Supplemental Information

Detailed Experimental Procedures:

O. carmela, Illumina library construction

A paired-end genomic library for Illumina sequencing was constructed using *Oscarella carmela* DNA prepared by whole genome amplification (WGA, (1)). To reduce contamination and polymorphism that could complicate genome assembly and analysis, a single sponge larva was isolated, washed five times in sterile-filtered seawater and lysed using the REPLI-g Mini kit for WGA (Qiagen, Valencia, CA). The lysate was divided and used to conduct four separate WGA reactions that were pooled to reduce the effects of stochastic amplification bias. Paired-end library construction was performed using the Illumina PE Adapter Oligo Mix and PCR primers (Illumina Inc., San Diego, CA) in combination with protocol modifications suggested by Quail and colleagues (2). Additionally, during each spin-column purification step, residual ethanol was pipetted out of the column prior to elution to prevent ethanol carry-over. Library quality was determined using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA) to confirm fragment size and concentration.

O. carmela Illumina sequencing and draft genome assembly

A total of 388,627,652 reads were generated from two separate paired-end Illumina runs on the same library: 39,460,320 reads from 2 lanes of 76 cycle sequencing (hereafter called “run 1”), and 349,167,332 reads from 7 lanes of 101 cycle sequencing (“run 2”). Before assembly, low frequency “noise” k-mers were corrected in the reads using the Corrector tool version 1.00 from the Beijing Genomics Institute [http://soap.genomics.org.cn/down/correction.tar.gz] with default parameter values. The two lanes from run 1 were corrected together using a frequency cutoff of 5 per k-mer, and each lane from run 2 was corrected individually using a frequency cutoff of 10 per k-mer. After correction, 33,249,809 reads from run 1 and 298,166,837 reads from run 2 remained, for a total of 331,416,646 reads. Genome assembly was performed iteratively using

SOAPdenovo version 1.04 (3) with default parameter values (unless otherwise noted), as follows: an initial assembly was created using a k-mer size of 31, with both runs used for building contigs and only run 1 used for building scaffolds. To close gaps in the initial assembly, we ran GapCloser version 1.10 (4) with default parameter values using only reads from run 1. We found that the processes of building scaffolds and gap closing were more successful using fewer reads, and thus we chose run 1 for both tasks; using the reads from any single lane of run 2 produced similar results. After running SOAPdenovo and GapCloser, we mapped all corrected reads back to the assembly using Bowtie version 0.12.1 (5) with default parameter values. We then created a final assembly using only the reads that mapped to the initial gap-closed assembly. We ran SOAPdenovo followed by GapCloser, repeating the initial assembly process but instead using at each step the set of reads mapping to the initial assembly. Assembly statistics are shown in Tables S1-S3.

O. carmela gene prediction

Gene prediction was performed *de novo* on the final assembly using Augustus version 2.3 (6) with the autoAug script and the 6,235 assembled Sanger ESTs (7) as prediction aids. Gene prediction was only performed on sequences with a minimum length of 500 (9,823 genes were predicted).

O. carmela genome: assembly statistics for scaffolds

Assemblies	Number of Scaffolds	Total Assembly Size (bp)	Number of Scaffolds + Contigs	Longest (bp)	N50 (bp)	N90 (bp)
Pilot assembly	29,148	57,006,393	70,595	49,630	3,324	416
Initial assembly	22,699	60,727,654	77,270	84,460	4,699	351
Final assembly	17,451	56,386,309	67,767	108,178	5,897	368

O. carmela genome: assembly statistics for contigs

Assemblies	Number of Reads	Total Assembly Size (bp)	Longest (bp)	N50 (bp)	N90 (bp)	Average Coverage (x)
Pilot assembly	39,460,320	46,779,956	6,153	339	124	22
Initial assembly	331,416,646	54,313,237	28,111	890	132	568
Final assembly	239,209,057	54,193,990	43,946	1,158	142	562

O. carmela genome: scaffold GC content, paired end insert size, and gap information

Assemblies	GC Content (percent)	Estimated Insert Size (bp)	Estimated Insert Size Standard Deviation (bp)	Number of Gaps Before GapCloser	Total Size of Gaps Before GapCloser (bp)	Number of Gaps Remaining After GapCloser	Total Size of Gaps Remaining After GapCloser (bp)
Pilot assembly	43.7	390	78	85,376	13,410,978	-	-
Initial assembly	43.5	397	71	55,768	7,991,639	21,580	5,733,376
Final assembly	43.5	395	79	39,994	4,765,458	8,105	2,452,188

Discovery and annotation of novel cadherins

The stand-alone BLAST search algorithm was used to search the best predicted protein set from the draft genomes of *S. rosetta, C. owczarzaki*, and *O. carmela* using the 23 predicted cadherins from the *M. brevicollis* genome (8) as a query. As a complement to this approach, Pfam (9), SMART (10) and Phobius (11) domain prediction programs were run on all predicted *S. rosetta* proteins. Every protein predicted to have at least one extracellular cadherin (EC) domain was annotated and categorized according to whether its overall domain composition and architecture matched known cadherins from *M. brevicollis* or any metazoan. The *S. rosetta* gene models are supported by 33-fold sequence coverage suggesting that we have identified most, if not all cadherins in the genome (12). Accurate abundance data for O. carmela could not be determined due to the early draft status of the genome. Therefore, cadherin abundance in sponges was determined from the genome of *Amphimedon queenslandica* (11). Cadherin abundance estimates for eumetazoans were derived from Hulpiau and van Roy (13) and references therein. Taxonomic data from SMART were used to conclude that no EC domains are present in any annotated plant or fungus.

