LETTER

OPEN

doi:10.1038/nature22033

The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest

Michael R. Hall[1]*, Kevin M. Kocot[2]*†, Kenneth W. Baughman[3]*, Selene L. Fernandez-Valverde[2]†, Marie E. A. Gauthier[2], William L. Hatleberg[2], Arunkumar Krishnan[2], Carmel McDougall[2], Cherie A. Motti[1], Eiichi Shoguchi[3], Tianfang Wang[4], Xueyan Xiang[2], Min Zhao[2,4], Utpal Bose[1,4], Chuya Shinzato[3], Kanako Hisata[3], Manabu Fujie[5], Miyuki Kanda[5], Scott F. Cummins[4], Noriyuki Satoh[3], Sandie M. Degnan[2] & Bernard M. Degnan[2]

The crown-of-thorns starfish (COTS, the *Acanthaster planci* species group) is a highly fecund predator of reef-building corals throughout the Indo-Pacific region[1]. COTS population outbreaks cause substantial loss of coral cover, diminishing the integrity and resilience of reef ecosystems[2–6]. Here we sequenced genomes of COTS from the Great Barrier Reef, Australia and Okinawa, Japan to identify gene products that underlie species-specific communication and could potentially be used in biocontrol strategies. We focused on water-borne chemical plumes released from aggregating COTS, which make the normally sedentary starfish become highly active. Peptide sequences detected in these plumes by mass spectrometry are encoded in the COTS genome and expressed in external tissues. The exoproteome released by aggregating COTS consists largely of signalling factors and hydrolytic enzymes, and includes an expanded and rapidly evolving set of starfish-specific ependymin-related proteins. These secreted proteins may be detected by members of a large family of olfactory-receptor-like G-protein-coupled receptors that are expressed externally, sometimes in a sex-specific manner. This study provides insights into COTS-specific communication that may guide the generation of peptide mimetics for use on reefs with COTS outbreaks.

COTS (Fig. 1a–c) are extremely fecund mass spawners[7], which predisposes them to population outbreaks that result in a pronounced loss of live coral cover and associated biodiversity. These outbreaks have a higher impact on reef health and resilience than the combined effects of coral bleaching and disease, and increase the susceptibility of reefs to other potentially detrimental events, such as severe storms[2–6] (Supplementary Note 1).

Although a range of local *in situ* control measures have been applied with some success (Supplementary Note 1), mitigation of COTS outbreaks on the necessary regional scale requires mass-deployed, species-specific strategies. In this context, genome-encoded COTS-specific attractants that underpin spawning aggregations have substantial potential as biocontrol agents. To identify attractants, we sequenced the genomes of two wild-caught individuals separated by over 5,000 km, one from the Great Barrier Reef (GBR), Australia and the other from Okinawa (OKI), Japan (Fig. 1c, d and Extended Data Fig. 1). We also sequenced transcriptomes from external organs, and proteins released into the seawater by COTS that were aggregating or were in the presence of their main predator, the giant triton *Charonia tritonis* (Fig. 1b).

We generated separate 384 megabase (Mb) draft assemblies for the GBR and OKI genomes (Extended Data Table 1 and Supplementary Note 2), both of which have unexpectedly low levels of heterozygosity (0.88 and 0.92%, respectively; Extended Data Table 1, Extended Data Fig. 2 and Supplementary Note 3). Reciprocal BLAST analysis of scaffolds longer than 10 kilobases (kb) revealed 98.8% nucleotide identity between GBR and OKI genomes, providing evidence of high

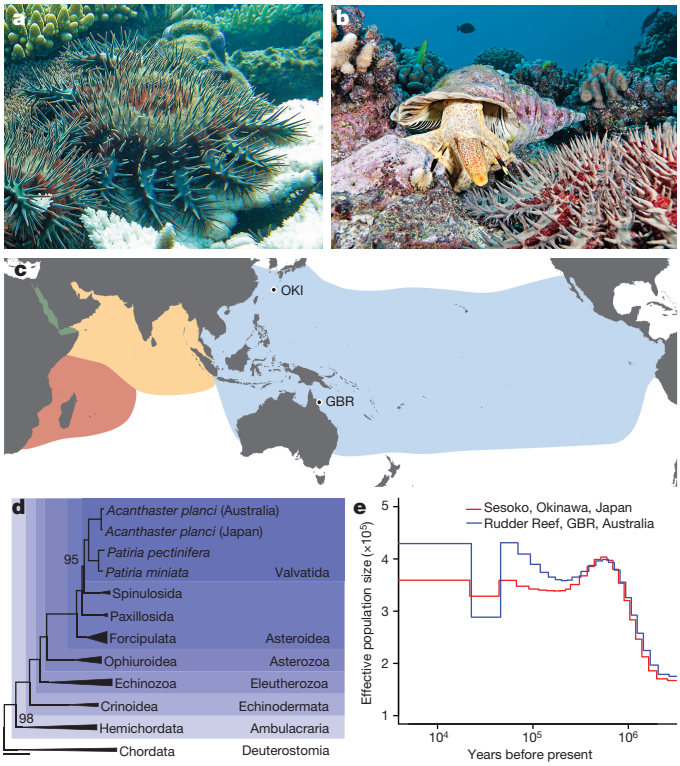


Figure 1 | The crown-of-thorns starfish. a, Adult COTS predating on coral. White coral skeleton (foreground), unconsumed coral (background). Photo by the Australian Institute of Marine Science. **b**, A COTS (foreground) and its predator, the giant triton. Photo by Oceanwide Images. **c**, Global distribution of COTS[8] and the collection sites of the two individuals sequenced. Blue, yellow, pink and green, Pacific Ocean, north Indian Ocean, south Indian Ocean and Red Sea clades, respectively. **d**, Phylogeny of Deuterostomia showing placement of *Acanthaster*. A partially condensed maximum likelihood topology is shown. Scale bar, 0.1 substitutions per site. Bootstrap support values below 100 are shown. **e**, Historical effective population sizes inferred from OKI and GBR genomes using multiple sequential Markovian coalescent analysis[9], assuming a generation time of 3 years and a substitution mutation rate of 1.0×10^{-8} per generation.

[1]Australian Institute of Marine Science (AIMS), Cape Ferguson, Townsville, Queensland 4810, Australia. [2]Centre for Marine Science, School of Biological Sciences, The University of Queensland, Brisbane, Queensland 4072, Australia. [3]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan. [4]Genecology Research Centre, University of the Sunshine Coast, Maroochydore DC, Queensland 4558, Australia. [5]DNA Sequencing Section, Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan. †Present addresses: Department of Biological Sciences and Alabama Museum of Natural History, The University of Alabama, Tuscaloosa, Alabama 35487, USA (K.M.K.); CONACYT, Laboratorio Nacional de Genómica para la Biodiversidad, Centro de Investigación y de Estudios Avanzados del IPN, Irapuato, Guanajuato, Mexico (S.L.F.-V.).
*These authors contributed equally to this work.

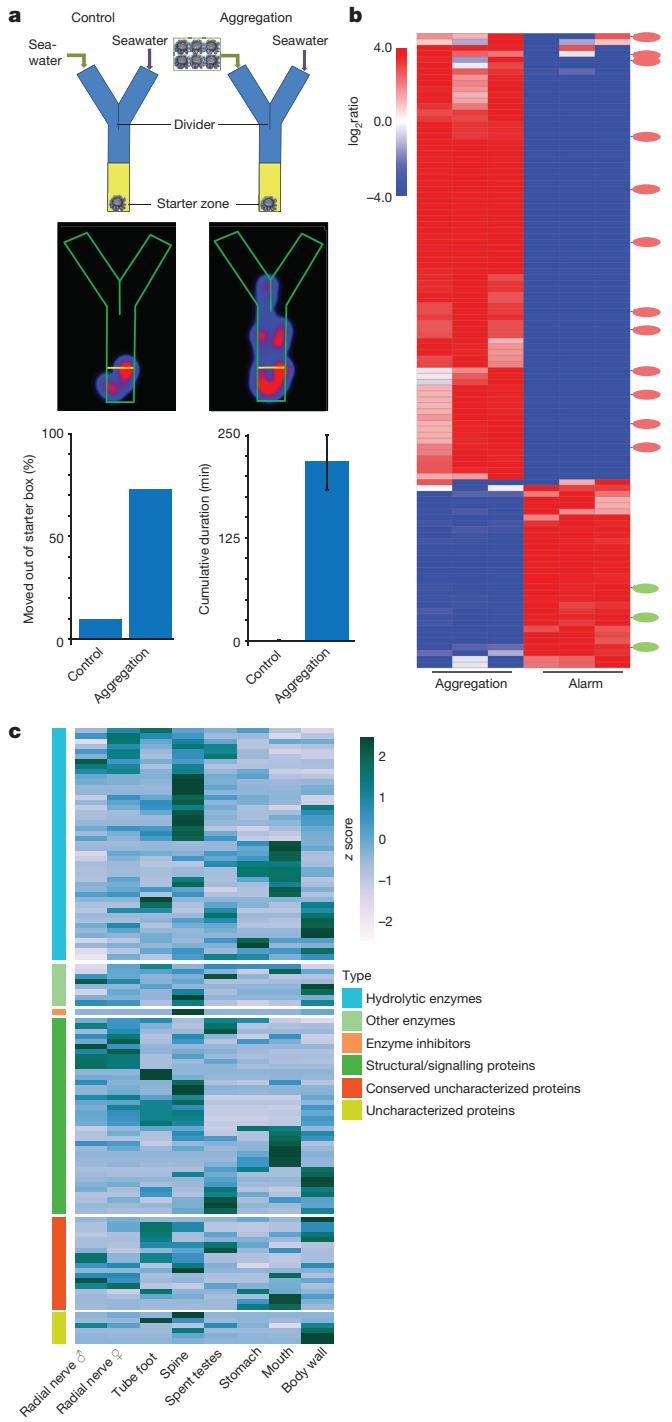


Figure 2 | Exoproteome of aggregating and alarmed COTS. a, Top, Y-maze experimental design showing arm dividers and starter zones (yellow). Middle, cumulative response of COTS over the first 45 min to seawater conditioned with six aggregating COTS (right, $n = 22$) and ambient seawater (left, control; $n = 32$). Red, the area COTS spent the most time; blue, the least time; black, no presence. Y mazes, green outline; starter zones are demarcated with yellow lines (see Supplementary Video 1). Bottom, response of COTS in a Y maze to water conditioned with aggregating COTS and ambient seawater. Movement of COTS out of the starter box ($P < 0.05$; tested with the Freeman–Halton extension of the Fisher's exact test) and the cumulative duration of movement ($P < 0.05$) over 45 min. Mean ± s.e.m. **b**, Detection of 108 secreted proteins in triplicate water samples taken around aggregating and giant triton-alarmed COTS, first three and last three lanes, respectively. EPDRs detected exclusively from aggregating COTS are marked with red ovals; EPDRs secreted from both aggregating and alarmed COTS, but more prevalent from alarmed COTS, are marked with green ovals. **c**, Tissue expression of genes encoding the 108 secreted proteins, divided into general protein classes.

