Elucidation of the speciation history of three sister species of crown-of-thorns starfish (Acanthaster spp.) based on genomic analysis

Hideaki Yuasa1, Rei Kajitani1, Yuta Nakamura1, Kazuki Takahashi1, Miki Okuno1, Fumiya Kobayashi1, Takahiro Shinoda1, Atsushi Toyoda2, Yutaka Suzuki3, Nalinee Thongtham4, Zac Forsman5, Omri Bronstein6,7, Davide Seveso8,9, Enrico Montalbetti8,9, Coralie Taquet10, Gal Eyal11,12, Nina Yasuda13*, and Takehiko Itoh1*

1School of Life Science and Technology, Tokyo Institute of Technology, Meguro-ku, Tokyo 152-8550, Japan, 2Department of Genomics and Evolutionary Biology, National Institute of Genetics, Mishima City, Shizuoka 411-8540, Japan, 3Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa-shi, Chiba 272-8562, Japan, 4Phuket Marine Biological Center, Box 60, Phuket, 83000, Thailand, 5Hawai’i Institute of Marine Biology, School of Ocean & Earth Sciences & Technology, University of Hawai’i at Mānoa, Coconut Island, Kāne’ohe, HI, USA, 6George S. Wise Faculty of Life Sciences, School of Zoology, Tel Aviv University, Tel Aviv 6997801, Israel, 7The Steinhardt Museum of Natural History, Israel National Center for Biodiversity Studies, Tel-Aviv University, Tel-Aviv 6997801, Israel, 8Department of Earth and Environmental Sciences (DISAT), University of Milano-Bicocca, 20126 Milan, Italy, 9Marine Research and High Education Center (MaRHE Center), 12030 Faafu Magoodhoo, Republic of Maldives, 10Collège Zéna M’Dégré, 97610 Pamandzi, Mayotte, 11ARC Centre of Excellence for Coral Reef Studies, School of Biological Sciences, University of Queensland, St. Lucia, QLD 4072, Australia, 12The Mina & Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan 5290002, Israel, and 13Faculty of Agriculture, University of Miyazaki, Miyazaki, Miyazaki 889-2192, Japan

*To whom correspondence should be addressed. Tel. +81 3 5734 3430. Fax. +81 3 5734 3630. Email: takehiko@bio.titech.ac.jp (T.I.); Tel. +81 985 58 7233. Email: nina27@cc.miyazaki-u.ac.jp (N.Y.)

Received 21 December 2020; Editorial decision 29 July 2021

Abstract

The crown-of-thorns starfish (COTS) is a coral predator that is widely distributed in Indo-Pacific Oceans. A previous phylogenetic study using partial mitochondrial sequences suggested that COTS had diverged into four distinct species, but a nuclear genome-based analysis to confirm this was not conducted. To address this, COTS species nuclear genome sequences were analysed here, sequencing Northern Indian Ocean (NIO) and Red Sea (RS) species genomes for the first time, followed by a comparative analysis with the Pacific Ocean (PO) species. Phylogenetic analysis and ADMIXTURE analysis revealed clear divergences between the three COTS species. Furthermore, within the PO species, the phylogenetic position of the Hawaiian sample was further away from the other Pacific-derived samples than expected based on the mitochondrial data, suggesting that it may be a PO subspecies. The pairwise sequentially Markovian coalescent model showed that the trajectories of the population size diverged by region during the Mid-Pleistocene transition when the sea-level was dramatically decreased, strongly suggesting that the three COTS species experienced allopatric speciation.
the orthologues indicated that there were remarkable genes with species-specific positive selection in the genomes of the PO and RS species, which suggested that there may be local adaptations in the COTS species.

**Key words:** three distinct sister species, common ancestor, coral predator, phylogenetic analysis, population demographic history

### 1. Introduction

Coral reefs have remarkably high levels of biodiversity, but they cover less than 0.2% of the marine surface. They are estimated to harbour approximately one-third of all described marine species,

suggesting that speciation has occurred within them. It is important to elucidate the processes that have generated the high levels of biodiversity coral reef ecosystems for their long-term conservation and management. In particular, the Indo-Pacific Ocean has many coral reef areas with relatively high levels of biodiversity, suggesting that the question of why marine biodiversity is concentrated in this region has been widely discussed in many previous investigations.

It has been hypothesized that the speciation of coral organisms in the Indo-Pacific was caused by geographical isolation during the glacial period and positive selection in complex habitats. However, due to the difficulty of collecting samples from the Indo-Pacific region, these hypotheses are based on a limited number of DNA markers, such as partial gene regions, and have not yet been confirmed at the whole genome level for non-model coral reef organisms.