HMM searches for Hh-N domain-containing proteins

We used the HMMER 3.0 suite of tools (14) to build custom models of the Hh-N signaling domain in order to increase sensitivity for searches of choanoflagellates and other opisthokonts. We used hmmsearch (14) with the Pfam domain Hh_signal [PF01085, Pfam version 24.0 (9)] to detect Hh-N domains in the predicted protein sets from the genomes of the sponge *A. queenslandica* (15), the sea anemone *N. vectensis* (16), and the choanoflagellates *S. rosetta* (12) and *M. brevicollis* (17). Using the sequences of all domains predicted by hmmsearch with an E value below the gathering threshold for the model in Pfam, we built a multiple alignment using the FSA web server version 1.15.2 (18). We used the resulting alignment to build a custom model with hmmbuild (14), and ran hmmsearch with the custom model against the predicted protein sets from *O. carmela*, *S. rosetta* and *M. brevicollis* in order to detect previously unidentified instances of the Hh-N domain.

Cloning full-length Ocar_bcat

Tissue of *O. carmela* was flash frozen and ground to a powder using a mortar and pestle containing liquid nitrogen. Messenger RNA was isolated using Trizol Reagent (Invitrogen Corp., Carlsbad, CA) followed by the Oligotex mRNA Mini Kit (Qiagen, Valencia, CA). The unknown 5' sequence of Ocar_bcat was cloned and sequenced using GeneRacer (Invitrogen Corp., Carlsbad, CA) in combination with an antisense primer (SN33R: 5' CCCAAGGGCAAGTCTTCGCTGGAT 3') corresponding to the known 3' EST sequence (7). The full-length sequence is deposited in GenBank (HQ234356).

Ocar_bcat structural predictions

The full-length sequence of Ocar_bcat was translated from the cloned mRNA transcript using NCBI ORF Finder. The predicted protein was analyzed for its homology to known beta-catenin sequences by comparing its primary sequence to the non-redundant Genbank database (nr) via blastp (19) and by searching for conserved structural domains (arm repeats) using Pfam (9) and SMART (10). Each predicted arm repeat in beta-catenin-related proteins from human, *O. carmela, M. brevicollis*, *S. rosetta*, *Dictyostelium discoideum* and *Arabidopsis thaliana* was subjected to pair-wise reciprocal blast (9). For example, arm repeat 1 from Ocar_bcat was used to perform a Blastp (19) search against a database of all arm repeats from all sampled proteins. We expected that orthologous sequences from different species would exhibit a co-linear sequence of arm repeat homology with human beta-catenin [Fig.S5; method modified from (20)]. In the example of O. carmela arm repeat 1, only a best-reciprocal blast with arm repeat 1 from human beta-catenin would be interpreted support homology of these two proteins.

To identify conserved functional residues and motifs within Ocar_bcat, multiple sequence alignment was performed using MUSCLE (21). Additionally, the three-dimensional structure of Ocar_bcat was analyzed using alignment-based fold-prediction as implemented by LOOPP (22). Predicted structures were visualized with PyMOL (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.).

Yeast two-hybrid screen

A yeast two-hybrid screen was conducted to identify candidate binding-partners of full-length Ocar_bcat. To construct a yeast expression library representative of the expressed genes of *O. carmela*, mRNA was isolated from pooled adult and embryonic tissues (from many individuals to maximize transcript diversity) and cloned into pDONR222 using the CloneMiner cDNA Library Construction Kit (Invitrogen Corp., Carlsbad, CA). Inserts from this library were shuttled into the yeast two-hybrid prey plasmid, pDEST22 using LR Clonase II enzyme mix (Invitrogen Corp., Carlsbad, CA) and transformed for storage and amplification into ElectroMAX DH10B T1 Phage Resistant Cells (Invitrogen Corp., Carlsbad, CA). Likewise, full-length Ocar_bcat was modified using PCR to incorporate Gateway compatible attB1/attB2 recombination sites and cloned into pDONR221 using BP Clonase II enzyme mix (Invitrogen Corp., Carlsbad, CA). This insert was shuttled into the yeast two-hybrid bait-plasmid, pDEST32 using LR Clonase II enzyme mix.

Yeast transformation and screening was performed at the yeast two-hybrid facility at Indiana University (23). Full-length Ocar_bcat and positive clones were tested for autoactivation on his- media. E-Amino-1,2,4-Triazol (3AT), which acts as a quantitative inhibitor of the HIS3 reporter gene, was used to control autoactivation by Ocar_bcat. After a <10 day screen, positive clones were retested on his- media, ura- media, and in LacZ assays. Inserts from positive clones were rescued and sequenced at the University of California DNA sequencing facility. Insert sequences from positive clones were compared against the draft assembly of the *O. carmela* genome using blastn (19) and predicted proteins were annotated using blastp (19), Pfam (9) and SMART (10) to test for homology with known proteins.

Seventeen unique candidate binding-partners of Oc_bcat were detected (Table S2), including three clones encoding the CCD region of OcCdh1 and an additional well-known beta-catenin binding protein, Axin. These detected interactions could not be independently validated using in vitro binding assays because recombinant forms of Ocar_bcat proved to be highly insoluble. Nevertheless, the conserved structural features of Ocar_bcat and Ocar_Cdh1, coupled with the fact that this is a widely conserved interaction in metazoans, suggest that the yeast two-hybrid result represents a *bona fide* interaction.

