouse, and 1 in human. We did not find the chitin synthase family to be significantly expanded in octopus. The suckers show strong enrichment of both chitobiase and chitinase expression (Figure S8.6.1). Interestingly, only the main chitinase family shows expression in the viscera (Figure S8.6.1b). Members of both classes show high expression in the retina as well.

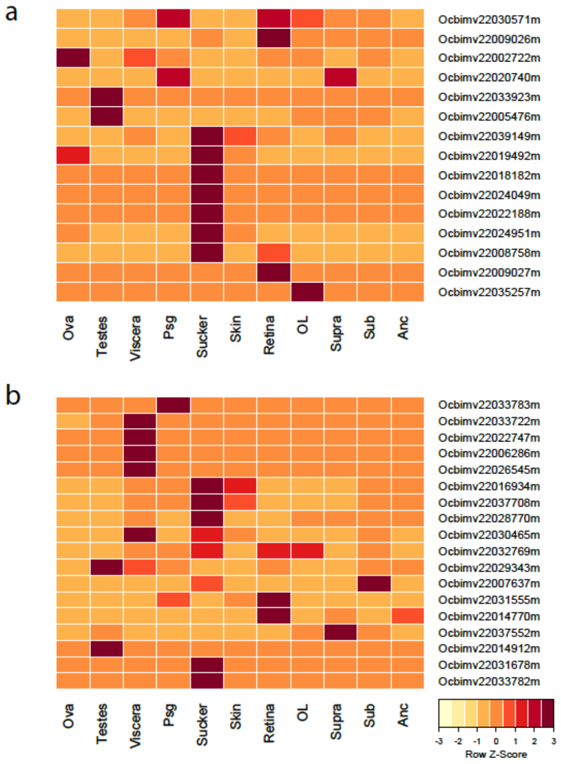


Figure S8.6.1. Expression of di-N-acetylchitobiase (a) and chitinase (b) in octopus.

8.7 IL17s and IL17 receptors

In vertebrates, chemokines and cytokines mediate the inflammatory response by orchestrating the migration of leukocytes and other immune cells to the site of infection. These small proteins share 4 cysteine residues that form disulfide bonds (Baggiolini, 1998). The interleukin 17 proteins are a family of proinflammatory cytokines involved in both acute and chronic inflammation, thus serving as an important mediator between adaptive and innate immune functions (Shabgah et al., 2014). Humans have 6 members of this family, IL17A-F. Each member has a different function: for example, IL17A plays an important role in mechanical hyperalgesia through the upregulation of TRPV4 channels (Segond von Banchet et al., 2013).

We found 31 interleukin genes in the octopus genome, all of which show similarity to mammalian *IL17* and are thus named "*IL17-like*" (Extended Data Figure 5). The 17 *IL17*-like genes found in our transcriptomes are highly expressed in suckers and skin, as well as the PSG and viscera (Extended Data Figure 5b). These tissues all directly contact the external environment. Similarly, the 6 *IL17-like* genes in *C. gigas* were found to be highly expressed in gill and digestive gland tissues (Li et al., 2014). Different *C. gigas* IL17-like proteins are inducible by different pathogenic agents and are proposed to play distinct roles in the acute phase of infection, just as human IL17 proteins do. These data raise the possibility that along with primary sequence similarity, tissue-specificity and function have been conserved between mammalian and molluscan IL17s.

IL17 family members have been shown to be able to homo- and heterodimerize (Chang and Dong, 2007). The 17 octopus *IL17-like* genes expressed in our transcriptomes could generat up to 171 dimers. We also searched for IL receptors in the octopus genome. We found 2 IL receptors that bear closest similarity to vertebrate IL1 and IL25 receptors, and an additional homolog to the vertebrate IL17 receptor (data not shown). Our findings substantially augment the nascent body of research on molluscan IL17 expression and function.

9. HOX COMPLEMENT AND DISTRIBUTION ON SCAFFOLDS

Hox genes encode a family of homeodomain transcription factors that play important roles in animal development, such as specifying anterior-posterior identity. In many bilaterians, the Hox genes cluster in the genome, and the order of genes in the cluster reflects the timing and pattern of their expression along the anterior-posterior axis in developing embryos; this feature is called collinearity. Clustering has been described in a wide range of animals, including at least one mollusc, *L. gigantea* (Simakov et al., 2013)*.* Nine of eleven known lophotrochozoan Hox genes have been isolated by PCR from the Hawaiian bobtail squid, *Euprymna scolopes* (Callaerts et al., 2002). The expression of these Hox genes in *E. scolopes* was examined using *in situ* hybridization, but a conventional Hox cluster was not demonstrated. The combinatorial expression patterns found did not conform to collinearity under the assumption of a conventionally organized cluster (Lee et al., 2003).

We searched the octopus genome and transcriptome assemblies for homeodomains using BLAST searches with sequences from HomeoDB as bait. Candidate Hox genes were verified using BLAST and Pfam. The identified octopus sequences were aligned with those from other bilaterians using MUSCLE, followed by manual curation and adjustment of the alignments. Phylogenetic trees were constructed with RAxML and FastTree using either the homeodomain or the full-length sequences.

We identified eight Hox genes in *O. bimaculoides*: *LAB*, *SCR*, *LOX5*, *ANTP*, *LOX2*, *LOX4*, *POST2* and *POST1*, as well as the ParaHox genes *GSX* and *CDX* (Figure S9.1). We did not find genes resembling Hox paralog groups (PG) 2, 3, or 4. While no PG2 (pb) Hox gene was identified in either *E. scolopes* (Callaerts et al., 2002) or *S. officinalis* (Pernice et al., 2006), one has been described for *Nautilus pompilius* (Iijima, 2006), which diverged from octopus ~400 mya (Kroger et al., 2011). Genes for *ZEN* (PG3) and *DFD* (PG4) were identified in *E.*

scolopes, while *Lox2*, a lophotrochozoan-specific central class gene, was not found (Callaerts et al., 2002). These findings may indicate that there are different patterns of Hox gene retention or loss across cephalopod lineages.

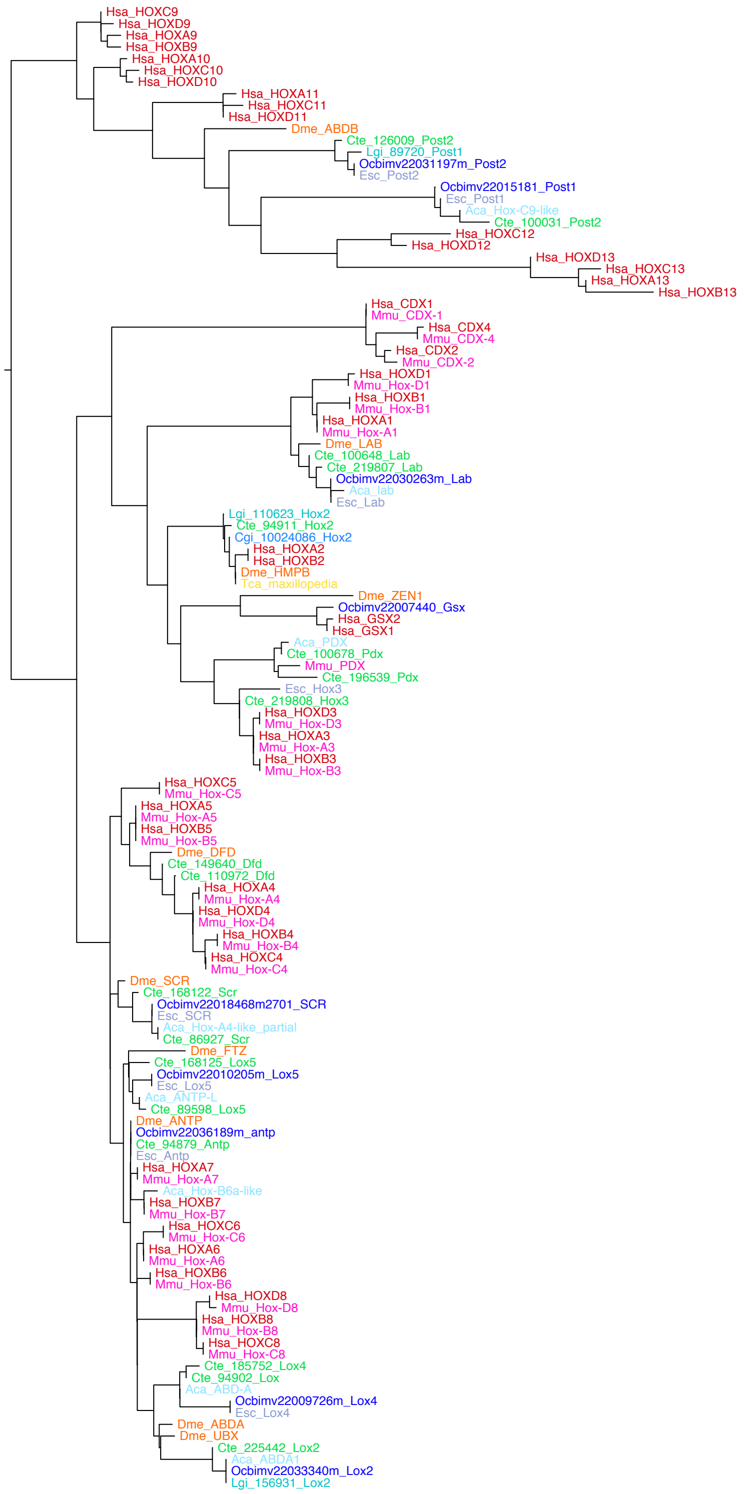


Figure S9.1. Phylogeny of bilaterian Hox and ParaHox genes based on homeodomain sequence, constructed using FastTree.

All eight *O. bimaculoides* Hox genes are detected in the transcriptome of stage 15 embryos, as would be expected for developmental regulatory genes. Hox genes were also expressed in adult tissues, particularly the subesophageal brain, axial nerve cord, skin and suckers (Figure S9.2).

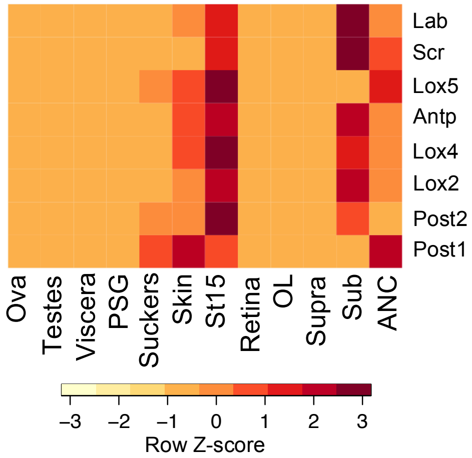


Figure S9.2. Expression of *O. bimaculoides Hox* genes detected in development, and in adult skin, suckers and neural tissues. Absence of supraesophageal *Hox* gene expression was also reported for *E. scolopes* (Lee et al., 2003).

In the *O. bimaculoides* genome assembly no two Hox genes are found on the same scaffold (Table S9.1). Many of these scaffolds are several hundred kb long, suggesting that the Hox complex has been fully atomized (Extended Data Figure 2). Even if the two smallest scaffolds (those containing *ANTP* and *LOX2*) were linked, the shortest distance between these Hox genes would be at least 50kb if they were transcribed on different strands, and 100kb if they were transcribed on the same strand as documented in other Hox clusters. With the exception of the short scaffold containing *ANTP*, all of the scaffolds containing Hox genes also have a number of short open reading frames outside of the Hox locus.

Paralog Group	Gene ID	Scaffold	Scaffold length (bp)	Coding start site	Coding stop site
Hox1	Ocbimv22030263	Scaffold5409	421,457	158,935	160,310
Scr	Ocbimv22018468	Scaffold2701	474,802	380,581	380,220
Lox5	Ocbimv22010205	Scaffold17471	751,982	487,766	487,551
Antp	Ocbimv22036189	Scaffold79555	53,356	41,620	41,390
Lox4	Ocbimv22009726	Scaffold169723	137,412	91,093	90,466
Lox2	Ocbimv22033340	Scaffold66266	423,253	192,815	193,791
Post2	Ocbimv22031197	Scaffold582	231,632	151,858	151,619
Post1	Ocbimv22015181	Scaffold22588	187,962	30,898	31,261

Table S9.1. Location of Hox genes in *O. bimaculoides* genome assembly.