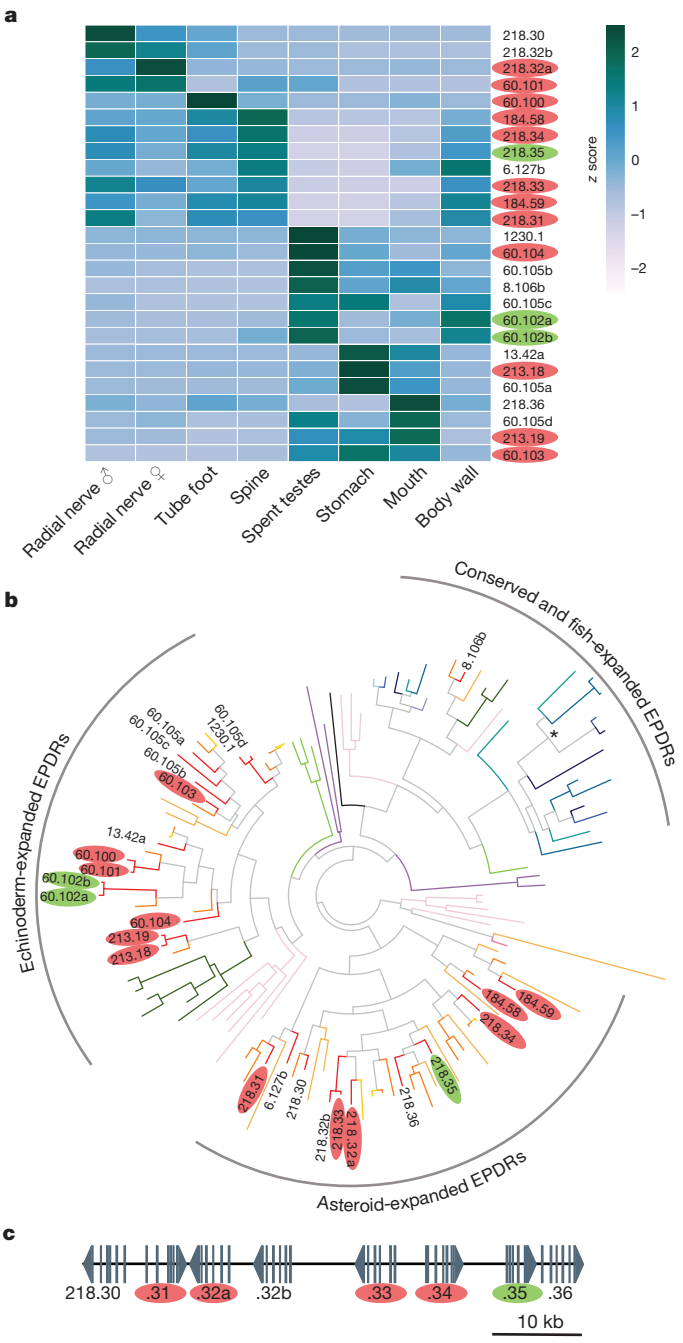


Figure 3 | Ependymin-related gene expansion and expression. a, Tissue expression of the COTS EPDR genes. **b**, Phylogeny of EPDR proteins. COTS genes are labelled and are marked with red lines; other asteroids, two shades of orange and yellow lines; sea urchins, dark green; hemichordates, light green; molluscs, pink; annelids, purple; cnidarians, black; and vertebrates, blue. The three clades to which COTS sequences belong are indicated by the outer circle. The asterisk denotes the fish-specific true ependymin clade. **c**, One of the COTS EPDR gene clusters on scaffold 218, with exons (grey bars and arrowheads), intergenic regions and introns (thin black lines) and direction of transcription (arrowhead at end of coding sequence) shown. Scale bar, 10 kb. In all panels, EPDRs secreted by COTS into the seawater are highlighted by red or green ovals as in Fig. 2b.

similarity in the approximately 24,500 predicted protein-coding genes. These results, and our phylogenetic analysis (Fig. 1d, Extended Data Fig. 1 and Supplementary Note 4), support that these geographically separated populations are the same species, *A. solaris*, within the *A. planci* species group[8] (Supplementary Note 4). Multiple sequential Markovian coalescent analysis[9] of GBR and OKI genomes suggests that

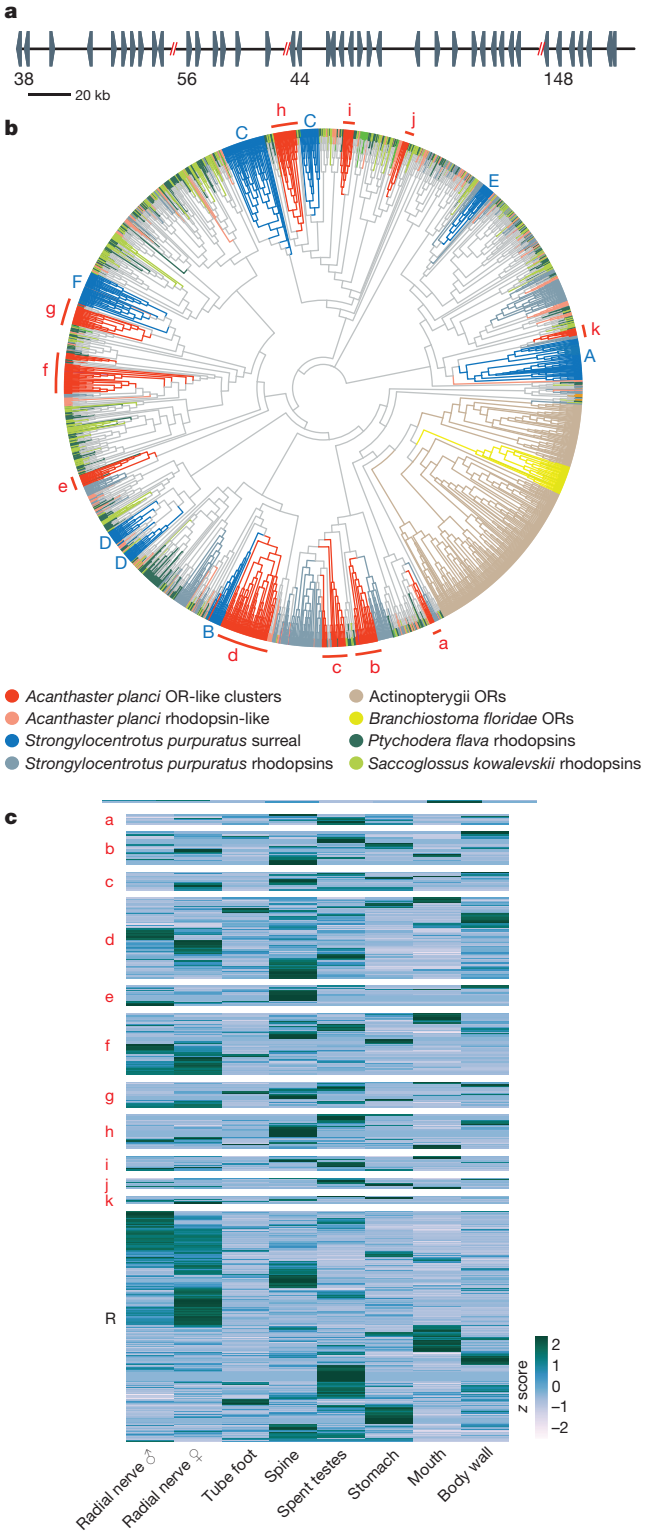


Figure 4 | Olfactory-receptor-like GPCR genes. a, Organization and orientation of single exon genes in clade c (see panel **b**) on scaffolds 38, 56, 44 and 148. Genes, grey arrowheads pointing in direction of transcription; black lines, intergenic regions. Scale bar, 20 kb. **b**, Phylogeny of ambulacrarian rhodopsin GPCRs, and *Branchiostoma* and Actinopterygii olfactory receptors (OR). The 6 sea urchin GPCRs (surreal) and 11 COTS OR-like gene clades are highlighted in blue and red, respectively. **c**, Expression of olfactory-receptor-like and rhodopsin GPCRs (R) in COTS tissues, grouped based on clades defined in **b**.

both populations declined and recovered in a similar manner during the late Pleistocene epoch (Fig. 1e and Supplementary Note 4). Given that GBR and OKI genomes are nearly identical, we treat them as one in subsequent analyses.

The COTS genome is a noteworthy addition to the existing suite of deuterostome genomes[10]. It shares many gene family and domain expansions with hemichordates and the sea urchin, although the COTS genome has fewer lineage-restricted gene and domain expansions (Extended Data Fig. 3 and Supplementary Notes 5, 6). The genome also has extensive microsynteny with other deuterostomes, including conserved Hox (Extended Data Fig. 4), ParaHox and Nkx gene clusters[11,12].

To identify candidate factors for the development of biocontrol mitigation technologies, we targeted genes potentially involved in conspecific chemical communication. COTS, like many other marine invertebrates, rely primarily on their chemosensory system to detect environmental signals including those from prey, predators, and conspecifics during reproduction[13,14]. Water-borne signals probably guide adults to form aggregations before a synchronised spawning event[15,16].

Proteins and peptides released by aggregating or alarmed COTS into the surrounding seawater were sequenced using mass spectrometry. By mapping these sequences to the genome, we identified gene products released by COTS when aggregating (244) or alarmed (77) or in both situations (73) (Supplementary Note 7). When exposed in a Y-maze assay to seawater containing putative aggregation factors, naive, normally sedentary COTS become highly active and move in the direction of the source of the conditioned seawater (Fig. 2a and Supplementary Videos 1, 2). This is consistent with water-borne factors being released during aggregation and detected by conspecifics at a distance. These released factors provide a potential basis for future biocontrol measures that include mass attraction to facilitate efficient collection and removal of COTS. This starfish also reacts rapidly and adversely to *C. tritonis*-conditioned seawater (Extended Data Fig. 5, Supplementary Note 7 and Supplementary Video 3).

Of the exoproteins identified, 108 contain signal peptides and are probably secreted in a regulated manner. 71 of these were secreted from aggregating COTS, 14 from alarmed COTS and 23 were secreted under both conditions (Fig. 2b and Supplementary Note 7). The genes encoding these secreted proteins are expressed in external tissues, including the spines, body wall and mouth, consistent with their release into the surrounding environment (Fig. 2c). Of the secreted proteins, 48 are enzymes, of which 83.3% (40) are hydrolyases, including plancitoxin-1, a type II DNase present in COTS venom[17]. In addition, 37 proteins are related to known secreted signalling and structural proteins, including 15 ependymin-related proteins (EPDRs), five lectins and four proteins related to deleted in malignant brain tumours 1. There are also 21 uncharacterized secreted proteins, 14 of which have substantial shared identity with proteins in other animals (Extended Data Fig. 6).

The detection of 15 EPDRs in the secretome of aggregating COTS (Fig. 2b) suggest that they potentially have a role in conspecific communication; an additional 11 EPDR genes in the COTS genome are also highly expressed, mostly in externally connected organs and tissues (Fig. 3a). Except for the signal peptide and a small number of spatially conserved cysteine residues, these 26 EPDRs share very low sequence similarity. Although EPDRs are present in most metazoans[18,19], this family has uniquely expanded in asteroids into at least eight orthology groups comprising two larger clades (Fig. 3b, Extended Data Fig. 7 and Supplementary Note 8). Identified asteroid orthologues share little sequence similarity, suggesting that these EPDRs have rapidly evolved to give rise to species-specific repertoires of putative communication factors. Nineteen of the COTS EPDR genes comprise two compact tandem arrays that vary markedly in sequence and expression (Fig. 3 and Supplementary Note 8). The concomitant secretion of a variety of proteases by aggregating COTS (Fig. 2c) lends support for a role of EPDRs in starfish communication; protease-released ependymin peptides act as signalling molecules in vertebrates[18,20]. Thus, EPDRs

provide a potential basis for the development of peptide mimetics for COTS biocontrol.