Fig. S1

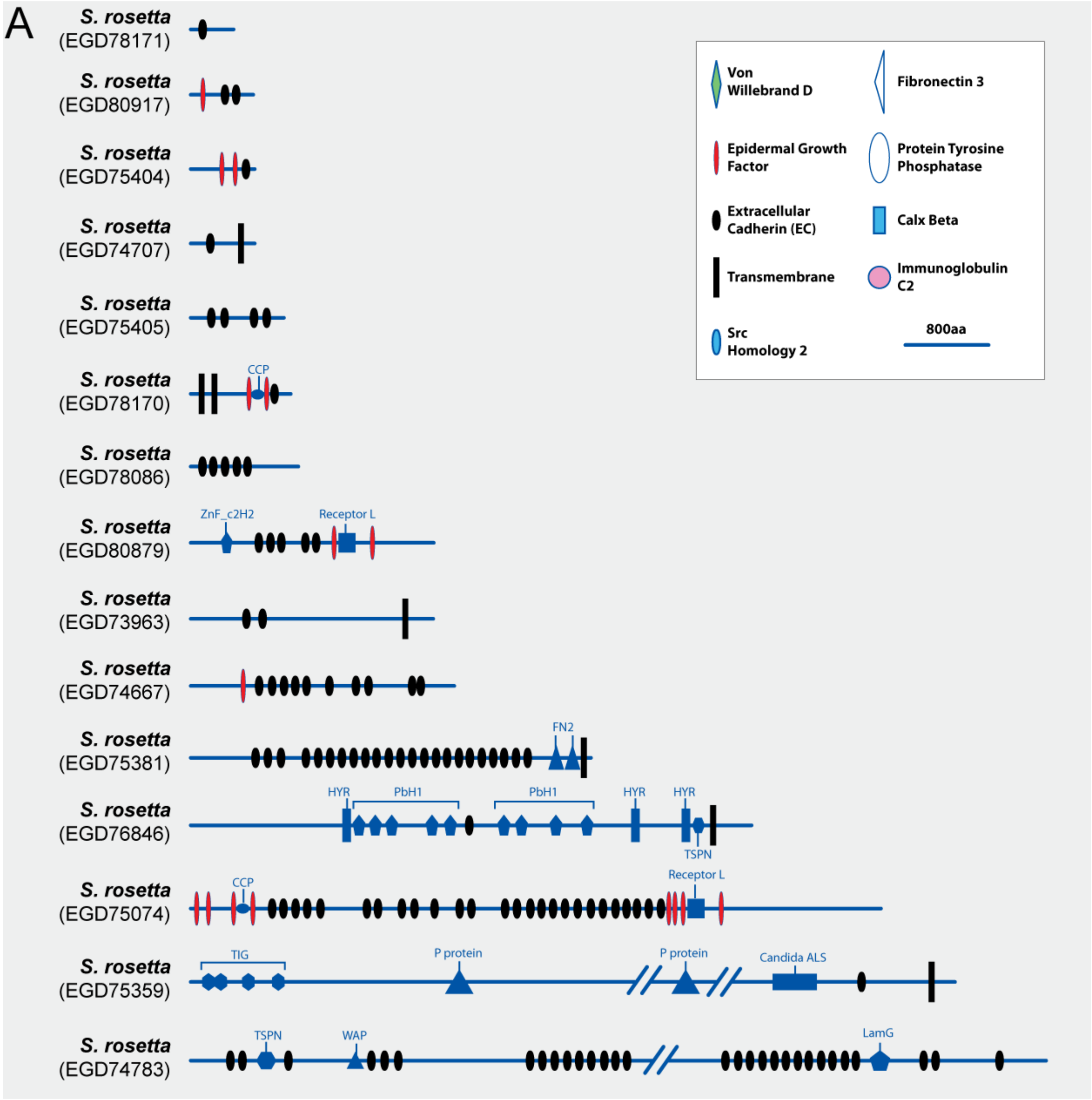

A
S. rosetta (EGD78171)
S. rosetta (EGD80917)
S. rosetta (EGD75404)
S. rosetta (EGD74707)
S. rosetta (EGD75405)
S. rosetta (EGD78170)
CCP
S. rosetta (EGD78086)
S. rosetta (EGD80879)
ZnF_c2H2
Receptor L
S. rosetta (EGD73963)
S. rosetta (EGD74667)
S. rosetta (EGD75381)
FN2
S. rosetta (EGD76846)
HYR
PbH1
PbH1
HYR
HYR
TSPN
S. rosetta (EGD75074)
CCP
Receptor L
S. rosetta (EGD75359)
TIG
P protein
P protein
Candida ALS
S. rosetta (EGD74783)
TSPN
WAP
LamG
Von Willebrand D
Fibronectin 3
Epidermal Growth Factor
Protein Tyrosine Phosphatase
Extracellular Cadherin (EC)
Calx Beta
Transmembrane
Immunoglobulin C2
Src Homology 2
800aa

Fig. S1, continued.