Crown-of-thorns starfish (COTS; excluding *Acanthaster brevispinus*) are well-known coral predators in the Indo-Pacific Ocean and they are highly associated with coral reef ecosystems. Recent mitochondrial DNA barcoding studies have shown that COTS diverged into four distinct sister-species, Pacific Ocean (PO) species (*Acanthaster solaris*), Northern Indian Ocean (NIO) species (*Acanthaster planci*), Southern Indian Ocean (SIO) species (*Acanthaster mauritensis*), and Red Sea (RS) species (not yet named, *Acanthaster sp.*) that corresponded to their geographical distributions. These divergences are thought to be due to sea-level changes during the Pliocene–Early Pleistocene (1.95–3.65 million years ago) according to the COI (Cytochrome c Oxidase subunit I) genes divergence rates.

However, phylogenetic inference when determined using the mitochondrial genome is not always concordant with that determined using the nuclear genome, suggesting that classifications based only on the results of mtDNA may not be reliable. Furthermore, the mitochondria do not necessarily reflect the phylogeny of a population, when considering evolutionary processes such as migration and admixture. To properly address these issues, nuclear DNA (nDNA)-based analysis among the COTS lineages is required. While the nuclear genome sequences of PO species were previously published by Hall 2017, the nuclear genome sequences of other species have not yet been published. In this study, we have sequenced and assembled the genomes of NIO and RS species to conduct a comparative genomic analysis of crown-of-thorns starfish (*Acanthaster spp.*) based on their nuclear genome sequences. Furthermore, using the obtained nuclear genome sequences, we aimed to clarify the taxonomy of the extant COTS species and to reveal their speciation process using the history of their population demographics.

### 2. Materials and methods

#### 2.1. Sample preparation and sequencing

Twenty individual COTS from seven locations throughout the Indo-Pacific Ocean (Fig. 1, Supplementary Table S1) were collected, and genomic DNA was extracted from their tube feet. Paired-end (PE) libraries of the 20 COTS samples and mate pair (MP) libraries of two of the COTS samples (Phuket01 of the NIO species: accession No. SAMD00229334, and Israel04 of the RS species: accession No. SAMD00229345) were prepared and sequenced using Illumina platforms (see Supplementary Table S2). RNA samples were extracted from the gonads, digestive glands, spines, stomach, and tube feet (accession No. SAMD00231968, SAMD00231971–SAMD00231974) of PO species that were sampled in a previous study (Miyazaki01: accession No. SAMD00056692) and sequenced using an Illumina HiSeq 2500 sequencer (see Supplementary Methods for details).

#### 2.2. Mitochondrial genome assembly, phylogenetic analysis, and comparative analysis

A total of 23 mitochondrial genome sequences were constructed using NovoPlasty v.3.7 from the Illumina PE sequences of 20 the COTS samples that were described in the previous section and three PO species samples previously published (Miyazaki01 and Okinawa, GBR, Supplementary Table S1) and sequenced using Illumina PE and MP reads of the PO species. The mitochondrial genome sequences that were reported in a previous study were also included. One was a PO species (accession No. NC007788), and the other was a COTS species (*Acanthaster brevispinus* (accession No. NC007789)) which is a species closely related to COTS. Finally, a total of 25 mitochondrial genome sequences were aligned using Mafft v.7.310. Pairwise sequence identities between the whole mitochondrial genomes were calculated for each pair from the multiple alignment results. Single-nucleotide polymorphism (SNP) sites were also extracted from the multiple alignment results, and a maximum likelihood (ML) phylogenetic tree was reconstructed using IQ-TREE v.1.6.12. A neighbour-joining (NJ) phylogenetic tree was constructed using MEGA X (see Supplementary Methods for details).

#### 2.3. Nuclear genome assembly and k-mer analysis

Raw Illumina PE and MP reads from two COTS (Phuket01 of NIO species and Israel04 of RS species) were adapter trimmed and quality filtered using Platanus_trimm and they were then initially assembled into scaffolds using Platanus v.1.2.5. Contaminated and mitochondrial sequences were removed from the assembled scaffolds using homology searches with BLASTN v.2.7.19 to construct the respective nuclear genome-derived sequences. The nuclear genome sequence and raw Illumina PE reads of the PO species were downloaded from the NCBI database (Miyazaki01). For these three nuclear genome sequences (Phuket01 of NIO species, Israel04 of RS species, and Miyazaki01 of PO species), genome completeness was assessed using
2.4. Nuclear genome phylogenetic analysis and comparative analysis

The nuclear genome sequence identities between the three species were investigated using Minimap2 v.2.17.23 Sequence identities between all the interspecies combinations were calculated from the local pairwise alignments of the reciprocal best hits. To survey the phylogenetic relationships between the COTS samples, SNP information in the nuclear genomes was extracted according to the following procedure. First, trimmed reads of the 23 COTS samples (20 in this study, 3 from Miyazaki01,12 and Hall et al. 2017,11 see Supplementary Table S1) were aligned to the nuclear genome sequences of PO species (Miyazaki01) using BWA-MEM v.0.7.17.24 In this study, species-specific genomes were used as a reference to reduce the reference bias which influence population genomic analysis to estimate ancestor.25 Mapped reads with AS/C20 XS as potential errors. Realignment and variant calling were then conducted using IndelRealigner and HaplotypeCaller in the Genome Analysis Toolkit (GATK) v.3.8.1,26 respectively. Then, the variants with read-depths >3 times or <1/3 of the average read-depth were filtered out using GATK-VariantFiltration. The variants obtained from the 23 samples were combined and common SNPs among all samples were extracted. Sequences were generated from extracted SNPs using vcflib0.1027 and an ML phylogenetic tree was constructed using IQ-TREE. The NJ phylogenetic tree was constructed using MEGA X (see Supplementary Methods for details).