The COTS genome encodes approximately 950 G-protein-coupled receptor (GPCR) genes, similar to other deuterostomes[21,22] (Extended Data Table 2 and Supplementary Note 9). Many of the approximately 750 COTS rhodopsin-class GPCRs—the class comprising chemoreceptors[23]—are organized in species-specific tandem arrays of unique, single exon genes, akin to putative olfactory receptors in other deuterostomes[21,22] (Fig. 4a, b and Extended Data Fig. 8). The enrichment of COTS olfactory-receptor-like GPCR transcripts in external and sensory tissues, including the radial nerve, spine and body wall (Fig. 4c), is consistent with their role in the detection of water-borne signals. A similar enrichment is observed in sea urchins, which also appear to change behaviours because of olfactory signals[13,14,21]. The COTS radial nerve is in direct contact with the external environment[13,14,24] and displays sexually dimorphic expression of rhodopsin GPCRs (Fig. 4c), suggesting that conspecific signals are perceived differently by male and female starfish.

Sequencing of the COTS genome and proteomic analyses have enabled the identification of species-specific secreted factors associated with aggregating starfish, such as EPDRs, which can lead to the development of peptide mimetics for biocontrol measures. The high similarity of GBR and OKI genomes indicates that genome-based mitigation strategies developed for one locale can be applied throughout the species' range. These genomic data will also be useful in ecological and population studies into the causes of COTS outbreaks, contributing to regional-scale management of this coral reef pest. Further, this study suggests that species-specific secreted factors involved in conspecific communication in marine animals[25]—identified through combined genomic and proteomic approaches—have the potential to revolutionise mitigation technologies for aquatic pests more generally.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 June 2016; accepted 5 March 2017.
Published online 5 April 2017.

1. Birkeland, C. & Lucas, J. Acanthaster planci: *Major Management Problems of Coral Reefs* (CRC Press, 1990).
2. Pratchett, M. S., Caballes, C. F., Rivera-Posada, J. A. & Sweatman, H. P. A. in *Oceanography and Marine Biology: an Annual Review* Vol. 52 (eds Hughes, R. N., Hughes, D. J. & Smith, I. P.) 133–200 (CRC Press, 2014).
3. Nakamura, M., Okaji, K., Higa, Y., Yamakawa, E. & Mitarai, S. Spatial and temporal population dynamics of the crown-of thorns starfish, *Acanthaster planci*, over a 24-year period along the central west coast of Okinawa Island, Japan. *Mar. Biol.* **161,** 2521–2530 (2014).
4. De'ath, G., Fabricius, K. E., Sweatman, H. & Puotinen, M. The 27-year decline of coral cover on the Great Barrier Reef and its causes. *Proc. Natl Acad. Sci. USA* **109,** 17995–17999 (2012).
5. Uthicke, S. *et al.* Climate change as an unexpected co-factor promoting coral eating seastar (*Acanthaster planci*) outbreaks. *Sci. Rep.* **5,** 8402 (2015).
6. Kayal, M. *et al.* Predator crown-of-thorns starfish (*Acanthaster planci*) outbreak, mass mortality of corals, and cascading effects on reef fish and benthic communities. *PLoS One* **7,** e47363 (2012).
7. Babcock, R. C. & Mundy, C. N. Reproductive biology, spawning and field fertilization rates of *Acanthaster planci*. *Aust. J. Mar. Freshw. Res.* **43,** 525–533 (1992).
8. Haszprunar, G. & Spies, M. An integrative approach to the taxonomy of the crown-of-thorns starfish species group (Asteroidea: *Acanthaster*): a review of names and comparison to recent molecular data. *Zootaxa* **3841,** 271–284 (2014).
9. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46,** 919–925 (2014).
10. Cameron, R. A., Kudtarkar, P., Gordon, S. M., Worley, K. C. & Gibbs, R. A. Do echinoderm genomes measure up? *Mar. Genomics* **22,** 1–9 (2015).
11. Baughman, K. W. *et al.* Genomic organization of Hox and ParaHox clusters in the echinoderm, *Acanthaster planci*. *Genesis* **52,** 952–958 (2014).
12. Simakov, O. *et al.* Hemichordate genomes and deuterostome origins. *Nature* **527,** 459–465 (2015).
13. Dale, J. Coordination of chemosensory orientation in the starfish *Asterias forbesi*. *Mar. Freshw. Behav. Physiol.* **32,** 57–71 (1999).
14. Heinzeller, T. & Welsch, U. in *Brain Evolution and Cognition* (eds Roth, G. & Wullimann, M. F.) 41–75 (John Wiley & Sons, Inc., 2001).
15. Beach, D. H., Hanscomb, N. J. & Ormond, R. F. G. Spawning pheromone in crown-of-thorns starfish. *Nature* **254,** 135–136 (1975).
16. Campbell, A. C., Coppard, S., D'Abreo, C. & Tudor-Thomas, R. Escape and aggregation responses of three echinoderms to conspecific stimuli. *Biol. Bull.* **201,** 175–185 (2001).
17. Shiomi, K., Midorikawa, S., Ishida, M., Nagashima, Y. & Nagai, H. Plancitoxins, lethal factors from the crown-of-thorns starfish *Acanthaster planci*, are deoxyribonucleases II. *Toxicon* **44,** 499–506 (2004).
18. Shashoua, V. E. Ependymin, a brain extracellular glycoprotein, and CNS plasticity. *Ann. NY Acad. Sci.* **627,** 94–114 (1991).
19. Suárez-Castillo, E. C. & García-Arrarás, J. E. Molecular evolution of the ependymin protein family: a necessary update. *BMC Evol. Biol.* **7,** 23 (2007).
20. Adams, D. S., Hasson, B., Boyer-Boiteau, A., El-Khishin, A. & Shashoua, V. E. A peptide fragment of ependymin neurotrophic factor uses protein kinase C and the mitogen-activated protein kinase pathway to activate c-Jun N-terminal kinase and a functional AP-1 containing c-Jun and c-Fos proteins in mouse NB2a cells. *J. Neurosci. Res.* **72,** 405–416 (2003).
21. Raible, F. *et al.* Opsins and clusters of sensory G-protein-coupled receptors in the sea urchin genome. *Dev. Biol.* **300,** 461–475 (2006).
22. Krishnan, A., Almén, M. S., Fredriksson, R. & Schiöth, H. B. Remarkable similarities between the hemichordate (*Saccoglossus kowalevskii*) and vertebrate GPCR repertoire. *Gene* **526,** 122–133 (2013).
23. Rosenbaum, D. M., Rasmussen, S. G. & Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **459,** 356–363 (2009).
24. Franco, C. F., Santos, R. & Coelho, A. V. Exploring the proteome of an echinoderm nervous system: 2-DE of the sea star radial nerve cord and the synaptosomal membranes subproteome. *Proteomics* **11,** 1359–1364 (2011).
25. Hay, M. E. Marine chemical ecology: chemical signals and cues structure marine populations, communities, and ecosystems. *Ann. Rev. Mar. Sci.* **1,** 193–212 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank R. C. Wyeth for assistance in behavioural and statistical analyses, and P. Thomas-Hall for assistance in COTS husbandry. This study was supported by funds from the Australian Research Council (B.M.D. and S.M.D.), the Australian Government Department of Environment Reef Rescue Program (M.R.H. and S.F.C.), and internal funds of OIST for Marine Genomics Unit (N.S.). K.M.K. was funded by a NSF International Postdoctoral Research Fellowship.

Author Contributions M.R.H. and B.M.D. conceived and designed the project. N.S. and B.M.D. coordinated genome and transcriptome sequencing undertaken by K.W.B., E.S., S.L.F.-V., M.E.A.G., K.M.K., C.M., C.S., K.H., M.F. and M.K. S.M.D. and B.M.D. coordinated genome and transcriptome analyses undertaken by K.M.K., C.M., W.L.H., A.K., X.X. and M.Z. S.F.C. coordinated proteome analyses undertaken by T.W., M.Z. and U.B. M.R.H., S.F.C and C.A.M. undertook the behavioural studies. B.M.D., S.M.D., M.R.H., K.M.K. and K.W.B. wrote the manuscript with contributions from other all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to B.M.D. (b.degnan@uq.edu.au).

Reviewer Information *Nature* thanks M. Matz, M. Medina, C. Vogel and K. Worley for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Genome sequencing and assembly. Genomic DNA was extracted from testes and sperm of two individuals. One was collected from Rudder Reef on the northern Great Barrier Reef (GBR), Australia (16° 11′ 46.4″ S 145° 41′ 48.7″ E) and the second from Motobu, Okinawa (OKI), Japan (26° 40′ 46.1″ N 127° 52′ 46.1″ E) (Fig. 1c). These two genomes were sequenced, assembled and annotated separately using standard methods[11,26] (Extended Data Table 1, Supplementary Fig. 2.1 and Supplementary Note 2). Paired-end libraries of 40× (GBR) and 46× (OKI) coverage were sequenced on an Illumina MiSeq sequencer, generating 250 nucleotide reads. An approximately 800 bp paired-end library for GBR was sequenced in three MiSeq runs, and two paired-end libraries of around 600 bp and around 1,000 bp for OKI were each sequenced in two MiSeq runs. For the GBR genome, 3 mate-pair libraries with insert sizes of 3, 8 and 12 kb were sequenced by Macrogen, Inc. For the OKI genome, 4 mate-pair libraries with insert sizes of 1.5–4, 4–6, 6–8 and 8–12 kb were sequenced on an Illumina HiSeq 2500 sequencer. Overall read coverage including both mate-pair and paired-end libraries was 152× for GBR and 139× for OKI (Extended Data Table 1 and Supplementary Note 2).

Paired-end reads were assembled by GS *De novo* Assembler version 2.3 (Newbler, Roche) and mate-pair sequencing data were scaffolded with SSPACE 3.0 (ref. 27).

Genome size estimation. COTS genome size was estimated by three methods (Extended Data Table 1 and Extended Data Fig. 2). (1) Genome size was estimated using kmergenie[28] to determine the optimal *k*-mer length, and JELLYFISH[29] to estimate genome size (http://koke.asrc.kanazawa-u.ac.jp/HOWTO/kmergenomesize.html). (2) DNA from COTS sperm was compared with DNA from *Takifugu rubripes* and *Danio rerio* sperm[26,30] by flow cytometry. All nuclei were treated with a DAPI flow cytometry kit and a BD Cycletest Plus DNA Reagent Kit (BD Biosciences), and analysed on a BD FACSAria II cell sorter (BD Biosciences)[31] (Supplementary Note 2). (3) Genome length was estimated on the basis of total scaffold length of the assembled genomes.

Transcriptome sequencing and assembly. RNA was extracted from testes, podia, spines and stomach/mouth tissues from the same individuals from which GBR and OKI genomic DNA was obtained; RNA from the other tissues was isolated from different OKI and GBR individuals (Supplementary Note 2). Tissue-specific RNA-seq libraries were generated and sequenced on an Illumina HiSeq 2500 using standard methods[26]. *De novo* transcriptomes were assembled using Trinity version r20131110 (ref. 32). Genome-guided RNA transcripts were generated using Tuxedo[33] (Supplementary Note 2).