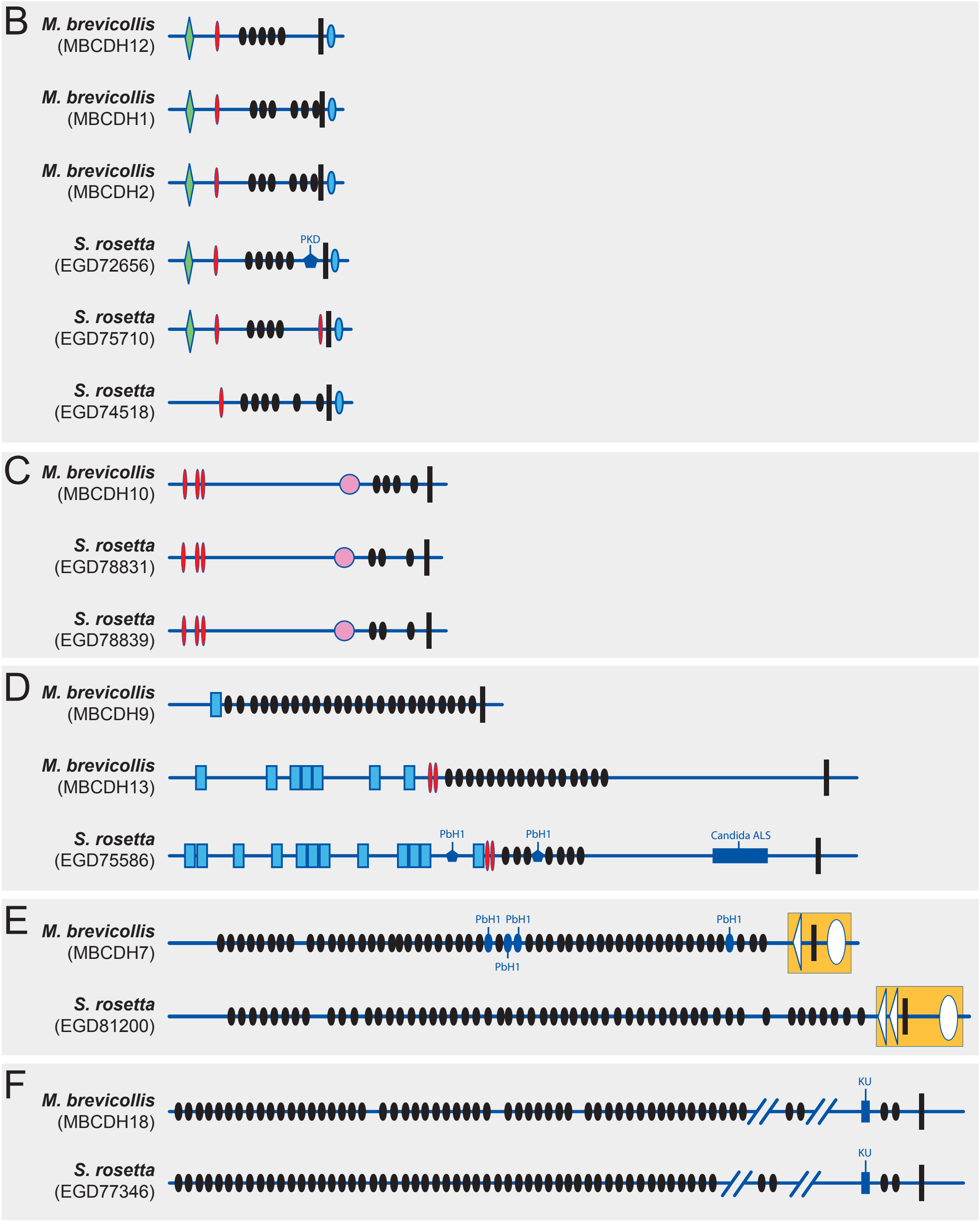


Fig. S1. Domain architecture of *S. rosetta* cadherins without orthologs in Metazoa or *C. owczarzaki*. (A) 16 out of 29 predicted *S. rosetta* cadherin proteins have no clear orthology to any cadherins known from other species,

whereas five protein families (B-F) can be identified as shared between and exclusive to *S. rosetta* and *M. brevicollis* based upon similarities in their domain composition and arrangement. Of these, one family (E) has partial homology to the lefftyrin family that is found in choanoflagellates and sponges. However, genes in this family differ from choanoflagellate lefftyrins in that they are predicted to have catalytically active cytoplasmic PTPase domains.

(Abbreviations: Candida ALS = Candida Agglutinin-like sequence; CCP = domain abundant in complement control proteins; FN2 = fibronectin 2; HYR = Hyalin Repeat; KU = BPTI/Kunitz family of serine protease inhibitors; LamG = laminin G domain; P protein = Proprotein convertase P-domain; PbH1 = parallel beta-helix repeats; PKD = polycystic kidney disease; TIG = transcription factor immunoglobulin-like domain; TSPN = Thrombospondin N-terminal-like domain; WAP = whey acidic protein; ZnF_c2h2 = zinc-finger, c2h2 type).

Fig. S2.