2.5. Population demographic history analysis

To infer the effective population size history, PSMC', a modified version of the pairwise sequentially Markovian coalescent (PSMC) analysis implemented in the MSMC (multiple sequential Markovian coalescent) package was used. The reason for this was as follows: (i) MCMC in multiple haplotypes requires haplotype phasing for accurate analysis, which is difficult to perform in non-model organisms; and (ii) the number of samples from each location was different. For these reasons, we applied PSMC' to avoid the effects of phasing error and sample size differences.

First, PE reads of the 21 samples (20 in this study and Miyazaki01,12 see Supplementary Table S1) were aligned to the nuclear genome sequences of the corresponding species (NIO: Phuket01, RS: Israel04, and PO: Miyazaki01) using BWA-MEM, and mapped reads with AS XS as potential errors. Mapping results were realigned using GATK-IndelRealigner, and the genotype of each position was determined using the Bcftools v.1.9 of the SAMtools package. Finally, the fluctuation of the population size was estimated using PSMC' implemented in the MSMC2 v.2.1.131 package. For this analysis, a generation time of 3 years was assumed, since it has been reported in COTS laboratory studies that the onset of gametogenesis is related to both body size and age, and the gonad development in the small individuals did not occur until after the latter part of their 2nd year.32 Furthermore, the mutation rates were estimated from the genome sizes (NIO: 419 Mbp, RS: 378 Mbp, and PO: 454 Mbp) based on previous studies33 and they were set to 9.4e−9, 8.8e−9, and 9.9e−9 bp per generation in the NIO, RS, and PO species, respectively (see Supplementary Methods for details).

2.6. Admixture analysis

Biallelic SNPs were extracted from the SNP dataset used in the phylogenetic analysis of the nuclear genomes using Bcftools. To avoid the negative effects of the SNPs being close to each other, SNPs under linkage disequilibrium, which exhibited a correlated ($r^2 = 0.5$) pair within a window of 200 kb were removed using PLINK v.2.0.0a2.3LM (--indep-pairwise 200 kb 1 0.5).34 Population genetic structures were inferred using ADMIXTURE v.1.3.0.35

2.7. Gene annotation

Protein coding genes were predicted for the three COTS species genomes by combining the results of the RNA-seq-based, homology-
based (searched for related-species proteins), and *ab initio*-based prediction methods using in-house pipelines. For RNA-seq-based predictions, RNA-seq reads of PO species were used for all three COTS species, and both assembly-first and mapping-first methods were employed. For the assembly-first method, RNA-seq reads were assembled using Trinity v.2.8.436 and Oases v.0.2.0936 and assembled contigs were splice-mapped to the genome sequences using GMAP v.2018-07-04.38 For the mapping-first method, RNA-seq reads were mapped to the genome sequences using HISAT2 v.2.1.0.39 and transcripts were predicted by StringTie v.2.0.40 from the mapping results. For homology-based predictions, amino acid sequences of *Strongylocentrotus purpuratus*, *Apostichopus japonicus*, *Saccoglossus kowalevskii*, and *Homo sapiens* (GRCh38) were splice-mapped to the genome sequences using Spaln v.2.3.3,44 and gene models were predicted. For *ab initio* prediction, AUGUSTUS v.3.3.245 and SNAP v.2006-07-28.56 were used. First, 1,000 genes were randomly selected as a training set from the results of the RNA-seq mapping-first method, and then, gene models were trained and gene predictions were performed. Finally, all the predicted gene candidates were merged using an in-house tool and the final predicted results were assessed by BUSCO in protein mode.

For functional annotation, predicted gene sequences were subjected to BLASTP homology searches against the SwissProt database (downloaded in December 2018) using DIAMOND v.0.9.14.11547 (cut-off: e-value \( \leq 10^{-5} \)). Gene Ontology (GO) terms were assigned using InterProScan v.5.39-77.0 with the -goterms option. The KEGG Orthology (KO) identifiers (K number) were also assigned using the webserver service of BlastKOALA v.2.249 and KofamKOALA v.2020-01-06.50 They were then translated to KEGG PATH and KEGG BRITE according to the KEGG ORTHOLOGY (KO) database (downloaded in September 2020 from https://www.genome.jp/kegg-bin/get_htext?ko00001),51 which is a database of molecular functions representing functional orthologues (see Supplementary Methods for details).