10. NEURONAL GENES

10.1 Neurotransmitter-related enzymes

Neurons and neuron subtypes can be characterized by detection of specific proteins, such as those that synthesize and degrade neurotransmitters. For example, the presence of tyrosine hydroxylase in a neuron identifies that neuron as catecholaminergic. We identified genes coding for elements of neuronal identity in the *O. bimaculoides* genome (Supplemental Table S10.1). We found an expansion in the number of catecholamine beta-hydroxylases, which is not surprising since invertebrates utilize catecholamines, including octopamine, broadly for neurotransmission. We also found five genes coding for acetylcholinesterase and three homologs of *ELAV*, which is frequently employed as a general cell-type marker for neurons.

	Obi	Lgi	Cte	Dme	Cel	Hsa
Choline Acetyltransferase	1	1	1	1	1	1
Acetylcholinesterase	5	2	1	1	4	2
Glutamine Synthetase	1	1	1	2	5	1
Tyrosine Hydroxylase	1	1	1	1	1	1
Catecholamine Beta-Hydroxylase	6	12	2	1	1	1
Glutamate Decarboxylase	1	1	2	1	1	2
ELAV	3	2	2	1	1	4

Table S10.1. Counts of identified genes coding for elements of neuronal identity.

10.2 Neurotransmitter vesicular transporters: SLC 17, 18 and 32

Solute carrier (SLC) genes represent a large class of membrane transport proteins. In humans, the *SLC17* family is composed of 9 sodium/anion cotransporters: 4 sodium-dependent phosphate transport proteins (NPTs), 1 vesicular nucleotide transporter (*VNUT*), 3 vesicular glutamate transporters (*VGLUT*), and sialin. These proteins are involved in a wide range of cellular processes, such as packing neurotransmitter vesicles and urate metabolism (Reimer, 2013). For example, sialin is implicated in at least two functions: it acts as a proton-coupled sialic acid transporter in lysosomes and as a vesicular transporter of the excitatory amino acids aspartate and glutamate in the nervous system (Miyaji et al., 2008). The human genome has one sialin gene (*SLC17A5*). Recent analysis of a ctenophore genome identified eight sialin-like genes (Moroz et al., 2014). Using previously sequenced *SLC17* genes as bait, we discovered 45 *SLC17* genes in the octopus genome (Figure S10.2.1). Octopus has 1 *VGLUT*, 1 *VGLUT-like* gene, 1 *VNUT*, 2 *VNUT-like* genes, 4 *NPT-like* genes, 1 sialin, and an expansion of 35 sialin-like genes that show enriched expression in the central brain, skin, suckers, viscera, and testes (Figure S10.2.2). Twenty-eight of these sialin-like genes are clustered along 3 scaffolds: Scaffold 14699 (9 genes), Scaffold 165330 (12 genes), and Scaffold 74738 (7 genes), indicating that tandem duplication may have contributed to this expansion. The non-clustered, non-sialin-like members of the SLC17 family are strongly expressed in peripheral tissues and the central nervous system.

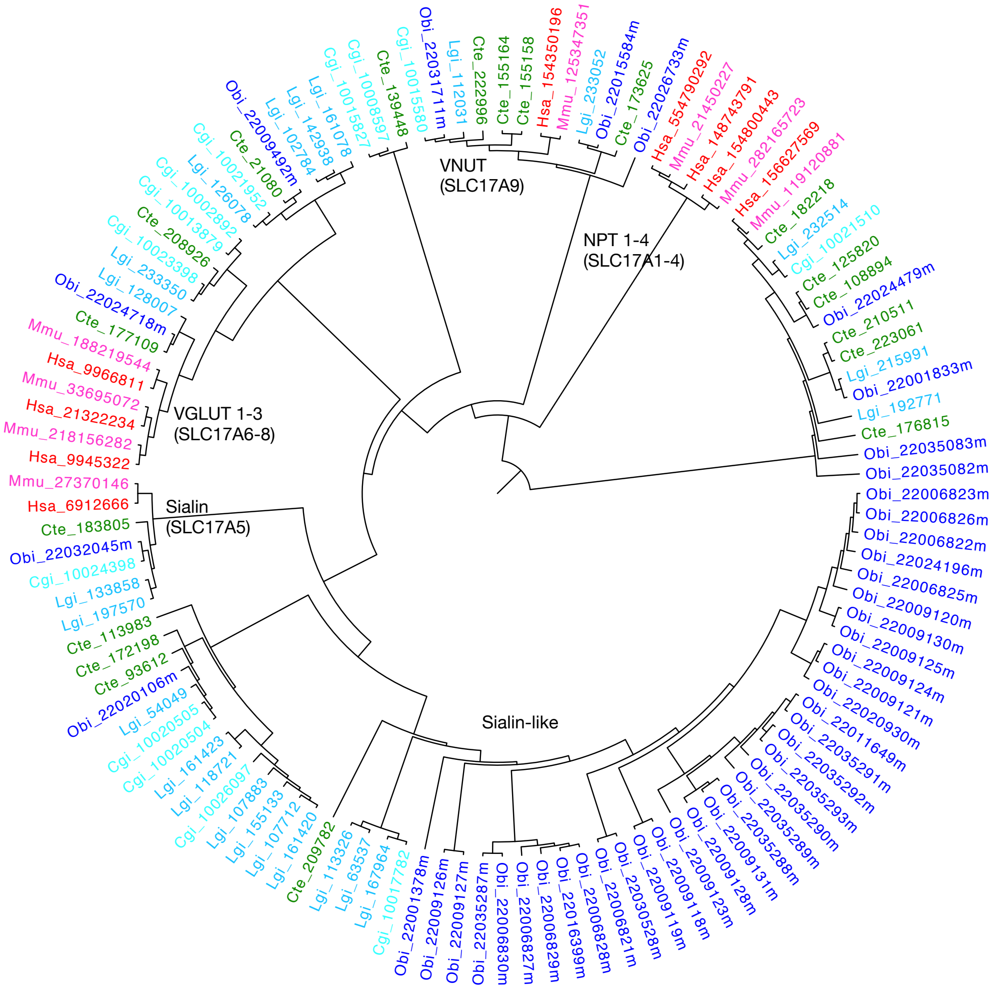


Figure S10.2.1. Phylogenetic tree of SLC17 genes.

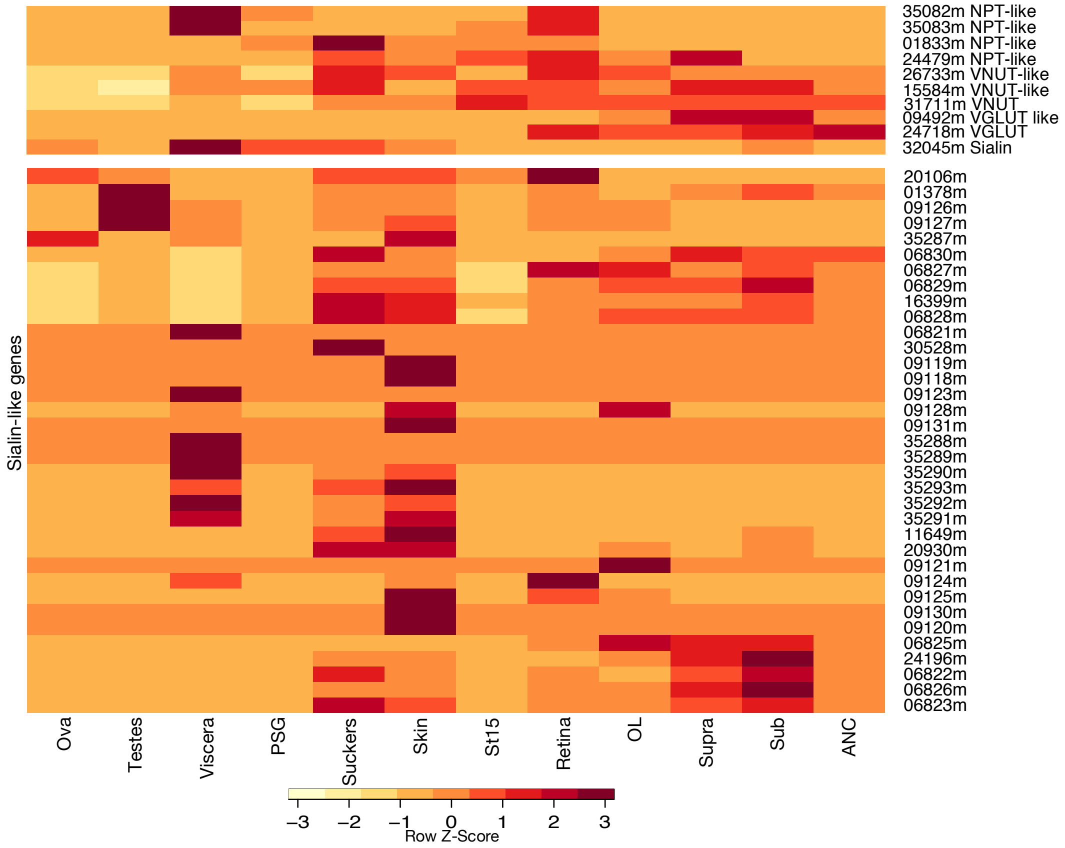


Figure S10.2.2. Expression profile of SLC17 genes in 12 octopus tissues.

We also investigated the *SLC18* and *SLC32* families. In mammals, the *SLC18* family consists of vesicular amine transporters: two monoamine transporters, a vesicular acetylcholine transporter, and a newly characterized vesicular polyamine transporter (Hiasa et al., 2014). Since many amines directly bind and modulate the activity of postsynaptic receptors, the proper storage of amines in the nervous system is important for the regulation of neurotransmission. We found 5 members of the *SLC18* gene family in octopus: 1 *VACHT* gene and 4 that show similarity to mammalian *SLC18A1* and *SLC18A2* (*VMAT* subfamily). As expected for genes that play a role in the packaging and release of amines in the nervous system, all 5 octopus *SLC18*s show enriched expression in nervous tissues (Figure S10.2.3).

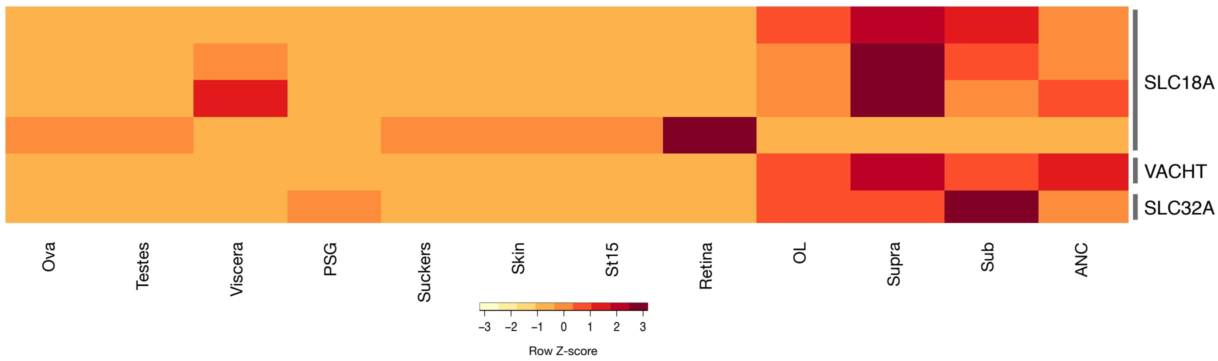


Figure S10.2.3. Expression profile of SLC18 and SLC32 gene families in 12 octopus tissues.