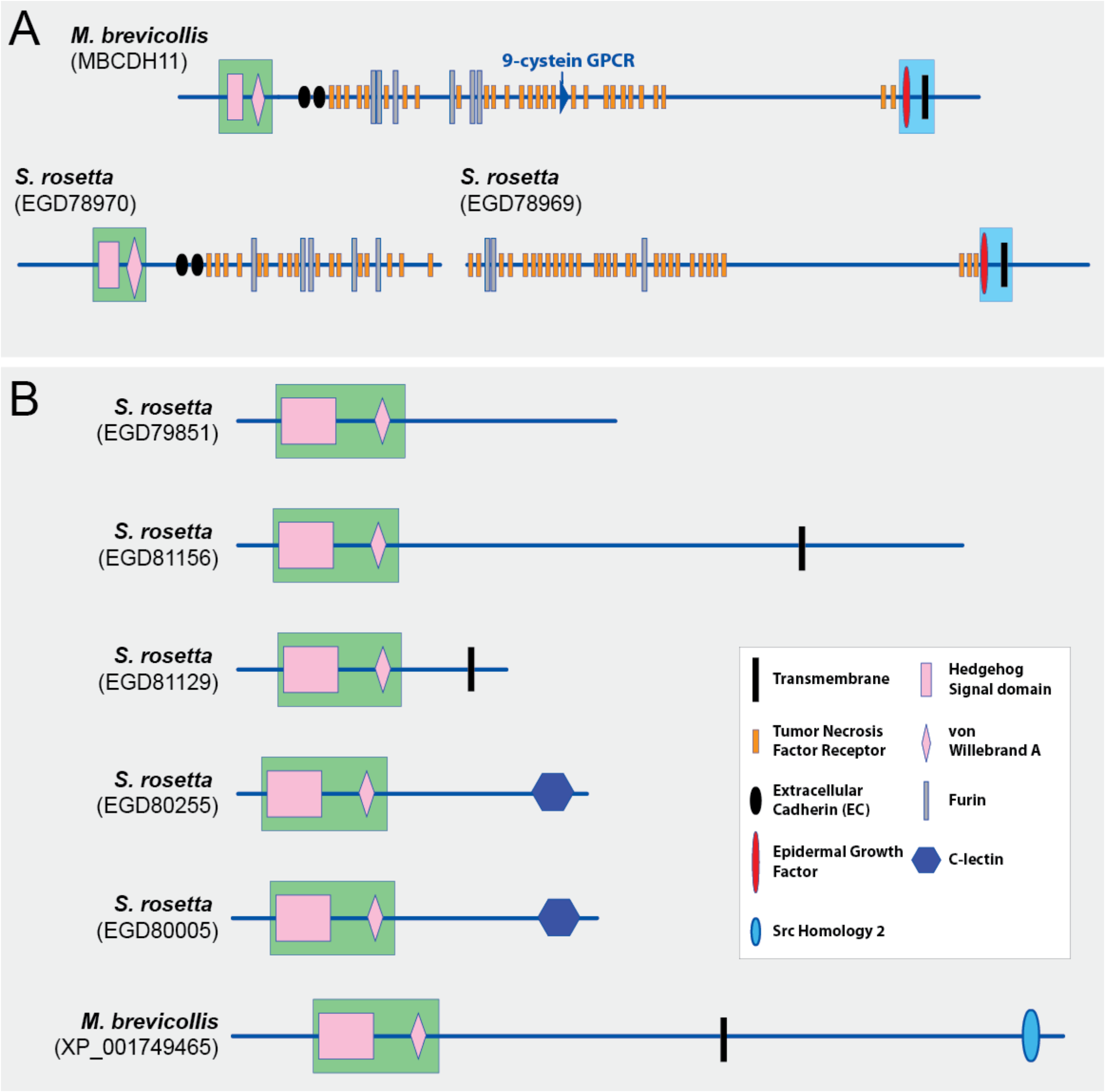
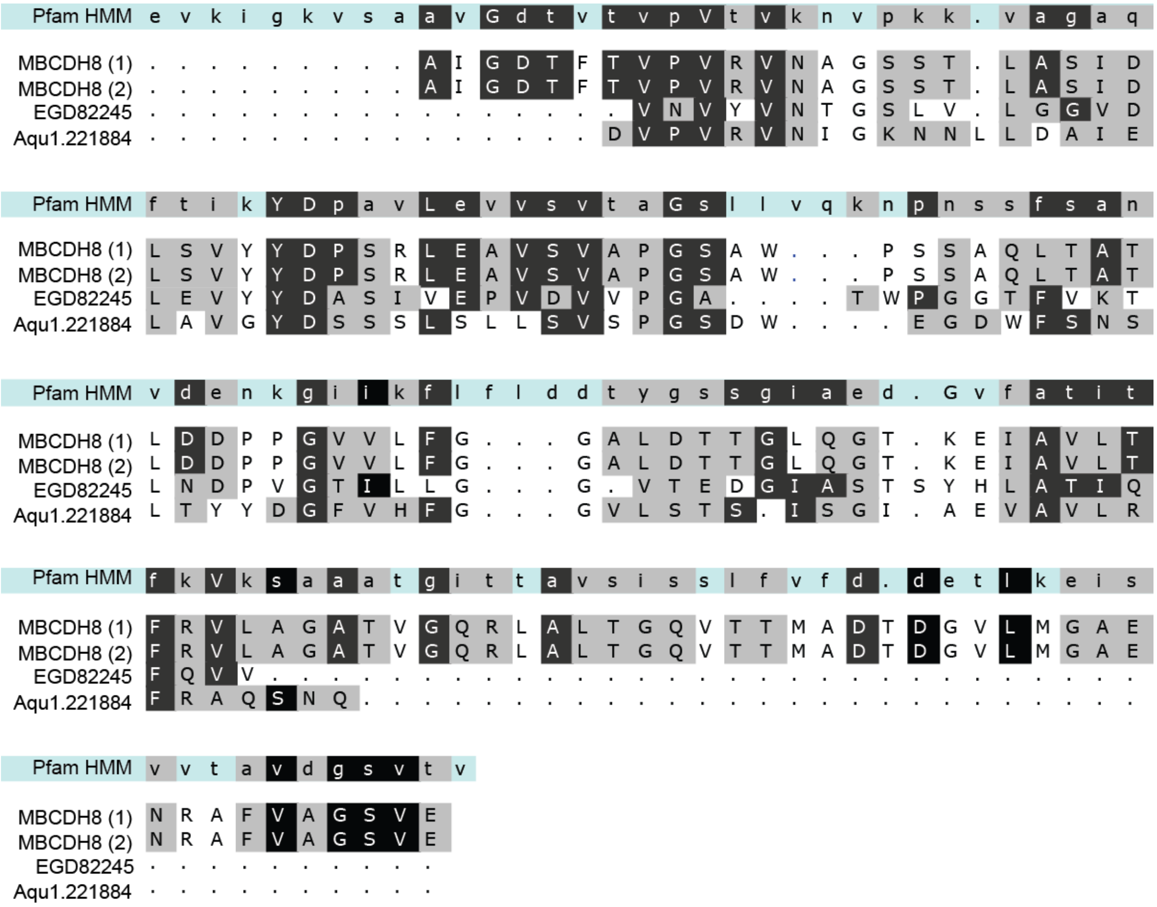


Fig. S2. Additional detected Hh-N domain containing proteins from *S. rosetta* and *M. brevicollis*. (A) In *S. rosetta*, two adjacent gene models on a single scaffold have close homology to parts of *M. brevicollis* hedgling (MBCDH11). Both gene models are supported by RNAseq expression data, but there is a predicted stop codon between them and there are no RNAseq reads that span the divide. We infer either that the stop codon that splits *S. rosetta* hedgling evolved following the divergence of the *M. brevicollis* and *S. rosetta* lineages, or that it is the result of a genome assembly error. Further interpretation

will require experimental investigation of these gene models. (B) Using a custom HMM created against the Hh-N domain of known hedgling proteins we also identified five *S. rosetta* proteins and one *M. brevicollis* protein that have a conserved Hh-N domain, but lack EC domains. In each case, as in all known hedglings, the Hh-N domain is adjacent to a von Willebrand A domain. Therefore, we hypothesize that the association of these two domains in diverse proteins and in diverse organisms reflects an ancestral function that has been lost in eumetazoans.