### 2.8. Orthologue analysis and detections of positive selection signatures in genes

To understand the evolutionary history of the three COTS species, positive selection analysis was performed for single-copy orthologue genes among the three COTS species. First, orthologue relationships among the three COTS species were identified by SonicParanoid v.1.3.4.52 Multiple alignments were then generated using MAFFFT and PAL2NAL v.14.53 for each single-copy gene group (14,203) and TrimAl v.1.4. rev2254 and the option of `-strict` was applied to eliminate low-quality regions in the alignments. Individual gene trees were generated using the IQ-TREE. Finally, every gene was tested for positive selection at all branches, without specified foreground branch, using the adaptive branch-site random effects likelihood (aBSREL) model35 with HyPhy v.2.5.1.156 (P-value \( < 0.05 \), with the Holm-Bonferroni correction). GO and KEGG enrichment analyses were performed based on the Fisher’s exact test using an in-house script (see Supplementary Methods for details).

### 2.9. Re-annotation of plexin B genes and interspecies comparisons

The protein sequences of the plexin-related genes in a previous study,41 including the PO species of COTS from the OKI genome, were downloaded as GenPept files (Supplementary File S1 sequences.zip) and manually converted to FASTA files. These COTS sequences were aligned to draft genomes using Spaln v.2.4.0 with the option of ‘-Q4 ‘MI -00’. The top-hit for each protein was used for the prediction of an exon-intron structure.

The genomic regions of the plexin B clusters were aligned using a nucler in the MUMmer package v.3.1.38 with the options of -maxmatch -nosimplify and dot plots were generated using the mummer-plot in the same package.

Phylogenetic analysis was conducted for plexin-related genes annotated with no in-frame stop codons. Plexin-1 of *Amphimedon queenslandica* was obtained from the previous study57 and used as the outgroup in this analysis. Multiple alignments were then generated using MAFFFT, then TrimAl with the option of ‘-automated1’, was applied to eliminate low-quality regions in the alignments. An ML phylogenetic tree was constructed using IQ-TREE.

### 3. Results and discussion

#### 3.1. Genome sequence construction

From the Illumina PE and MP data, we successfully obtained 400.4 Mb and 385.6 Mb for the genome assemblies, with scaffold N50s of 3.90 Mb and 1.66 Mb, for the NIO and RS species, respectively. Furthermore, with the BUSCO evaluation, 98.0% and 97.7% of the Metazoan genes were completely detected in these scaffolds (Table 1). The genome of the downloaded PO species was 420.0 Mb.

#### 3.2. Estimation and comparison of genome sizes based on k-mers

Assembly results (draft genomes) often lack assembly-hard regions such as repetitive ones, and this can deviate an assembly size from the true genome size. In contrast, it is expected that k-mer-based estimations of genome sizes can take such assembly-hard regions into account and produce more accurate values.22 Since the sizes of draft genomes are substantially influenced by assembly methods, we discuss genome-size variations only based on the k-mer-based estimations.

According to the k-mer analysis, the estimated total genome sizes in the NIO, RS, and PO species were 418.7 Mb, 378.3 Mb, and 454.1 Mb, and the duplicated sequence sizes were 147.4 Mb, 114.4 Mb, and 182.9 Mb, respectively, suggesting an approximately 10% genome size difference among the species and a similar trend between estimated genome size and duplicated sequence size (Table 1). Statistically, the differences between the NIO and RS species and the PO species were significant (P-value \( < 0.001; t\)-test), while the differences between the NIO and RS species were not significant (P-value \( > 0.05 \)). This trend was confirmed in the analysis of 21 samples: 20 samples sequenced in this study and the Miyazaki01 sample (Supplementary Fig. S1 and Table S1). This suggests that the duplicated sequences contributed to the differences in genome size among the COTS species. Additionally, we verified the accuracy of GenomeScope using organisms with near-complete reference genomes: *Caenorhabditis elegans* (roundworm) and *Drosophila melanogaster* (fruit fly). Illumina PE reads of these organisms were downloaded from the public database and analysed with the same procedures as used for COTS. As a result, the differences between the estimated and reference genome sizes were less than 0.8% of the reference genome (Supplementary Table S3), suggesting high accuracy of the estimation.
Table 1. Summary of the genomes and genes of the three COTS species