There is only one member of the *SLC32* gene family in mammals, *SLC32A1*, which encodes the vesicular inhibitory amino acid transporter (Schioth et al., 2013). This gene encodes the only known protein capable of transporting GABA and glycine into neurotransmitter vesicles. Octopus also has only one gene in the *SLC32* family, which shows strongest expression in the subesophageal brain (Figure S10.2.3). Overall, the numbers of vesicular transporter genes in octopus are comparable to the numbers in mammals, with the exception of the dramatic expansion of the sialin-like genes in the *SLC17* gene family.

10.3 SNAREs

The release of vesicles is essential to the function of all cells. In neurons, vesicular exocytosis enables communication with other cells through the secretion of neurotransmitters. This process depends on high calcium concentration in the synaptic terminal and the coordination of many proteins, called Soluble NSF Attachment Protein REceptors (SNAREs), to dock neurotransmitter vesicles to the presynaptic membrane. SNAREs have traditionally been categorized as t-SNAREs (associated with the target, or presynaptic membrane) or v-SNAREs (associated with the synaptic vesicle). The formation of the SNARE complex is initiated by the activity of synaptotagmin, a putative calcium-sensing protein. Calcium-bound synaptotagmin has a higher affinity for syntaxin, a t-SNARE. Syntaxin, SNAP-25 (t-SNARE), and synaptobrevin (v-SNARE) form the core of the SNARE complex, which tethers vesicles to the membrane, allowing for fusion of the membranes and diffusion of cargo into the synaptic cleft. SNARE machinery is conserved among many cell types to facilitate exocytosis of different types of cargo (Burgoyne and Morgan, 2003). We examined SNARE proteins, as well as synaptotagmin, in octopus. We found 13 synaptotagmin genes, 10 syntaxin genes, 1 SNAP-25 gene, and 4 synaptobrevin genes in the octopus genome. These numbers closely resemble the human complement of 17 synaptotagmin genes, 10 syntaxin genes, 1 SNAP-25 gene, and 7 synaptobrevin genes. Octopus SNAREs showed their highest expression in the supra- and subesophageal brains, retina, optic lobes, and ANC. A few SNAREs were enhanced in peripheral tissues (data not shown).

10.4 Channel and receptor subfamilies

Ligand-gated ion channels (LGICs) mediate chemical cell-to-cell signaling and determine the postsynaptic effects of ligand binding. The necessity of these channels for complex neural signaling is thought to have driven their evolution in early multicellular organisms. LGICs vary in voltage sensitivity, ion permeability, activation time, and response duration. The wide diversity of LGIC isoforms underlies their great variation in function.

The number of glutamate receptor subunits identified within the *O. bimaculoides* genome is reported in Table 2 and Figure S10.4.1. While there is expansion of the kainate- and AMPA-type glutamate receptors, this is not unusual for invertebrates. Expression levels of these subunits and those of NMDA receptors were highest in transcriptomes of neural tissues (OL, Supra, Sub, ANC). Expression of four of the AMPA-like receptor subunits in the sucker and gonad transcriptomes warrants further investigation.

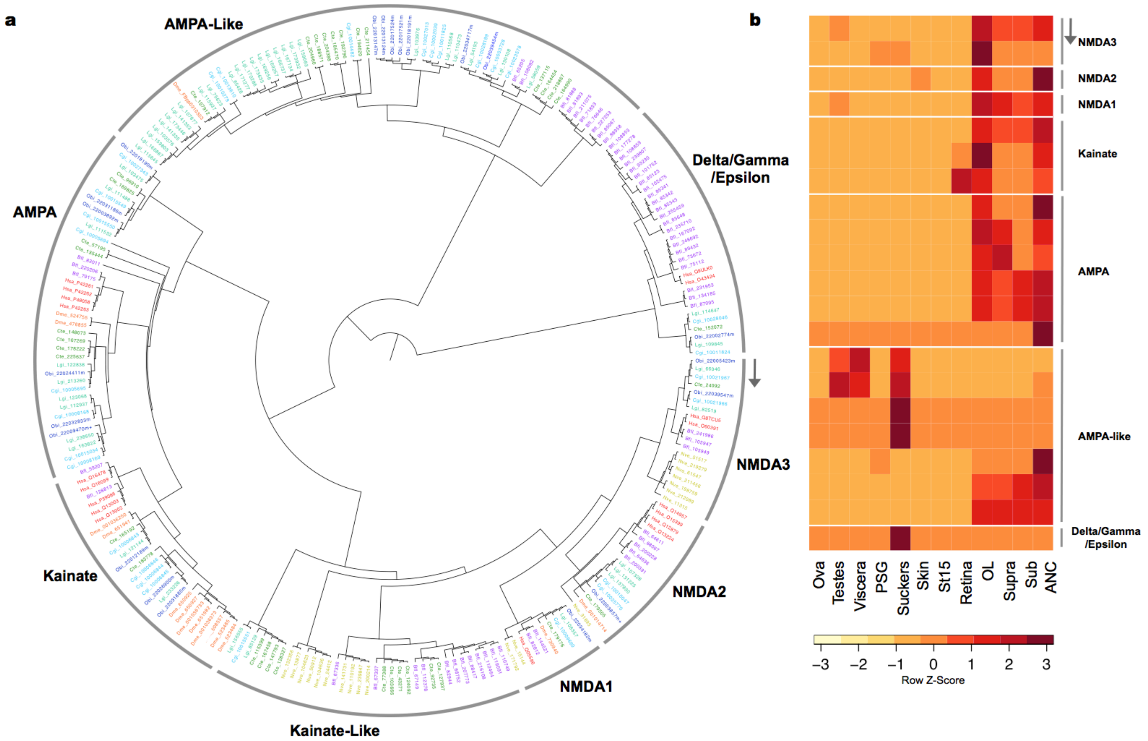


Figure S10.4.1. Glutamate receptor subunits in *O. bimaculoides*. **a**, Phylogenetic tree of glutamate receptor subunit genes identified across multiple taxa. Putative identities based on sequence homology are indicated along the arcs surrounding the tree. **b**, Heatmap of the expression profile of glutamate receptor subunits across 12 transcriptomes. Genes are shown in the order in which they appear in the tree, starting at the gray arrow indicated in **a** and continuing clockwise.

Vertebrates possess receptors that are gated only by glycine, but such receptors have not been previously reported in invertebrates (Dent, 2010). Here, we report the presence of receptors in *O. bimaculoides* (Figure S10.4.2) and other invertebrate taxa that show similarity to vertebrate glycine receptors (Table S10.4.1). These invertebrate subunits lack the amino acid residues identified as critical to glycine binding in vertebrates (Pless et al., 2008), so further characterization of their function is needed.

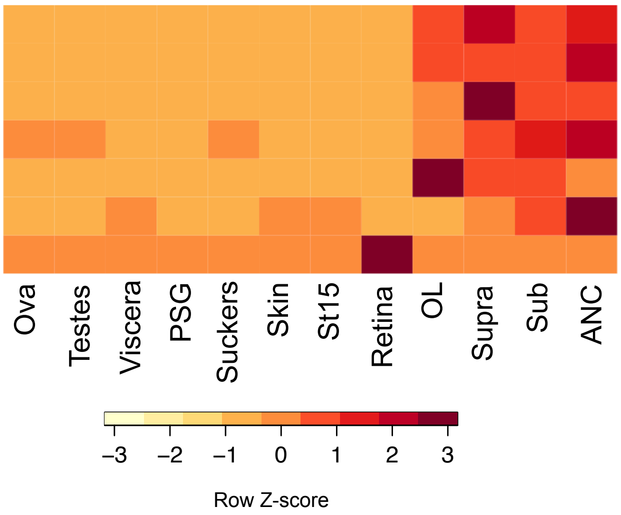


Figure S10.4.2. Glycine receptor-like subunit expression in *O. bimaculoides*.

	Obi	Aca	Lgi	Cte	Dme	Cel	Hsa
Glycine Receptor-Like	7	6	3	16	1	0	4
Voltage Gated Hydrogen Channels	3	5	5	2	0	0	1
Voltage Gated Chloride Channels	9	4	8	11	3	6	9
Cyclic Nucleotide Gated Potassium Channels, Hyperpolarization Activated (HCN)	2	1	1	3	1	0	4
Calcium Activated Chloride Channels	9	7	7	9	5	29	14
Intracellular Chloride Channels	1	3	2	2	1	2	6
Cyclic Nucleotide Gated Potassium Channels, Non-Voltage Gated (CNG)	8	4	9	4	4	6	6
Calcium, 2 Pore	3	4	3	3	0	0	2
Sodium, Leak/Non-selective	2	1	2	1	1	2	1
Acid Sensing Ion Channels	4	9	29	68	24	22	4
ATP P2X Receptors	1	2	3	1	0	0	7
Aquaporins	13	10	14	19	6	8	13
Zinc Activated Channels	0	0	1	0	0	1	1

Table S10.4.1. Counts of identified genes coding for selected channel and receptor subfamilies.

10.5 Innexins

Gap junctions are intercellular connections mediated by homo- or heteromeric protein hemichannel complexes. Open channel complexes provide cytoplasmic continuity between cells, allowing two cells to become chemically and electrically coupled. This mode of intercellular communication is used in many physiological processes, including the formation electrical synapses in the central nervous system. In invertebrates, gap junctions are formed by innexin proteins. A previous report identified independent expansions of the innexin family in *Lottia*, *Capitella*, and *Helobdella* (Simakov et al., 2013). We found 8 innexin genes in the octopus genome. All 8 octopus innexins show greatest similarity to *Lottia* innexins (Figure S10.5.1). Our phylogenetic analysis indicates that annelid-specific and molluscan-specific innexin subtypes may exist. Innexins are widely and differentially expressed across our transcriptomes, including tissues known to have gap junctions (the central nervous system, the viscera, and the ova; Figure S10.5.2). Some innexins are also enriched in the suckers.