Gene modelling and annotation. Both GBR and OKI genomes were masked using RepeatMasker version 4.0.3 (parameters: -qq -pa 8 -gff -species 'fungi/metazoa group' -no_is). Assembled transcripts were then mapped back to either the GBR or the OKI masked genome and used to generate a consensus transcript set via PASA (version 20140417). Only transcripts with over 90% transcript coverage (parameter: min_percent_aligned) and 95% identity (parameter: min_avg_per_id) were merged. Open reading frames (ORFs) predicted from PASA-assembled transcripts using TransDecoder (https://sourceforge.net/p/transdecoder/) were used to train Augustus to generate gene predictions for each genome (Supplementary Note 2). Additionally, all core eukaryotic genes were mapped to each genome using CEGMA (version 2.4). CEGMA predictions were used to train SNAP (version 2013-11-29). Unsupervised genes were also predicted using GeneMarkES (version 20120203).

Final gene predictions were generated using EVM[34] by combining (1) *ab initio* predictions by Augustus, SNAP and GeneMarkES, (2) consensus transcripts generated by PASA based on combined transcriptomes of both populations and (3) TransDecoder best ORF predictions based on PASA consensus transcripts. A genome browser is available at: http://marinegenomics.oist.jp/cots/.

Estimation of intra- and inter-genome heterozygosity. Overall genome heterozygosity was estimated by single-nucleotide polymorphism (SNP) analysis, by mapping paired-end reads onto the scaffolds using BWA[35]. SNPs were called and analysed using Stools[36]. Further SNP analysis was done by mapping OKI reads to the GBR genome, and vice versa. OKI and GBR COTS genomic assemblies were aligned by reciprocal BLASTN+[37]. Scaffolds >10 kb or alignments with *E* values $<1\times10^{-5}$ were analysed. SNP distribution across each genome was also compared (Supplementary Note 3).

LiftOver analysis between GBR and OKI genomes. Comparison of OKI and GBR gene models was performed using batch coordinate conversion (LiftOver) from the UCSC Genome Browser Utilities[38]. LiftOver settings were optimised to generate the maximal number of significant gene model matches between the two genomes (Supplementary Note 3).

Phylogenomic analysis. Phylogenomic methods followed the general approach of ref. 39 (Supplementary Note 4). Transcriptomes from refs 40 and 41 were assembled as described previously. Other publicly available Illumina transcriptomes were digitally normalized and assembled using the 13April 2014 release of Trinity[42]. Contigs were translated with TransDecoder (https://sourceforge.net/p/transdecoder/) using Pfam 27 as a guide. Predicted proteins and translated transcriptomes were combined for each of the COTS.

For orthology inference, we employed HaMStR 13 (ref. 43) using the 1,032 'model organisms' profile hidden Markov models (pHMMs). Sequences matching the pHMM of an orthology group were then compared to the proteome of *Homo sapiens* using BLASTP with the default settings implemented by HaMStR. If the *H. sapiens* amino acid sequence contributing to the pHMM was the best BLASTP hit in each of these back-BLASTs, the sequence was then assigned to that orthology group.

Sequences in orthology groups that were shorter than 50 amino acids were discarded. Redundant identical sequences were removed with UniqHaplo (http://raven.iab.alaska.edu/~ntakebay/) leaving only the most complete, unique sequences for each taxon. In cases where one of the first or last 20 characters of an amino acid sequence was an X (corresponding to a codon with an ambiguity, gap or missing data), all characters between the X and that end of the sequence were deleted and treated as missing data. Each orthology group was then aligned with MAFFT (mafft–auto–localpair–maxiterate 1000)[44]. Alignments were trimmed with Aliscore[45] and Alicut[46] to remove ambiguously aligned regions. Subsequently, any putatively mistranslated sequence regions were deleted; these were regions of 20 or fewer amino acids in length surrounded by ten or more gaps on either side. Next, alignments that were shorter than 50 amino acids in length were discarded. Last, we deleted sequences that did not overlap with all other sequences in the alignment by at least 20 amino acids, starting with the shortest sequences. Finally, orthology groups, which were sampled for fewer than 15 taxa after these filters, were discarded.

To screen putative orthology groups or evidence of paralogy, an 'approximately-maximum likelihood' tree was inferred for each remaining alignment using FastTree 2 (refs 47, 48) with the 'slow' and 'gamma' options. PhyloTreePruner[48] was then used to generate a tree-based approach to screen each candidate orthology group for evidence of paralogy. Nodes with support values below 0.95 were collapsed into polytomies and the maximally inclusive subtree was selected where all taxa were represented by no more than one sequence or, in cases where more than one sequence was present for any taxon, all sequences from that taxon formed a monophyletic clade or were part of the same polytomy. Putative paralogues (sequences falling outside of this maximally inclusive subtree) were then deleted. In cases where multiple sequences from the same taxon formed a clade or were part of the same polytomy, all sequences except the longest were deleted.

Phylogenetic analysis was conducted using ML with RAxML 7.7.6 (ref. 49). Matrices were partitioned by gene and the PROTGAMMALG model was used for all partitions. For each analysis, the tree with the best likelihood score after 10 random addition sequence replicates was retained and topological robustness (that is, nodal support) was assessed with 100 replicates of nonparametric bootstrapping (the -f a command line option was used).

Multiple sequential Markovian coalescent analysis. The multiple sequential Markovian coalescent analysis method[9], using default parameters, was used to infer the historical effective population sizes of the OKI and GBR COTS. All paired-end reads were aligned to each soft-masked genome using Bowtie 2 (ref. 50). SAMtools(version 1.19)[36] was used to filter the unmapped reads and reads with minimum base and mapping quality scores of 20. Bcftools (version 1.19)[36], of the SAMtools package, was then used to call genotype for each position. Real historical time and effective population size were estimated assuming a generation time of 3 years and a substitution mutation rate of 1.0×10^{-8} per generation, which was based on an estimated genome size of 431 Mb[51] (Supplementary Note 4).

Protein domain annotations. Protein domains were downloaded from *Mnemiopsis leidyi*, *Amphimedon queenslandica*, *Trichoplax adhaerens*, *Nematostella vectensis*, *Lottia gigantea*, *Lingula anatina*, *Capitella teleta*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Brachiostoma floridae*, *Ciona intestinalis*, *D. rerio*, *Xenopus tropicalis*, *H. sapiens*, *Saccoglossus kowalevskii*, *Ptychodera flava*, *Strongylocentrotus purpuratus* and *A. planci* (Supplementary Note 5) and annotated using HMMER of all known protein domains in the Pfam database (version 29.0)[52]. If a domain occurred multiple times in a protein sequence, it was counted only once (Supplementary Note 5). To exclude transposon-derived domains, mispredictions or unknown domains, we removed Pfams that were categorised as 'unknown', 'not named', 'uncharacterized', 'transposase', 'helitron', 'helicase', 'DUF', or 'DDE_Tnp'. We then iteratively conducted a Fisher exact test using R[53], comparing the number of counts in Pfam families found in species, to the background, defined as the average of the counts in the remaining species (Supplementary Note 5).

To assess differences in protein domains across metazoan genomes, we examined protein domain expansion and contraction in each species, based on the total number of unique genes that each Pfam domain contained. We used the scaled value for each individual Pfam domain as a proxy for expansion, whereby any value greater than the mean was considered a domain expansion (Extended Data Fig. 3).

Analysis of tissue transcriptomes. Seven and ten tissue transcriptomes were sequenced from GBR and OKI, respectively (Supplementary Notes 2 and 6). Trimmed reads from all transcriptomes were mapped to GBR and OKI gene models, and fragments per kb of transcript per million mapped reads (FPKM) were calculated for all genes in all transcriptomes. Relationships between tissue-type and geographic location (that is, GBR versus OKI) were determined using Euclidean distance and principle component analyses (PCAs) of FPKM values for all genes shared between the two genomes based on the LiftOver analysis using the R package, DESeq2 (ref. 54). On the basis of the PCAs, we selected the following eight tissue transcriptomes for further analysis: male and female radial nerves, tube foot (podia), spine, body wall, stomach, mouth and spent testes (Supplementary Note 6). The presented order of these tissues was derived from the Euclidean distance[54] of these transcriptomes on the basis of both expression and Pfam[52] similarity (Supplementary Note 6).

***In silico* prediction of secreted and cleaved proteins.** *A. planci* protein N-terminal signal sequences were predicted using SignalP 4.1 (ref. 55) (neural network and hidden Markov model algorithms), Predisi[56] and Phobius[57], while transmembrane domains were determined using TMHMM[58] and HMMTOP[59]. A protein was designated as secreted only when it met the criteria of both SignalP and Predisi, and did not have a transmembrane domain. Proteolytic cleavage sites and post-translational modifications (PTMs) were determined on the basis of homology to other known proteins and predicted with the Neuropred tool (http://neuropro teomics.scs.illinois.edu/neuropred.html) with a cleavage probability >0.8.

Collection of samples and exoproteome analysis. COTS aggregation-conditioned seawater was produced from three adult COTS (300–350 mm diameter) that were kept in a single 60 l flow-through glass tank (780 × 380 mm) for 24 h. Water flow was stopped for 1 h before draining 51 l; the COTS remained *in situ* to minimise release of alarm-related chemistry. Conditioned water was acidified with 0.1% trifluoroacetic acid (TFA), then filtered through a 0.45-μm PVDF membrane (Millipore) and absorbed onto a C18 Sep-Pak, 5 g sorbent per cartridge, 37–55 μm particle size (Waters). Filter cartridges were washed with 100% methanol between samples to remove any carryover. Biomolecules were eluted with 70% acetonitrile:0.5% acetic acid, and then lyophilised and stored at −20 °C. COTS alarm-conditioned seawater was produced when a single adult COTS (300–350 mm diameter)—kept in a single 60 l flow-through glass tank for 24 h before the water flow was stopped for 1 h—was then exposed to a giant triton (separated by a mesh divider) for 1 h. The giant triton was removed and 51 l was drained; the COTS remained *in situ*. Conditioned seawater was acidified (0.1% TFA), then biomolecules purified and lyophilised as described above. Procedures for collection of both treatments were repeated three times ($n = 3$). Reconstituted samples containing about 1 mg exoproteins (evaluated by NanoDrop 2000, Thermo Scientific) in 100 μl extraction buffer (8 M urea, 0.8 M NH_4HCO_3, pH 8.0), were processed by reduction, alkylation, trypsin digestion, SCX-HPLC and then NanoHPLC-ESI-Triple Time-of-Flight mass spectrometry (see Supplementary Note 7 for details).

Protein identification and quantification. The protein database used for MS/MS data analysis was derived from both GBR and OKI, (Supplementary Note 2, Supplementary Table 7.2a–n). A composite target−decoy database was built with the forward and reverse sequences for calculating the false discovery rate (FDR). MS/MS data were imported to the PEAKS studio (Bioinformatics Solutions Inc., version 7.0) with the assistance of the MS Data Converter (Beta 1.3, http://sciex.com/software-downloads-x2110). *De novo* sequencing of peptides, database search and characterization of unspecified PTMs were used to analyse the MS/MS data; the FDR was set to ≤ 1%, and ($-10\log P$) was calculated accordingly where P is the probability that an observed match is a random event. The PEAKS studio parameters are defined in Supplementary Note 7.