| Species | NIO | RS | PO |
|---------|-----|----|----|
| Specimen Source | Phuket01 | Israel04 | Miyazaki01 |
| Genome assembly | Assembly size (bp) | 400,438,766 | 385,630,944 | 420,016,408 |
| | Scaffold number | 3,642 | 3,552 | 4,526 |
| | Scaffold NS50 (bp) | 3,897,460 | 1,662,177 | 2,941,429 |
| | Max Scaffold length (bp) | 11,313,316 | 8,528,531 | 10,098,247 |
| | Total number of Ns (bp) | 7,197,764 | 7,359,702 | 4,264,832 |
| | GC content (%) | 41.4 | 41.4 | 41.3 |
| Genome estimation | Genome size (bp) | 418,678,427 | 378,256,433 | 454,083,373 |
| | (17-mer analysis) | Duplicated size (bp) | 147,396,297 | 114,220,755 | 182,858,213 |
| | Heterozygosity (%) | 0.72 | 1.24 | 0.97 |
| Genome evaluation | Complete (%) | 98.0 | 97.7 | 98.1 |
| | Complete and single-copy (%) | 97.6 | 97.4 | 97.6 |
| | Complete and duplicated (%) | 0.4 | 0.3 | 0.5 |
| | Fragmented (%) | 0.7 | 0.9 | 1.4 |
| | Missing (%) | 1.3 | 1.4 | 0.5 |
| Predicted Proteins | Gene number | 18,045 | 18,879 | 19,397 |
| Protein evaluation | Complete (%) | 98.2 | 98.0 | 98.1 |
| | Complete and single-copy (%) | 97.8 | 97.8 | 97.6 |
| | Complete and duplicated (%) | 0.4 | 0.2 | 0.5 |
| | Fragmented (%) | 1.3 | 1.3 | 1.3 |
| | Missing (%) | 0.5 | 0.7 | 0.6 |

NIO, RS, and PO represent COTS species, NIO: Northern Indian Ocean species, RS: Red Sea species, and PO: Pacific Ocean species.

3.3. Genome sequence comparison and phylogenetic analysis
The sequence identities between the mitochondrial genomes of the three COTS species were calculated using the aligned region of 15,663 bp among the 24 COTS and one Acanthaster brevispinus, which is a species closely related to COTS, resulting in 91.31–92.48% (Supplementary Table S4). The sequence identities of the nuclear genomes between the species were calculated using regions corresponding to reciprocal best hits of pairwise alignments for the three genome assemblies (Phuket01, Israel04, and Miyazaki01), resulting in 97.52–98.11% (Supplementary Table S5). Although the estimated heterozygosity rates of all samples (Table 1) were higher than that for the human genome (~0.1%), pairwise sequence identities of the mitochondrial and nuclear genomes between the COTS species were similar to those between humans and chimpanzees (mitochondria: 91.1%, nuclear: 98.77%), indicating that each population had a common ancestral population. It has previously been suggested that the population of Acropora corals, the preferred food of COTS, increased up to the MPT and then declined, suggesting that the fluctuations in the abundance of COTS around the MPT may have been influenced by the amount of available food. Similarly, we can see from the figures that during the MPT, the trajectory of the effective population size started to diverge by region. This implies that each region started to diverge to an independent population at that time. The MPT is a period that included sea-level oscillations due to dramatic increases in glaciation, resulting, for the first time in millions of years, in the sea level dropped to 120 m, exposing the shelf break and upper-slope deposits. This indicates that the divergence of the COTS populations during the MPT may have been caused by physical isolation due to the unparalleled dropping of the sea level.
3.5. Admixture analysis
ADMIXTURE analysis was performed using the 262,994 independent SNPs in the nuclear genome. It was found that $K = 1$, which assumes that a single ancestor had the lowest cross-validation error (Supplementary Table S6). This indicates that individuals diverged from a common ancestral population, which supports the results of the populations demographic history in the previous section. However, at $K = 3$, the NIO, RS, and PO species were divided into three distinct ancestor compositions (Supplementary Fig. S4). This indicated genetic divergence among the three COTS species, which was concordant with the results of the phylogenetic analysis. It should be noted that the composition of the PO from Hawaii seemed to be different from the other PO species at $K = 3$ and formed an independent structure at $K = 4$. This further showed the uniqueness of the Hawaiian PO species, and was concordant with the results of the phylogenetic analysis.

3.6. Orthologue and enrichment analysis for local adaptation exploration
We focused on the differences in the genes between species. Specifically, we searched for genes that could be found only in certain

![Figure 2](https://academic.oup.com/dnaresearch/article/28/4/dsab012/6350483)

**Figure 2.** Maximum likelihood (ML) phylogeny in Acanthaster spp. NIO, RS, and PO represent COTS species (NIO: Northern Indian Ocean species, RS: Red Sea species, and PO: Pacific Ocean species). The blue frame indicates the position of a Hawaiian sample. (A) ML tree based on SNPs in the mitochondrial genome of 24 COTS. Twenty COTS samples were sequenced in this study plus four additional samples, which included four samples of PO species (NC_007788 and 3 sequences assembled using data from Miyazaki0112 and Hall et al. 201711 in this study) and Acanthaster brevispinus (NC_007789)13 as an outgroup. (B) ML tree based on SNPs in the nuclear genome. In addition to the 20 samples used in this study, we used the data of 3 PO species from Miyazaki0112 and Hall et al. 2017.13

![Figure 3](https://academic.oup.com/dnaresearch/article/28/4/dsab012/6350483)

**Figure 3.** Population demographic history of three COTS species in seven regions. Effective population sizes inferred using PSMC with a generation time of 3 years and a mutation rate of 9.9e-9, 9.4e-9, and 8.8e-9 bp per generation in Northern Indian Ocean (NIO), and Red Sea (RS), and Pacific Ocean (PO) species, respectively. A mutation rate was assumed from the estimated genome size (NIO: 419 Mbp, RS: 378 Mbp, and PO: 454 Mbp).31 Grey shade during 0.7–1.25 Mya: Mid-Pleistocene Transition (MPT).
species and had undergone positive selection in each marine environment.