Fig. S3



```
   Pfam HMM  e v k i g k v s a a v G d t v t v p V t v k n v p k k . v a g a q

MBCDH8 (1)   . . . . . . . . . A I G D T F T V P V R V N A G S S T . L A S I D
MBCDH8 (2)   . . . . . . . . . A I G D T F T V P V R V N A G S S T . L A S I D
  EGD82245   . . . . . . . . . . . . . . . V N V Y V N T G S L V . L G G V D
Aqu1.221884  . . . . . . . . . . . . . . D V P V R V N I G K N N L L D A I E

   Pfam HMM  f t i k Y D p a v L e v v s v t a G s l l v q k n p n s s f s a n

MBCDH8 (1)   L S V Y Y D P S R L E A V S V A P G S A W . . . P S S A Q L T A T
MBCDH8 (2)   L S V Y Y D P S R L E A V S V A P G S A W . . . P S S A Q L T A T
  EGD82245   L E V Y Y D A S I V E P V D V V P G A . . . . T W P G G T F V K T
Aqu1.221884  L A V G Y D S S S L S L L S V S P G S D W . . . . E G D W F S N S

   Pfam HMM  v d e n k g i i k f l f l d d t y g s s g i a e d . G v f a t i t

MBCDH8 (1)   L D D P P G V V L F G . . . G A L D T T G L Q G T . K E I A V L T
MBCDH8 (2)   L D D P P G V V L F G . . . G A L D T T G L Q G T . K E I A V L T
  EGD82245   L N D P V G T I L L G . . . G . V T E D G I A S T S Y H L A T I Q
Aqu1.221884  L T Y Y D G F V H F G . . . G V L S T S . I S G I . A E V A V L R

   Pfam HMM  f k V k s a a a t g i t t a v s i s s l f v f d . d e t l k e i s

MBCDH8 (1)   F R V L A G A T V G Q R L A L T G Q V T T M A D T D G V L M G A E
MBCDH8 (2)   F R V L A G A T V G Q R L A L T G Q V T T M A D T D G V L M G A E
  EGD82245   F Q V V . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
Aqu1.221884  F R A Q S N Q . . . . . . . . . . . . . . . . . . . . . . . . . .

   Pfam HMM  v v t a v d g s v t v

MBCDH8 (1)   N R A F V A G S V E
MBCDH8 (2)   N R A F V A G S V E
  EGD82245   . . . . . . . . . .
Aqu1.221884  . . . . . . . . . .
```

Fig. S3. Cohesin domains from Coherin family proteins aligned against the Cohesin Hidden Markov Model from Pfam. Residues that exactly match Pfam HMM (highlighted in blue) are indicated with black shading whereas residues that are considered to be a conservative substitution with respect to what the model expects are indicated with gray shading. Cohesin domains 1 and 2 from *Monosiga brevicollis* (MBCDH8) are identical to each other. Protein identifiers correspond to Fig. 2c. (Abbreviations: HMM: Hidden Markov Model).

