Pervasive gene duplications as a major evolutionary driver of coral biology

Benjamin Noel[1,‡], France Denoeud[1,‡], Alice Rouan[2], Carol Buitrago-López[3], Laura Capasso[4,5], Julie Poulain[1], Emilie Boissin[6], Mélanie Pousse[2], Corinne Da Silva[1], Arnaud Couloux[1], Eric Armstrong[1], Quentin Carradec[1,7], Corinne Cruaud[8], Karine Labadie[8], Julie Lê-Hoang[1], Sylvie Tambutté[4], Valérie Barbe[1], Clémentine Moulin[9], Guillaume Bourdin[10], Guillaume Iwankow[6], Sarah Romac[11], *Tara Pacific Consortium Coordinators*[Δ], Denis Allemand[4], Serge Planes[6], Eric Gilson[2,12], Didier Zoccola[4], Patrick Wincker[1,7], Christian R Voolstra[3], Jean-Marc Aury[1,*]

‡ Equal contribution

* Corresponding author: jmaury@genoscope.cns.fr

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

[2] Université Côte d'Azur, CNRS, Inserm, IRCAN, France

[3] Department of Biology, University of Konstanz, Konstanz, Germany

[4] Centre Scientifique de Monaco, 98000 Monaco,

[5] Sorbonne Université, Collège Doctoral, F-75005 Paris, France

[6] PSL Research University: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Laboratoire d'Excellence CORAIL, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France

[7] Research Federation for the Study of Global Ocean Systems Ecology and Evolution, R2022/Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016, Paris, France

[8] Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

[9] Fondation Tara Océan, Base Tara, 8 rue de Prague, 75 012 Paris, France

[10] School of Marine Sciences, University of Maine, USA

[11] Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, ECOMAP, Roscoff, France

[12] Department of Human Genetics, CHU Nice, Nice, France

[Δ] The *Tara Pacific Consortium Coordinators* and their affiliations are listed at the end of this manuscript

Abstract

Over the last decade, several coral genomes have been sequenced allowing a better understanding of these symbiotic organisms threatened by climate change. Scleractinian corals are reef builders and are central to these ecosystems, providing habitat and food to a great diversity of species. In the frame of the Tara Pacific expedition, we generated two coral genomes, *Porites lobata* and *Pocillopora meandrina* with vastly improved contiguity that allowed us to study the functional organisation of these genomes. We annotated their gene catalog and reported a relatively higher gene number than that found in other public coral genome sequences. This finding is explained by a high number of tandemly duplicated genes, which are generally difficult to assemble and annotate using short-read sequencing technologies. These duplicated genes, which we show to originate from multiple and distinct duplication events in the coral lineage, belong to gene-families linked to the immune system and disease-resistance that we suggest to be functionally linked to coral host resilience. At large, we show the importance of previously missed duplicated genes to inform the biology of reef-building corals and provide novel avenues to understand and screen for differences in stress resilience.

Introduction

Coral reefs are one of the most diverse ecosystems on the planet. Although covering less than 0.2% of the ocean floor, coral reefs are home to over 25% of all described marine species[1,2]. Coral reefs also provide coastal protection and services to human societies. They support the livelihoods of millions of people through fishing or tourism[3]. Reef-building corals are the foundation species of coral reefs with critical roles in their function and maintenance. At the heart of this complex ecosystem, coral holobionts are meta-organisms composed of three main components: the coral host (cnidaria), photosynthetic Symbiodiniaceae (dinoflagellates), and associated prokaryotes, among other organismal entities[4,5].

For several decades now, coral reefs have been declining, impacted by global ocean warming, besides local anthropogenic impacts[6–9]. This temperature increase disrupts the coral and Symbiodiniaceae symbiosis leading to massive coral bleaching and mortality[10]. In addition, ocean acidification due to the increased levels of atmospheric CO_2 concentration reduces the ability of coral to produce its calcium carbonate skeleton and lowers its resilience[11,12]. Recently, the Intergovernmental Panel on Climate Change (IPCC) reported a projected decrease of 70% to 90% of the coral reefs coverage even if global warming is constrained to

1.5°C[13]. This will drastically affect reef ecosystems and spurs incentives to develop mitigation strategies besides the curbing of CO2 emissions[14–16].

Marine sessile species have a wide range of lifespans, ranging from weeks to thousands of years for some deep-sea corals and sponges[17]. Habitat depth appears to play a key role : indeed, at greater depths, living organisms are protected from issues that affect species in shallower waters, such as climatic temperature changes, climate events, and most importantly human activity[18]. Despite their fragility, corals are resilient organisms and several species have colonies with an extreme longevity, of the order of hundreds or thousands of years[17,19]. This could be seen as a paradox for those sessile species that cannot evade external threats and environmental changes. However, corals form colonies of multiple genetically identical and independent individuals, called polyps, and clonal organisms can escape age-related deterioration[20,21]. In addition, the colony can remain functional over time, even though parts may die.

The genome of *Acropora digitifera* was published ten years ago and was the first scleractinian coral genome available[22]. This first opportunity to uncover the architecture of a coral genome revealed a general absence of gene transfer with the endosymbiont, despite their long evolutionary relationship, a contradictory result with a more recent study[23]. It also highlighted the capacity to synthesise ultraviolet-protective compounds, the presence of genes with putative roles in calcification and a complex innate immunity repertoire with a putative role in the coral-Symbiodiniaceae symbiosis. With the further development of short-read sequencing technologies, a large number of coral genomes have been sequenced over the past decade[24–31]. The analysis of short-read assemblies from two corals of the highly diverged complex and robust clade[32] suggests, on the one hand, that many gene families exhibit expansion in corals (in particular genes having a role in innate immunity), and on the other hand that these gene family expansions have occurred independently in complex and robust corals[24]. Tandem organisation of these expanded gene families was suggested, as some amplified genes of a given family are localised on the same scaffolds[26,33]. Coral genomes are diploid and often highly heterozygous, which represents a major difficulty in generating high quality genomes[31,34]. Owing to the circumstance of short-read sequencing, many available coral genomes exhibit low contiguity and incomplete assemblies. Even though it has been generally accepted that short-reads assemblies are exhaustive for genes, repetitive regions are generally underrepresented[25,35], and in particular tandemly duplicated genes are a special case of repetitive regions which may be missed[35,36].

Here we report high-quality genome assemblies of two globally prevalent corals: the complex coral *Porites lobata* and the robust coral *Pocillopora meandrina*, based on long-reads generated using the Oxford Nanopore technology (ONT). In addition, we sequenced a second

genome of a morphologically similar *Porites* (*Porites evermanni*) with divergent stress susceptibility using short-reads[37]. On the basis of these three genomes and other available cnidarian genomes, we carried out a broad comparative analysis. Our results expose the vast presence of duplicated gene families in both coral genomes mapping to functions associated with the innate immune system, which escaped previous analyses based on fragmented and incomplete genomes assemblies due to sequencing method constraints. We posit that gene duplication is an evolutionary driver of coral biology that contributes to their longevity.

Results

Coral genome sequence assemblies and gene catalogs

The *P. lobata* and *P. meandrina* genomes were sequenced using a combination of ONT long reads and Illumina short reads (Tables S1 and S2). Using kmer distributions, *P. lobata* and *P. meandrina* genome sizes were estimated to be 543 Mb and 315 Mb respectively, and a high level of heterozygosity was detected, 2.3% and 1.14% respectively (Figure S1). As cumulative size of the two genome assemblies was almost twice as large as expected, and subsequent analyses revealed the presence of allelic duplications, Haplomerger2 was used on both assemblies to generate a haploid version (Figures S2 and S3, Table S3). The resulting haploid assemblies, with a cumulative size of 646 Mb for *P. lobata* and 347 Mb for *P. meandrina*, contained 1,098 contigs and 252 contigs with N50 of 2.15 Mb and 4.7 Mb, respectively (Table 1). In addition, we sequenced the genome of *Porites evermanni* using short-reads. Although more fragmented, this assembly has been used to perform comparative genomic analyses. Despite the fact that a large fraction of the repetitive elements is still unknown in the here-sequenced coral genomes (Table S4), DNA transposons were detected as the most abundant in the *P. lobata* genome (representing 17.4%), and in contrast, the most abundant repeat type was retroelements in the *P. meandrina* genome (10.5%). We annotated the three genomes using transcriptomic data (for *P. lobata* and *P. meandrina*) and 25 cnidarian proteomes, resulting in 42,872, 40,389 and 32,095 predicted genes for *P. lobata*, *P. evermanni* and *P. meandrina* respectively (Table 1). During the annotation process, we identified alignments of known proteins and transcripts that span large genomic regions (Figure S4) and further investigations indicated that these regions contain tandemly duplicated genes (TDG). These particular genes are generally difficult to assemble and to predict accurately. Here we developed a new annotation process to systematically improve the annotation of these TDG.

Finally, gene completeness was estimated using BUSCO and was 97.7%, 98.4% and 94.5%, respectively (Table 1).

Telomeric sequences

Telomeres are composed of short repeated DNA sequences located at the end of linear eukaryotic chromosomes. The telomeric repeat motif TTAGGG is highly conserved among metazoans[38]. This type of sequence can also be found within the chromosome and is therefore called interstitial telomeric sequence (ITS). We identified ITSs in our three genome assemblies and the previously sequenced *Stylophora pistillata*[24], along with a low proportion of contigs with telomeric repeats at their ends (5 and 3 for *P. meandrina* and *P. lobata*, respectively), suggesting the absence of contigs representing complete chromosomes and the quasi-absence of terminal chromosome fragments. This absence of telomeric repeats at contig ends may be the consequence of a technical issue during the basecalling of nanopore data[39]. Noteworthy, we noticed in the three *Porites* species (*P. lobata*, *P. evermanni* and *P. lutea*) the presence of a 188-nt length satellite DNA sequence containing a palindromic telomeric sequence (Figure S5). These satellites are found tandemly repeated in intergenic regions. Attempts to search for this sequence failed to find it outside the *Porites* genus.

Comparison with available coral genome assemblies

To date, only three other scleractinian genomes, i.e. *Montipora capitata, Acropora millepora*, and *Acropora tenuis* have been assembled using long read sequencing technologies[28,40,41], and the genome sequences of *P. lobata* and *P. meandrina* we have generated are the most contiguous and complete coral genome assemblies so far (Figure 1). Likewise, we observed that several genomes have a high number of duplicated BUSCO genes, indicating that they still contain allelic duplications, potentially due to the aforementioned high levels of heterozygosity (Figure 1G). Coupled with a fragmented assembly, these remaining duplications are detrimental for subsequent analysis, as it is then complicated to differentiate true duplications from allelic copies of a given gene. In our two high-quality assemblies of *P. lobata* and *P. meandrina*, we predicted more genes than in other coral species of the same genus (Figure 1D) with a proportion containing conserved domains comparable to other corals (78% and 80% respectively, Figure 1E). To further compare coral genomes, orthologs with 22 other cnidarian species were calculated (Table S5). As expected, conservation between orthologous genes inside the *Porites* and *Pocillopora* genera was high. Surprisingly, however, the conservation of orthologous genes between *P. lobata* and other robust corals was as low as the conservation to other complex corals that are at least 245 mya apart[42] (Figure S6).

Notably, although an initial matter of debate, the classification of coral species into two evolutionary divergent clades (complex and robust) is recognised as real[26,32] and confirmed in our analyses (Figure 1A). As few morphological or biological criteria resolve the two groups[26], we suggest here that guiding the analyses by splitting into robust and complex clades does not always make sense, and alternative grouping could be sometimes more relevant to compare coral genomes. In addition, these orthologous relationships allowed us to examine the conservation of gene order within corals. As already reported, synteny across complex and robust coral lineages is highly conserved[26,33] (Figure 2). With their higher contiguity, the two long-read assemblies better resolve the macro- and micro-synteny within each of the two lineages. Interestingly, the synteny between complex and robust corals, which was previously described as conserved[26], is not conserved at the scale of large genomic regions. Indeed, fragmented assemblies give only a partial insight into the synteny between organisms. Here, we observed only a conservation at the micro-synteny level between *Porites* and *Pocillopora*, and despite the 245 Mya that separate these species, the syntenic blocks nevertheless cover at least 75% of both genomes (Figure 2 and Table S6). In comparison, only 40% of the assemblies are covered if comparing the two short-read assemblies of *P. lutea* and *P. verrucosa*, showing the shortcomings associated with analysing fragmented genomes.

Tandemly duplicated genes

Tandem duplications are an important mechanism in the evolution of eukaryotic genomes, notably allowing the creation of unconstrained genes that can lead to new functions[43–45], in particular for genes clustered into gene families[46,47]. We detected tandemly duplicated genes (TDG) in the available Cnidaria genome assemblies and annotations, and found a high proportion of TDG in *P. lobata* and *P. meandrina*, 29.9% (12,818 genes) and 32.6% (10,449 genes) of their respective gene catalog. In comparison, the proportion of TDG is lower in short-read assemblies (Figure 3A), except for *Orbicella faveolata* which also displays a high proportion of allelic duplications making TDG detection confusing (Figure 1G). Clusters of TDG are scattered on all contigs (Figure S7) and contain on average two genes in both coral genomes, but some clusters contain more genes up to a maximum of 64 genes for *P. lobata* and 48 genes for *P. meandrina* (Figure 3B). In general, we found larger clusters of TDG in genome assemblies with higher N50, which is expected as larger genomic sequences contain more candidate genes.