The quantitative analysis of proteins was carried out using the label-free quantification module (PEAKS Q[60]) of PEAKS studio version 7.0, which is based on the relative intensities of featured peptides detected in multiple samples. The detection of features was separately performed on each sample and the expectation–maximisation algorithm[61,62] was used to identify overlapping features. Then, an alignment algorithm[63] was used to align the features of the same peptide from different samples. The extracted proteins in different replicate samples were quantified as described above; for each sample, 1.5 μg of protein was analysed using LC–MS/MS. Biological triplicate samples of aggregation and alarm were used in tandem repeats for LC–MS/MS procedure as described above, and the relative concentrations of proteins were compared and presented as the final results. The mass shift between different runs was set to 50 p.p.m., and 1.0 min was used for evaluating the retention time shift tolerance. Featured peptides with a FDR threshold of 1%, including PTMs mentioned above, were included in the quantitative analysis. Validation of quantitative analysis was performed as described in Supplementary Note 7.

Behavioural response of COTS to signals from starfish aggregations. Adult COTS were collected from various regions of the central GBR by the Australian Marine Parks Tourist Operators (AMPTO) Crown-of-Thorns Starfish Control Program and transported to the Australian Institute of Marine Science (AIMS) SeaSim aquarium precinct (www.aims.gov.au/seasim). COTS were housed in outdoor flow-through seawater aquaria at ambient conditions with temperatures of 26–29 °C and salinity averaging 35 p.p.t. When moved to indoor experimental systems, water temperature and photoperiod were simulated as per ambient natural outdoor conditions. COTS were not fed in captivity; therefore they were replaced fortnightly with freshly collected specimens to minimise behavioural changes owing to partial starvation.

Behavioural responses of starfish were examined in black fibreglass tanks (4.4 × 2.3 m) containing 6,070 l of seawater and a Y-maze (main channel 1.75 m long, channel width of 0.6 m, each arm 2.35 m long) to test for behavioural changes, such as motivation for, or direction of, movement. Seawater supply to each arm of the Y-maze was balanced to give a flow of 5 cm s^{-1} moving towards the main channel, and a 0.8 m divider at the base of the arms ensured no backflow from one arm into the other. One arm of the Y-maze was fed with ambient seawater directly from a pipe (control). The water to the second arm passed through a 250 l header tank (1 × 0.5 × 0.5 m) that was either empty (control) or contained six adult starfish that had been in place a minimum of 24 h before the experiment. The COTS in the header tank formed aggregations. At the start of the experiment, a test subject starfish was placed into the distal end of the main channel in a 'starter box', which was 0.6 m^2, and its movement recorded for up to 8 h on video. As COTS are primarily nocturnal, experiments were conducted at the end of the daylight illumination period and filmed during the nocturnal period. As the aggregating starfish were in an inaccessible header tank, the test subject could not visually detect or physically join the aggregation.

The tanks were illuminated with a bank of 850 nm infrared LED lamps (CMVision, Model IR-200LF/WF) filmed with an infrared acA1300-60gmNIR camera (Basler AG) fitted with a 4.4−11 mm/F1.6 1/1.8″ manual C-mount CCTV lens (Kowa Optical Products Ltd) and 850 nm cut-off filter (Helipan ES43). The infrared spectrum is beyond the detection range of starfish photoreception (425–580 nm) and therefore does not interfere with overlying photoperiod lighting[64,65]. The tanks were exposed to regional photoperiod changes (19.25° S, 146.8° E) with full sunlight spectrum plasma units (Luxim Model GRO-41-01, Luma America) with crepuscular twilight ramping. In this arrangement, only the reflection of infrared from the body of the starfish is detected by the infrared sensitive camera. Video footage was captured and analysed with Ethovision XT (http://www.noldus.com/animal-behavior-research/products/ethovision-xt). The experiments were conducted for $n \geq 10$. The results were summed to determine the overall typical behaviour. Statistical analyses were performed using Ethovision XT and SPSS (version 20, IBM)[66].

Motivation was determined if the test subject moved out of the original starter box. Changes in motivation were graphically represented in heat maps where the frequency of a specific position in a 2D space was visualized as a colour representing the minimum and maximum per-pixel frequency over the duration of the experiment. The spectrum variation was set from dark blue (minimum) to dark red (maximum); heat maps are primarily qualitative. Activity, how long and how frequent a subject has been active, was determined by the number of changed pixels for a current sample divided by the total number of pixels in the arena. Activity is not necessarily an indication of total distance moved, as anxiety movement will be detected as activity and such behaviour is typically triggered by a stimulus. A threshold of >60% active time was imposed as a measure of 'highly active'. For example, test subject starfish were clearly agitated when exposed to seawater conditioned with aggregated COTS and their behaviour was indicative of searching for the source; even though starfish entered a particular Y-maze arm, few remained or settled within that arm but rather exhibited continued mobility.

Analysis of EPDR genes. Potential COTS EPDR genes were identified from (1) transcriptomes via BLAST searches using partial exoprotein sequences, (2) the genome assembly via HMMER3.1 (ref. 67) searches using the ependymin pHMM (Pfam[52] accession PF00811.14), and (3) HMMER searches on ORFs exceeding 50 nucleotides extracted from genome scaffolds using the getorf tool from the EMBOSS package 6.5.7 (ref. 68). EPDR transcripts were then used as queries to identify the correct intron/exon architecture of the genes in the genome assemblies (schematic created using FancyGene[69]). An alignment of the manually curated GBR EPDRs was created using AliView[70] (see Supplementary Note 8) and assessed for the possession of signal peptides using SignalP 4.1 (ref. 55) (sequences with

a HMMER *E*-value greater than 1×10^{-5} were found to align poorly and were removed). Sequence logos were created using WebLogo[71].

EPDR pHMM searches were performed on predicted protein datasets from whole-genome data from select species and on ambulacrarian transcriptomes used in the phylogenomic analyses (see above) using an *E*-value cutoff of 1×10^{-5} (Supplementary Note 8). Sequences were aligned and identical or very similar sequences within a species were removed (Supplementary Note 8). Maximum likelihood trees were performed as described above using RAxML 7.7.6, automatic model selection (VT) and 1,000 bootstrap replicates. Bayesian analysis was performed using MrBayes 3.2.6, with automatic (mixed) model selection (BLOSUM) and sampling every 10,000 generations until convergence (standard deviation of split frequencies <0.01, 2.3 million generations). Topologies of the resulting phylogenetic trees were largely congruent. A sequence logo was created for each subclass of EPDRs using WebLogo[71].

Identification and analysis of GPCRs and olfactory receptor-like genes. Methods for GPCR identification followed the general approach of ref. 72. We screened the protein models of both OKI and GBR genomes, *B. floridae*, *H. sapiens*, *S. kowalevskii*, *P. flava*, and *S. purpuratus* using PFAM-scan.pl (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools/) against version 27 of the Pfam-A database. Sequences annotated by PFAM_scan.pl with domains in the GPCR_A Pfam clan (CL0192), and with at least 5 transmembrane regions according to HMMTOP[59], were considered to be GPCRs and were further annotated with InterProScan 5.8–49.0 (ref. 73).

Many sequences were annotated as rhodopsin-like and therefore sequences annotated with PFAM 00001 were trimmed specifically to the region annotated as '7 transmembrane receptor (rhodopsin family)' by InterProScan and subsequently parsed into subfamilies using FastOrtho (http://enews.patricbrc.org/fastortho/), a modified version of OrthoMCL[74] with an inflation parameter of 1.5. This resulted in the identification of 957 groups of at least two GPCRs in the rhodopsin family (7tm_1) (Supplementary Note 9). The number of rhodopsin genes in each group for each species was visualized using Pretty Heatmap (https://cran.r-project.org/web/packages/pheatmap) in R[53].

Other GPCRs were similarly trimmed to the transmembrane receptor region for phylogenetic analysis. The annotations used for trimming each of these GPCRs were as follows: 7TM_3/Glutamate (PF00003); Dicty_CAR (PF05462) 'G-protein coupled receptors family 2 profile 2'; Frizzled (PF01534) 'Frizzled/Smoothened family membrane region'; GpcrRhopsn4 (PF10192) 'rhodopsin-like GPCR transmembrane domain'; Lung_7-TM_R (PF06814) 'Lung seven transmembrane receptor'; and Ocular_alb (PF02101) 'Ocular albinism type 1 protein'. Phylogenetic analyses were conducted on the transmembrane receptor region for each GPCR family using FastTree 2 (ref. 56) with the slow and gamma model options (Supplementary Note 9).

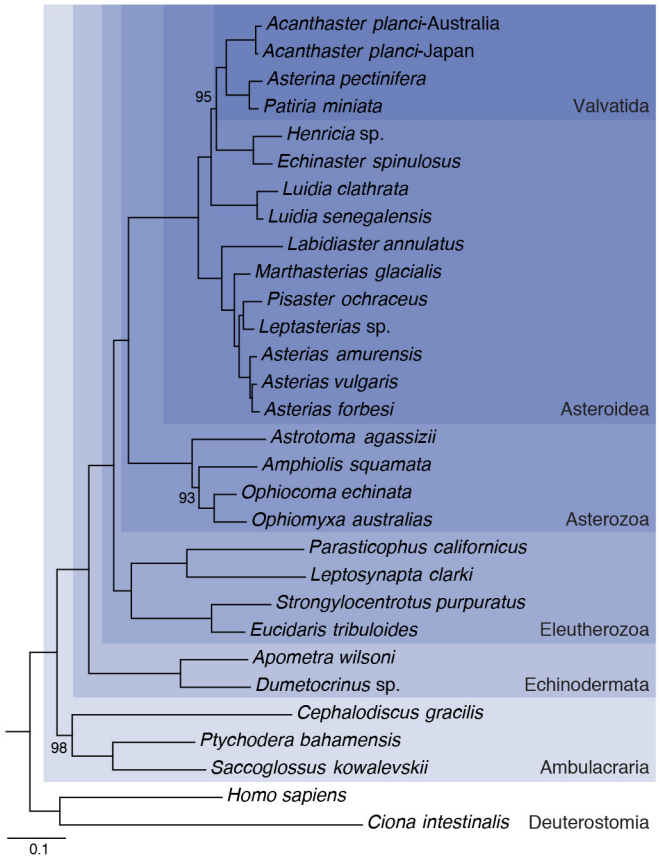
To identify putative olfactory-receptor-like genes in the COTS genomes, we followed the approach of refs 75 and 76 with modifications to incorporate the approaches of ref. 21 (Supplementary Note 9). We built 13 distinct pHMMs from previously curated olfactory-receptor repertoires comprising fishes (fugu, medaka, pufferfish, zebrafish and stickleback), amphioxus, sea urchin ('Specific rapidly expanded lineages of rhodopsin family' GPCRs (surreal GPCRs) groups 1–6) and manually curated olfactory receptors from Swiss-Prot. All non-redundant hits were retrieved from the combined results of all pHMM searches (Supplementary Note 9).

To distinguish olfactory receptors from the other 12 rhodopsin subfamilies (non-olfactory receptors), we conducted a BLASTP search (default settings) against a local database containing all class A or rhodopsin-like GPCRs from the Swiss-Prot database, followed by an all-against-all BLASTP comparison of COTS rhodopsin-like GPCRs. To determine if these COTS paralogue clusters of class A GPCRs are species-specific, and to resolve their relationship to other class A deuterostome GPCRs, we conducted a phylogenetic analysis. The dataset included class A rhodopsin-like GPCRs from *S. purpuratus*, which includes the surreal GPCRs, and two hemichordates (*P. flava* and *S. kowalevskii*), as well as olfactory receptors from fish (fugu, medaka, pufferfish, zebrafish and stickleback) and amphioxus. All sequences that contained 5 to 7 transmembrane helices were considered complete and were included in the phylogenetic analysis. The final dataset (2,615 sequences) was aligned using MAFFT version 7 with the FFT-NS-2 progressive method[77] and the alignment was manually trimmed to conserved blocks of transmembrane regions for phylogenetic tree reconstruction. The maximum likelihood phylogenetic tree was built using MEGA7 using a Poisson model with rate uniformity across sites[78].