Fig. S4

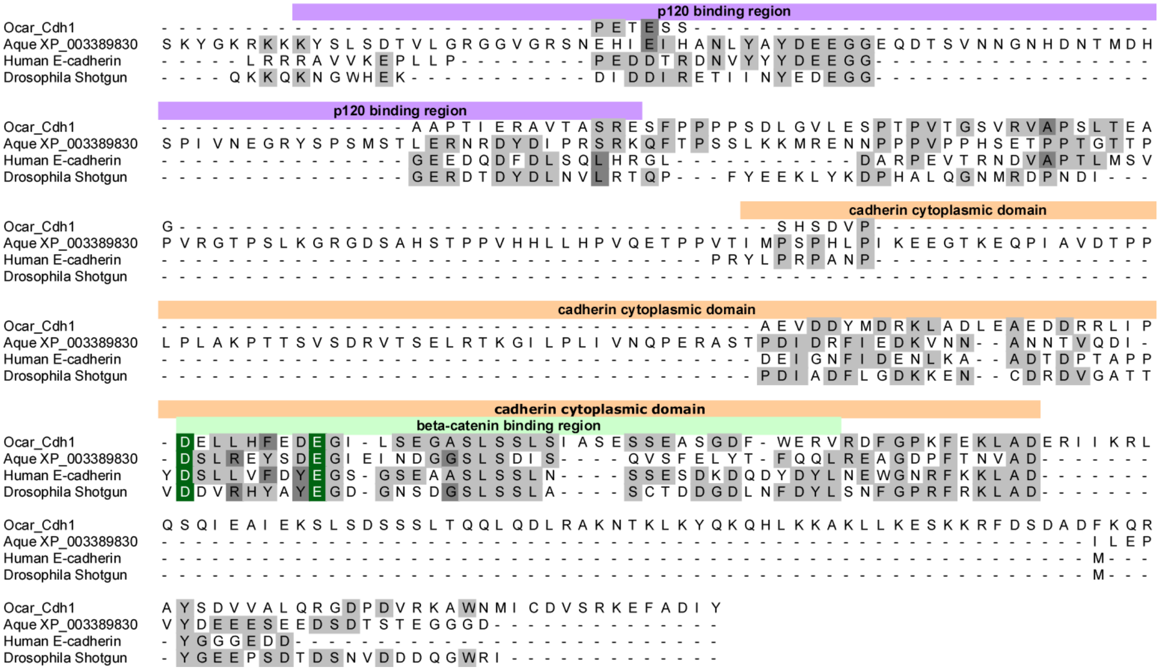

p120 binding region
Ocar_Cdh1 --------------------------PETESS----------------------------
Aque XP_003389830 SKYGKRKKKYSLSDTVLGRGGVGRSNEHIEIHANLYAYDEEGGEQDTSVNNGNHDNTMDH
Human E-cadherin -----LRRRAVVKEPLLP--------PEDDTRDNVYYYDEEGG-----------------
Drosophila Shotgun ----QKKQKNGWHEK-----------DIDDIRETIINYEDEGG-----------------
p120 binding region
Ocar_Cdh1 ---------------AAPTIERAVTASRESFPPPPSDLGVLESPTPVTGSVRVAPSLTEA
Aque XP_003389830 SPIVNEGRYSPSMSTLERNRDYDIPRSRKQFTPSSLKKMRENNPPPVPPHSETPPTGTTP
Human E-cadherin ---------------GEEDQDFDLSQLHRGL-----------DARPEVTRNDVAPTLMSV
Drosophila Shotgun ---------------GERDTDYDLNVLRTQP---FYEEKLYKDPHALQGNMRDPNDI---
cadherin cytoplasmic domain
Ocar_Cdh1 G------------------------------------SHSDVP-----------------
Aque XP_003389830 PVRGTPSLKGRGDSAHSTPPVHHLLHPVQETPPVTIMPSPHLPIKEEGTKEQPIAVDTPP
Human E-cadherin ---------------------------------PRYLPRPANP-----------------
Drosophila Shotgun --
cadherin cytoplasmic domain
Ocar_Cdh1 ------------------------------------AEVDDYMDRKLADLEAEDDRRLIP
Aque XP_003389830 LPLAKPTTSVSDRVTSELRTKGILPLIVNQPERASTPDIDRFIEDKVNN--ANNTVQDI-
Human E-cadherin ------------------------------------DEIGNFIDENLKA--ADTDPTAPP
Drosophila Shotgun ------------------------------------PDIADFLGDKKEN--CDRDVGATT
cadherin cytoplasmic domain
beta-catenin binding region
Ocar_Cdh1 -DELLHFEDEGI-LSEGASLSSLSIASESSEASGDF-WERVRDFGPKFEKLADERIIKRL
Aque XP_003389830 -DSLREYSDEGIEINDGGSLSDIS----QVSFELYT-FQQLREAGDPFTNVAD-------
Human E-cadherin YDSLLVFDYEGS-GSEAASLSSLN----SSESDKDQDYDYLNEWGNRFKKLAD-------
Drosophila Shotgun VDDVRHYAYEGD-GNSDGSLSSLA----SCTDDGDLNFDYLSNFGPRFRKLAD-------
Ocar_Cdh1 QSQIEAIEKSLSDSSSLTQQLQDLRAKNTKLKYQKQHLKKAKLLKESKKRFDSDADFKQR
Aque XP_003389830 --ILEP
Human E-cadherin --M---
Drosophila Shotgun --M---
Ocar_Cdh1 AYSDVVALQRGDPDVRKAWNMICDVSRKEFADIY
Aque XP_003389830 VYDEEESEEDSDTSTEGGGD--------------
Human E-cadherin -YGGGEDD--------------------------
Drosophila Shotgun -YGEEPSDTDSNVDDDQGWRI-------------

Fig. S4. Annotated alignment of classical cadherin cytoplasmic tails. The juxtamembrane domain (purple box) that constitutes the binding site for p120 catenin is partially conserved between human and *Drosophila* and *Amphimedon*, but is divergent in Ocar_Cdh1. In contrast, the beta-catenin binding domain (light green box) of the predicted CCD (light orange box) of Ocar_Cdh1 is conserved, including at residues that are required for the interaction (dark green). The sponge sequences are predicted to be longer than their bilaterian counterparts, complicating alignment of all but the most highly conserved residues.

Fig. S5.

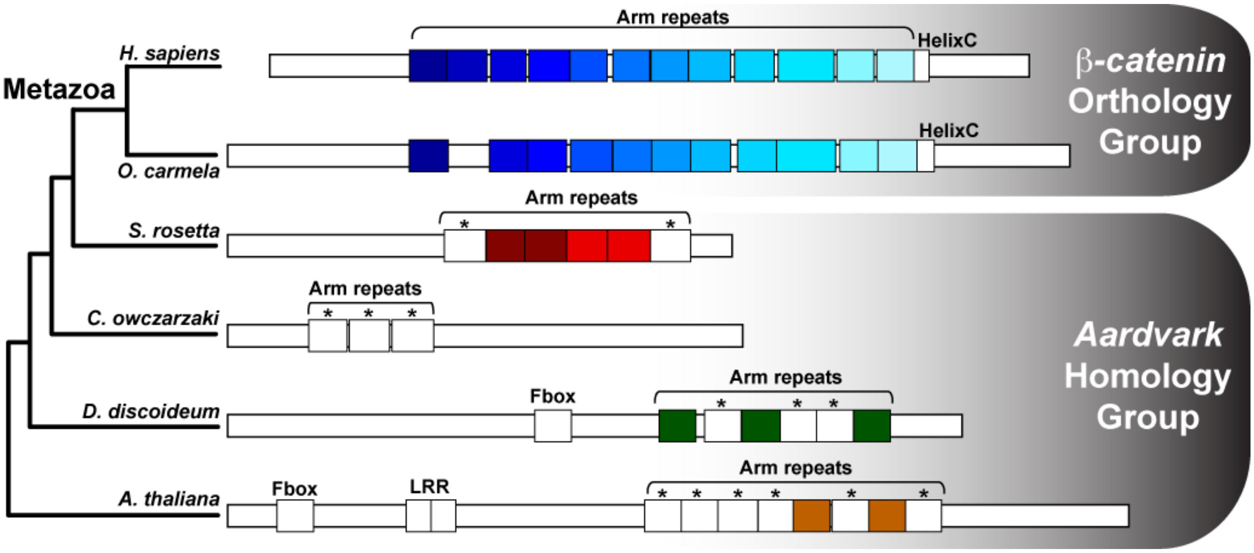


Fig. S5. Domain organization and phylogenetic distribution of proteins with homology to beta-catenin.

Protein diagrams are mapped onto a previously determined phylogenetic tree (24) with arm domains colored to indicate their similarity. Repeats of the same color are best-reciprocal Blast pairs. Arm repeats without close identity to any other are uncolored and indicated with an asterisk. Linear conservation of homologous arm repeats is restricted to metazoan beta-catenin orthologs, suggesting that the metazoan roles of beta-catenin evolved in the metazoan stem lineage and have been highly conserved throughout metazoan evolution.

Tables.

Table S1. *S. rosetta* cadherin expression levels.