The fact that other coral genomes have a lower proportion of TDG than *P. lobata* and *P. meandrina* was surprising, and we investigated whether this difference could be due to biases in the genome assembly or gene prediction workflows. To be independent from such biases, the number of members in gene families was estimated using short read-based data and

conserved genes. Orthologous genes computed with the 22 other Cnidaria species were grouped into orthogroups (OG) and a consensus sequence was built for each OG. The number of gene-copies per OG was estimated for each species by aligning short-read sequencing data of the corresponding species to each OG consensus. Normalisation was performed using 705 coral-specific and single-copy genes. The estimated gene copy number based on mapping of short reads is similar among all *Porites* and *Pocillopora* species, whereas the number of annotated gene copies is higher for *P. lobata* as well as for *P. meandrina*. We found that *Porites* gene catalogs of short-read assemblies (*P. lutea* and *P. evermanni*) lack a high number of copies when compared to *P. lobata*, and the same trend was observed when comparing *Pocillopora* short-read assemblies with *P. meandrina* (Figure 3C). Genome assemblies based on short reads thus appear to lack a substantial number of gene copies, particularly in TDG clusters. Indeed, *P. lobata* and *P. meandrina* have a higher number of genes but also a higher proportion of genes linked to an OG (respectively 96.6% and 98.1% of the genes are in an OG composed of genes from at least two different species, Figure 1F), suggesting that their gene annotation is a better representation of the gene catalog of coral genomes. Interestingly, the *P. evermanni* assembly is also based on short reads but our improved gene prediction method appears to exonerate the number of missing gene copies (Figure 3C).

Amplified gene families in corals

To assess whether TDG contributed to gene family expansions, we searched for orthogroups (OG) with significant gene number differences between corals and sea anemones. We identified 192 OG that were expanded in corals (Extended data 1) in comparison to only 28 OG in sea anemones (Figure 4). Most of the expanded gene families contained a high ratio of TDG (Figure 4C), which suggests that tandem duplication is an important mechanism for gene family amplification in corals. The functions of amplified OG, based on InterProscan domain identification and blastP searches, correspond in vast majority to transmembrane receptors, cell adhesion and extracellular signal transduction. The most abundant domain is G-protein coupled receptor, rhodopsin-like (GPCR), that corresponds to 34/192 OG amplified in corals. Among the receptors that were identified, some are involved in innate immunity and possibly coral/Symbiodiniaceae symbiotic relationships[48]. As previously reported in other coral species[24,25], we observe a high heterogeneity of copy numbers in gene families among coral genera: each coral genus displays specific gene family expansions, with similar profiles among *Porites* species and among *Pocillopora* species (Extended data 1 and Figure 4B). Additionally, we detect more pronounced amplifications in five genome species, i.e. *Porites lobata*, *Porites evermanni*, *Pocillopora meandrina*, *Goniastrea aspera* and *Fungia sp*. However, the lack of

high-quality assemblies and annotations for some of the genomes did not allow us to ascertain the biological significance of these observations. These results confirm on a larger range of species and at a higher scale, that although the same functional categories (extracellular sensing, cell adhesion, signalling pathways) are amplified in all corals, individual amplified gene families diverge among genera, corroborating previous notions[24].
To relax constraints of the analysis, we looked at a broader scale considering all genes and using PFAM domain annotations. Similar to the above analysis, heterogenic profiles were observed in that corals were clearly distinct from other cnidarians but diverse among themselves. It is especially noteworthy that domain abundances were more consistent between Actiniara (*Aiptasia*) and Corallimorpharia (Afen, Disco), than within Scleractinia (Figure S8). Nevertheless, corals of the genus *Porites* and family Pocilloporidae clustered together. Thus, even at the domain level, corals were diverse with regard to amplified functional families. However, in a broader view, the gene functions of the amplified domains could be assorted majoritively (Extended data 2) to signalling pathways. Glycosyl transferase domains stand out as very enriched in corals. Such domains were shown to be amplified in various coral species[26] and to be associated with NACHT domains (NLR proteins) and TIR domains (TLR proteins) in *Acropora*[49].

Comparison of gene family amplifications within corals

To date, only two biosynthetic differences between robust and complex have been reported[22,26] and few biological criteria resolve the two groups. Here, based on the same method as for the coral vs sea anemones comparison, we identified OG that are differentially expanded in robust vs complex species. We obtained a list of 38 OG amplified in robust species and 43 OG amplified in complex species (Extended data 3 and 4). Again, we observe that the differences between gene abundance occur mainly at the genus level rather than transcend to the level of robust or complex lineages (Figure S9), *i.e.* that gene expansions are genus- or species- specific. For instance, OG0000628 (G protein-coupled receptor, rhodopsin-like) is strongly abundant in *Porites lobata* but not in other complex species and OG0000316 (C-type lectin-like) is abundant in *Pocillopora meandrina* but not in other robust species. Considering all genes and taking PFAM domain annotation into account, analysis of enriched PFAM domains shows that although corals broadly separate into complex and robust clades, coral species within their respective clades exhibit differences with regard to domain abundance. Notably, *Pocilloporidae* corals within the robust clade as well as *Porites* corals within the complex clade cluster together, suggesting that besides the substantial differences between species, conserved patterns that align with phylogeny at various levels are perceptible (Figure S10 and Extended data 5).

Additionally, we tried to discriminate coral species based on the morphological distinction of massive and branched coral colonies, since it has been described that massive colonies are more tolerant to bleaching than branched colonies[50–52]. We obtained a list of 65 OG amplified in massive species and 20 OG amplified in branched species (Figure S11, Extended data 6 and 7). Even if the amplified gene families have similar functions, the situation is much more imbalanced compared to the robust/complex comparison. This observation may suggest the functional link between the observed gene amplification (especially disease-resistance and immune gene-pools) and the resilience of massive species.

Since the number of genes annotated in all species might not be comparable (especially for TDG in genomes sequenced with short reads), we compared the two genomes that we sequenced with long reads and annotated with the same procedure, *Porites lobata* and *Pocillopora meandrina*. We show that among the 192 OG amplified in corals, most display higher gene copy numbers in *Porites lobata* than *Pocillopora meandrina* (respectively 117 and 64; Extended data 1), which is in agreement with the observation that more TDG are detected in *Porites lobata* than *Pocillopora meandrina*. However, OG containing EGF-like domains, and especially EGF_CA (calcium-binding EGF domain) are more abundant in *Pocillopora meandrina* than *Porites lobata* (Figure 5A). This domain has previously been shown to be abundant in the extracellular matrix[53]. But again, the lack of long-read assemblies makes it difficult to determine whether these differences are a general trait in complex and robust corals or massive and branched corals.

Mechanisms of gene amplification and evolution of amplified genes

We investigated evolutionary rate variation of expanded gene families using CAFE[54]. We reconstructed the number of genes in each OG for internal nodes of the phylogenetic tree and identified significant increases or decreases in gene copy numbers. It is notable that more events have occurred on branches leading to coral species than to other hexacorallia. These events are occurring on various branches of the phylogenetic tree and each OG has a different history (Figure S12), which is in accordance with the variety of gene abundance profiles already observed (Figure 4B). We also detected a large number of gene amplifications that are species specific. However, the species for which the highest number of gene amplifications is detected by CAFE is *Porites lobata*, which likely reflects the higher number of annotated TDG compared to other *Porites* (Figure S13). This result highlights the difficulty to conduct such analyses of amplification/reduction with species having heterogeneous gene prediction exhaustivity.

To trace the evolutionary history of these duplicated genes, we calculated the synonymous substitution rates (Ks) between pairs of tandemly duplicated genes in *Porites lobata* and

Pocillopora meandrina. Ks are used to represent the divergence time between duplicated copies (lower Ks reflect higher divergence). Although a single peak in the Ks distribution is distinguishable for the two species (Figure 5B), the degrees of conservation are highly variable within orthogroups (Figure 5C) or even between tandemly duplicated gene clusters inside one orthogroup (Figure 6E). When looking at an example of gene family amplified in corals (TIR domain-containing) (Figure 6), we observe that some TDG predate the scleractinian/non-scleractinian divergence, and others are specific of robust or complex clades, or of *Porites* (Figure 6B) or *Pocilloporidae* (Figure 6C). As expected, for TDG shared by more species (more ancient), Ks are higher (Figure 6E). Gene family amplification by tandem duplication thus appears to be a dynamic process that has been occurring for a long time and is still at play. We propose that, in corals, the main gene family expansion mechanism is birth-and-death evolution[55]. Birth of new copies occurs by tandem duplications at the genomic level, which is in accordance with the observation that more distant gene pairs on the genome show lower conservation (Figure S14). This mechanism of duplication at the genomic level is consistent with the existence of duplications of groups of two or three adjacent genes together (Figure S15). Tandem duplication is followed by divergence of the gene copies, especially in intronic sequences. Indeed, when comparing structures of duplicated gene pairs, the vast majority show conserved exon lengths unlike introns (only 11 *Porites lobata* and 21 *Pocillopora meandrina* gene pairs show perfectly conserved exon and intron structures), and the sequence conservation of introns is lower than that of exons (Figures S16 and S17).

Transcription profile of amplified genes in environmental samples

Gene duplication plays a key role in the creation of novelty[43] (including new gene functions) but can also, when associated with regulatory mutations, affect gene expression and lead to new expression patterns. These changes in gene expression may underlie much of phenotypic evolution[56]. Therefore, comparing the expression patterns of both old and new duplicate genes can provide information about their functional evolution. Here, we quantified the abundance of *Pocillopora meandrina* transcripts in 103 available environmental samples coming from 11 different islands of the Tara Pacific expedition (Figure 7A). Of the 32,095 genes of *P. meandrina*, 94.3% are expressed in at least one environmental sample. This proportion is even higher for TDG (97.4%) underlining the strong support of these particular genes. We then examined the expression of the duplicated genes across the 103 samples and found highly correlated expression patterns in the case of recent duplicated genes (average Pearson 0.6) compared to old gene duplications (average Pearson 0.2, Figure 7B). This observation highlights the importance of mutations in gene expression changes after gene duplications. The subfunctionalization of expression is described as a rare event[57], and usually recent

duplicates are down-regulated in accordance with the dosage-sharing hypothesis. Data generated in the frame of the Tara Pacific expedition did not allow us to investigate the expression dosage of these duplicate genes at the colony level, because one cannot exclude the existence of differences in the number of copies of a given gene between the colonies, but we suggest that the high number of duplicated genes in corals makes them interesting models to study these questions. Additionally, we investigated the expression profiles of the genes from the 192 amplified OG in coral genomes. We did not find significant differences between gene function or provenance of transcriptomic samples, except for the island of Rapa Nui. Indeed, amplified genes appear to exhibit increased expression in all the samples of this island (Figure 7C), which is also the case for non-amplified genes (Figure S18). Interestingly, the Rapa Nui island exhibits a unique association of coral and symbiont lineages (Armstrong et al, in preparation) as well as short telomeric DNA associated to an overexpression of several telomere maintenance genes (Rouan et al, in preparation).

Amplification of gene families functionally important for corals

Innate immune system receptors

Most of the gene families that are amplified in coral genomes correspond to receptors that are likely related to self/non-self recognition and play a role in innate immunity or host-symbiont interactions[58,59]. Membrane pattern recognition receptors (PRRs) play a central role in immune recognition in cnidarians[59,60]. Toll-like receptors (TLRs) are transmembrane receptors containing TIR (Toll/Interleukin-1 receptor) domain, leucine rich repeats (LRRs), and a cysteine-rich domain[61,62]. Lectins, including tachylectins[63] also act as PRRs and are involved in activation of the complement cascade. NOD-like receptors function as cytosolic receptors and contain NACHT/NB-ARC domains[49]. These gene families have previously been shown to be amplified in corals compared to other cnidaria[61,25,22]. We identified TIR, NACHT and lectin containing gene families based on their domain composition and counted the number of corresponding genes in each coral species (Extended data 8). As seen for the 192 OG amplified in corals (Figure 4B), each species/genus displays specific amplified innate immunity gene families (Figure S19). This observation is in agreement with earlier studies that have shown that coral species diverge on which innate immunity-related genes are expanded, and have also suggested that innate immune pathways might play diverse adaptive roles[24,25]. Strikingly, *Porites* species display the highest number of gene copies when cumulating the three innate immune system receptor categories studied here (Figure 8A). It is tempting to speculate that this huge repertoire of innate immune system genes could play a role in the notably long lifespan and high resilience of *Porites* species. Interestingly, we also identified a

high number of GPCR-like (G-protein coupled receptor) proteins among coral-amplified gene families: the function of this very abundant transmembrane receptor family in corals will require further investigation, but it is likely to be involved in innate immunity and/or host-symbiont interactions.

We studied the domain composition of TIR containing and NACHT/NB-ARC containing proteins in 5 coral species, and confirmed the high number of domain combinations already observed in *Acropora digitifera*[49]. Interestingly, our high-quality assembly of *Porites lobata* allowed the annotation of a high number of IL-1R-like proteins (containing TIR domains associated with Immunoglobulin domains). We also identified SARM-like proteins shared between all corals, and confirmed the expansion of TIR-only proteins in corals[61] (Figure 8B).

Porites lobata contains a much lower proportion of truncated (lacking N- and/or C-terminal regions) NACHT/NB-ARC containing genes than *Porites lutea* or *Acropora digitifera*, where a high number of NACHT/NB-ARC-only proteins were described[49] : we hypothesise that a substantial fraction of these proteins corresponds to truncated annotations (Figure S20). As already observed for NOD-like receptors in *Acropora digitifera*, the effector and C-terminal repeated domains found are distinct between NACHT containing (mostly LRR and WD40 repeats) and NB-ARC containing (mostly tetratricopeptide repeats) proteins (Figure S21). Distinct domain combination abundances are observed in the different coral species, and some are species-specific (Figure 8C). A high number of N-terminal RNA-binding “DZIP3-like HEPN” domains are observed. HEPN domains are found in a variety of defence and stress response systems across the tree of life and could play a role in antiviral, antitransposon, apoptotic systems or RNA-level response to unfolded proteins[64].