First, we successfully predicted 18,045, 18,879, and 19,397 protein-coding genes from the draft genomes of the NIO, RS, and PO species, respectively, and evaluated the quality of these gene sets using the BUSCO tool in the protein mode. The complete BUSCO (single copy and duplicate) rate was extremely high at over 98% for all three species (Table 1), indicating that the genes were predicted with high accuracy. At least 51.8% of the genes of each species had hits in the SwissProt or KEGG databases (Supplementary Table S7). We then performed an orthologue analysis of the predicted genes using SonicParanoid and found 16,652 orthologous groups in total. For each species, 16,054–16,495 genes were clustered in 15,923–16,058 orthologous groups, and 1,902–2,902 genes remained as non-orthologous genes (Supplementary Table S8). The orthologue results obtained included 14,203 common single-copy orthologue genes/groups formed from one gene each from the three COTS species. To detect species-specific signatures, positive selection analysis genes/groups formed from one gene each from the three COTS species. We then performed an orthologue analysis of the predicted genes using SonicParanoid and found 16,652 orthologous groups in total. For each species, 16,054–16,495 genes were clustered in 15,923–16,058 orthologous groups, and 1,902–2,902 genes remained as non-orthologous genes (Supplementary Table S8). The orthologue results obtained included 14,203 common single-copy orthologue genes/groups formed from one gene each from the three COTS species. To detect species-specific signatures, positive selection analysis using HyPhy was performed for single-copy orthologues, which showed that 194, 213, and 185 genes showed species-specific positive selection for NIO, RS, and PO, respectively (Supplementary Fig. S5). GO and KEGG enrichment analyses were performed on species-specific non-orthologous genes (Supplementary Tables S9 and S10; columns of ‘Non-ortholog genes’) and single-copy orthologous genes with positive selection (Supplementary Tables S11 and S12; columns of ‘Target genes’) to investigate function bias (P-value < 0.05). The most striking result from the GO enrichment analysis of the non-orthologous genes was that the genes assigned to GO: 0017154 (semaphorin receptor activity) and GO: 0071526 (semaphorin-plexin signalling pathway) were highly enriched in the non-orthologous genes from each lineage. According to the results (Supplementary Table S9), nearly half of the genes belonging to these GOs were enriched as non-orthologues. However, while the evaluation of the gene set by BUSCO showed a high degree of completeness, there were also inevitable errors in the automatic gene structure prediction, and some genes may have been classified as non-orthologues due to these errors; thus results had to be carefully verified. Consequently, we found that the genes assigned to GO:0017154 and GO:0071526 belonged to the plexin family, which regulates cellular interactions in a wide range of developmental and physiological contexts. This family was previously reported to have expanded in the COTS species, and there were several plexins with significantly lower identities to the reference gene sequences of the PO species (Supplementary Table S13). Figure 4 shows an interspecies dot plot of the genomic regions where the plexin B2–B8 were clustered. The figure clearly shows that independent genomic duplications and insertions were observed in the corresponding genomic regions of each of the three species, and the synteny relationships are collapsed in this region. Here, we verified the accuracy of the plexin B regions in the draft genomes based on physical coverage (Supplementary Methods; filtering of read mapping results were also described). For each site, physical coverage is defined as the number of read pairs spanning that site. All sites between a pair of mapped positions of reads are considered to be spanned even if not covered by reads. In contrast to read coverage, physical coverage can properly reflect the information of long-insert libraries. As a result, physical coverages were ≥24 for all sites in the plexin B regions, suggesting that there was not clear misassembly. Additionally, we confirmed that read coverages were ≥9 for all sites except for gaps and their adjacent regions (±600bp, which corresponds to the PE insert-sizes). Phylogenetic analysis based on the annotated Plexin-family genes was also conducted, and we observed incongruence between the topology and labels of the genes for the plexin B in the regions with structural variations (Supplementary Fig. S6). For example, genes with ‘Plexin-B3’ labels were not monophyletic. Although it is difficult to accurately predict the gene structures in the duplicated regions of the genome and phylogenetic analysis requires further scrutiny, it is of interest that structural changes to regions of the genome seem to have occurred independently in each species that were found, and that the plexin family genes existed in these regions. Prior studies have suggested that the plexins have extremely ancient origins, predating the common ancestors of the metazoans and that plexins regulate multiple development processes, with functions in controlling cytoskeletal dynamics and cell adhesion, prior to their role as axon guidance molecules. Considering this, it is possible that differences in the plexins may cause the phenotypic differences such as those related to body colours and spine shapes among the COTS species, but further analysis is required since the functions of the plexin genes can be extensive.