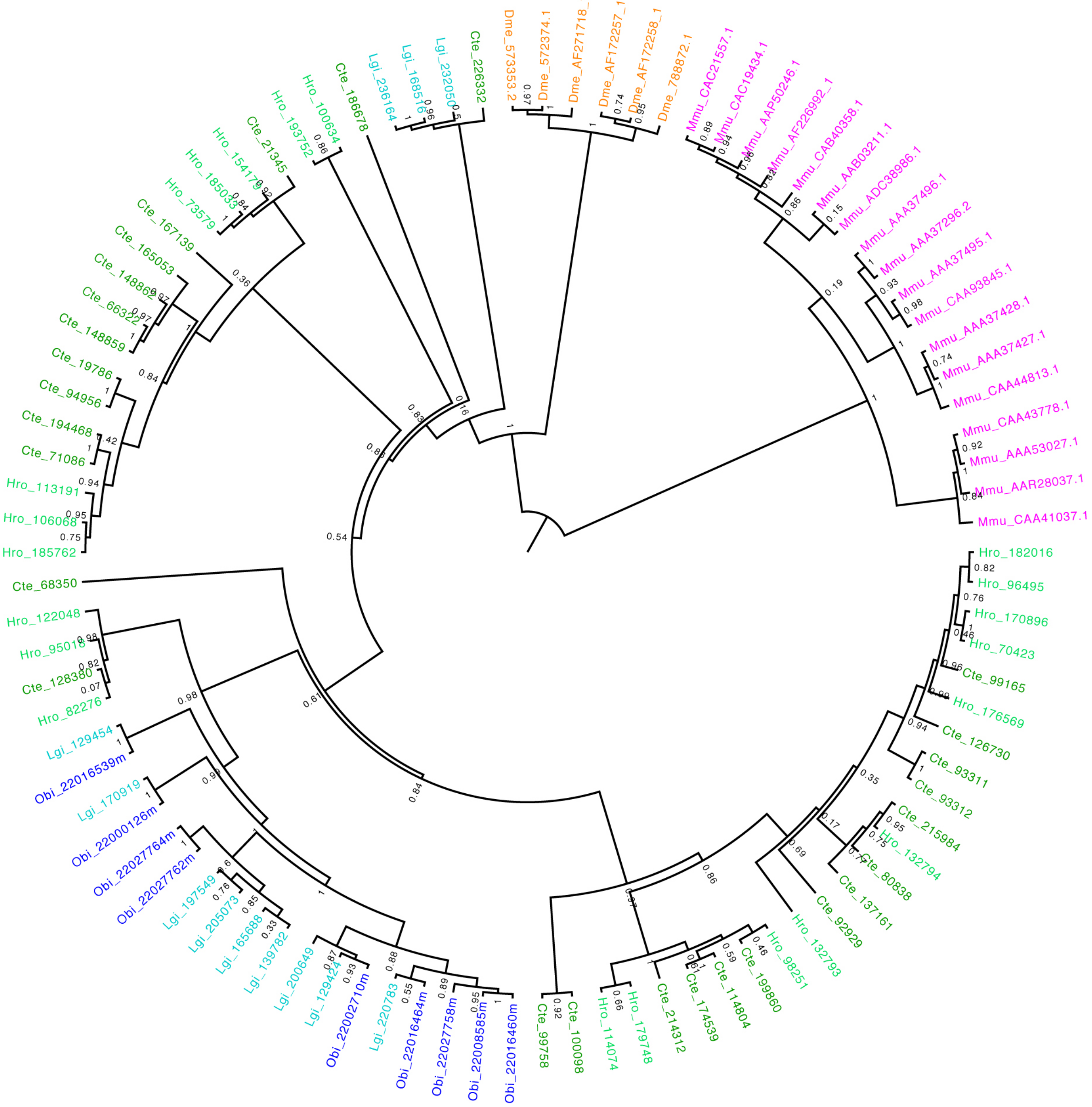


Figure S10.5.1. Innexin phylogenetic tree. Hro: *Helobdella robusta*. Mouse connexin sequences used as outgroup.

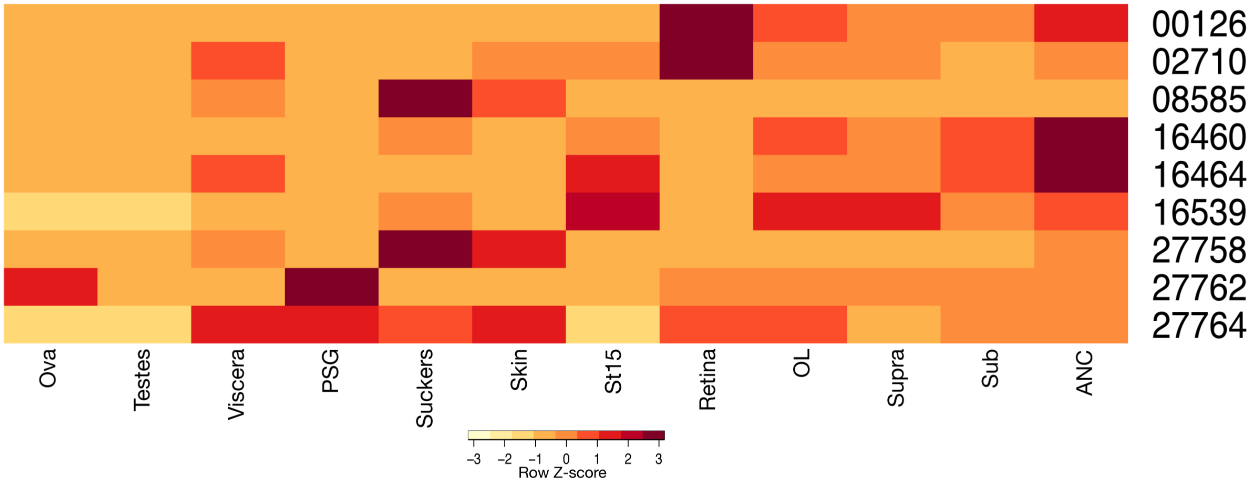


Figure S10.5.2. Expression profiles of 8 innexin genes in 12 octopus transcriptomes.

10.6 Discs, Large (DLG) proteins

The postsynaptic terminal contains signaling complexes that play a crucial role in processing incoming information. These multiprotein complexes are composed of receptors, adhesion proteins, and signaling enzymes held together by scaffold proteins, such as the membrane-associated guanylate kinases (MAGUK) superfamily of proteins. Molecular diversity of signaling complexes can produce synapses with different specificities. One such complex, called MASC (MAGUK-associated signaling complex), uses the scaffold protein Discs, Large (DLG). Previously, only two invertebrate (*D. melanogaster*) DLG proteins have been described: one shows similarity to the vertebrate DLG1-4 cluster and one shows similarity to the vertebrate DLG5. Using the 5 human DLG protein sequences as bait, we searched the octopus genome and the sequenced genomes of other invertebrates. We found multiple *DLG* genes in several invertebrate genomes, with 2-4 *DLG* genes identified in the non-mammalian non-cephalopod genomes examined (Figure S10.6.1). We identified the greatest number of invertebrate *DLG* homologs (5) in *O. bimaculoides,* 2 of which are *DLG5-like* and 3 of which are similar to the *DLG1-4* cluster. Three of these genes (Obi_00352-3m, Obi_32639m, Obi_32640m) are broadly expressed in octopus central nervous system tissues and the other two (Obi_25516m, Obi_25518m) are particularly enriched in each of two octopus specializations: the suckers and the ANC (Figure S10.6.2).

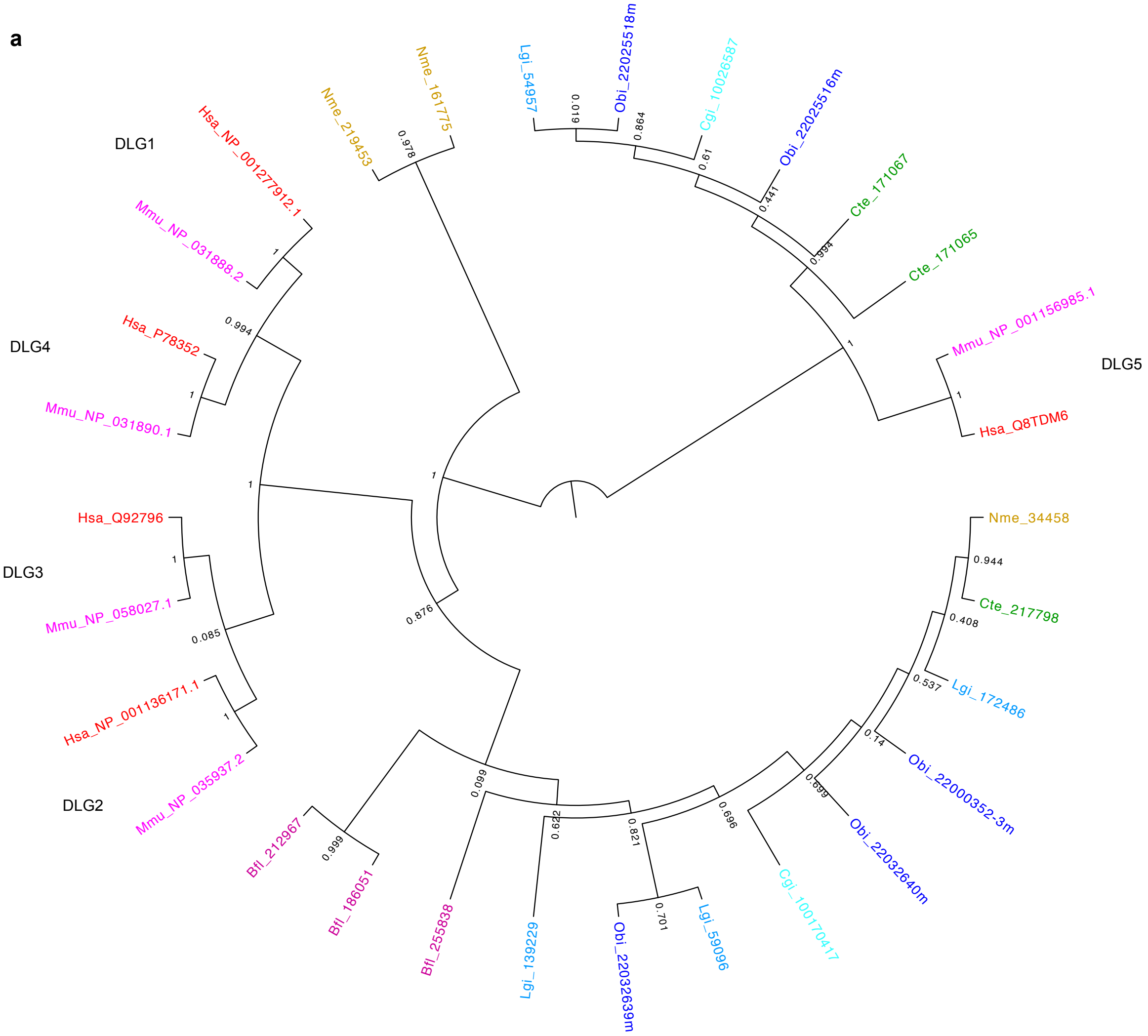


Figure S10.6.1. Phylogenetic tree of *DLG* proteins.

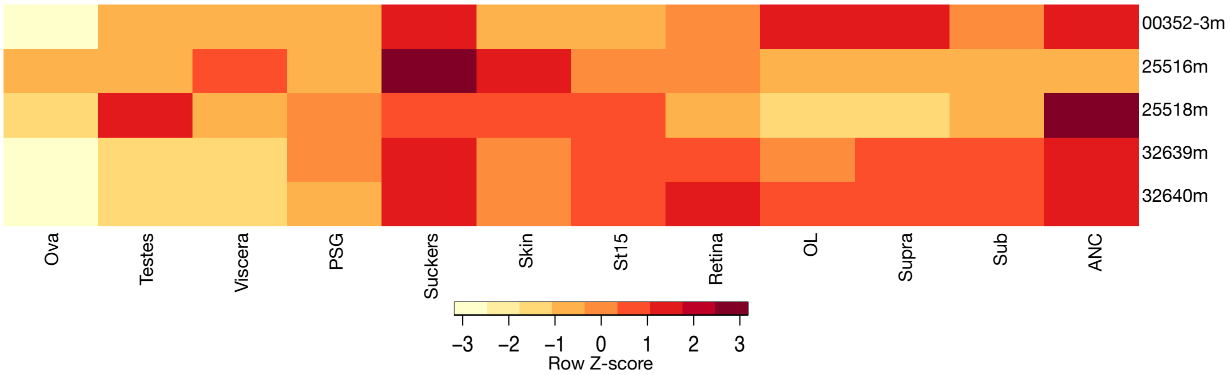


Figure S10.6.2. Expression profiles of 5 *DLG* genes in 12 octopus transcriptomes.

11. CEPHALOPOD NOVELTIES

11.1 Cephalopod-specific gene identification

Taxonomically restricted genes can be important in the evolution of lineage-specific traits (Johnson and Tsutsui, 2011; Tautz and Domazet-Loso, 2011). To identify cephalopod-specific genes, we conducted extensive BLASTP searches of the octopus gene set against the NR database (Pruitt et al., 2005), isolating protein sequences that have a hit to a cephalopod sequence (e-value better than 1E-5) but do not have a hit to any other organism at an e-value cutoff of 1E-3. As there are relatively few published cephalopod sequences available, we also conducted TBLASTX searches against *de novo* transcriptome and genome assemblies from *D. pealeii* (Brown et al., 2014) and deposited ESTs for *E. scolopes* (Chun et al., 2006) and *S. officinalis* (Bassaglia et al., 2012). To ensure that we had as close to full-length sequence as possible, we extended proteins predicted from octopus genomic sequence with our *de novo* assembled transcriptomes, using the longest match for each putative cephalopod-specific gene to query NR, transcriptome and EST sequences. Together, this process identified 1,811 candidate cephalopod novelties. To characterize these genes further, alignments were constructed with MUSCLE (Edgar, 2004) using *O. bimaculoides* sequences and their hits in other cephalopods (at least two sequences per alignment), clustering 585 octopus genes in 558 cephalopod gene families.