Calcification-related genes

Corals build the structural foundation of coral reefs through calcification. Genes and their associated functions involved in this process are already well characterised[65]. Here, we examined seven families of candidate genes encoding a set of proteins involved in coral calcification[53,66]. Among these families, proteins can be divided into 2 categories on the basis of their role and localization in calcification: ion membrane transporters/enzymes and skeletal organic matrix proteins (Figure S22). The first category comprises: ammonium transporters Amt1 (Figures S23 and S24, Capasso et al, submitted); the Bicarbonate Anion Transporter SLC4 gamma[66,67] (Figure S25); plasma-membrane calcium ATPases[53,67] (PMCAs); and carbonic anhydrases[65,68,69] (CAs). CAs also fall within the category of organic matrix proteins since one isoform is found in coral skeletons[70]. Organic matrix proteins comprise coral acid rich proteins[71,72] (CARPs) and neurexin[73].

We compared these seven families in the genomes of 12 scleractinian (6 Robust and 6 Complex) and 3 non-scleractinian species (two Corallimorphs and one Actiniaria) and found different evolutionary histories (Table S7). As observed previously[53], calcification related genes are often clustered in tandem in coral genomes (Figures S24 and S25). Scleratinian divide into Robust and Complex clades, which diverged about 245 Mya and show different skeletal properties[53]. They diverged from Actiniaria approximately 506 Mya and from Corallimorpharia about 308 Mya at the time they acquired the ability to calcify[53]. First, our results show that a set of proteins (PMCA, neurexin) are present in the proteomes of all *Pocillopora* and *Porites* as well as other Hexacorallia with no significant difference in number. Second, they possess the SCL4γ, which is scleractinian specific and is a tandem duplication of the SLC4β gene. Finally, *Pocillopora* possesses a higher number of orthologs of CARPs, Amt1 and CAs than *Porites* (Table S7 and Figure S22). Our results clearly confirm that gene amplification for this set of calcification related gene families differ between Robust and Complex but show co-option of genes and neofunctionalization for calcification that occured during evolutionary history of species.

Discussion

The introduction of long-read sequencing is a game changer for obtaining complete reference genome sequences. Short-read sequencing has long been considered sufficient to access the gene catalog of a given species. Herein we have shown, using two coral genomes from two different groups, that long-read technologies are essential to obtain an accurate view of the gene content and the functional landscape of genomes. Coupled with a dedicated annotation method, we highlighted pervasive and species-specific gene duplications in coral genomes. This peculiarity was previously underestimated due to the difficulty of assembling repetitive regions with short reads and the greater fragmentation of the resulting assemblies which did not allow the detection of large blocks of TDG. Moreover, the high level of heterozygosity in coral genomes complicates the detection of duplicated genes. Indeed, the remaining allelic duplications isolated on small contigs can be confused with true duplicated genes.

Based on these complete gene catalogs of corals, we observe large (and ancient) arrays of TDG that remained clustered on the genome over a long time. This situation is quite unusual compared to what is observed in plant genomes for instance, where only recent TDG are clustered together and ancient copies have been translocated to other chromosomes[46,74]. Translocation of duplicated genes has been proposed to be a means to escape concerted evolution, that is homogenisation of gene copies through gene conversion[75]. In corals,

however, gene conversion does not appear to be a common mechanism, since very few gene pairs happen to show strongly conserved exon/intron structures, and even genes that are tightly linked on the genome accumulate mutations (mostly in introns). Accordingly, in the last decades, the growing availability of genomic data revealed that most multigene families display high levels of intraspecific diversity, which is not consistent with a homogenising mechanism[76]. The currently accepted model of multigene family evolution is the birth-and-death evolution model, first proposed by Nei and colleagues[55,77]. This model is in accordance with what is observed in coral gene families, where tandem duplication appears to be a dynamic process that has been taking place for a long time and is still ongoing. The fact that tandemly repeated genes have remained clustered together could be related to the high synteny conservation between coral species. This observation suggests that gene translocation (leading to loss of synteny) is not needed to allow genes to diverge and evolve new functions, since gene conversion is very rare.

The high number of TDG in corals and their maintenance in arrays suggest that these species are excellent models to study tandem duplications in genomes, although this complicates the generation of contiguous genome assemblies. Sequencing and annotating more coral genomes at chromosome-scale and a high quality level, especially outside *Porites* and *Pocillipora* genera, is important. This will first give a higher resolution of the scleractinia order, then will be necessary to decipher how and where genomic duplications occur, and the evolution of these gene duplications. Additional transcriptomic resources will also be needed to make it possible to investigate the fate of duplicated copies, in particular the impact of duplicated genes on the expression dosage. Global gene expression analysis of *Pocillopora* has revealed a high transcriptomic plasticity dependent on both the genetic lineage and the environment (Armstrong et al, in preparation). It is tempting to draw a parallel with our observation of divergent expression profiles between duplicated genes and suggest that these gene expression patterns may play a role in the acclimatisation capacities of *Pocillopora* to the environment.

Our comparative analysis based on available coral genomes reveals a high number of amplified gene families compared to sea anemones. We show that these gene families are functionally related to signal reception and transduction, especially innate immunity. Several studies have shown that some of these families play a key role in maintaining the symbiosis of corals with their associated Symbiodiniaceae[48,78]. We also noted that these families are mainly amplified through tandem duplications, and their retention in the genome through evolution underlines their functional importance in corals. Gene duplication has already been

described as playing an important role in phenotypic evolution: in particular, a link with long lifespan has already been reported in other organisms, such as trees and fishes. Indeed, a convergent expansion of disease-resistance gene families across several tree species suggests that the immune system contributes to the survival of long-lived plants[46]. Similarly, gene expansion of immunoregulatory genes in rockfishes may have facilitated adaptations to extreme life span[79]. We hypothesise that these amplified gene families in corals, related to innate immunity and disease-resistance, may have contributed to the resilience and long lifespan of these sessile organisms. Our analyses also revealed a short sequence of 188-bp tandemly repeated in several intergenic regions that is uniquely found in *Porites* genomes. This satellite-like sequence is intriguing because it contains a palindromic telomeric sequence. Since telomeres are key ageing hallmarks in numerous organisms[80], it is tempting to speculate that it is involved in the stress resistance and extreme longevity features shared by *Porites* species[81].

Precise identification of the timing of duplications for each individual gene is hampered by the fact that current coral genome assemblies and annotations have missed some duplicated genes. However, we found ancient duplicated genes that are shared by all corals, but also recently duplicated genes that are species- or genus-specific. Even though functions of amplified gene families are common across coral species, the individual genes that are amplified can be different. This striking pattern is a neat example of convergent evolution and can be seen as an important evolutionary advantage. Differences between coral species, namely which genes have been amplified as well as which new expression patterns have emerged within duplicated genes, could provide them different abilities to overcome environmental changes. This gene-content plasticity may represent an opportunity for coral species and involved genes are potential targets for assisted evolution of more resilient corals[48,82,83].

Methods

Coral material

Porites evermanni, *Porites lobata* and *Pocillopora meandrina* colonies were collected in French Polynesia by the CRIOBE at Moorea. Fifty grams of each coral colony were flash frozen using liquid nitrogen and stored at -80 °C until further use. In addition, a small fragment of *Porites lobata* was placed in 2 ml of Lysing Matrix A beads (MP Biomedicals, Santa Ana, CA, USA) in presence of 1.5 ml of DNA/RNA shield (Zymo Research, Irvine, CA, USA) and stored at -20 °C until RNA purification.

DNA extraction

DNA extractions from flash frozen tissues were based on a nuclei isolation approach to minimise contamination with Symbiodiniaceae DNA. Briefly, cells were harvested using a Waterpik in 50 ml of 0.2 MEDTA solution refrigerated at 4°C. Extracts were passed sequentially through a 100 µm then a 40 µm cell strainer (Falcon) to eliminate most of the Symbiodiniaceae. Then extracts were centrifuged at 2,000 g for 10 min at 4°C and the supernatants were discarded. Two different protocols of DNA purification were used as follows. For *Porites evermanni* and *Pocillopora meandrina*, the resulting pellets were homogenised in lysis buffer (G2) of the Qiagen Genomic DNA isolation kit (Qiagen, Hilden, Germany). The DNA were then purified following manufacturer instructions using genomic tip 100/G. For *Porites lobata* the pellet was homogenised in CTAB lysis buffer followed by a Chloroform/isoamyl alcohol purification.

RNA extraction

Total RNA of *Pocillopora meandrina* was extracted from flash frozen tissue. For *Porites lobata*, the fragment placed in 2 ml of Lysing Matrix A beads (MP Biomedicals, Santa Ana, CA, USA) in presence of 1.5 ml of DNA/RNA shield (Zymo Research, Irvine, CA, USA) was thawed and disrupted by the simultaneous multidirectional striking using a high-speed homogenizer FastPrep-24 5G Instrument (MP Biomedicals, Santa Ana, CA, USA) (following conditions: speed: 6.0 m/s, time: 30 s, pause time: 60 s, cycles: 3). Total RNA was then purified from one aliquot of 500 µl of homogenised suspension, following the instruction of the commercial Quick-DNA/RNA Kit (Zymo Research, Irvine, CA, USA).

Approximately 5 µg of total purified RNA were then treated with the Turbo DNA-free kit (Thermo Fisher Scientific, Waltham, MA, USA), according to the manufacturer's protocol. Quantity were assessed on a Qubit 2.0 Fluorometer using a Qubit RNA HS Assay kit and the quality were checked by capillary electrophoresis on an Agilent Bioanalyzer using the RNA 6,000 Pico LabChip kit (Agilent Technologies, Santa Clara, CA, USA). Purified RNA was stored at -80 °C until further use.

Illumina library preparation and sequencing

Genome sequencing of *Porites evermanni* was performed using Illumina reads from both paired-end and mate-pair (MP) libraries of different insert sizes. The paired-end library was prepared using the NEBNext DNA Modules Products (New England Biolabs, MA, USA) with a 'on beads' protocol developed at the Genoscope, as previously described[84] in Alberti et al. (doi: 10.1038/sdata.2017.93.). The library was quantified by qPCR (MxPro, Agilent Technologies) using the KAPA Library Quantification Kit for Illumina Libraries (Roche), and its profile was assessed using a DNA High Sensitivity LabChip kit on an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). The library was paired-end sequenced on the Illumina HiSeq2500 (Illumina, USA) sequencing platform (2 x 251bp). The MP libraries were prepared using the Nextera Mate Pair Sample Preparation Kit (Illumina, San Diego, CA). Briefly, genomic DNA (4 µg) was simultaneously enzymatically fragmented and tagged with a biotinylated adaptor. Tagmented fragments were size-selected (3-5; 5-8; 8-11 and 11-15 Kb) through regular gel electrophoresis, and circularised overnight with a ligase. Linear, non-circularized fragments were digested and circularised DNA was fragmented to 300-1000-bp size range using Covaris E220. Biotinylated DNA was immobilised on streptavidin beads, end-repaired, and 3`-adenylated. Subsequently, Illumina adapters were ligated. DNA fragments were PCR-amplified using Illumina adapter-specific primers and then purified. Finally, libraries were quantified by qPCR and library profiles were evaluated using an Agilent 2100 Bioanalyzer. Each library was sequenced using 151 base-length read chemistry on a paired-end flow cell on the Illumina HiSeq4000 sequencing platform (Illumina, USA).

The genomes of *Porites lobata* and *Pocillopora meandrina* were sequenced using a combined approach of short and long reads. The short reads were obtained by preparing Illumina PCR free libraries using the Kapa Hyper Prep Kit (Roche). Briefly, DNA (1.5µg) was sonicated to a 100- to 1500-bp size range using a Covaris E220 sonicator (Covaris, Woburn, MA, USA). Fragments were end-repaired, 3′-adenylated and Illumina adapters were ligated according to the manufacturer's instructions. Ligation products were purified with AMPure XP beads (Beckman Coulter Genomics, Danvers, MA, USA) and quantified by qPCR. Library profiles were assessed using an Agilent High Sensitivity DNA kit on the Agilent 2100 Bioanalyzer.

Libraries were paired-end sequenced on an Illumina HiSeq2500 instrument (Illumina, San Diego, CA, USA) using 251 base-length read chemistry.

Illumina RNAseq libraries were prepared for both *Porites lobata* and *Pocillopora meandrina* using the TruSeq Stranded mRNA kit (Illumina, San Diego, CA, USA), according to the manufacturer's protocol starting with 500 ng total RNA. Briefly, poly(A)+ RNA were selected with oligo(dT) beads, chemically fragmented and converted into single-stranded cDNA using random hexamer priming. Then, the second strand was generated to create double-stranded cDNA. The resulting cDNAs were subjected to A-tailing, adapter ligation and PCR-amplification. Ready-to-sequence Illumina libraries were then quantified by qPCR and library profiles evaluated with an Agilent 2100 Bioanalyzer. Each library was sequenced using 151 bp paired-end reads chemistry on a HiSeq4000 Illumina sequencer. Short Illumina reads were bioinformatically post-processed sensu Alberti et al[84] to filter out low quality data. First, low-quality nucleotides (Q < 20) were discarded from both read ends. Then remaining Illumina sequencing adapters and primer sequences were removed and only reads ≥ 30 nucleotides were retained. These filtering steps were done using in-house-designed software based on the FastX package[85]. Finally, read pairs mapping to the phage phiX genome were identified and discarded using SOAP aligner[86] (default parameters) and the Enterobacteria phage PhiX174 reference sequence (GenBank: NC_001422.1).