Genbank ID	Min FPKM[1]	Max FPKM	Mean FPKM	Median FPKM
EGD80879	27.617977	113.246096	56.93163613	48.3590645
EGD80917	2.25581	6.049739	3.860855875	3.533781
EGD78831	7.874201	40.03944	19.4444355	15.1492895
EGD78839	0.109114	26.26796	11.104367	9.8600525
EGD79002	1.87756	6.256101	3.839630625	3.370277
EGD79017	29.325694	128.553899	80.03320963	89.619573
EGD82245	3.403667	15.877104	9.775254375	10.434421
EGD82557	0.85664	8.627106	4.377121	3.1091385
EGD72656	168.694501	984.67225	624.6796178	621.626123
EGD73963	2.017099	8.457588	4.626551625	3.828202
EGD74518	46.138224	267.075716	159.4101904	161.7252545
EGD74707	1.962277	15.002993	8.222477875	8.487063
EGD75381	0.133699	51.162787	18.63580838	15.35265
EGD75404	3.990599	9.91731	7.319792125	7.684626
EGD75405	2.37142	9.804725	6.56574025	6.3694265
EGD75586	0.087914	6.21004	2.840604125	2.17229
EGD75074	2.197013	6.533185	4.66290875	4.722256
EGD74783	0.026136	3.799631	1.556376	1.0930275
EGD75710	71.962177	626.409101	259.4577603	220.060925
EGD76846	5.967787	85.11871	33.87954975	17.35544
EGD77346	7.357232	20.994633	12.16801713	10.4218215
EGD78086	0	7.934519	2.76326325	1.801815
EGD78170	18.746381	50.514396	28.3529975	26.7736315
EGD78171	23.099038	61.605291	35.97480775	33.790346
EGD81200	0.053023	20.10651	9.513880375	8.764376
EGD78969	9.214266	59.752132	31.85870863	33.0626255
EGD78970	5.89713	31.968329	15.953967	15.8754105
EGD74667	2.071066	14.728778	7.26199325	6.9755495
EGD75359	0.023944	7.275513	3.04185575	2.4945935
EGD79249	0.020866	3.374047	1.3963085	1.166545

[1] The number of fragments per kilobase per million sequenced reads (FPKM) mapping to each identified S. rosetta cadherin from RNA-seq of eight growth conditions is summarized as evidence of gene expression.

Table S2. *O. carmela* binding partners predicted from yeast two-hybrid screen of beta-catenin.

gene ID	Tentative Identification	Predicted domain architecture (Pfam)	Predicted domain architecture (Smart)
g4908.t1	none	none	none
g9583.t1	none	death	none
g6098.t1	Upstream binding protein	CP2	none
g8349.t1	40S ribosomal protein S11	Ribosomal S17	none
g6246.t1	Tenascin	EGF 2 (x9); EGF Ca (x2)	VWD; EGF like; EGF (x10); EGF Ca (x2)
g6719.t1	none	EIF4E-T	coiled coil
g8701.t1	Transcription factor AP-1/c-Jun	bZIP 1	BRLZ
g2054.t1	Calumenin	SPARC Ca bdg; efhand (x2)	EFh (x2)
g6285.t1	E74-like factor	Ets	ETS
g10012.t1	Chromosomal segregation protein SMC	none	coiled-coil
g4744.t1	GTPase Rab2	Ras	RAB
g8915.t1	Baculoviral IAP repeat-containing protein 4	BIR (x4)	BIR (x4); RING
g6056.t1	Ribosomal protein L13	Ribosomal L13e	none
g2979.t1	Ral	Ras	RAS
g3724.t1	Choline-phosphate cytidylyltransferase	none	coiled-coil
g6554.t1	Axin	RGS; DIX	RGS; DAX
AEC12441	Ocar_Cdh1	EC; EGF; Lam-G; CCD	EC; EGF; Lam-G

References

1. Hosono S, *et al.* (2003) Unbiased whole-genome amplification directly from clinical samples. *Genome Res* 13(5):954-964.
2. Quail MA, *et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5(12):1005-1010.
3. Li R, *et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 20(2):265-272.
4. http://soap.genomics.org.cn/down/GapCloser.tar.gz
5. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 10(3):R25.
6. Stanke M, Diekhans M, Baertsch R, & Haussler D (2008) Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 24(5):637-644.
7. Nichols SA, Dirks W, Pearse JS, & King N (2006) Early evolution of animal cell signaling and adhesion genes. *Proc Natl Acad Sci U S A* 103(33):12451-12456.
8. Abedin M & King N (2008) The premetazoan ancestry of cadherins. *Science* 319(5865):946-948.
9. Finn RD, *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res* 38(Database issue):D211-222.
10. Schultz J, Milpetz F, Bork P, & Ponting CP (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* 95(11):5857-5864.
11. Kall L, Krogh A, & Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res* 35(Web Server issue):W429-432.
12. http://www.broadinstitute.org/annotation/genome/multicellularity_project/MultiHome.html
13. Hulpiau P & van Roy F (2011) New insights into the evolution of metazoan cadherins. *Mol Biol Evol* 28(1):647-657.
14. http://www.hmmer.janelia.org
15. Srivastava M, *et al.* (2010) The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* 466(7307):720-726.
16. Putnam NH, *et al.* (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* 317(5834):86-94.
17. King N, *et al.* (2008) The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans. *Nature* 451(7180):783-788.
18. Bradley RK, *et al.* (2009) Fast statistical alignment. *PLoS Comput Biol* 5(5):e1000392.
19. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389-3402.
20. Oda H, Tagawa K, & Akiyama-Oda Y (2005) Diversification of epithelial adherens junctions with independent reductive changes in cadherin form: identification of potential molecular synapomorphies among bilaterians. *Evol Dev* 7(5):376-389.

21. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32(5):1792-1797.
22. Tobi D & Elber R (2000) Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins* 41(1):40-46.
23. http://sites.bio.indiana.edu/~michaelslab/yeast_two_hybrid_facility.html
24. Ruiz-Trillo I, Roger AJ, Burger G, Gray MW, & Lang BF (2008) A phylogenomic investigation into the origin of metazoa. *Mol Biol Evol* 25(4):664-672.