We performed further searches for any similarity between members of these gene families and sequences in GenBank NR, Pfam-A, and Pfam-B (Finn et al., 2014). We additionally performed BLASTP searches against a collection of transcriptomes from non-cephalopod molluscs (Table S11.1). By removing gene families with hits to NR or to these molluscan transcriptomes, we have limited our set of cephalopod novel genes to 174 gene families, including 180 octopus sequences. We also used multiple sequence alignments to build Hidden Markov Models (HMMs) using the hmmbuild tool from hmmer3 (Finn et al., 2011), and used these HMMs to search for novel domains. We searched for any similarity in these cephalopod HMMs to sequences in UNIREF90 (Suzek et al., 2007) using hmmsearch (Eddy, 2011). This extensive filtering identified 126 gene families with members in octopus and at least one other cephalopod that have no hits to non-cephalopod sequences in NR, Pfam or UNIREF even using sensitive HMM screens. A number of these genes are predominantly expressed in tissues associated with cephalopod innovations, including in the retina (6 gene families), suckers (4 gene families), and in neural tissues (Extended Data Figure 10).

Species Name	**Reference**
Antalis entalis	Smith et al. (2011)
Aplysia californica	Moroz et al. (2006)
Cadulus tolmeiei	Smith et al. (2011)
Chaetopleura apiculata	Smith et al. (2011)
Crepidula fornicata	Sadamoto et al. (2012)
Laevipilina hyalina	Smith et al. (2011)
Lingula anatina	Smith et al. (2011)
Littorina littorea	Smith et al. (2011)
Lymnea stagnalis	Henry et al. (2010)
Neomenia megatrapezata	Smith et al. (2011)
"Neomeniomorph"	Smith et al. (2011)
Nucula expansa	Smith et al. (2011)
Perotrochus lucaya	Smith et al. (2011)
Siphonaria pectinata	Smith et al. (2011)
Yoldia limatula	Smith et al. (2011)

Table S11.1. Non-cephalopod mollusc transcriptomes used to identify candidate cephalopod-specific genes.

Using our HMM-based approach, we also identified 48 gene families that have weak hits to UNIREF90 sequences previously known to be cephalopod-specific genes, including reflectins and the visual GTP-binding protein gamma subunit NP2. We also found a retina-specific gene (Ocbimv22013664m) with a weak similarity to a PDZ domain and a TTD non-photosensitive 1 protein-like from the elephant fish *Callorhinchus milii* (UniRef90 V9LJX1) in a broadly expressed gene family. Those distant similarities suggest the possible origins of some cephalopod "novel" genes as highly divergent members of older gene families, and hint at potential function of some of the cephalopod-specific genes.

11.2 Reflectins

A hallmark of coleoid cephalopod species is their ability to alter skin color and reflectance rapidly and reversibly. Though reflective tissues are common throughout the animal kingdom, cephalopods are unique in having proteinaceous platelet structures called iridophores. Squid iridophores are composed of a family of proteins called reflectins, which allow for dynamic tuning of iridescence (Crookes et al., 2004). We found six reflectin genes in the *O. bimaculoides* genome. Five of these genes are clustered along one scaffold (Scaffold 57337, Figure S11.2.1); an additional gene resides on Scaffold 210828 (data not shown).

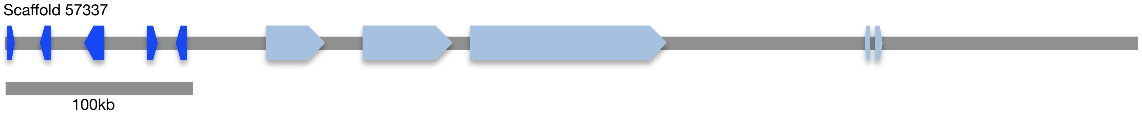


Figure S11.2.1. Scaffold 57337 contains 5 reflectins, shown in dark blue. Other genes on this scaffold are colored in light blue.

Two distinguishing features of the reflectin proteins are the presence of repeating domains and specific tissue deposition. The reflectin domain, initially characterized by Crookes et al. in 2004, appears 4-6 times in *S. officinalis/E. scolopes* reflectins and 1-7 times in the *O. bimaculoides* reflectins: $[M/F]DX_5MDX_5MDX_{3/4}$. The octopus reflectins expand this domain to $[M/F/Y]DX_5MDX_5M[N/D]X_{3/4,}$ and also contain the N-terminal conserved peptide characterized by Izumi et al. in 2010 (Figure S11.2.2).

```
>Ocbimv22030998m_Scaffold57339:45977-54246
MNRSRNMFRNSSRKHRGVMEPMTRMTMDFQGRYLDSSGRLVEPRCNDYYGRNSNYDRYRPMQNTGIYDNDKFQKYGRFMHFPERQMDMSGYQMDMRGRY
MDKYGRHCNPYSRRHMNYPNNNYDNYHMYNPEKLMDMSNFQMDMHGRWMDSNGRYSSPFSNYGSRHHQNYPHFNYNWGQRGFNYPDRFFDMSNYQMDLD
GKWMDTYGRHCHPFYDNSNYYGKQYNYNMYPHYNYNWGQKYYHYPERYFDMSNYQMDFDGRWMDMFGRSHSPFNGYNNNQGRQHHGQPHNSFSYGQRYQ
DRNFDIGNYQMDFDGRWMDMYGRYSHPFYGYNNFQSRYQHNLPQNFNWGQRSFHNPERLFDMGNYQMDFDGHWMDMDDRHCQPFTGNYNHSNRYQQNCN
HSPSQNFNWSQRYQDNPEKFFDMSGYQMEFDGRWMDSNNYNSDNFW
>Ocbimv22030996m_Scaffold57339:1379-1806
YYRRYMYCPYMNFRHMYHPERFFDMSHYQMDFNGHWMDMYNHDAHPFFGHNHYYRESNYYNYYPYSNYSWGHRFYNYPERYFDMSHYQMDFNGHWMNMY
DHGYHPFHSFSGYHGYHHSGYYPYHSYSQGRRYHNYYDMFYDMSHYQMDFDGNWMDMYNHYSHPFFGYDHYHRGSHYYNHYPYHNYSWGHRFYDYPERF
FDMSHYEMDFNGRWMNMHRF
>Ocbimv22031000m_Scaffold57339:101048-107169
MNRLMNKFRHHFGRKYRGIMEPMSVMSMDFQGRYMDSYGRMVDPRFYEFYGRYSDNDRYYGKSMYNYYGFYDNDRFHRYGNFMDFPERFMDMSSYQMDM
YGRWMDMHGHHSSPYWYMFNSSRHGHYPGYRYGRNWFYPERFMDMSHYQMDMYGRYMDRYGRQCNPYYNYYRRYMYYPYMNFYYMHYPERFMDMSGYQM
DMYGRWMDMYGRHSTPFYTNYGRYYHNYPYYNYSWGQRYYNYPERYFDMTNYQMDFDGRWMDMYSRHCTPFYSYHGRYHHYYPYHSYSWGQRFYNNPER
WYDMDYEFHSMSPYNYHSRYHYFNYSPYFSGGHRWFDMSNYQMDFGGQWMDMNGRYMNHFDHWNEYFF
>Ocbimv22030997m_Scaffold57339:15001-19853
MNRYMNRRNNFSRRYRGIMEPMSRMTMDFQGRYMDSYGRMVDPRFYGFYGRYSDNDRYYGRSMYNYYGFYDNDRFHRYGNFMDFPERFMDMSGYQMDMS
GRWMDMHGHYSSPHWHMFNSSRQGYYPGYHYGRNWFYPERFMDMSHYQMDMYGRYMDRNGRHCNPYYNYYRRYMYYPYMNFYHMYYPERFMDMSGYQMD
MYGRWMDMSGRHSSPFYSYHSRFHHNYPYYNYSWGQRYYNYPERYFDMGNYQMDFDGRWMDMYSRHCTPFYNYHGKFHHNYPYYNYSWGQRYYNYPERY
FDMGNYQMDFDGRWMDNYGRYSSPFNSYHGRFHHNYPYYNYSWGQRYYNYPERYFDMGNYQMDFDGRWMDNYGRYSSPFYNYHGRYHNYPYYNYSWGQR
YYNYPEGNYQMDFDGRWMDNYGRYFHGYNYHNRHYYNSYPNSYNYNWGQRYYDYPERNFDMFNYQMDFDSRWMDGQNFHYYGDNYNY
>Ocbimv22030999m_Scaffold57339:81227-85638
MNRFMNRFRPQFNRKYRGFMEPMNMMSMVFQGRYMDSYGKMVDPKLYEFYGKYSDNDRYYGKSMYNYYGFYDNDRFHRNGNFMDFPERFMDMSGYQMDM
NGKWMDTQGQNSHPYWNMFSSSRQGCYPGYSYGRNWFFPERFMDMSHYQMDMNGRYMDKSGRHCNPYYSYYRRYMSHPQMNFNQMHYPERFMDMSSYQM
DIGGRYMDKWGCHINPFSTYYFGKQSYFPHNYWSQRKYMDMSSYQMDMQYNSMDMNSRNCDQLHYFRNFDMWNNQMDFDGHWMNMNNQSYHPSSFIRNQ
VYYNPYHFYTWMSRYYNHPEKFYDTSNYQVEFGGKWPSQEYECITQE
>Ocbimv_skin_comp51140_c0_seq1_Scaffold210828
RKQKLRLFFSLSLLVSSMNRYMNRRNNFSRRYRGIMEPMSRMTMDFQGRYMDSYGRMVDPRFYGFYGRYSDNDRYYGRSMYNYYGFYDNDRFHRYGNFM
DFPERFMDMSGYQMDMSGRWMDMHGHYSSPHWHMFNSSRQGYYPGYHYGRNWFYPERFMDMSHYQMDMYGRYMDRNGRHCNPYYNYYNFYHMYYPERFM
DMSGYQMDMYGRWMDMSGRHSSPFYSYHSRFHHNYPYRRYMYYPYM
```

Figure S11.2.2. Octopus reflectin sequences. Canonical reflectin domain highlighted in yellow; N-terminal peptide highlighted in magenta.

The octopus reflectins are expressed in a highly tissue-specific manner; all octopus reflectin genes show extremely high levels of expression in the skin transcriptome (Figure S11.2.3). We also detected reflectins in the Stage 15 embryo, retina, viscera, and sucker transcriptomes. Reflectins are present in both dynamically (skin) and statically (viscera) reflective tissues of the octopus and are present in embryonic stages, as previously reported for *Sepia* reflectins (Bassaglia et al., 2012).

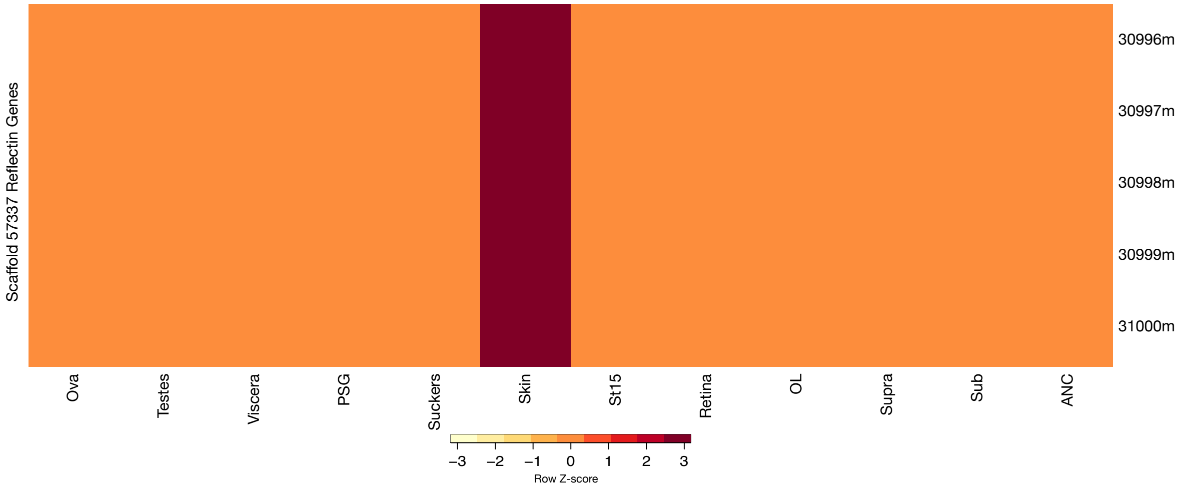


Figure S11.2.3. Expression profile of 5 reflectins in 12 octopus transcriptomes.