MinION and PromethION library preparation and sequencing

Genomic DNA fragments of *Pocillopora meandrina* ranging from 20 to 80 Kb were first selected using the Blue Pippin (Sage Science, Beverly, MA, USA) then repaired and 3'-adenylated with the NEBNext FFPE DNA Repair Mix and the NEBNext® Ultra™ II End Repair/dA-Tailing Module (New England Biolabs, Ipswich, MA, USA). Sequencing adapters provided by Oxford Nanopore Technologies (Oxford Nanopore Technologies Ltd, Oxford, UK) were then ligated using the NEBNext Quick Ligation Module (NEB). After purification with AMPure XP beads, the library was mixed with the Sequencing Buffer (ONT) and the Loading Beads (ONT). For *Pocillopora meandrina*, a first library was prepared using the Oxford Nanopore SQK-LSK108 kit and loaded onto a R9.5 MinION Mk1b flow cell. Three other libraries were prepared using the same kit and following the same protocol, but without the size selection. They were loaded onto R9.4.1 MinION Mk1b flow cells. An additional library was prepared using the Oxford Nanopore SQK-LSK109 kit and loaded onto a R9.4.1 PromethION flow cell. Reads were basecalled using Albacore version 2. For *Porites lobata*, eight libraries were prepared using the Oxford Nanopore SQK-LSK109 kit. Four libraries were loaded onto R9.4.1 MinION Mk1b flow cells and the other four onto PromethION R9.4.1 flow

cells. Reads were basecalled using Guppy version 2. In both cases, the resulting raw nanopore long reads were directly used for the genome assembly.

Short reads-based genome assembly

The genome of *Porites evermanni* was assembled with megahit[87] using the filtered high-quality Illumina technology paired-end reads from shotgun libraries (Table S1). The resulting assembly was scaffolded using mate-pair libraries and SSpace[88] and gap-filled with gapcloser[89] and the paired-end reads. This process generated an assembly of 604Mb composed of 8,186 scaffolds with an N50 of 171Kb (Table 1).

Long reads-based genome assemblies

To generate long-reads based genome assemblies we generated three samples of reads: i) all reads, ii) 30X coverage of the longest reads and iii) 30X coverage of the filtlong[90] highest-score reads that were used as input data for four different assemblers, Smartdenovo[91], Redbean[92], Flye[93] and Ra (Table S2). Smartdenovo was launched with -k 17 and -c 1 to generate a consensus sequence. Redbean was launched with '-xont -X5000 -g450m' and Flye with '-g 450m'. The resulting assemblies were evaluated based on the cumulative size and contiguity, with the Smartdenovo and all reads combination producing the best assembly. This assembly was polished three times using Racon[94] with Nanopore reads, and two times with Hapo-G[95] and Illumina PCR-free reads.

Reconstruction of allelic relationships and haploid assembly

The cumulative size of both assemblies was higher than expected due to the high heterozygosity rate (Figures S1). As indicated by BUSCO[96] and KAT[97], we observed the two alleles for many genes and a significant proportion of homozygous kmers were present twice in the assembly. We used Haplomerger2 with default parameters and generated a haploid version of the two assemblies. Haplomerger2 detected allelic duplications through all-against-all alignments and chose for each alignment the longest genomic regions (parameter --selectLongHaplotype), which may generate haplotype switches but ensure to maximise the gene content. At the end of the process, the *P. lobata* and *P. meandrina* diploid assemblies have a cumulative size of 650Mb and 350Mb respectively, closer to the expected ones. BUSCO and KAT analysis showed a reduction of the allelic duplications (Table S3 and Figures S2 and S3). Final assemblies were polished one last time with Hapo-G[95] and Illumina short reads to ensure that no allelic regions present twice in the diploid assembly have remained unpolished.

Transcriptome assembly

First, ribosomal RNA-like reads were detected using SortMeRNA (Kopylova et al., 2012) and filtered out. Illumina RNA-Seq short reads from *Porites lobata* and *Pocillopora meandrina* were assembled using Velvet[98] 1.2.07 and Oases[99] 0.2.08, using a k-mer size of 89 bp and 81 bp respectively. Reads were mapped back to the contigs with BWA-mem and only consistent paired-end reads were kept. Uncovered regions were detected and used to identify chimeric contigs. In addition, open reading frames (ORF) and domains were searched using respectively TransDecoder and CDDsearch. Contigs were broken in uncovered regions outside ORF and domains. At the end, the read strand information was used to correctly orient RNA-Seq contigs.

Repeat detection

Libraries of genomic repeats were first detected using RepeatModeler (v.2.0.1, default parameters) on both genomes. Then, these libraries were annotated with RepeatMasker[100] (v.4.1.0, default parameters) and RepBase (from RepeatMasker v4.0.5). Finally, the *P. lobata* and *P. meandrina* genomes were masked using their respective libraries. The numbers of bases were counted according to the classification of repeat overlapping each base. In case of overlapping repeated fragments, the longest annotated one was selected.

Gene prediction

Gene prediction was done using proteins from 18 Cnidarian species, *Acropora digitifera*, *Acropora millepora*, *Aiptasia*, *Aurelia aurita* from Atlantic, *Aurelia aurita* from Pacific, *Clytia hemisphaerica*, *Fungia* spp., *Galaxea fascicularis*, *Goniastrea aspera*, *Hydra vulgaris*, *Montipora capitata*, *Morbakka virulenta*, *Nematostella vectensis*, *Orbicella faveolata*, *Pocillopora damicornis*, *Porites lutea*, *Porites rus* and *Stylophora pistillata*. The proteomes were aligned against *Porites lobata* and *Pocillopora meandrina* genome assemblies in two steps. Firstly, BLAT[101] (default parameters) was used to quickly localise corresponding putative genes of the proteins on the genome. The best match and matches with a score ≥ 90% of the best match score were retained. Secondly, the alignments were refined using Genewise[102] (default parameters), which is more precise for intron/exon boundary detection. Alignments were kept if more than 50% of the length of the protein was aligned to the genome. In order to reduce mapping noise, for each proteome mapping alignments without introns are removed if they represent more than 40% of the number of alignments. Moreover, alignments containing at least one unique intron (*i.e.* intron detected

using only one proteome alignements) are removed if they cover at least 10 exons detected in all alignments using all proteomes.

To allow the detection of expressed and/or specific genes, we also aligned the assembled transcriptomes of each species on their respective genome assembly using BLAT[101] (default parameters). For each transcript, the best match was selected based on the alignment score. Finally, alignments were recomputed in the previously identified genomic regions by Est2Genome[103] in order to define precisely intron boundaries. Alignments were kept if more than 80% of the length of the transcript was aligned to the genome with a minimal identity percent of 95%.

To proceed to the gene prediction for both species, we integrated the protein homologies and transcript mapping using a combiner called Gmove[104]. This tool can find CDSs based on genome located evidence without any calibration step. Briefly, putative exons and introns, extracted from the alignments, were used to build a simplified graph by removing redundancies. Then, Gmove extracted all paths from the graph and searched for open reading frames (ORFs) consistent with the protein evidence. Single-exon genes with a CDS length smaller or equal to 100 amino acids were filtered out. From the remaining genes, only genes with homologies against more than one species (Diamond[105] v0.9.24, blastp, e-value ≤ 10^{-10}) or spliced genes with a ratio CDS length / UTR length greater or equal to 0.75 were kept. Then, putative transposable elements (TEs) set were removed from the predicted gene using three different approaches : (i) genes that contain a TE domain from Pfam[106]; (ii) transposon-like genes detected using TransposonPSI (http://transposonpsi.sourceforge.net/, default parameters); (iii) and genes overlapping repetitive elements detected using RepeatMasker[100] and RepeatModeler[107] (v2.0.1, default parameters, repetitive sequence detected by RepeatModeler were annotated using RepeatMasker) on the genome assembly. Also, InterProScan[108] (v5.41-78.0, default parameters) was used to detect conserved protein domains in predicted genes. So, predicted genes without conserved domain covered by at least 90% of their cumulative exonic length, or matching TransposonPSI criteria or selected Pfam domains, were removed from the gene set.

Completeness of the gene catalogs was assessed using BUSCO[96] version 4.0.2 (eukaryota dataset odb10 and default parameters).

Adaptation of gene prediction workflow for tandemly duplicated genes

The presence of highly conserved genes in the same genomic regions can hinder the gene prediction, mostly if based on the alignments of conserved proteins. Indeed, during the spliced

alignment step, individual exons of a given protein sequence can be distributed over several genes. Therefore, in these specific genomic regions, alignments of proteins or RNA-Seq data generally span several genes (with larger introns), which lead to the prediction of chimeric genes and the underestimation of the gene number (Figure S4). To avoid these gene fusions, we added a step in our workflow. Namely, large introns of spliced alignments obtained with BLAT were post-processed. For each intron having a size greater than 5kb, the corresponding alignment was splitted in two inside the intron, and the query sequence was realigned on the two new genomic regions. If the sum of the two alignment scores was greater than the score of the previous alignment, then the two new alignments were kept in place of the alignment that contained the large intron. This process was recursively applied until the sum of the two alignment scores did not satisfy the previous condition. At the end, alignments were refined (with dedicated alignment tools) as described in the Gene Prediction section.

Telomeric sequences

The interstitial telomeric sequences (ITS) were specifically searched as follows. First, the motif (TTAGGG)4 was searched in the coral genomes using blastn92. The blast result was filtered to keep hits with an identity percentage above 75%, a minimum coverage of 75% and two mismatches maximum were allowed. Distant hits of less than 400 bp were gathered to form a single ITS. The 188 bp satellite sequence was searched in the different coral genomes and in the NT database using Blastn. All the hits had an evalue < 1e-13 and an identity percentage > 80%. Then, the matching sequences were used to build a HMM profile, using hmmbuild from HMMER suite[109] (v3.3).

Detection of tandemly duplicated genes

Protein sets of *Porites lobata*, *Pocillopora meandrina, Porites evermanni, Acropora millepora, Acropora digitifera, Montipora capitata, Galaxea fascicularis, Porites lutea, Pocillopora verrucosa, Pocillopora damicornis, Stylophora pistillata, Goniastrea aspera and Orbicella faveolata* (see references in Orthogroups and orthologous genes section) were aligned against themselves using Diamond[110] (v0.9.24). Only matches with an e-value ≤ 10^{-20} and 80% of the smallest protein aligned were kept. Two genes were considered as tandemly duplicated if they were co-localized on the same genomic contig and not distant from more than 10 genes to each other. Then, all tandemly duplicated genes were clustered using a single linkage clustering approach.

Functional assignment of predicted genes

The derived proteins of *Porites lobata* and *Pocillopora meandrina* predicted genes were functionally assigned by aligning them against nr from the BLAST Databases distributed by NCBI (version 25/10/2019) using Diamond[110] (v0.9.24, e-value ≤ 10^{-5}). Then, for each predicted protein, the best match based on bitscore against RefSeq proteins is selected. If there is no match against RefSeq proteins, then the best match is kept from other matches.

Orthogroups and orthologous genes

First, we selected the proteins of 24 Cnidarian species: *Acropora millepora, Acropora digitifera, Aiptasia sp., Amplexidiscus fenestrafer, Aurelia aurita* from Pacific*, Aurelia aurita from Atlantic, Clytia hemisphaerica, Dendronephthya gigantea, Discosoma sp., Fungia* spp.*, Galaxea fascicularis, Goniastrea aspera, Hydra vulgaris, Montipora capitata, Morbakka virulenta, Nematostella vectensis, Orbicella faveolata, Pocillopora damicornis, Pocillopora verrucosa, Pocillopora acuta, Porites evermanni, Porites lobata, Porites lutea, Porites rus, Stylophora pistillata*. The proteomes were aligned against each other using Diamond[110] (v0.9.24, e-value ≤ 10^{-10}, -k 0). Matches were kept only if 50% of the smallest protein length of each pair is aligned. Then, orthogroups (OG) and orthologous genes were built with OrthoFinder[111] (v2.3.11, default parameters). Additionally, OrthoFinder built gene trees for each OG and used them to reconstruct a rooted species tree, that is in agreement with the currently accepted phylogeny of cnidarians (Figure 1A and 4B). Based on quality metrics, we noticed that genome assemblies of *Pocillopora acuta*, *Acropora digitifera* and *Montipora capitata* are of lower quality and decided to exclude them from comparative analysis.

For each Orthogroup, we listed the 5 most abundant domains detected with InterProScan on proteins from 25 Cnidarian species (Extended data 9), the 5 most abundant BLASTP hits on nrprot for *Pocillopora meandrina* proteins (Extended data 10) and the 5 most abundant BLASTP hits for *Porites lobata* proteins (Extended data 11). We inspected these lists manually to assign the most likely function for the 192 orthogroups amplified in corals (Extended data 1).

Coral synteny

For each genome comparison, OG were used to build syntenic clusters using orthodotter (https://www.genoscope.cns.fr/orthodotter/). Only genomic contigs containing at least 5 genes with orthologs are selected. Co-localised orthologous genes less than 15 other orthologous genes apart are considered as belonging to the same syntenic cluster. A cluster has to be formed by at least 5 syntenic genes. Dot-plots for the analysis of synteny in coral genomes

were built using orthodotter, and circular views of syntenic regions were generated using Circos[112].