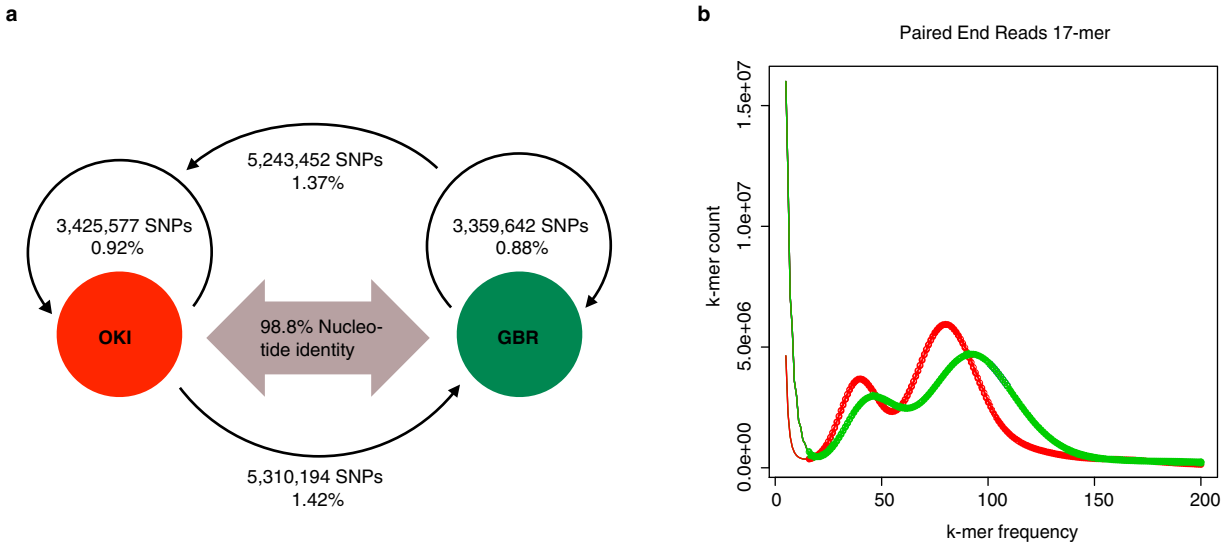
Data availability. The *Acanthaster planci* genome sequence can be accessed at DDBJ and Bioproject (NCBI) as PRJDB3175, which links to the Sequence Read Archive for all genome raw and assembled scaffold (nucleotide) data for GBR and OKI, under BioSamples SAMD00020546 and SAMD00054104, respectively. All tissue transcriptome data are available in the NCBI Sequence Read Archive database under accession DRA005145. A genome browser is available at http://marinegenomics.oist.jp/cots/. Proteomic data are available through ProteomeXchange with identifier PXD005409 (http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD005409). Data files for EPDRs, olfactory receptors and phylogenomic analyses are on FigShare https://figshare.com/s/f3b5caefbba303b99349, https://figshare.com/s/8418f468219eb598a306 and https://figshare.com/s/b58c9a71fbea8ed7268d, respectively.

26. Shoguchi, E. *et al.* Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Curr. Biol.* **23,** 1399–1408 (2013).
27. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).
28. Chikhi, R. & Medvedev, P. Informed and automated *k*-mer size selection for genome assembly. *Bioinformatics* **30,** 31–37 (2014).
29. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27,** 764–770 (2011).
30. Ciudad, J. *et al.* Flow cytometry measurement of the DNA contents of G0/G1 diploid cells from three different teleost fish species. *Cytometry* **48,** 20–25 (2002).
31. Davies, D. & Allen, P. in *Flow Cytometry* (ed. Macey, M. G.) 165–179 (Springer, 2007).
32. Haas, B. J. *et al. De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols* **8,** 1494–1512 (2013).
33. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protocols* **7,** 562–578 (2012).
34. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).
35. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26,** 589–595 (2010).
36. Li, H. *et al.* The sequence alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
37. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).
38. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
39. Kocot, K. M. *et al.* Phylogenomics reveals deep molluscan relationships. *Nature* **477,** 452–456 (2011).
40. Cannon, J. T. *et al.* Phylogenomic resolution of the hemichordate and echinoderm clade. *Curr. Biol.* **24,** 2827–2832 (2014).
41. O'Hara, T. D., Hugall, A. F., Thuy, B. & Moussalli, A. Phylogenomic resolution of the class Ophiuroidea unlocks a global microfossil record. *Curr. Biol.* **24,** 1874–1879 (2014).
42. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).
43. Ebersberger, I., Strauss, S. & von Haeseler, A. HaMStR: profile hidden Markov model based search for orthologs in ESTs. *BMC Evol. Biol.* **9,** 157 (2009).
44. Katoh, K., Kuma, K., Toh, H. & Miyata, T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33,** 511–518 (2005).
45. Misof, B. & Misof, K. A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Syst. Biol.* **58,** 21–34 (2009).
46. Kück, P. ALICUT: a Perlscript which cuts ALISCORE identified RSS version 2.0. *Zoologisches Forschungsmuseum Alexander Koenig (ZFMK)* https://www.zfmk.de/en/research/research-centres-and-groups/aliscore (2009).
47. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5,** e9490 (2010).
48. Kocot, K. M., Citarella, M. R., Moroz, L. L. & Halanych, K. M. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol. Bioinform. Online* **9,** 429–435 (2013).
49. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
50. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
51. Lynch, M. Evolution of the mutation rate. *Trends Genet.* **26,** 345–352 (2010).
52. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–D285 (2016).
53. R Development Core Team. R: a language and environment for statistical computing. http://www.R-project.org/ (Foundation for Statistical Computing, Vienna, 2008).
54. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 550 (2014).
55. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8,** 785–786 (2011).
56. Hiller, K., Grote, A., Scheer, M., Münch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32,** W375–W379 (2004).

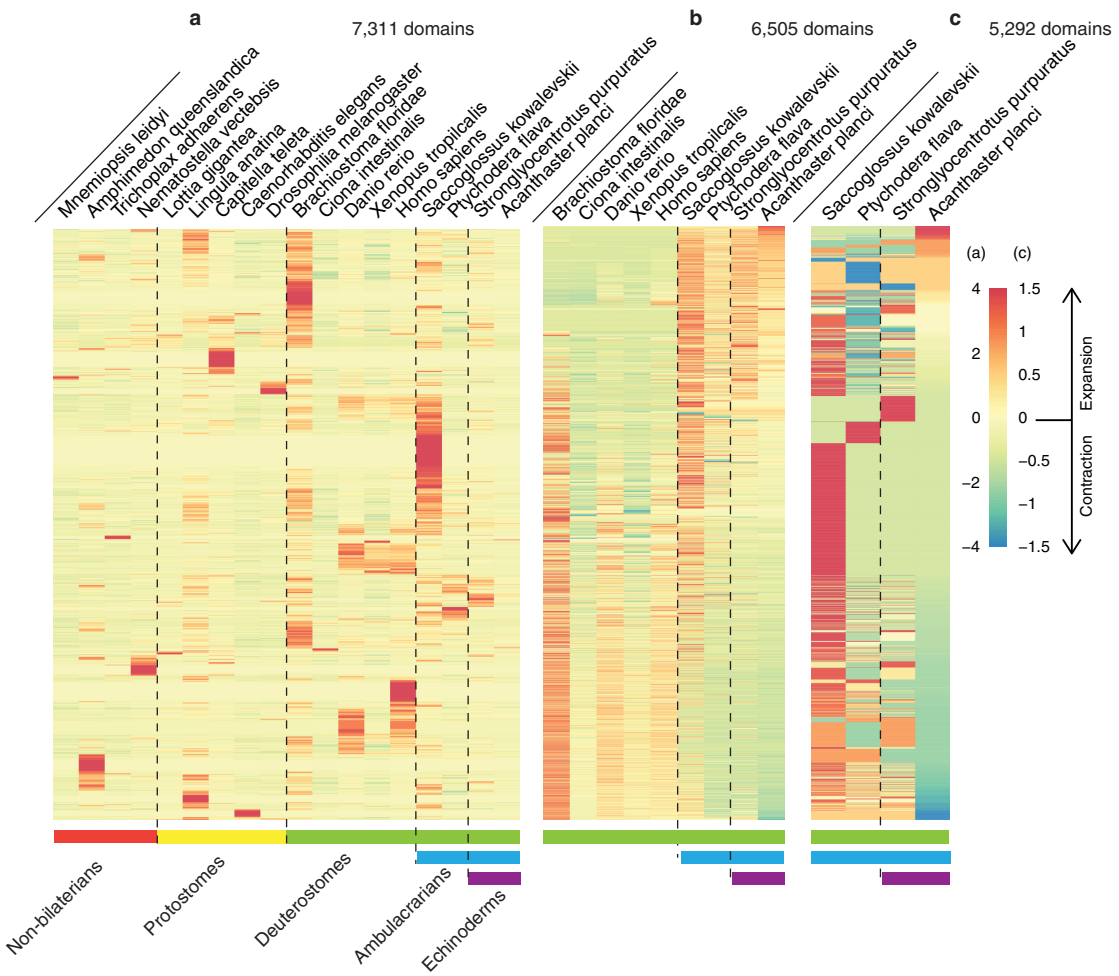
57. Käll, L., Krogh, A. & Sonnhammer, E. L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* **338,** 1027–1036 (2004).
58. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
59. Tusnády, G. E. & Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics* **17,** 849–850 (2001).
60. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26,** 1367–1372 (2008).
61. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75,** 4646–4658 (2003).
62. Nesvizhskii, A. I., Vitek, O. & Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods* **4,** 787–797 (2007).
63. Lin, H., He, L. & Ma, B. A combinatorial approach to the peptide feature matching problem for label-free quantification. *Bioinformatics* **29,** 1768–1775 (2013).
64. Petie, R., Hall, M. R., Hyldahl, M. & Garm, A. Visual orientation by the crown-of-thorns starfish (*Acanthaster planci*). *Coral Reefs* **35,** 1139–1150 (2016).
65. Garm, A. & Nilsson, D.-E. Visual navigation in starfish: first evidence for the use of vision and eyes in starfish. *Proc. R. Soc. B* **281,** 20133011 (2014).
66. Gerber, S. B. & Finn, K. V. *Using SPSS for Windows: Data Analysis and Graphics* (Springer, 2013).
67. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14,** 755–763 (1998).
68. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the european molecular biology open software suite. *Trends Genet.* **16,** 276–277 (2000).
69. Rambaldi, D. & Ciccarelli, F. D. FancyGene: dynamic visualization of gene structures and protein domain architectures on genomic loci. *Bioinformatics* **25,** 2281–2282 (2009).
70. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30,** 3276–3278 (2014).
71. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome Res.* **14,** 1188–1190 (2004).
72. Krishnan, A. *et al.* The GPCR repertoire in the demosponge *Amphimedon queenslandica*: insights into the GPCR system at the early divergence of animals. *BMC Evol. Biol.* **14,** 270 (2014).
73. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30,** 1236–1240 (2014).
74. Li, L., Stoeckert, C. J., Jr & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
75. Niimura, Y. On the origin and evolution of vertebrate olfactory receptor genes: comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* **1,** 34–44 (2009).
76. Niimura, Y. Identification of chemosensory receptor genes from vertebrate genomes. *Methods Mol. Biol.* **1068,** 95–105 (2013).
77. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
78. Kumar, S., Stecher, G. & Tamura, K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* **33,** 1870–1874 (2016).