11.3 Octopus-specific gene identification

Many of the predicted proteins identified in the octopus genome without a match to other animal proteins also did not have a match to EST or transcriptome sequences from any other cephalopod with an e-value cutoff of 1e-5. The available cephalopod sequences that we searched are predominantly from the decapodiforms *E. scolopes* and *S. officinalis*, so many of these sequences could represent octopod-specific genes. We found 3,557 putative octopus-specific protein-coding genes expressed in the transcriptomes, 1,020 of which have a match to an *Octopus vulgaris* transcriptome (Smith et al., 2011). Of these candidate octopus-specific transcripts, 1,520 are expressed in a tissue-specific manner, which we defined as having more than 75% of the total expression, when normalized to transcriptome size, in a single tissue (Figure S11.3). Because many of these genes are expressed in the central nervous system (ANC, OL, supraesophageal and subesophageal brain), expression values for these four transcriptomes were combined for this analysis. Many octopus-specific genes are expressed primarily in the testes, which have been described as a site for orphan gene expression in other animals (Begun et al., 2007; Levine et al., 2006; Palmieri et al., 2014).

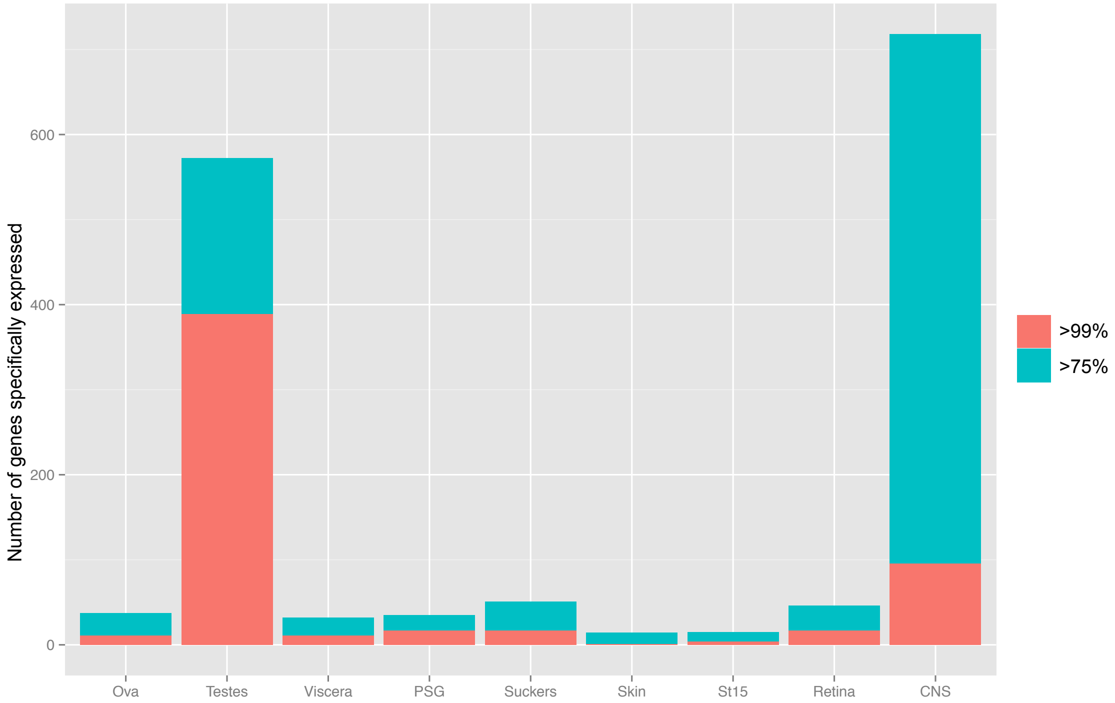


Figure S11.3. Tissue-specific expression of octopus-specific genes. Genes with >75% of total expression in a single tissue are represented. The subset with >99% expression in a tissue is shown in red. A large number of octopus-specific genes are found in the central nervous system and the testes.

12. REFERENCES

Adachi K, Ohnishi K, Kuramochi T, Yoshinaga T, Okumura S-I. 2014. Molecular cytogenetic study in Octopus (Amphioctopus) areolatus from Japan. Fisheries Science **80**:445-450.

Adrangi S, Faramarzi MA. 2013. From bacteria to human: a journey into the world of chitinases. Biotechnology advances **31**:1786-1795.

Akasaki T, Nikaido M, Nishihara H, Tsuchiya K, Segawa S, Okada N. 2010. Characterization of a novel SINE superfamily from invertebrates: "Ceph-SINEs" from the genomes of squids and cuttlefish. Gene **454**:8-19.

Alon S, Garrett SC, Levanon EY, Olson S, Graveley BR, Rosenthal JJ, Eisenberg E. 2015. The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. eLife **4**.

Amemiya CT, Alfoldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N, Organ C, Chalopin D, Smith JJ, Robinson M, Dorrington RA, Gerdol M, Aken B, Biscotti MA, Barucca M, Baurain D, Berlin AM, Blatch GL, Buonocore F, Burmester T, Campbell MS, Canapa A, Cannon JP, Christoffels A, De Moro G, Edkins AL, Fan L, Fausto AM, Feiner N, Forconi M, Gamieldien J, Gnerre S, Gnirke A, Goldstone JV, Haerty W, Hahn ME, Hesse U, Hoffmann S, Johnson J, Karchner SI, Kuraku S, Lara M, Levin JZ, Litman GW, Mauceli E, Miyake T, Mueller MG, Nelson DR, Nitsche A, Olmo E, Ota T, Pallavicini A, Panji S, Picone B, Ponting CP, Prohaska SJ, Przybylski D, Saha NR, Ravi V, Ribeiro FJ, Sauka-Spengler T, Scapigliati G, Searle SM, Sharpe T, Simakov O, Stadler PF, Stegeman JJ, Sumiyama K, Tabbaa D, Tafer H, Turner-Maier J, van Heusden P, White S, Williams L, Yandell M, Brinkmann H, Volff JN, Tabin CJ, Shubin N, Schartl M, Jaffe DB, Postlethwait JH, Venkatesh B, Di Palma F, Lander ES, Meyer A, Lindblad-Toh K. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature **496**:311-316.

Baggiolini M. 1998. Chemokines and leukocyte traffic. Nature **392**:565-568.

Bassaglia Y, Bekel T, Da Silva C, Poulain J, Andouche A, Navet S, Bonnaud L. 2012. ESTs library from embryonic stages reveals tubulin and reflectin diversity in Sepia officinalis (Mollusca — Cephalopoda). Gene **498**:203-211.

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics **176**:1131-1137.

Bolognesi R, Beermann A, Farzana L, Wittkopp N, Lutz R, Balavoine G, Brown SJ, Schroder R. 2008. Tribolium Wnts: evidence for a larger repertoire in insects with overlapping expression patterns that suggest multiple redundant functions in embryogenesis. Development genes and evolution **218**:193-202.

Bonaldo MF, Lennon G, Soares MB. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome research **6**:791-806.

Brown CT, Graveley B, Rosenthal JJ. 2014. *Loligo pealeii* (squid) data dump.

Burglin TR, Kuwabara PE. 2006. Homologs of the Hh signalling network in C. elegans. WormBook: the online review of C elegans biology:1-14.

Burgoyne RD, Morgan A. 2003. Secretory granule exocytosis. Physiological reviews **83**:581-632.

Callaerts P, Lee PN, Hartmann B, Farfan C, Choy DW, Ikeo K, Fischbach KF, Gehring WJ, de Couet HG. 2002. HOX genes in the sepiolid squid Euprymna scolopes: implications for the evolution of complex body plans. Proceedings of the National Academy of Sciences of the United States of America **99**:2088-2093.

Chang SH, Dong C. 2007. A novel heterodimeric cytokine consisting of IL-17 and IL-17F regulates inflammatory responses. Cell research **17**:435-440.

Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: de novo genome assembly with short paired-end reads. PloS one **6**:e23501.

Chapman JA, Kirkness EF, Simakov O, Hampson SE, Mitros T, Weinmaier T, Rattei T, Balasubramanian PG, Borman J, Busam D, Disbennett K, Pfannkoch C, Sumin N, Sutton GG, Viswanathan LD, Walenz B, Goodstein DM, Hellsten U, Kawashima T, Prochnik SE, Putnam NH, Shu S, Blumberg B, Dana CE, Gee L, Kibler DF, Law L, Lindgens D, Martinez DE, Peng J, Wigge PA, Bertulat B, Guder C, Nakamura Y, Ozbek S, Watanabe H, Khalturin K, Hemmrich G, Franke A, Augustin R, Fraune S, Hayakawa E, Hayakawa S, Hirose M, Hwang JS, Ikeo K, Nishimiya-Fujisawa C, Ogura A, Takahashi T, Steinmetz PR, Zhang X, Aufschnaiter R, Eder MK, Gorny AK, Salvenmoser W, Heimberg AM, Wheeler BM, Peterson KJ, Bottger A, Tischler P, Wolf A, Gojobori T, Remington KA, Strausberg RL, Venter JC, Technau U, Hobmayer B, Bosch TC, Holstein TW, Fujisawa T, Bode HR, David CN, Rokhsar DS, Steele RE. 2010. The dynamic genome of Hydra. Nature **464**:592-596.

Chen CX, Cho DS, Wang Q, Lai F, Carter KC, Nishikura K. 2000. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. RNA **6**:755-767.

Chen WV, Maniatis T. 2013. Clustered protocadherins. Development **140**:3297-3302.

Cho SJ, Valles Y, Giani VC, Jr., Seaver EC, Weisblat DA. 2010. Evolutionary dynamics of the wnt gene family: a lophotrochozoan perspective. Molecular biology and evolution **27**:1645-1658.

Chun CK, Scheetz TE, Bonaldo Mde F, Brown B, Clemens A, Crookes-Goodson WJ, Crouch K, DeMartini T, Eyestone M, Goodson MS, Janssens B, Kimbell JL, Koropatnick TA, Kucaba T, Smith C, Stewart JJ, Tong D, Troll JV, Webster S, Winhall-Rice J, Yap C, Casavant TL, McFall-Ngai MJ, Soares MB. 2006. An annotated cDNA library of juvenile Euprymna scolopes with and without colonization by the symbiont Vibrio fischeri. BMC genomics **7**:154.

Crookes WJ, Ding LL, Huang QL, Kimbell JR, Horwitz J, McFall-Ngai MJ. 2004. Reflectins: the unusual proteins of squid reflective tissues. Science **303**:235-238.

DeGiorgis JA, Cavaliere KR, Burbach JPH. 2011. Identification of Molecular Motors in the Woods Hole Squid, *Loligo pealei*: an EST Approach. Unpublished.

Dent JA. 2010. The evolution of pentameric ligand-gated ion channels. Advances in experimental medicine and biology **683**:11-23.

Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P. 2004. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science **304**:304-307.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS computational biology **7**:e1002195.

Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics **5**:113.