OG consensus construction

For each of the 27826 orthogroups (OG) containing at least one coral species, the following steps were applied. Coral proteic sequences were extracted, and aligned using Muscle[113] (version 3.8.1551, default parameters). Then, the multiple alignment was filtered using OD-Seq[114] (version 1.0) to remove outlier sequences, with parameter --score -*i.e.* threshold for outliers in numbers of standard deviations- set to 1.5. For large gene families, where the consensus obtained contained a high proportion of gaps -*i.e*. where (consensus length - median length of sequences in the input) >15% of median length of sequences-, a second run of OD-Seq was performed. After the filtering of outlier protein sequences, the consensus was extracted from the multiple alignment using hmmemit in the HMMER3 package[109] (version 3.1b1, default parameters).

Then, OG consensus were aligned against each other using Blat[101] (version 36, default parameters), in order to detect unspecific regions, *i.e.* regions of 30 amino acids having a hit (with >=85% identity) with another consensus (they typically correspond to common domains). The threshold of 85% was chosen because it corresponds to the average %identity observed when mapping reads from each coral species on consensus proteic sequences. 2039 OG that contain unspecific domains were tagged. Additionally, transposon-like domains were looked for in the consensus and interproscan outputs from all genes in each OG, and 577 OG that were likely to correspond to transposable elements were also tagged. OG tagged as TE or unspecific domain-containing were not used for subsequent analyses.

Construction of gene trees

Orthofinder provides gene trees for each OG, containing the 25 cnidarian species that were used as input. We also needed to generate trees for various subsets of species (for instance, only *Pocillopora meandrina*, only *Porites lobata*, only 11 coral species…). For each set of species needed, we extracted the corresponding proteic sequences after removal of outlier sequences (as described in “OG consensus construction” paragraph), and aligned them with MAFFT[115] (v7.464). Then we used FastTree[116] (2.1.11) with default parameters for construction of approximately-maximum-likelihood phylogenetic trees. Trees were edited and visualised using Itol[117] and R ggtree package[118].

Calculation of Ks between *P. lobata* and *P. meandrina* gene pairs

Ks (rates of synonymous substitutions) were respectively calculated between pairs of *P. lobata* and *P. meandrina* paralogous genes after aligning protein sequences with muscle[113] (default parameters), and generating codon alignments with pal2nal[119] (V13). Then codeml (PAML package[120] version 4.8) was used to calculate dS values (*i.e.* Ks).

Mapping of short reads on OG consensus to estimate gene copy numbers

In order to estimate gene family copy number independently from assembly and annotation processes, we downloaded short reads datasets (illumina paired end) for 14 coral and 4 sea anemone species (Table S5). We extracted 50 million sequences from pair1 files, trimmed to 100nt for consistency between analyses, and mapped those using diamond[110] on orthogroup coral consensus sequences. Unique hits were retained for each read, and depth of coverage was calculated on each consensus OG (25210 OG after filtering TE and unspecific regions). Then the depth obtained for each OG was normalised for each species by dividing by the depth obtained on a set of conserved single copy genes, in order for the final value obtained to be representative of the gene copy number. Indeed, the ratio obtained for single copy genes is close to 1 (Figure S26A).

Identification of a set of single copy genes present in all corals

In order to normalise depths of mapping of short reads on each OG consensus, we needed a set of single copy genes. We made a first attempt with BUSCO[96] version 4.0.2 (metazoa odb10 ancestral genes: 877 among the 954 consensus sequences were used, after discarding the ancestral genes that were not present in at least 10 coral species) but due to the divergence of corals with BUSCO metazoa ancestral sequences, few reads could be mapped and the depths obtained were low. Alternatively, we generated a set of genes that are present in exactly 1 copy in at least 13 coral species (among the 14 species listed in Table S5). This conservative threshold avoids discarding genes that may have been missed or duplicated in the annotation of only one of the 14 genomes. The coral monocopy gene set contains 705 genes after removing OG with dubious transposon related domains. For each species, after discarding a few outlier OG (with unexpectedly high or low coverages), we calculated the overall depth on the remaining single-copy OG. The values obtained were used to normalise the depths obtained by mapping the reads on each consensus OG. As expected, depths obtained after normalisation on the set of 705 coral monocopy genes are tightly grouped around a value of 1. Contrastively, depths obtained on the same set of OG normalised with BUSCO metazoa are higher than 1 (due to the low coverage of mapping on BUSCO

consensus), more heterogeneous between species (probably reflecting their distance to the consensus), and more variable inside species (Figure S26). The set of 705 coral monocopy OG consensus could be used for other applications in coral comparative genomics. The consensus sequences are available in Extended data 12.

Detection of amplified/reduced gene families between clades

To detect gene families with significantly expanded/reduced gene numbers between corals and sea anemones we calculated, for each OG (total=25,210), the total number of genes in the groups of species to compare (11 coral species : *Fungia* spp., *Orbicella faveolata*, *Goniastrea aspera*, *Stylophora pistillata*, *Pocillopora meandrina*, *Pocillopora verrucosa*, *Porites evermanni*, *Porites lutea*, *Porites lobata*, *Galaxea fascicularis*, *Acropora millepora*) vs 3 non coral hexacorallia (*Aiptasia sp*, *Amplexidiscus fenestrafer*, *Discosoma sp*). We then performed a binomial test with a parameter of 11/14 to identify OG that display a proportion of coral genes that is significantly different from the expected proportion (11/14, under the null hypothesis where there are equal gene numbers in all species). Benjamini-Hochberg FDR correction for multiple testing was then applied and we selected the OG with corrected P-values < 0.001. The same procedure was performed to compare corals within the complex and robust clades, keeping only one species per genus (3 complex species: *Porites lobata*, *Galaxea fascicularis*, *Acropora millepora*, vs 5 robust species: *Fungia spp*, *Orbicella faveolata*, *Goniastrea aspera*, *Stylophora pistillata*, *Pocillopora meandrina*) with a parameter of 3/8 for the binomial test. Finally, to compare 4 massive (*Orbicella faveolata*, *Goniastrea aspera, Porites lobata*, *Galaxea fascicularis*) and 3 branched (*Acropora millepora, Pocillopora meandrina, Stylophora pistillata*) corals species, we used a parameter of 4/7. All calculations were performed with R (version 4.1.0).

History reconstruction of gene family copy number variations

The CAFE5 software[54] was used to estimate gene copy numbers in internal nodes of the phylogenetic tree, and identify branches with significant amplification or reductions of gene families. It was applied on the OG detected as amplified in corals in comparison to sea anemones, on the 11 corals and 3 sea anemone species. Among the 192 OG detected, 120 were present in the common ancestor of all species and could thus be analysed with CAFE, but one (OG0000004) was removed because CAFE failed to identify parameters, probably because of the very large number of genes (n=3038). After testing several sets of parameters, the following parameters were used: -p (poisson distribution for the root frequency distribution) -k 2 (number of gamma rate categories to use). *NB:* using the parameter -e (to estimate the global error model) provided almost identical results.

Detection of enriched PFAM domains between clades

To assess putative differences in the proteome (amino acid translated genes) based on Pfam protein domains, genomic gene sets were annotated using InterProScan (v5.41-78.0 with default parameters). The following species were included in the analysis: Corals: *Acropora millepora* (Amil), *Porites lutea* (Plut), *Porites lobata* (Plob) , *Porites evermanni* (Peve), *Galaxea fascicularis* (Gfas), *Stylophora pistillata* (Spis), *Pocillopora verrucosa* (Pver), *Pocillopora meandrina* (Pmea), *Fungia* sp. (Fungia), *Goniastrea aspera* (Gasp), *Orbicella faveolata* (Ofav). Others: *Aiptasia* sp. (Aiptasia), *Amplexidiscus fenestrafer* (Afen), *Discosoma* sp. (Disco). Based on the proportions for each Pfam domain per species (# pfam occurrences/# of all pfam occurrences for the genomic gene set), we determined standard deviations for each of the Pfam domains per species and performed non-parametric Mann-Whitney U testing for each Pfam domain comparing pairs of groups. FDR was applied using qvalue[121] (version 2.24.0) for multiple test adjustment. Heatmaps were generated based on the top100 most significant domains for each comparison. Heatmaps were generated using the R package pheatmap[122] (version 1.0.12), with Pfam domain proportions being scaled across species by means of subtracting the overall mean (centering) and dividing by the overall standard deviation (scaling).

Abundance quantification of *Pocillopora meandrina* tandem duplicated genes in meta-transcriptomic samples

Meta-transcriptomic reads of 103 samples coming from 11 islands were mapped on predicted transcript of *P. meandrina* using RSEM[123] (version 1.3.3, default parameters with bowtie2 option). This tool and its underlying model have been designed to properly take read mapping uncertainty into account, which is an important feature when dealing with duplicated genes. However, we checked for the potential impact of ambiguous read assignment to transcripts and found only 313 transcripts that contained no unique 31-mer and, on average, approximately 93% of reads from a sample were assigned uniquely to a given transcript. Uniq 31-mers were extracted using UniqueKMER[124] and quantifications analyses were performed using the TPM metric provided by RSEM. Pearson correlations of TPM distribution between all TDG were computed using the cor function of R (version 3.6.0) and the associated p-value with rcorr function of Hmisc package of R. Heatmap showing the z-score of the mean TPM by OG was created using the heatmap.2 function of the R package gplots, and z-scores were computed with the scale function of heatmap.2.

Identification of innate immune system genes

TIR domain containing orthogroups (putative Toll-like receptors, TLR):

We annotated orthogroups (OG) where InterProScan detected TIR domains (IPR000157, IPR035897) in at least 20% of the genes in 14 cnidarian species, as TIR-domain containing orthogroups. We obtained 21 OG totalizing 643 genes in 14 cnidaria species. The threshold (20% of genes containing the domain required to annotate the OG) was set by inspection of a manually curated set of TIR containing OG. In *Porites lobata*, and *Pocillopora meandrina*, respectively 56 and 47 genes belong to the identified OG, among which 49 (87.5%) and 42 (89.4%) contain the TIR domain. The relatively low threshold is due to more difficult domain identification in other species, where gene annotations can be fragmented.

NACHT domain containing orthogroups (putative NOD-like receptors, NLR):

We annotated OG where InterProScan detected NACHT domain (IPR007111) in at least 5% of the genes as putative NLR orthogroups. We obtained 46 OG totalizing 2991 genes in 14 cnidaria species. As described above, the thresholds were set by inspection of a manually curated gene set, and the % of genes in retained OG that actually contain the NACHT domain is high in *P. lobata* and *P. meandrina*.

Lectin domain containing orthogroups:

We annotated OG where InterProScan detected lectin-like domains (IPR001304, IPR016186, IPR016187, IPR018378, IPR019019, IPR033989, IPR037221, IPR042808) in at least 50% of the genes as putative lectin orthogroups. We obtained 81 OG totalizing 2475 genes in 14 cnidaria species.

Domain composition:

We studied the domain composition of TIR and NACHT/NB-ARC containing proteins in 5 species: *A. digitifera*, *P. lobata*, *P. lutea*, *P. meandrina* and *P. verrucosa*. We used all genes containing TIR (IPR000157, IPR035897) and NACHT/NB-ARC (IPR007111/IPR002182) domains (not only the ones in OG fulfilling the criteria mentioned above). We derived domain compositions from InterProScan outputs after manual curation to discard redundant domains. The list of genes and domain compositions are available in Extended data 13 and 14. For NACHT/NB-ARC containing proteins, we identified truncated proteins when less than 250 aa were annotated and no domain was detected with InterProScan upstream and/or downstream of the NACHT/NB-ARC domain. When the upstream/downstream sequence contained more than 250 aa, and no domain was annotated, the gene was tagged as “noneDetected”. Simplified domain compositions obtained are displayed in Extended data 15.

Additional files

All the supporting data are included in two additional files: (a) A supplementary file which contains Supplementary Tables 1-7 and Supplementary Figures 1-26; (b) A supplementary file which contains extended data 1-15.

Competing interests

The authors declare that they have no competing interests. JMA received travel and accommodation expenses to speak at Oxford Nanopore Technologies conferences.

Funding

This work was supported by the Genoscope, the Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA) and France Génomique (ANR-10-INBS-09-08).

Table 1. Statistics of the three long-read coral genome assemblies from this study compared to representative, existing genomes of the same clades.