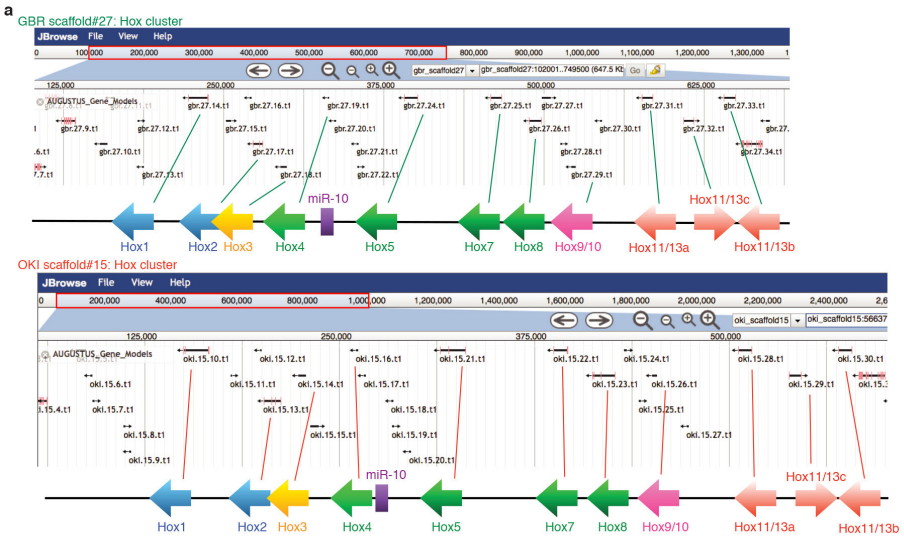
Extended Data Figure 1 | Deuterostome phylogeny showing placement of *Acanthaster* within asteroids. A concatenated supermatrix of 427 genes (95,585 amino acids, 45.16% missing data) recovering a fully resolved tree. With exception of support for hemichordate monophyly (bootstrap support value = 98%), we found maximal support for all phylum- and class-level taxa. Species sampled, annotations and characteristics of each gene analysed are presented in Supplementary Note 4. Bootstrap support values below 100 are shown. Scale bar: 0.1 substitutions per site.



Extended Data Figure 2 | *Acanthaster planci* heterozygosity. a, Single-nucleotide polymorphism (SNP) analysis showing the number of SNPs identified within and between OKI and GBR genomes. Percentage heterozygosity within these genomes and the level of nucleotide variance between genomes are shown. See Supplementary Note 2 for further details. **b**, *k*-mer (17-mer) plot. The GBR (green) and OKI (red) genomes were estimated to be 441 and 421 Mb, respectively.

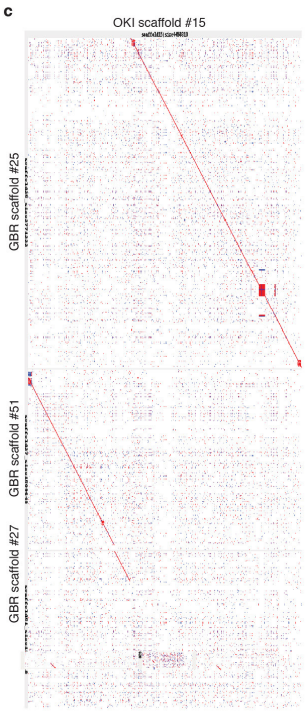


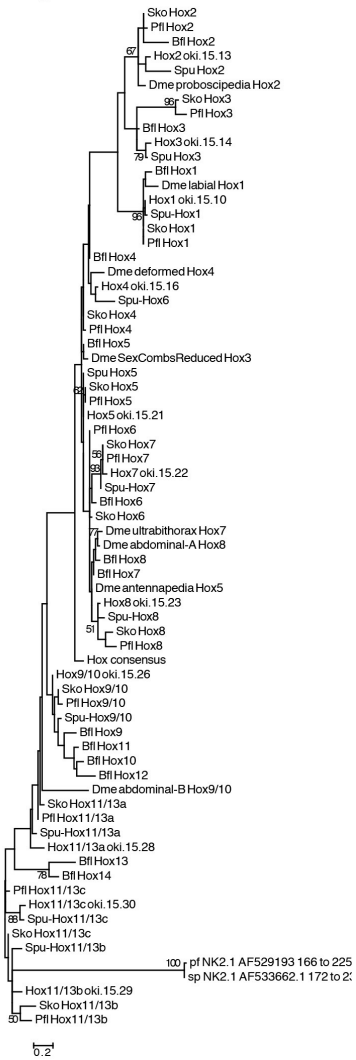
Extended Data Figure 3 | Pfam enrichment in the genomes of selected metazoans displayed as relative abundance heat maps. a, Comparison of metazoans. **b**, Comparison of deuterostomes. **c**, Comparison of ambulacrarians. See Supplementary Note 5 for further details of methods and analyses.



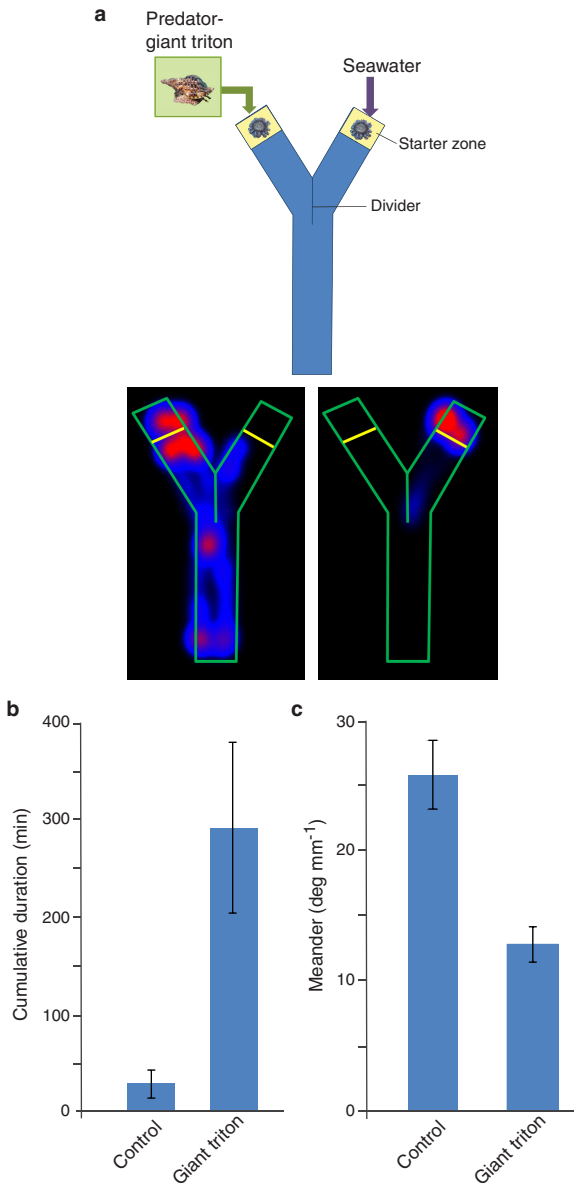
b

COTS Hox	OKI	GBR
HOX1	15.10	27.14
HOX2	15.13	27.17
HOX3	15.14	27.18
HOX4	15.16	27.19
HOX5	15.21	27.24
HOX6	-	-
HOX7	15.22	27.25
HOX8	15.23	27.26
HOX9/10	15.26	27.29
HOX11.13A	15.28	27.31
HOX11.13B	15.30	27.33
HOX11.13C	15.29	27.32