Eisenmann DM. 2005. Wnt signaling. WormBook : the online review of C elegans biology:1-17.

Fairclough SR, Chen Z, Kramer E, Zeng Q, Young S, Robertson HM, Begovic E, Richter DJ, Russ C, Westbrook MJ, Manning G, Lang BF, Haas B, Nusbaum C, King N. 2013. Premetazoan genome evolution and the regulation of cell differentiation in the choanoflagellate Salpingoeca rosetta. Genome biology **14**:R15.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic acids research **42**:D222-230.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic acids research **39**:W29-37.

Fitzgerald LW, Iyer G, Conklin DS, Krause CM, Marshall A, Patterson JP, Tran DP, Jonak GJ, Hartig PR. 1999. Messenger RNA editing of the human serotonin 5-HT2C receptor. Neuropsychopharmacology **21**:82S-90S.

Flot JF, Hespeels B, Li X, Noel B, Arkhipova I, Danchin EG, Hejnol A, Henrissat B, Koszul R, Aury JM, Barbe V, Barthelemy RM, Bast J, Bazykin GA, Chabrol O, Couloux A, Da Rocha M, Da Silva C, Gladyshev E, Gouret P, Hallatschek O, Hecox-Lea B, Labadie K, Lejeune B, Piskurek O, Poulain J, Rodriguez F, Ryan JF, Vakhrusheva OA, Wajnberg E, Wirth B, Yushenova I, Kellis M, Kondrashov AS, Mark Welch DB, Pontarotti P, Weissenbach J, Wincker P, Jaillon O, Van Doninck K. 2013. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga. Nature **500**:453-457.

Frank M, Kemler R. 2002. Protocadherins. Current opinion in cell biology **14**:557-562.

Freeman JL, Adeniyi A, Banerjee R, Dallaire S, Maguire SF, Chi J, Ng BL, Zepeda C, Scott CE, Humphray S, Rogers J, Zhou Y, Zon LI, Carter NP, Yang F, Lee C. 2007. Definition of the zebrafish genome using flow cytometry and cytogenetic mapping. BMC genomics **8**:195.

Garrett S, Rosenthal JJ. 2012. RNA editing underlies temperature adaptation in K+ channels from polar octopuses. Science **335**:848-851.

Gesualdi SC, Haerry TE. 2007. Distinct signaling of Drosophila Activin/TGF-beta family members. Fly **1**:212-221.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology **29**:644-652.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, Langton L, Perrimon N, Sandler JE, Wan KH, Willingham A, Zhang Y, Zou Y, Andrews J, Bickel PJ, Brenner SE, Brent MR, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Oliver B, Celniker SE. 2011. The developmental transcriptome of Drosophila melanogaster. Nature **471**:473-479.

Gumienny TL, Savage-Dunn C. 2013. TGF-beta signaling in C. elegans. WormBook: the online review of C elegans biology:1-34.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, Salzberg SL, White O. 2003. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research **31**:5654-5666.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols **8**:1494-1512.

Hallinan NM, Lindberg DR. 2011. Comparative analysis of chromosome counts infers three paleopolyploidies in the mollusca. Genome biology and evolution **3**:1150-1163.

Henry JJ, Perry KJ, Fukui L, Alvi N. 2010. Differential localization of mRNAs during early development in the mollusc, Crepidula fornicata. Integrative and comparative biology **50**:720-733.

Hiasa M, Miyaji T, Haruna Y, Takeuchi T, Harada Y, Moriyama S, Yamamoto A, Omote H, Moriyama Y. 2014. Identification of a mammalian vesicular polyamine transporter. Scientific reports **4**:6836.

Hooper JE, Scott MP. 2005. Communicating with Hedgehogs. Nature reviews Molecular cell biology **6**:306-317.

Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch GJ, White S, Chow W, Kilian B, Quintais LT, Guerra-Assuncao JA, Zhou Y, Gu Y, Yen J, Vogel JH, Eyre T, Redmond S, Banerjee R, Chi J, Fu B, Langley E, Maguire SF, Laird GK, Lloyd D, Kenyon E, Donaldson S, Sehra H, Almeida-King J, Loveland J, Trevanion S, Jones M, Quail M, Willey D, Hunt A, Burton J, Sims S, McLay K, Plumb B, Davis J, Clee C, Oliver K, Clark R, Riddle C, Elliot D, Threadgold G, Harden G, Ware D, Begum S, Mortimore B, Kerry G, Heath P, Phillimore B, Tracey A, Corby N, Dunn M, Johnson C, Wood J, Clark S, Pelan S, Griffiths G, Smith M, Glithero R, Howden P, Barker N, Lloyd C, Stevens C, Harley J, Holt K, Panagiotidis G, Lovell J, Beasley H, Henderson C, Gordon D, Auger K, Wright D, Collins J, Raisen C, Dyer L, Leung K, Robertson L, Ambridge K, Leongamornlert D, McGuire S, Gilderthorp R, Griffiths C, Manthravadi D, Nichol S, Barker G, Whitehead S, Kay M, Brown J, Murnane C, Gray E, Humphries M, Sycamore N, Barker D, Saunders D, Wallis J, Babbage A, Hammond S, Mashreghi-Mohammadi M, Barr L, Martin S, Wray P, Ellington A, Matthews N, Ellwood M, Woodmansey R, Clark G, Cooper J, Tromans A, Grafham D, Skuce C, Pandian R, Andrews R, Harrison E, Kimberley A, Garnett J, Fosker N, Hall R, Garner P, Kelly D, Bird C, Palmer S, Gehring I, Berger A, Dooley CM, Ersan-Urun Z, Eser C, Geiger H, Geisler M, Karotki L, Kirn A, Konantz J, Konantz M, Oberlander M, Rudolph-Geiger S, Teucke M, Lanz C, Raddatz G, Osoegawa K, Zhu B, Rapp A, Widaa S, Langford C, Yang F, Schuster SC, Carter NP, Harrow J, Ning Z, Herrero J, Searle SM, Enright A, Geisler R, Plasterk RH, Lee C, Westerfield M, de Jong PJ, Zon LI, Postlethwait JH, Nusslein-Volhard C, Hubbard TJ, Roest Crollius H, Rogers J, Stemple DL. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature **496**:498-503.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

Hulpiau P, van Roy F. 2009. Molecular evolution of the cadherin superfamily. The international journal of biochemistry & cell biology **41**:349-369.

Hulpiau P, van Roy F. 2011. New insights into the evolution of metazoan cadherins. Molecular biology and evolution **28**:647-657.

Iijima M. 2006. Evolution of Hox genes in molluscs: a comparison among seven morphologically diverse classes. Journal of Molluscan Studies **72**:259-266.

Iuchi S. 2001. Three classes of C2H2 zinc finger proteins. Cellular and molecular life sciences : CMLS **58**:625-635.

Izumi M, Sweeney AM, DeMartini D, Weaver JC, Powers ML, Tao A, Silvas TV, Kramer RM, Crookes-Goodson WJ, Mathger LM, Naik RR, Hanlon RT, Morse DE. 2010.

Changes in reflectin protein phosphorylation are associated with dynamic iridescence in squid. Journal of the Royal Society Interface **7**:549-560.

Johnson BR, Tsutsui ND. 2011. Taxonomically restricted genes are associated with the evolution of sociality in the honey bee. BMC genomics **12**:164.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and genome research **110**:462-467.

Kang D. 2003. A hedgehog homolog regulates gut formation in leech (Helobdella). Development **130**:1645-1657.

Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**:617-624.

King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I, Marr M, Pincus D, Putnam N, Rokas A, Wright KJ, Zuzow R, Dirks W, Good M, Goodstein D, Lemons D, Li W, Lyons JB, Morris A, Nichols S, Richter DJ, Salamov A, Sequencing JG, Bork P, Lim WA, Manning G, Miller WT, McGinnis W, Shapiro H, Tjian R, Grigoriev IV, Rokhsar D. 2008. The genome of the choanoflagellate Monosiga brevicollis and the origin of metazoans. Nature **451**:783-788.

Klug A. 2010. The discovery of zinc fingers and their development for practical applications in gene regulation and genome manipulation. Quarterly reviews of biophysics **43**:1-21.

Kocot KM, Cannon JT, Todt C, Citarella MR, Kohn AB, Meyer A, Santos SR, Schander C, Moroz LL, Lieb B, Halanych KM. 2011. Phylogenomics reveals deep molluscan relationships. Nature **477**:452-456.

Kolodkin AL, Tessier-Lavigne M. 2011. Mechanisms and molecules of neuronal wiring: a primer. Cold Spring Harbor perspectives in biology 3:a001727

Kroger B, Vinther J, Fuchs D. 2011. Cephalopod origin and evolution: A congruent picture emerging from fossils, development and molecules: Extant cephalopods are younger than previously realised and were under major selection to become agile, shell-less predators. BioEssays : news and reviews in molecular, cellular and developmental biology **33**:602-613.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: an information aesthetic for comparative genomics. Genome research **19**:1639-1645.

Kusserow A, Pang K, Sturm C, Hrouda M, Lentfer J, Schmidt HA, Technau U, von Haeseler A, Hobmayer B, Martindale MQ, Holstein TW. 2005. Unexpected complexity of the Wnt gene family in a sea anemone. Nature **433**:156-160.

Layden MJ, Meyer NP, Pang K, Seaver EC, Martindale MQ. 2010. Expression and phylogenetic analysis of the zic gene family in the evolution and development of metazoans. EvoDevo **1**:12.

Lee JJ, von Kessler DP, Parks S, Beachy PA. 1992. Secretion and localized transcription suggest a role in positional signaling for products of the segmentation gene hedgehog. Cell **71**:33-50.

Lee PN, Callaerts P, De Couet HG, Martindale MQ. 2003. Cephalopod Hox genes and the origin of morphological novelties. Nature **424**:1061-1065.

Lefebvre JL, Kostadinov D, Chen WV, Maniatis T, Sanes JR. 2012. Protocadherins mediate dendritic self-avoidance in the mammalian nervous system. Nature **488**:517-521.

Leveugle M, Prat K, Perrier N, Birnbaum D, Coulier F. 2003. ParaDB: a tool for paralogy mapping in vertebrate genomes. Nucleic acids research **31**:63-67.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. Proceedings of the National Academy of Sciences of the United States of America **103**:9935-9939.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079.

Li J, Zhang Y, Zhang Y, Xiang Z, Tong Y, Qu F, Yu Z. 2014. Genomic characterization and expression analysis of five novel IL-17 genes in the Pacific oyster, Crassostrea gigas. Fish & shellfish immunology **40**:455-465.

Liu H, Chang LH, Sun Y, Lu X, Stubbs L. 2014. Deep vertebrate roots for mammalian zinc finger transcription factor subfamilies. Genome biology and evolution **6**:510-525.

McGlinn E, Tabin CJ. 2006. Mechanistic insight into how Shh patterns the vertebrate limb. Current opinion in genetics & development **16**:426-432.

Miyaji T, Echigo N, Hiasa M, Senoh S, Omote H, Moriyama Y. 2008. Identification of a vesicular aspartate transporter. Proceedings of the National Academy of Sciences of the United States of America **105**:11720-11724.

Morishita H, Yagi T. 2007. Protocadherin family: diversity, structure, and function. Current opinion in cell biology **19**:584-592.

Moroz LL, Edwards JR, Puthanveettil SV, Kohn AB, Ha T, Heyland A, Knudsen B, Sahni A, Yu F, Liu L, Jezzini S, Lovell P, Iannucculli W, Chen M, Nguyen T, Sheng H, Shaw R, Kalachikov S, Panchin YV, Farmerie W, Russo JJ, Ju J, Kandel ER. 2006. Neuronal transcriptome of Aplysia: neuronal compartments and circuitry. Cell **127**:1453-1467.