	Complex				***Robust***		
	Porites lobata	*Porites evermanni*	*Porites lutea*	*Acropora millepora*	*Pocillopora meandrina*	*Pocillopora verrucosa*	*Pocillopora damicornis*
Publication	This study	This study	Robbins et al.	Fuller et al.	This study	Buitrago-López et al.	Cunning et al.
Estimated genome size	543 Mb	497 Mb	552 Mb*	?	315 Mb	407 Mb*	262 Mb*
# contigs|scaffolds	1,098	8,186	2,975	854	252	18,268	4,393
Cumulative size	646,152,978	603,805,388	552,020,673	475,381,253	347,233,126	381 Mb	234,335,492
N50 (L50)	2,154,615 (84)	171,385 (935)	660,708 (242)	19.8 Mb (9)	4,753,879 (23)	333,696 (326)	326,133 (198)
Max size	8,615,247	1,802,771	3,122,227	39,361,238	11,895,822	2,095,917	2,168,405
# of N's	0 (0%)	40,756,223 (6.75%)	48,123,166 (8.72%)	37,012 (0.01%)	0 (0%)	510,035 (0.13%)	8,607,682 (3.67%)
# contigs	1,098	32,888	47,330	1,234	252	54,131	53,036
N50 (L50)	2,154,615 (84)	33,681 (4,563)	19,557 (7,534)	1,091,365 (129)	4,753,879 (23)	23,429 (3,851)	25,941 (2,282)
Repeat coverage (% of assembly)	51.28	42.26	42.36	?	36.67	38.44	20.36
# number of genes	42,872	40,389	31,126	28,188	32,095	27,439	26,077
Genes density (genes/Mb)	66.4	66.7	56.4	59.3	92.5	72	111.4
% BUSCO (compl.; frag.; miss.) N= 255 genes	97.7; 1.2; 1.1	94.5; 3.9; 1.6	92.2; 4.3; 3.5	73.7; 16.5; 9.8	98.4; 0.4; 1.2	90.2; 5.1; 4.7	86.3; 9.0; 4.7

* data from publications

Figure 1. Comparison of available coral genome assemblies. Species from the complex clade are in orange, species from the robust clase are in red and the three genomes described in this study are in bold. **A.** Rooted species tree of 25 cnidarian species based on OrthoFinder. **B.** Genome assembly sizes are in Megabases, green bars indicate the estimated genome size based on kmers calculated from short-reads when available. **C.** Contig N50 values in Megabases. **D.** Number of annotated genes. **E.** Proportion of genes containing a functional domain. **F.** Proportion of genes in orthogroups (OG) that contain at least two different species. **G.** BUSCO scores computed with the Metazoan gene set (N=954 genes). Numbers in the blue bar represent the proportion of complete and single-copy genes in each genome assembly.

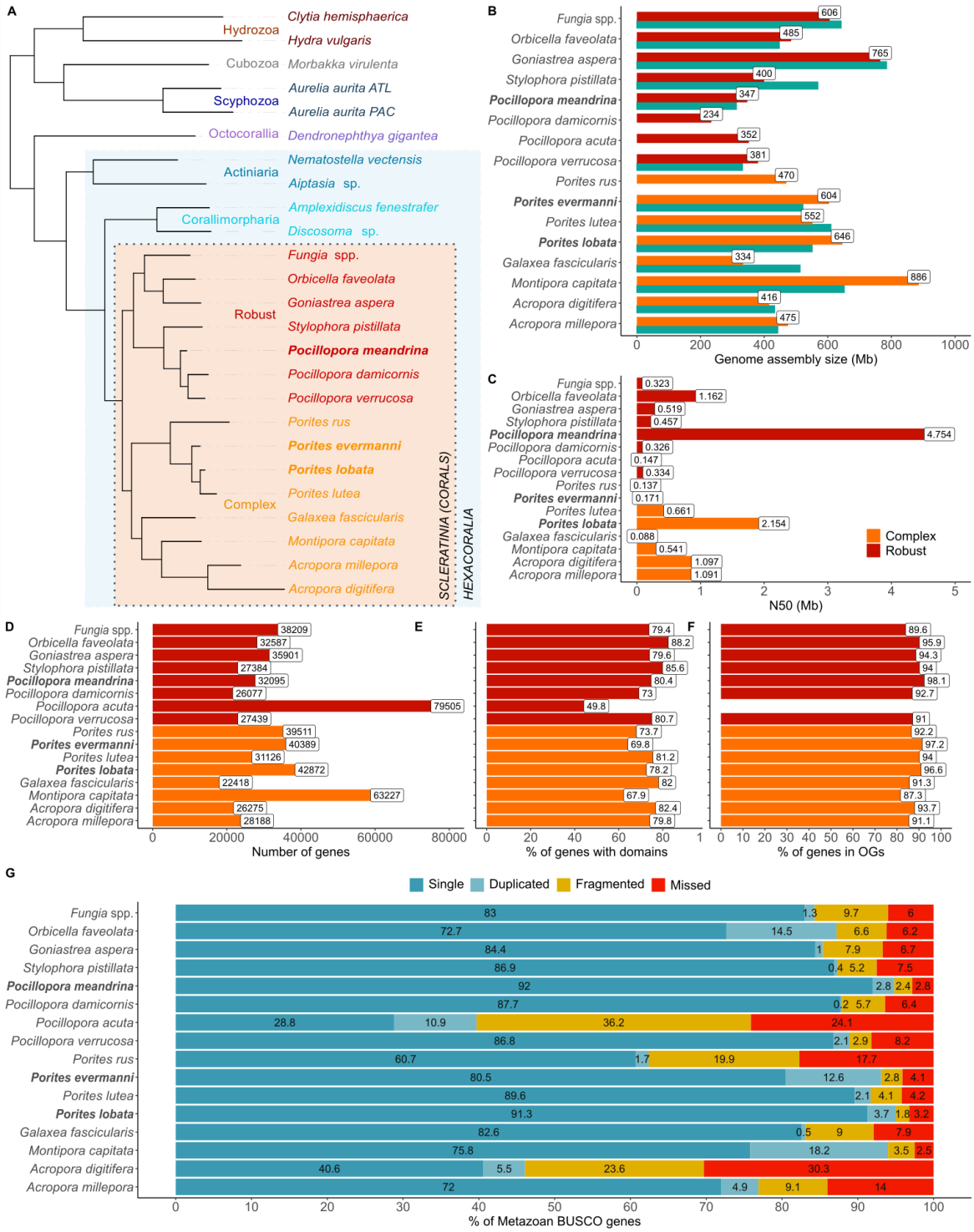


Figure 2. Coral synteny. Circular (left) and dotplot (right) representations of the synteny between the longest contigs. Each colored link represents linkage between two orthologous genes which are in a syntenic cluster. Colors of links represent syntenic clusters. Grey links connect orthologous genes that are not syntenic. Dotplots display only regions of contigs that contain orthologous genes. **A.** Synteny between the longest contig of *P. lobata* (blue) and its syntenic scaffolds in *P. lutea* (green). **B.** Synteny between the longest contig of *P. meandrina* (blue) and its syntenic scaffolds in *P. verrucosa* (green). **C.** Synteny between the longest contig of *P. meandrina* (blue) and its syntenic contigs in *P. lobata* (green).

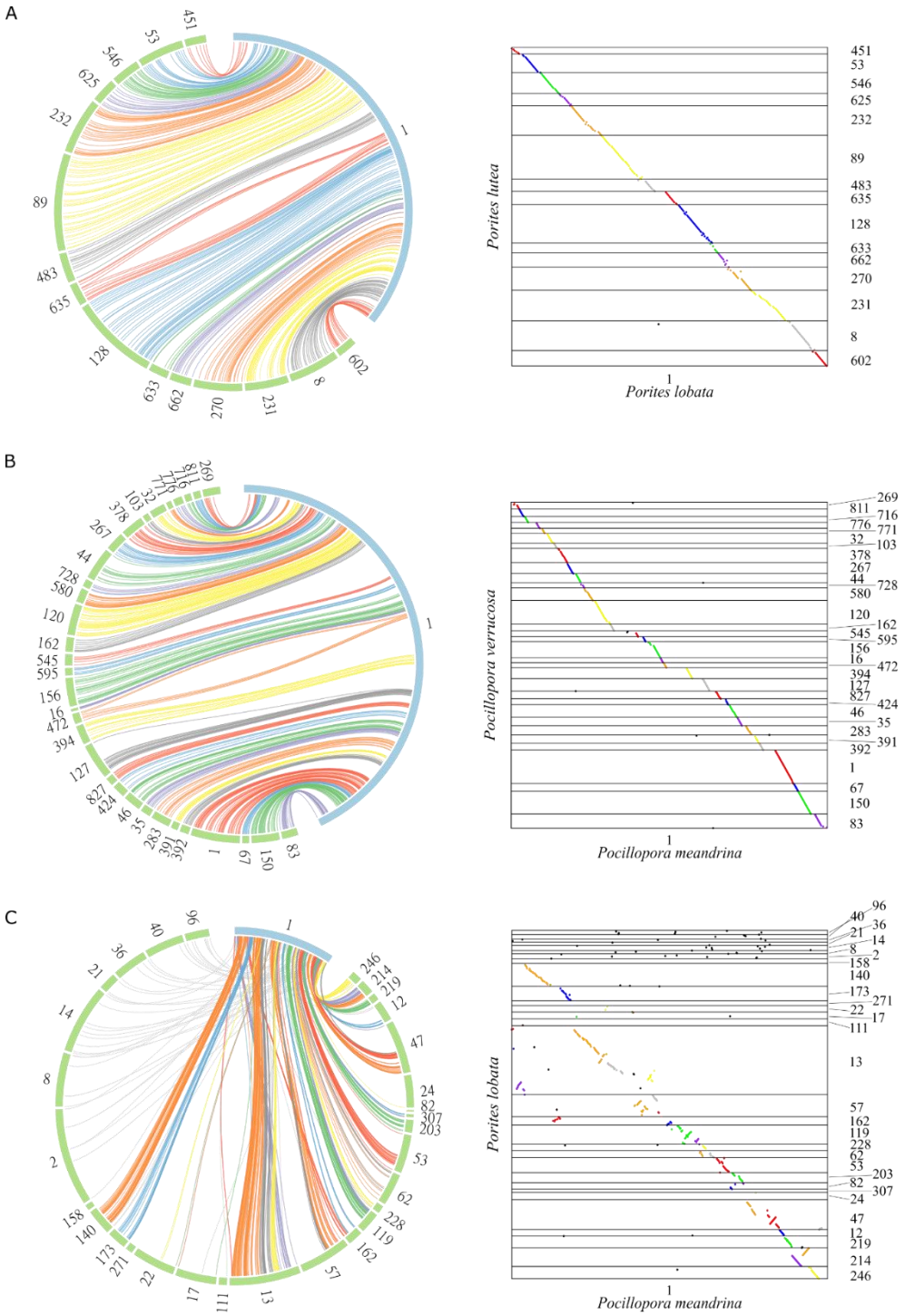


Figure 3. Quantification of tandemly duplicated genes (TDG) in coral genomes. A. Number of TDG for each species. **B.** Distribution of the number of genes per TDG cluster. **C.** For 499 gene families (orthogroups with >=10 genes in *P. meandrina* or *P. lobata*)**,** the number of genes in *Pocillopora* and *Porites* species is compared to the normalised depth of mapping of short reads on OG consensus (*i.e.* estimated gene copy number based on mapping of short reads). Pie charts represent the proportion of TDG genes in each species. For *Pocillopora damicornis*, no value of depth was computed since we were not able to identify a set of illumina short reads to download.

Figure 4. Amplified gene families in corals vs sea anemones. A. Average number of gene copies in corals vs sea anemones. Orthogroups colored in orange have significantly more gene copies in corals compared to sea anemones and orthogroups colored in blue have significantly less gene copies in corals compared to sea anemones (binomial test, adjusted p value < 0.001). Dot sizes correspond to the ratio of TDG for each OG in 11 coral genomes. **B.** Heatmap of gene copy numbers in 15 species for 192 OG amplified in corals and 28 OG amplified in sea anemones. The phylogenetic tree is the output of the OrthoFinder software. **C.** Proportion of TDG in 192 OG amplified in corals (orange), 28 amplified in sea anemones (blue), and not amplified OG (grey). The pie charts represent the proportion of TDG among the OG amplified in corals or sea anemones, in *Porites lobata* (orange), *Pocillopora meandrina* (red) and in sea anemones (blue).

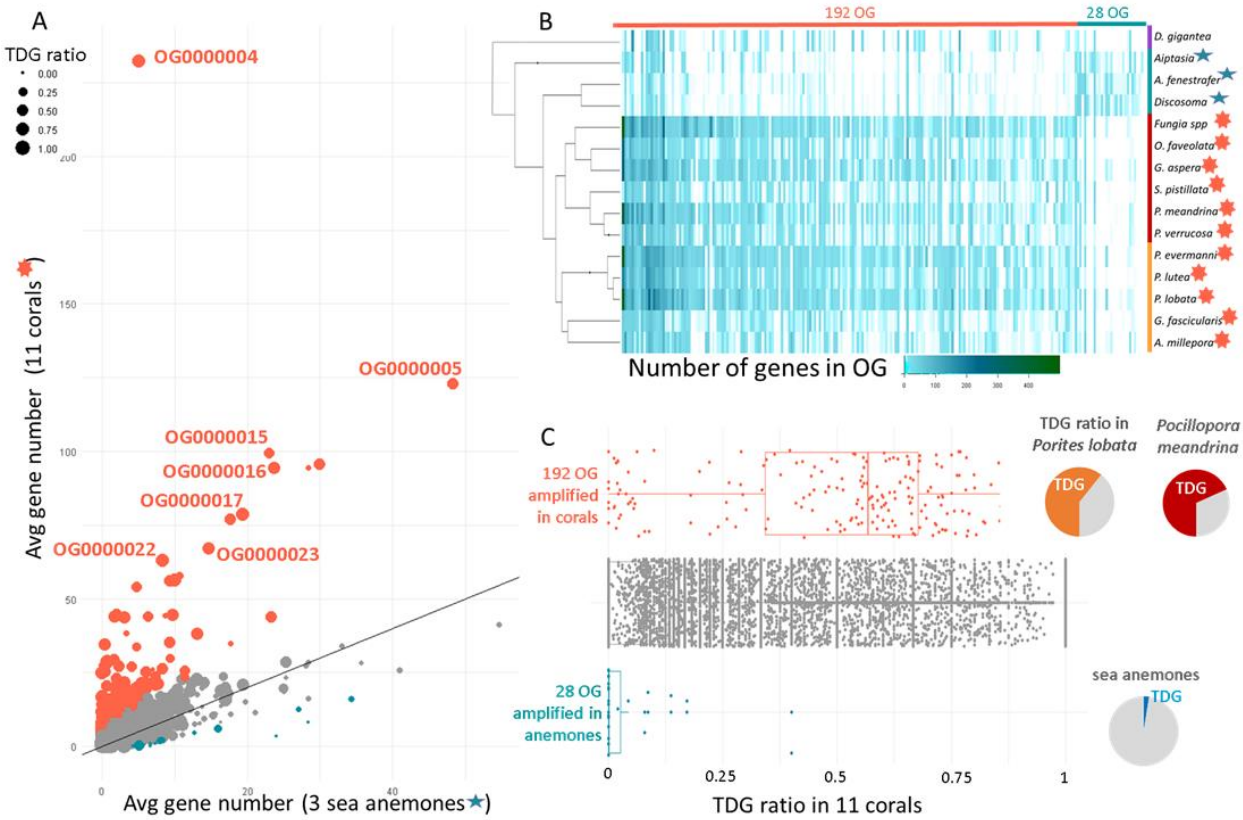


Figure 5. A. Comparison of the number of genes in amplified families of *Porites lobata* and *Pocillopora meandrina* genomes. Gene families are grouped by their functional annotation. The largest gene family is indicated by a coloured bar, for *P. meandrina* (red) and *P. lobata* (orange). **B.** Ks distribution of *Porites lobata* (Pl) and *Pocillopora meandrina* (Pm) tandemly duplicated gene pairs (TDG). **C.** Ks distributions for TDG pairs in Pl and Pm for 11 orthogroups (OG) that are amplified in corals and contain NACHT domains.