Since the draft genomes are the consensus sequences of paternal and maternal haplotypes, comparison of the haplotypes is effective to understand the details. The haplotype assemblator, Platanus-allee v.2.2.265 (see Supplementary Methods), allowed for structural variations such as indels and duplications to be observed, even within individuals (Supplementary Fig. S7). The results suggest that this region is a hotspot of structural variations for both inter- and intra-species populations. Not only the differences in the number of duplications, but the sequence identities between some pairs of duplication units were low (i.e. highly divergent). Although quantification has not yet been performed, such duplications and indels can contribute to the variation of genome sizes (Supplementary Fig. S1 and Table S1), and this region is a notable example. Note that the draft genomes represented one or mixed haplotypes. Such complicated structural variations hinder estimation of evolutionary history. In addition, the orders of the plexin B genes in Fig. 4 were determined simply based on the top-hits of the protein alignments. In addition to the complicated structural variations, the lack of transcriptome data for the NIO and RS species made curations difficult.

We then focused on single-copy orthologous genes with positive selection. From the KEGG enrichment analysis results, we found that the ko04927 (cortisol synthesis and secretion) and ko00430 (taurine and hypotaurine metabolism) pathways in RS species were enriched in single-copy orthologous genes. Cortisol is reported to be related to the osmoregulation of both high and low salinity in teleost fish,66 and taurine is reported to be a key organic osmolyte for cells.67 Given that the RS experiences salinity fluctuations,68 enrichment in these pathways may have been associated with adaptations to a saline environment. Another example is that in the PO species, ko4918 (thyroid hormone synthesis) and ko4919 (thyroid hormone signalling pathway) were enriched in single-copy orthologous genes, and thyroid hormones are involved in the regulation of body growth;69 thus enrichment in these pathways may have influenced the growth of PO species. Given that the COTS population outbreaks in the Pacific Ocean appear to be more large-scale and frequent than the other regions,6,7 the differences in the genes belonging to the thyroid hormone pathway are of interest. However, further experimental studies are required to verify the local adaptations observed in PO and RS species.
Finally, we focused on immune-related genes to analyse the possibility that each COTS species may have evolved their own immune system to respond to pathogens in each sea area; noting that examples of rapid evolution of the innate immune system under the influence of pathogens have previously been reported. If there are immune-related environmental adaptations, they may appear as non-orthologous genes or genes that have undergone positive selection, so we investigated whether genes with these functions were extracted. The results are shown in Supplementary Table S14.

Ko00512 (mucin-type O-glycan biosynthesis), ko04054 (pattern recognition receptor), and ko04091 (Lectins) were detected as species-specific gene groups that were significantly more abundant in PO species than in the other two species (P-value < 0.05). Since bacterial biodiversity is qualitatively similar to the biodiversity of plants and animals in the environment, it is possible that PO species in the Pacific Ocean, a coral reef region with high biodiversity, are exposed to a greater diversity of pathogens than other species and are at a higher risk of infection, and consequently changes in the immune-related genes of PO species may reflect local adaptations to the Pacific coral reef environment.

3.7. Conclusions and future perspectives
In this study, we constructed high-quality draft genome sequences of NIO and RS species for the first time and compared the sequences of the three COTS species. The results of the genome sequence comparisons, genome-wide phylogenetic analysis, and ADMIXTURE analysis strongly indicate that there are three distinct COTS species, which had previously been reported using mitochondrial gene comparison analyses. However, our phylogenetic analysis also showed that a Hawaiian individual had diverged more from other PO species in the nuclear genome than in the mitochondrial genome. Furthermore, the ADMIXTURE analysis indicated that the Hawaiian samples form an independent structure when the ancestral population was set to 4. Given the difference within Pacific populations based on alzyme
markers in the previous study, it may be necessary to consider the Hawaiian population as a subspecies of the PO species.

Given the trajectory of the effective population size started to diverge by region during the MPT when the dramatic fluctuations in sea level occurred, the physical isolation may have contributed to the speciation among COTS species. Additionally, some traces of local adaptation were found in the genomes of each species using the orthologous analysis. We have discovered that the genomic structural changes that seem to have occurred independently for each species occurred in genomic regions where plexin family genes existed in clusters.

Thus, we have succeeded in constructing new genome sequences for the NIO and RS species, unravelling the history of their speciation, and providing a basis for future genomic analysis. To further clarify the differences within and between COTS species, further genomic research is required that will include more samples, especially Southern Indian Ocean (SIO) species (A. mauritensis) that could not be included in this analysis. Individual regions of the genome where differentiation has been observed should also be investigated in-depth.

Acknowledgements

We thank members of the Itoh Laboratory and Toyoda Laboratory, T. Hayashi, Y. Obara, Y. Gotoh, D. Yoshimura, and K. Teshima for their logistical and technical support. We also thank members of Yasuda Laboratory, Y. Higashimura, M. Ueno, and M. Fukuda for local sampling.

Accession numbers

DDBJ/ENA/GenBank BioProject No. PRJDB9937.