Moroz LL, Kocot KM, Citarella MR, Dosung S, Norekian TP, Povolotskaya IS, Grigorenko AP, Dailey C, Berezikov E, Buckley KM, Ptitsyn A, Reshetov D, Mukherjee K, Moroz TP, Bobkova Y, Yu F, Kapitonov VV, Jurka J, Bobkov YV, Swore JJ, Girardo DO, Fodor A, Gusev F, Sanford R, Bruders R, Kittler E, Mills CE, Rast JP, Derelle R, Solovyev VV, Kondrashov FA, Swalla BJ, Sweedler JV, Rogaev EI, Halanych KM, Kohn AB. 2014. The ctenophore genome and the evolutionary origins of neural systems. Nature **510**:109-114.

Naef, A. 1928. Die Cephalopoden. Embryologie. *Fauna Flora Golf. Neapel* **35**, 1–357. (English translation available; von Boletzky, S. *The cephalopoda-embryology.* Smithsonian Institution Press, Washington DC, 2001).

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. Molecular biology and evolution **20**:1231-1239.

Noonan JP, Grimwood J, Schmutz J, Dickson M, Myers RM. 2004. Gene conversion and the evolution of protocadherin gene cluster diversity. Genome research **14**:354-366.

Nowick K, Hamilton AT, Zhang H, Stubbs L. 2010. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. Molecular biology and evolution **27**:2606-2617.

Ohshima K, Okada N. 1994. Generality of the tRNA origin of short interspersed repetitive elements (SINEs). Characterization of three different tRNA-derived retroposons in the octopus. Journal of molecular biology **243**:25-37.

Palavicini JP, O'Connell MA, Rosenthal JJ. 2009. An extra double-stranded RNA binding domain confers high activity to a squid RNA editing enzyme. Rna **15**:1208-1218.

Palmieri N, Kosiol C, Schlotterer C. 2014. The life cycle of Drosophila orphan genes. eLife **3**:e01311.

Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. Nucleic acids research **37**:289-297.

Pernice M, Deutsch JS, Andouche A, Boucher-Rodoni R, Bonnaud L. 2006. Unexpected variation of Hox genes' homeodomains in cephalopods. Molecular phylogenetics and evolution **40**:872-879.

Pless SA, Millen KS, Hanek AP, Lynch JW, Lester HA, Lummis SC, Dougherty DA. 2008. A cation-pi interaction in the binding site of the glycine receptor is mediated by a phenylalanine residue. Journal of neuroscience **28**:10937-10942.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics **21 Suppl 1**:i351-358.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PloS one **5**:e9490.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research **33**:D501-504.

Putnam NH, Srivastava M, Hellsten U, Dirks B, Chapman J, Salamov A, Terry A, Shapiro H, Lindquist E, Kapitonov VV, Jurka J, Genikhovich G, Grigoriev IV, Lucas SM, Steele RE, Finnerty JR, Technau U, Martindale MQ, Rokhsar DS. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317**:86-94.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics **26**:841-842.

Ramaswami G, Li JB. 2014. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic acids research **42**:D109-113.

Reimer RJ. 2013. SLC17: a functionally diverse family of organic anion transporters. Molecular aspects of medicine **34**:350-359.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends in genetics **16**:276-277.

Rosenthal JJ, Seeburg PH. 2012. A-to-I RNA editing: effects on proteins key to neural excitability. Neuron **74**:432-439.

Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Lewis S. 2000. Comparative genomics of the eukaryotes. Science **287**:2204-2215.

Ryan JF, Pang K, Schnitzler CE, Nguyen AD, Moreland RT, Simmons DK, Koch BJ, Francis WR, Havlak P, Program NCS, Smith SA, Putnam NH, Haddock SH, Dunn CW, Wolfsberg TG, Mullikin JC, Martindale MQ, Baxevanis AD. 2013. The genome of the ctenophore Mnemiopsis leidyi and its implications for cell type evolution. Science **342**:1242592.

Sadamoto H, Takahashi H, Okada T, Kenmoku H, Toyota M, Asakawa Y. 2012. De novo sequencing and transcriptome analysis of the central nervous system of mollusc Lymnaea stagnalis by deep RNA sequencing. PloS one **7**:e42546.

Salamov AA, Solovyev VV. 2000. Ab initio gene finding in Drosophila genomic DNA. Genome research **10**:516-522.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics **19**:301-302.

Savva YA, Jepson JE, Chang YJ, Whitaker R, Jones BC, St Laurent G, Tackett MR, Kapranov P, Jiang N, Du G, Helfand SL, Reenan RA. 2013. RNA editing regulates transposon-mediated heterochromatic gene silencing. Nature communications **4**:2745.

Savva YA, Rieder LE, Reenan RA. 2012. The ADAR protein family. Genome biology **13**:252.

Schioth HB, Roshanbin S, Hagglund MG, Fredriksson R. 2013. Evolutionary origin of amino acid transporter families SLC32, SLC36 and SLC38 and physiological, pathological and therapeutic aspects. Molecular aspects of medicine **34**:571-585.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18**:502-504.

Schreiner D, Weiner JA. 2010. Combinatorial homophilic interaction between gamma-protocadherin multimers greatly expands the molecular diversity of cell adhesion. Proceedings of the National Academy of Sciences of the United States of America **107**:14893-14898.

Segond von Banchet G, Boettger MK, Konig C, Iwakura Y, Brauer R, Schaible HG. 2013. Neuronal IL-17 receptor upregulates TRPV4 but not TRPV1 receptors in DRG neurons and mediates mechanical but not thermal hyperalgesia. Molecular and cellular neurosciences **52**:152-160.

Shabgah AG, Fattahi E, Shahneh FZ. 2014. Interleukin-17 in human inflammatory diseases. Postepy dermatologii i alergologii **31**:256-261.

Shannon M, Hamilton AT, Gordon L, Branscomb E, Stubbs L. 2003. Differential expansion of zinc-finger transcription factor loci in homologous human and mouse gene clusters. Genome research **13**:1097-1110.

Shigeno S, Nishimura O, Tarui H, Agata K. 2006. Molecular dissection of squid brains unpublished.

Shinzato C, Shoguchi E, Kawashima T, Hamada M, Hisata K, Tanaka M, Fujie M, Fujiwara M, Koyanagi R, Ikuta T, Fujiyama A, Miller DJ, Satoh N. 2011. Using the Acropora digitifera genome to understand coral responses to environmental change. Nature **476**:320-323.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology **7**:539.

Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, Savage R, Osoegawa K, de Jong P, Grimwood J, Chapman JA, Shapiro H, Aerts A, Otillar RP, Terry AY, Boore JL, Grigoriev IV, Lindberg DR, Seaver EC, Weisblat DA, Putnam NH, Rokhsar DS. 2013. Insights into bilaterian evolution from three spiralian genomes. Nature **493**:526-531.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC bioinformatics **6**:31.

Smit A, Hubley R. 2008-2010. RepeatModeler Open-1.0.

Smit A, Hubley R, Green P. 1996-2010. RepeatMasker Open-3.0.

Smith SA, Wilson NG, Goetz FE, Feehery C, Andrade SC, Rouse GW, Giribet G, Dunn CW. 2011. Resolving the evolutionary relationships of molluscs with phylogenomic tools. Nature **480**:364-367.

St Laurent G, Tackett MR, Nechkin S, Shtokalo D, Antonets D, Savva YA, Maloney R, Kapranov P, Lawrence CE, Reenan RA. 2013. Genome-wide analysis of A-to-I

RNA editing by single-molecule sequencing in Drosophila. Nature structural & molecular biology **20**:1333-1339.

Strugnell J, Norman M, Jackson J, Drummond AJ, Cooper A. 2005. Molecular phylogeny of coleoid cephalopods (Mollusca: Cephalopoda) using a multigene approach; the effect of data partitioning on resolving phylogenies in a Bayesian framework. Molecular phylogenetics and evolution **37**:426-441.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. 2007. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23**:1282-1288.

Tabata T, Kornberg TB. 1994. Hedgehog is a signaling protein with a key role in patterning Drosophila imaginal discs. Cell **76**:89-102.

Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, Shoguchi E, Fujiwara M, Shinzato C, Hisata K, Fujie M, Usami T, Nagai K, Maeyama K, Okamoto K, Aoki H, Ishikawa T, Masaoka T, Fujiwara A, Endo K, Endo H, Nagasawa H, Kinoshita S, Asakawa S, Watabe S, Satoh N. 2012. Draft genome of the pearl oyster Pinctada fucata: a platform for understanding bivalve biology. DNA research **19**:117-130.

Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic biology **56**:564-577.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nature reviews Genetics **12**:692-702.

Team RC. 2014. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. 2003. PANTHER: a library of protein families and subfamilies indexed by function. Genome research **13**:2129-2141.

Trapnell C, Pachter L, Salzberg SL. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**:1105-1111.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols **7**:562-578.

Ulbricht RJ, Emeson RB. 2014. One hundred million adenosine-to-inosine RNA editing sites: hearing through the noise. BioEssays **36**:730-735.

Vanhalst K, Kools P, Staes K, van Roy F, Redies C. 2005. delta-Protocadherins: a gene family expressed differentially in the mouse brain. Cellular and molecular life sciences **62**:1247-1259.

Vinogradov AE. 1998. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. Cytometry **31**:100-109.

Vinogradov AE. 2013. Density peaks of paralog pairs in human and mouse genomes. Gene **527**:55-61.

Wang IX, So E, Devlin JL, Zhao Y, Wu M, Cheung VG. 2013. ADAR regulates RNA editing, transcript stability, and gene expression. Cell reports **5**:849-860.

Werry TD, Loiacono R, Sexton PM, Christopoulos A. 2008. RNA editing of the serotonin 5HT2C receptor and its effects on cell signalling, pharmacology and brain function. Pharmacology & therapeutics **119**:7-23.

Wharton K, Derynck R. 2009. TGFbeta family signaling: novel insights in development and disease. Development **136**:3691-3697.

Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics **21**:1859-1875.

Yang Y, Lv J, Gui B, Yin H, Wu X, Zhang Y, Jin Y. 2008. A-to-I RNA editing alters less-conserved residues of highly conserved coding regions: implications for dual functions in evolution. RNA **14**:1516-1525.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences : CABIOS **13**:555-556.

Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Molecular biology and evolution **17**:32-43.

Yeh RF, Lim LP, Burge CB. 2001. Computational inference of homologous gene structures in the human genome. Genome research **11**:803-816.

Young JZ. 1971. The anatomy of the nervous system of Octopus vulgaris. Oxford,: Clarendon Press. xxxi, 690pp

Yu WP, Rajasegaran V, Yew K, Loh WL, Tay BH, Amemiya CT, Brenner S, Venkatesh B. 2008. Elephant shark sequence reveals unique insights into the evolutionary history of vertebrate genes: A comparative analysis of the protocadherin cluster. Proceedings of the National Academy of Sciences of the United States of America **105**:3819-3824.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, Xiong Z, Que H, Xie Y, Holland PW, Paps J, Zhu Y, Wu F, Chen Y, Wang J, Peng C, Meng J, Yang L, Liu J, Wen B, Zhang N, Huang Z, Zhu Q, Feng Y, Mount A, Hedgecock D, Xu Z, Liu Y, Domazet-Loso T, Du Y, Sun X, Zhang S, Liu B, Cheng P, Jiang X, Li J, Fan D, Wang W, Fu W, Wang T, Wang B, Zhang J, Peng Z, Li Y, Li N, Wang J, Chen M, He Y, Tan F, Song X, Zheng Q, Huang R, Yang H, Du X, Chen L, Yang M, Gaffney PM, Wang S, Luo L, She Z, Ming Y, Huang W, Zhang S, Huang B, Zhang Y, Qu T, Ni P, Miao G, Wang J, Wang Q, Steinberg CE, Wang H, Li N, Qian L, Zhang G, Li Y, Yang H, Liu X, Wang J, Yin Y, Wang J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature **490**:49-54.

Zipursky SL, Sanes JR. 2010. Chemoaffinity revisited: dscams, protocadherins, and neural circuit assembly. Cell **143**:343-353.