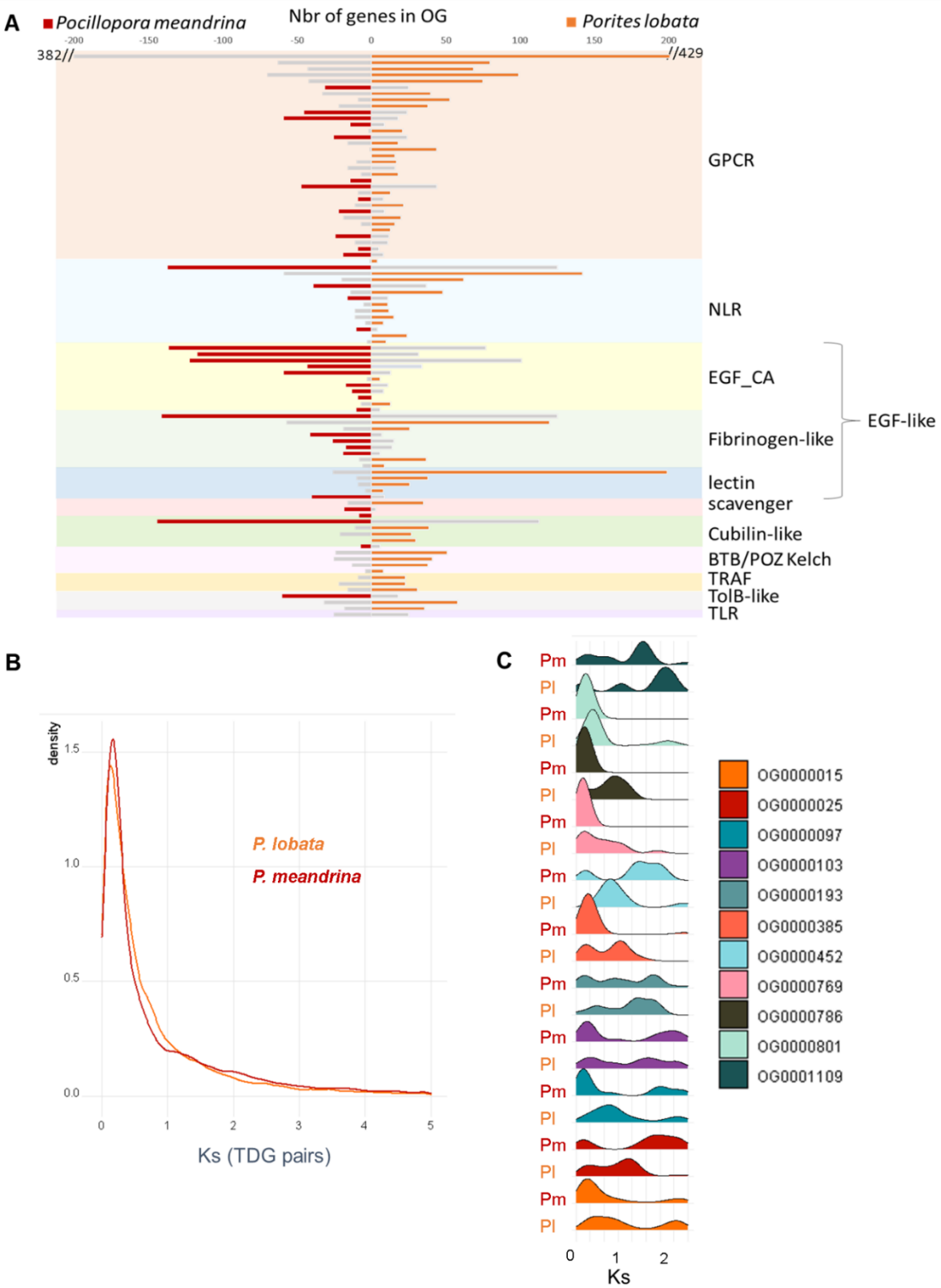


Figure 6. A. Approximately maximum-likelihood phylogenetic tree obtained with FastTree after aligning proteins from OG0000106 (TIR domain-containing orthogroup) in *Porites lobata* (orange dots) and *Pocillopora meandrina* (red dots). Colours correspond to tandemly repeated gene clusters (singletons are in red). **B,C**: Trees obtained for 15 coral species for two individual TDG clusters. Dots colours correspond to species displayed in the species tree in **D**. **E**: Distribution of Ks between pairs of genes in TDG clusters, in *P. lobata* and *P. meandrina*.

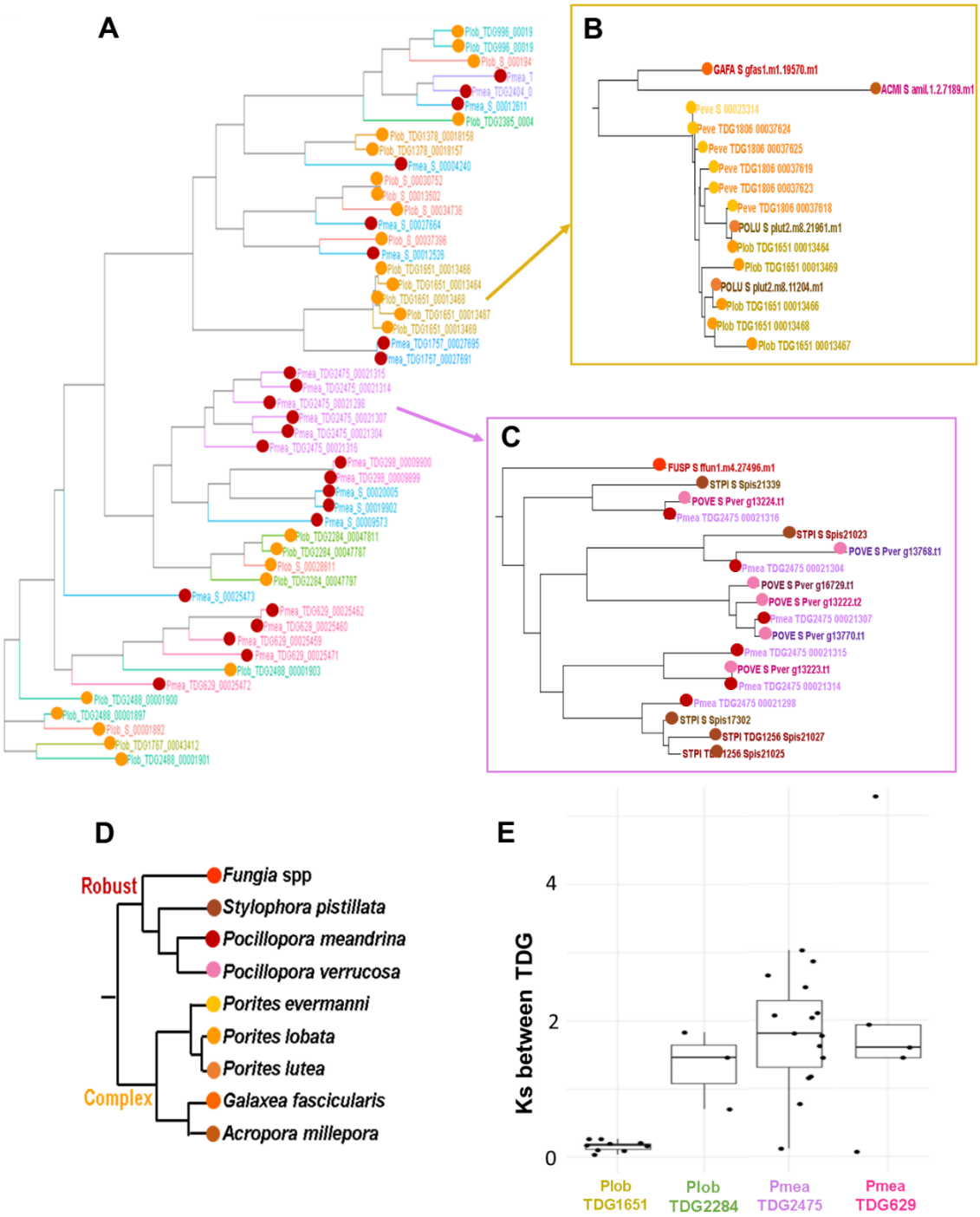


Figure 7. A. Map showing the 11 islands sampled during the Tara Pacific expedition. **B.** Pearson correlations of gene expression profiles across the 103 samples for different values of dS between pairs of TDG. **C.** Heatmap of expression quantification (z-score of mean TPM per OG) of amplified genes in coral genomes across the 103 samples.