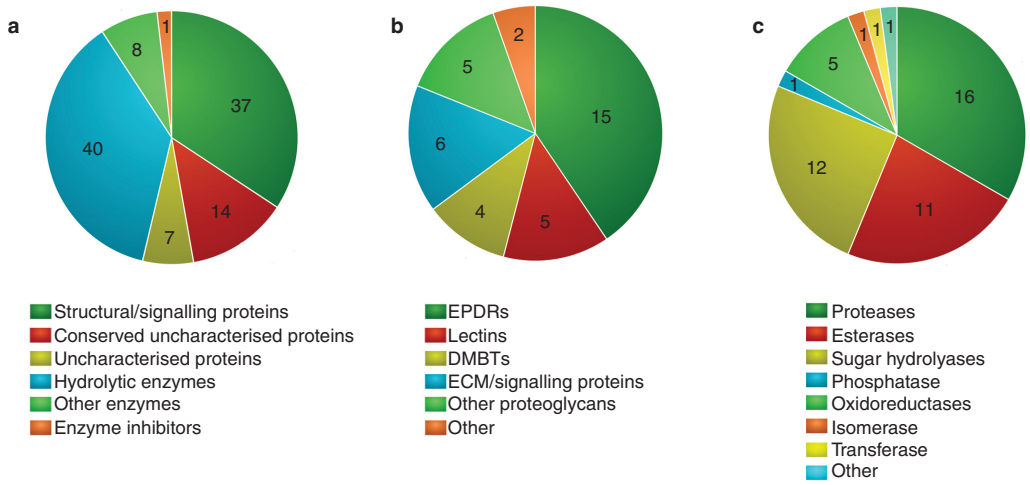




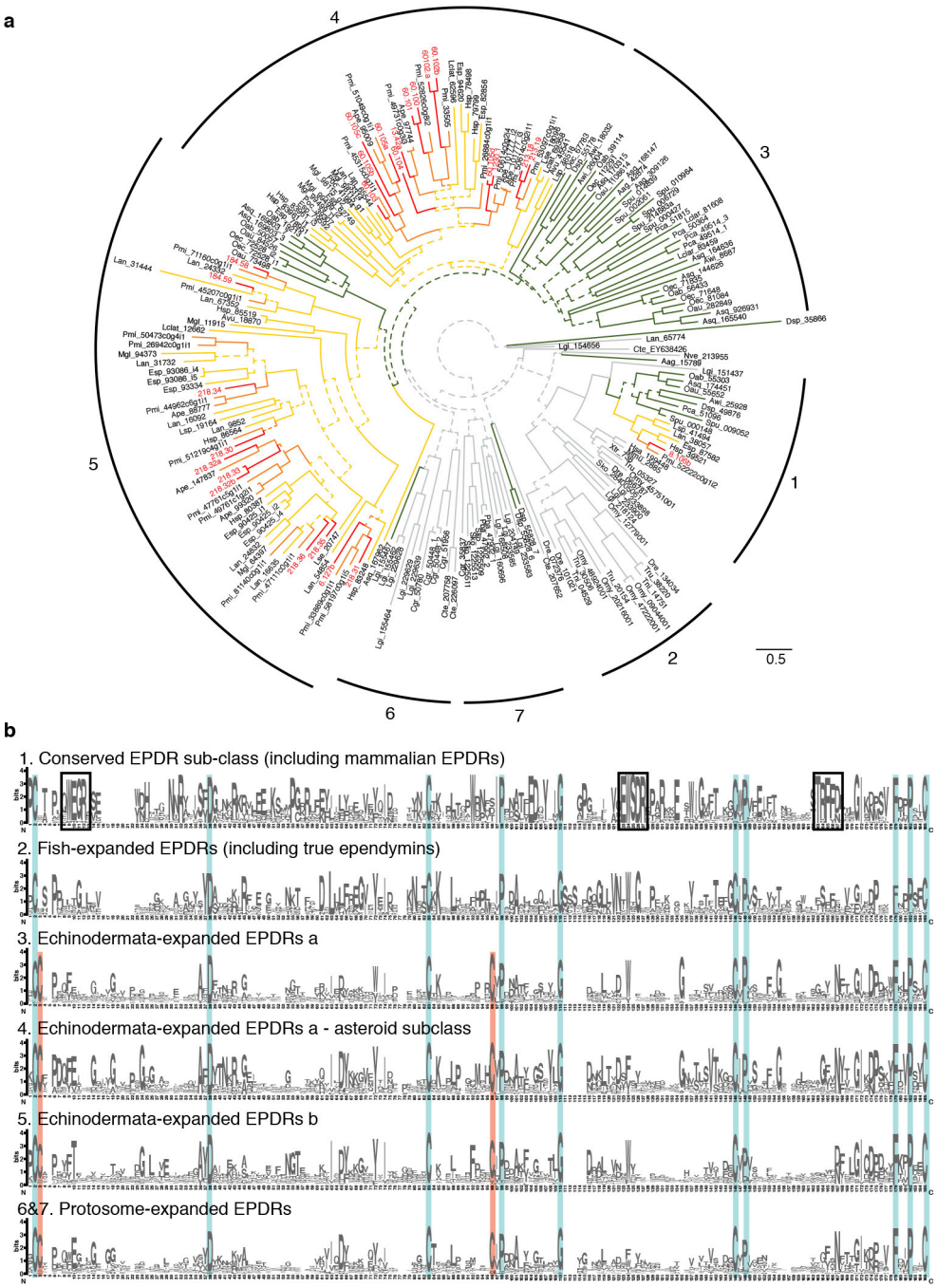
Extended Data Figure 4 | Comparison of Hox clusters. a, Genome browser views of the Hox cluster on GBR scaffold 27 and OKI scaffold 15. Stylised Hox clusters are shown below each scaffold with the corresponding gene model for each Hox gene identified on the scaffold. **b**, Table of OKI and GBR Hox gene models. Prefix corresponds to scaffold. **c**, Micro-synteny of Hox cluster-containing OKI scaffold 15 and GBR scaffolds 27, 51 and 25. **d**, Mapping of OKI and GBR scaffolds containing the Hox cluster to each other. **e**, Molecular phylogenetic analysis of select bilaterian Hox genes by the maximum-likelihood method. Bootstrap support values over 50% are shown. Scale bar: 0.2 substitutions per site. Species abbreviations: Bfl, *Branchiostoma floridae*; Dme, *Drosophila melanogaster*; oki.scaffold.genemodel, *A. planci* OKI; Pfl, *Ptychodera flava*; Sko, *Saccoglossus kowalevskii*; and Spu, *Stronglocentrotus purpuratus*.



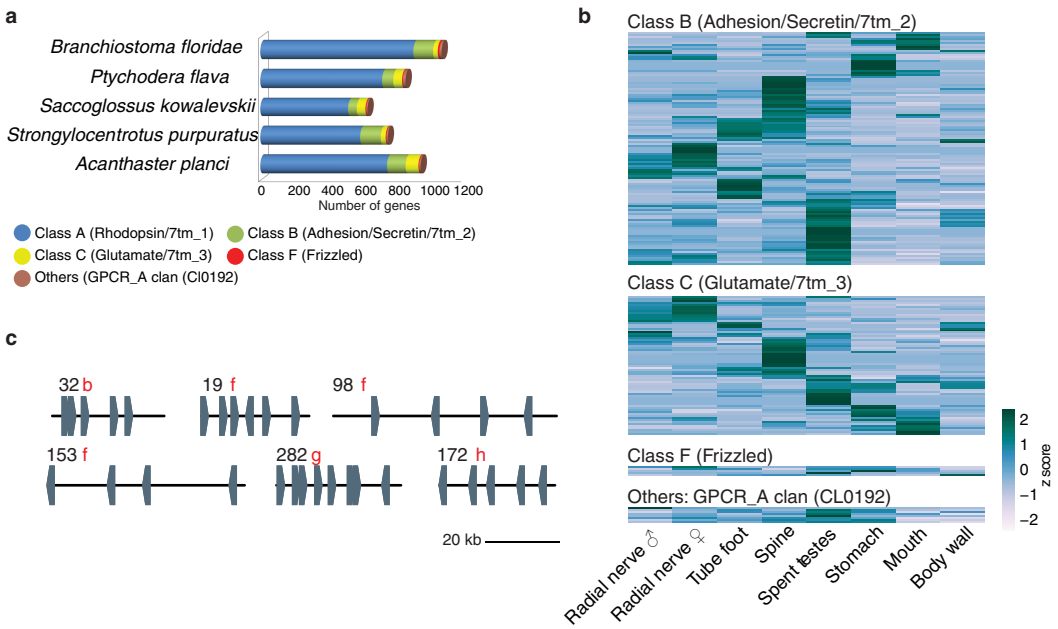
Extended Data Figure 5 | Response of crown-of-thorns starfish to seawater conditioned with its predator the giant triton, *Charonia tritonis*. **a**, Top, diagram showing Y-maze experimental design showing arm dividers and starter zones (yellow). Middle, heat maps showing the cumulative response of COTS over 45 min to water conditioned with a giant triton (left) and ambient seawater (right) ($n = 18$). Red, area in which COTS spent most of the time with descending time to blue; black, no presence. Green outline represents the Y-maze and arm divider that prevents recirculation of water into the opposite arm; starter zones are demarcated by yellow lines. **b**, The duration of movement (highly active threshold set at >60%; $t = -2.936$, $P = 0.006$, 2-tailed t-test). **c**, The meander (change in direction of movement) of active animals over 45 min ($t = 4.437$, $P = 0.000$, 2-tailed t-test). Control, ambient seawater only; giant triton, ambient seawater conditioned with giant triton exudate. Mean ± s.e.m. See Supplementary Video 3 and Supplementary Note 7 for further details.



Extended Data Figure 6 | Protein classes in the crown-of-thorns starfish secretome. a, Overall distribution of characterized secretome. **b**, Distribution of structural, signalling and unclassified proteins. **c**, Distribution of enzyme types.



Extended Data Figure 7 | Extended phylogeny of the EPDR proteins. **a**, Phylogenetic tree of EPDRs incorporating those identified from ambulacrarian transcriptomes. COTS genes are indicated in red, those from non-COTS taxa within the order Valvatida in orange, from non-valvatid taxa within the class Asteroidea in yellow, and from non-asteroid taxa within the phylum Echinodermata in green. Branches with maximum-likelihood bootstrap values >70 and Bayesian posterior probability values >0.9 are indicated by a solid line; those with lower values are indicated by a dashed line. The scale bar indicates the number of substitutions per site. Major EPDR clades are indicated by numbers on the outer circle. Sequences used in the alignment can be found in Supplementary Note 8. **b**, Sequence logos constructed from the conserved region of sequences from each of the seven clades identified in **a**. The height of the amino acid residues indicates the level of conservation, residues highlighted in blue are highly conserved across all clades. Clade 1 is the most highly conserved EPDR clade (ultraconserved motifs are boxed). Clades 3–7 show much lower sequence conservation overall, and possess an extra pair of cysteine residues (highlighted in red).



Extended Data Figure 8 | GPCR abundance, structure and expression in crown-of-thorns starfish. a, Abundance of GPCR genes in ambulacrarians and amphioxus, showing the distribution of the five GPCR classes in each species. See Supplementary Note 9 for further details on genes and analyses. **b**, Tissue expression of each non-rhodopsin class GPCRs in COTS tissues. **c**, Additional examples of GPCR gene clusters in COTS, with genes in clades b, and f–h shown in Fig. 4b. All genes have one exon and are depicted as grey arrowheads that point in the direction of transcription. GBR scaffold numbers are shown above the line; scale bar, 20 kb.

Extended Data Table 1 | Summary of GBR and OKI COTS genomes and transcriptomes

Specimen collection	**GBR**	**OKI**
Location	Rudder Reef, Australia 16°11'36.2"S 145°41'47.0"E	Sesoko, Okinawa, Japan 26°40'46.1"N 127°52'46.1"E
Genome Sequencing		
Paired-end read coverage (x)	66x (3x MiSeq runs)	66x (4x MiSeq runs)
Ave. PE read depth of genome by BWA	40x	46x
Mate-pair library insert sizes	3, 8, 12 kb (HiSeq)	1.5-4, 4-6, 6-8,8-12 kb (HiSeq)
% of properly paired mate-pairs by BWA	93.75%	92.79%
Total read coverage	152x	139x
Genome size estimates		
K-mer analysis (17mer)	441,650,244 bp	421,092,086 bp
Flow cytometry (FACS) of sperm sample	N/A	480 Mb
Genome contigs (Newbler V2.3)		
Contig total length (bp)	376,648,295	376,390,978
Contig number	17,868	17,265
Contig N50 (bp)	54,939	54,788
Genome scaffolding (SSPACE 3.0)		
Scaffold total length (bp)	383,525,304	383,843,944
Scaffold number	3274	1765
Scaffold N50 (bp)	916,880	1,521,119
GC Content (%)	41.31	41.30
Max scaffold size (bp)	3,939,554	11,795,022
Total number of Ns (bp)	10,051,337	9,873,513
Homozygosity	**GBR mapped to OKI**	**OKI mapped to GBR**
% by BLASTN of length >10 kb, mean	98.73	98.72
% by BLASTN of length >10 kb, median	98.78	98.77
% by BLASTN of BitScore >10,000, mean	98.67	98.66
Single nucleotide polymorphism (SNP)		
% internal heterozygosity (GBR to GBR; OKI to OKI)	0.875 1.366	0.917 1.422
% heterozygosity (GBR to OKI; OKI to GBR)		
Gene models		
Augustus gene models (number)	25,995	26,221
EVM Final gene models (number)	24,747	24,323
EVM Single copy orthologs (number)	977	976
Final gene models 'lifted over' (number)	20,328	20,441
RNA-Seq transcriptome analyses		
Tissues sampled (number)	5	10
de novo Trinity Components (number)	93,094	110,737
de novo Trinity Transcripts (number)	153,191	186,200
de novo Trinity contig N50 (number)	3,255	2,853
Trinity. transcript % genome cover by BLAT	97.94	97.43
(baa.pl)	33,036	29,635
Genome guided Cufflinks/Tuxedo Genes (number)	69,261	52,365
Genome guided Cufflinks/Tuxedo Isoforms (number)		

Extended Data Table 2 | The GPCR gene familiy in ambulacrarians and amphioxus

Species	Class A Rhodopsin	Class B Adhesion/ Secretin	Class C Glutamate	Class F Frizzled	Others
Acanthaster planci GBR	775	121	78	6	8
Acanthaster planci OKI	772	128	78	6	9
Strongylocentrotus purpuratus	630	130	31	6	8
Saccoglossus kowalevskii	527	53	51	6	8
Ptychodera flava	752	74	62	6	20
Branchiostoma floridae	1090	130	34	12	15