Funding

This work was supported by the Grant-in-Aid for JSPS Fellows (Grant Number 18J23317), JSPS KAKENHI (Grant Numbers 22150002, 16H06279 (PAGS), 25870563, 17H04996, 16H04719, and 19H03206).

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at DNARES online.

References

1. Reaka-Kudla, M.L. 2001, Known and unknown biodiversity, risk of extinction and conservation strategy in the sea. In: Bendell-Young, L. and Gallaugher, P. (eds.), Waters in Peril. Springer US: Boston, MA, pp. 19–33.
2. Veron, J., Stafford-Smith, M., DeVantier, L. and Turak, E. 2015, Overview of distribution patterns of zooxanthellate Scleractinia, Front. Mar. Sci., 1, 81.
3. Keyse, J., Crandall, E.D., Toonen, R.J., Meyer, C.P., Treml, E.A. and Riglos, C. 2014, The scope of published population genetic data for Indo-Pacific marine fauna and future research opportunities in the region, Bull. Mar. Sci., 90, 47–78.
4. Carpenter, K.E., Barber, P.H., Crandall, E.D., et al. 2011, Comparative phylogeography of the coral triangle and implications for marine management, J. Mar. Biol., 2011, 1–14.
5. Prise, S.A., Holzman, R., Near, T.J. and Wainwright, P.C. 2011, Coral reefs promote the evolution of morphological diversity and ecological novelty in labrid fishes, Ecol. Lett., 14, 462–9.
6. Vogler, C., Benzie, J., Lessios, H., Barber, P. and Worobiec, G. 2008, A threat to coral reefs multiplied? Four species of crown-of-thorns starfish, Biol. Lett., 4, 696–9.
7. Haszprunar, G., Vogler, C. and Worobiec, G. 2017, Persistent gaps of knowledge for naming and distinguishing multiple species of crown-of-thorns-seastar in the Acanthaster planci species complex, Diversity., 9, 22.
8. Haszprunar, G. and Spies, M. 2014, An integrative approach to the taxonomy of the crown-of-thorns starfish species group (Asteroida: Acanthaster): a review of names and comparison to recent molecular data, Zootaxa, 3841, 271–84.
9. Shaw, K.L. 2002, Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: what mtDNA reveals and conceals about modes of speciation in Hawaiian crickets, Proc. Natl. Acad. Sci. USA., 99, 16122–7.
10. Wiens, J.J., Kuczynski, C.A. and Stephens, P.R. 2010, Discordant mitochondrial and nuclear gene phylogenies in enyid turtles: implications for speciation and conservation, Biol. J Linn Soc., 99, 44–61.
11. Hall, M.R., Kocot, K.M., Baughman, K.W., et al. 2017, The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest, Nature, 544, 231–4.
12. Horowia, M., Nakamura, T., Yuasa, H., et al. 2021, Integrated population genomic analysis and numerical simulation to estimate larval dispersal of Acanthaster cf. solitarii between Ogasawara and other Japanese regions, bioRxiv, doi: 10.1101/2021.03.25.436911.
13. Dierckxsens, N., Mardulyn, P. and Smits, G. 2017, NOVOPlasty: de novo assembly of organelle genomes from whole genome data, Nucleic Acids Res., 45, e18.
14. Yasuda, N., Hamaguchi, M., Sasaki, M., Nagai, S., Saba, M. and Nadaoka, K. 2006, Complete mitochondrial genome sequences for Crown-of-thorns starfish Acanthaster planci and Acanthaster brevispina, BMC Genomics, 7, 17.
15. Katoh, K. and Standley, D.M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol., 30, 772–80.
16. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A. and Minh, B.Q. 2015, IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies, Mol. Biol. Evol., 32, 268–74.
17. Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. 2018, MEGA X: molecular evolutionary genetics analysis across computing platforms, Mol. Biol. Evol., 35, 1547–9.
18. Kajitani, R., Toshimoto, K., Noguchi, H., et al. 2014, Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads, Genome Res., 24, 1384–95.
19. Alschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. 1990, Basic local alignment search tool, J. Mol. Biol., 215, 403–10.
20. Seppey, M., Manni, M. and Zdobnov, E.M. 2019, BUSCO: assessing genome assembly and annotation completeness. In: Kollmar, M. (ed.), Gene Prediction: Methods and Protocols. Springer: New York, NY, pp. 227–45.
21. Marçais, G. and Kingsford, C. 2011, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers, Bioinformatics, 27, 764–70.
22. Ranallo-Benavidez, T.R., Jaron, K.S. and Schatz, M.C. 2020, GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes, Nat. Commun., 11, 1432.
23. Li, H. 2018, Minimap2: pairwise alignment for nucleotide sequences, Bioinformatics, 34, 3094–100.
24. Li, H. 2013, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, arXiv:1303.3997.
25. Günther, T. and Nettekoven, C. 2019, The presence and impact of reference bias on population genomic studies of prehistoric human populations, PLOS Genet., 15, e1007830.
26. McKenna, A., Hanna, M., Banks, E., et al. 2010, The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res., 20, 1297–303.