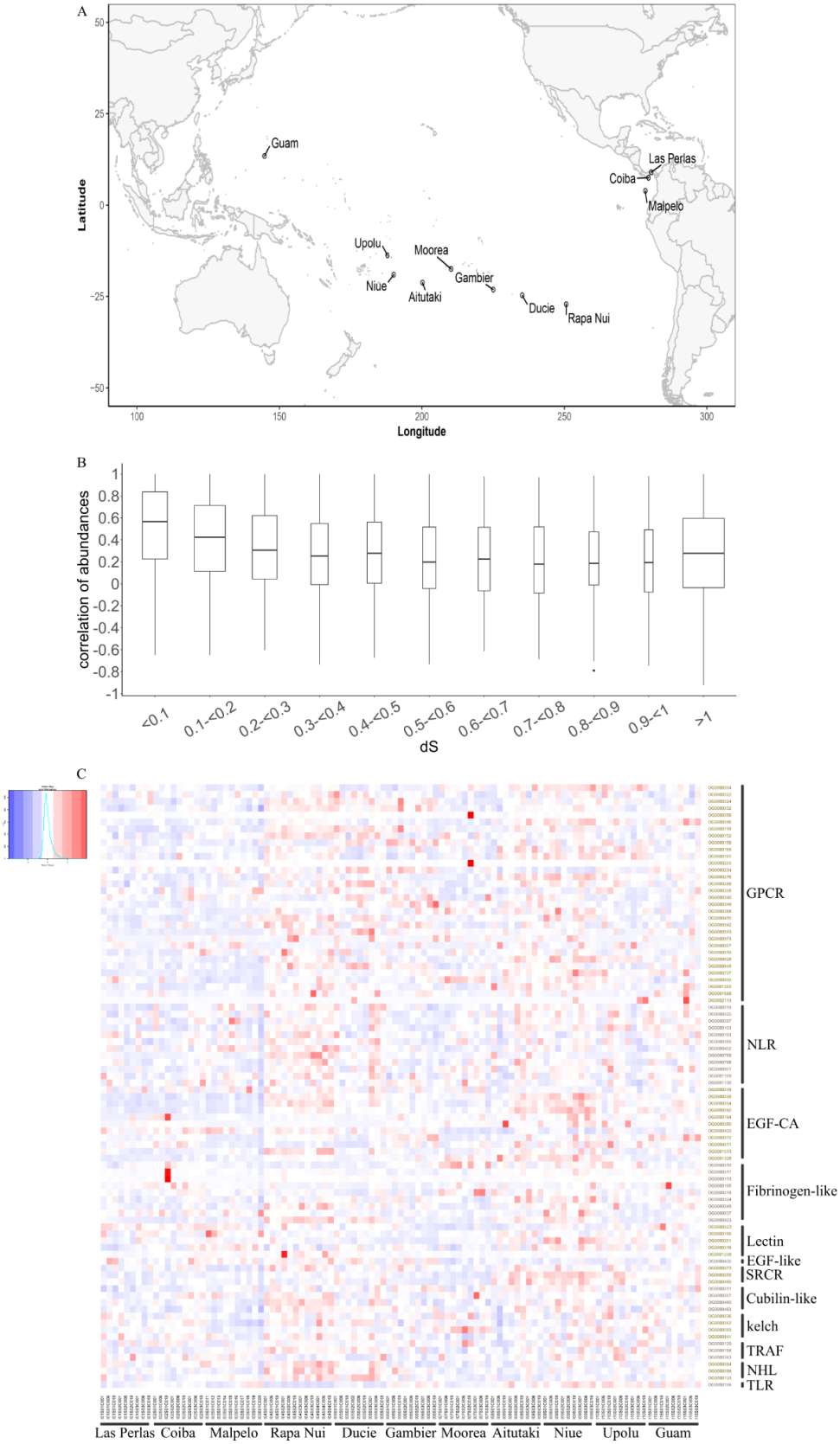
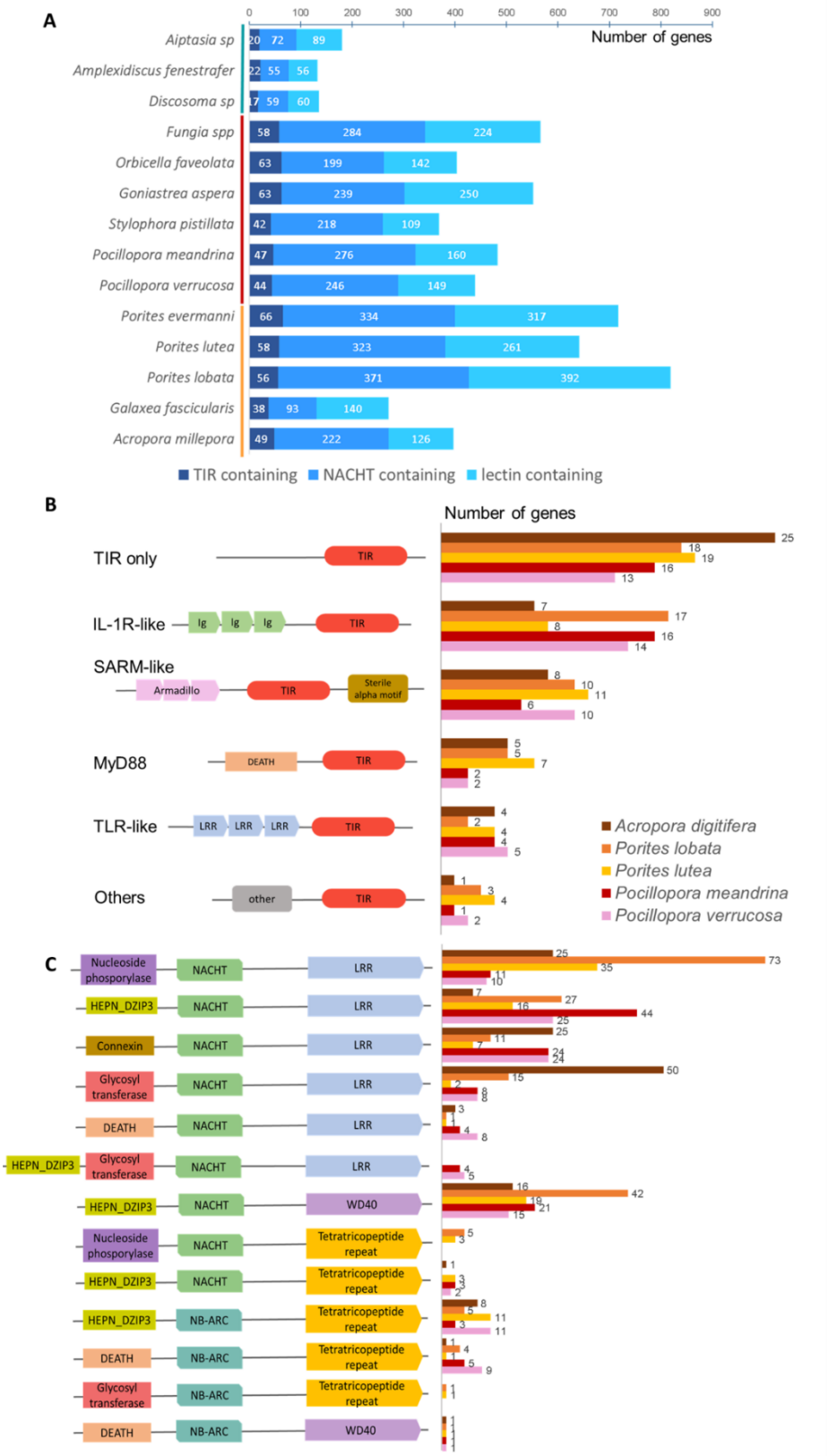


Figure 8. A. Cumulative bar plot representing the number of genes in innate immune receptor OG identified from domain annotation of 14 cnidarian gene sets, for three innate immune receptor categories. **B.C.** Domain composition of TIR-containing (**B**) and NACHT/NB-ARC (**C**) containing proteins in 5 coral species. Left panels: schematic view of domain composition;Right panels: number of genes in each species.



List of Tara Pacific CONSORTIUM COORDINATORS

Sylvain Agostini (orcid.org/0000-0001-9040-9296) Shimoda Marine Research Center, University of Tsukuba, 5-10-1, Shimoda, Shizuoka, Japan
Denis Allemand (orcid.org/0000-0002-3089-4290) Centre Scientifique de Monaco, 8 Quai Antoine Ier, MC-98000, Principality of Monaco
Bernard Banaigs (orcid.org/0000-0003-3473-4283) PSL Research University: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Université de Perpignan, France
Emilie Boissin (orcid.org/0000-0002-4110-790X) PSL Research University: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Laboratoire d'Excellence CORAIL, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France
Emmanuel Boss (orcid.org/0000-0002-8334-9595) School of Marine Sciences, University of Maine, Orono, 04469, Maine, USA
Chris Bowler (orcid.org/0000-0003-3835-6187) Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
Colomban de Vargas (orcid.org/0000-0002-6476-6019) Sorbonne Université, CNRS, Station Biologique de Roscoff, AD2M, UMR 7144, ECOMAP 29680 Roscoff, France & Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/ Tara Oceans-GOSEE, 3 rue Michel-Ange, 75016 Paris, France
Eric Douville (orcid.org/0000-0002-6673-1768) Laboratoire des Sciences du Climat et de l'Environnement, LSCE/IPSL, CEA-CNRS-UVSQ, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
Michel Flores (orcid.org/0000-0003-3609-286X) Weizmann Institute of Science, Department of Earth and Planetary Sciences, 76100 Rehovot, Israel
Didier Forcioli (orcid.org/0000-0002-5505-0932) Université Côte d'Azur, CNRS, INSERM, IRCAN, Medical School, Nice, France and Department of Medical Genetics, CHU of Nice, France
Paola Furla (orcid.org/0000-0001-9899-942X) Université Côte d'Azur, CNRS, INSERM, IRCAN, Medical School, Nice, France and Department of Medical Genetics, CHU of Nice, France
Pierre Galand (orcid.org/0000-0002-2238-3247) Sorbonne Université, CNRS, Laboratoire d'Ecogéochimie des Environnements Benthiques (LECOB), Observatoire Océanologique de Banyuls, 66650 Banyuls sur mer, France
Eric Gilson (orcid.org/0000-0001-5738-6723) Université Côte d'Azur, CNRS, Inserm, IRCAN, France
Fabien Lombard (orcid.org/0000-0002-8626-8782) Sorbonne Université, Institut de la Mer de Villefranche sur mer, Laboratoire d'Océanographie de Villefranche, F-06230 Villefranche-sur-Mer, France
Stéphane Pesant (orcid.org/0000-0002-4936-5209) European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
Serge Planes (orcid.org/0000-0002-5689-5371) PSL Research University: EPHE-UPVD-CNRS, USR 3278 CRIOBE, Laboratoire d'Excellence CORAIL, Université de Perpignan, 52 Avenue Paul Alduy, 66860 Perpignan Cedex, France
Stéphanie Reynaud (orcid.org/0000-0001-9975-6075) Centre Scientifique de Monaco, 8 Quai Antoine Ier, MC-98000, Principality of Monaco

Matthew B. Sullivan (orcid.org/0000-0003-4040-9831) The Ohio State University, Departments of Microbiology and Civil, Environmental and Geodetic Engineering, Columbus, Ohio, 43210 USA
Shinichi Sunagawa (orcid.org/0000-0003-3065-0314) Department of Biology, Institute of Microbiology and Swiss Institute of Bioinformatics, Vladimir-Prelog-Weg 4, ETH Zürich, CH-8093 Zürich, Switzerland
Olivier Thomas (orcid.org/0000-0002-5708-1409) Marine Biodiscovery Laboratory, School of Chemistry and Ryan Institute, National University of Ireland, Galway, Ireland
Romain Troublé (ORCID not-available) Fondation Tara Océan, Base Tara, 8 rue de Prague, 75 012 Paris, France
Rebecca Vega Thurber (orcid.org/0000-0003-3516-2061) Oregon State University, Department of Microbiology, 220 Nash Hall, 97331 Corvallis OR USA
Christian R. Voolstra (orcid.org/0000-0003-4555-3795) Department of Biology, University of Konstanz, 78457 Konstanz, Germany
Patrick Wincker (orcid.org/0000-0001-7562-3454) Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France
Didier Zoccola (orcid.org/0000-0002-1524-8098) Centre Scientifique de Monaco, 8 Quai Antoine Ier, MC-98000, Principality of Monaco

References

1. Connell, J. H. Diversity in Tropical Rain Forests and Coral Reefs. *Science* **199**, 1302–1310 (1978).
2. Spalding, M. D. *World Atlas of Coral Reefs*. (UNEP-WCMC, 2001).
3. Moberg, F. & Folke, C. Ecological goods and services of coral reef ecosystems. *Ecol. Econ.* **29**, 215–233 (1999).
4. Knowlton, N. & Rohwer, F. Multispecies Microbial Mutualisms on Coral Reefs: The Host as a Habitat. *Am. Nat.* **162**, S51–S62 (2003).
5. Pogoreutz, C. *et al.* The coral holobiont highlights the dependence of cnidarian animal hosts on their associated microbes. in *Cellular Dialogues in the Holobiont* (CRC Press, 2020).
6. Chen, P.-Y., Chen, C.-C., Chu, L. & McCarl, B. Evaluating the economic damage of climate change on global coral reefs. *Glob. Environ. Change* **30**, 12–20 (2015).
7. Hoegh-Guldberg, O. Climate change, coral bleaching and the future of the world's coral reefs. *Mar. Freshw. Res.* **50**, 839–866 (1999).
8. Pandolfi, J. M., Connolly, S. R., Marshall, D. J. & Cohen, A. L. Projecting Coral Reef Futures Under Global Warming and Ocean Acidification. *Science* **333**, 418–422 (2011).
9. Dixon, A. M., Forster, P. M., Heron, S. F., Stoner, A. M. K. & Beger, M. Future loss of local-scale thermal refugia in coral reef ecosystems. *PLOS Clim.* **1**, e0000004 (2022).
10. Anthony, K. R. N., Connolly, S. R. & Hoegh-Guldberg, O. Bleaching, energetics, and coral mortality risk: Effects of temperature, light, and sediment regime. *Limnol. Oceanogr.* **52**, 716–726 (2007).
11. Langdon, C. *et al.* Effect of calcium carbonate saturation state on the calcification rate of an experimental coral reef. *Glob. Biogeochem. Cycles* **14**, 639–654 (2000).
12. Anthony, K. R. N. *et al.* Ocean acidification and warming will lower coral reef resilience. *Glob. Change Biol.* **17**, 1798–1808 (2011).
13. Global Warming of 1.5 ºC —. https://www.ipcc.ch/sr15/.

14. Zoccola, D. *et al.* The World Coral Conservatory (WCC): A Noah's ark for corals to support survival of reef ecosystems. *PLOS Biol.* **18**, e3000823 (2020).

15. Kleypas, J. *et al.* Designing a blueprint for coral reef survival. *Biol. Conserv.* **257**, 109107 (2021).

16. Voolstra, C. R. *et al.* Extending the natural adaptive capacity of coral holobionts. *Nat. Rev. Earth Environ.* 1–16 (2021) doi:10.1038/s43017-021-00214-3.

17. Roark, E. B., Guilderson, T. P., Dunbar, R. B., Fallon, S. J. & Mucciarone, D. A. Extreme longevity in proteinaceous deep-sea corals. *Proc. Natl. Acad. Sci.* **106**, 5204–5208 (2009).

18. Montero-Serra, I., Linares, C., Doak, D. F., Ledoux, J. B. & Garrabou, J. Strong linkages between depth, longevity and demographic stability across marine sessile species. *Proc. R. Soc. B Biol. Sci.* **285**, 20172688 (2018).

19. Irwin, A. *et al.* Age and intraspecific diversity of resilient Acropora communities in Belize. *Coral Reefs* **36**, 1111–1120 (2017).

20. Gardner, S. N. & Mangel, M. When Can a Clonal Organism Escape Senescence? *Am. Nat.* **150**, 462–490 (1997).

21. Salguero-Gómez, R. Implications of clonality for ageing research. *Evol. Ecol.* **32**, 9–28 (2018).

22. Shinzato, C. *et al.* Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* **476**, 320–323 (2011).

23. Baumgarten, S. *et al.* The genome of Aiptasia, a sea anemone model for coral symbiosis. *Proc. Natl. Acad. Sci.* **112**, 11893–11898 (2015).

24. Voolstra, C. R. *et al.* Comparative analysis of the genomes of Stylophora pistillata and Acropora digitifera provides evidence for extensive differences between species of corals. *Sci. Rep.* **7**, 17583 (2017).

25. Cunning, R., Bay, R. A., Gillette, P., Baker, A. C. & Traylor-Knowles, N. Comparative analysis of the Pocillopora damicornis genome highlights role of immune system in coral

evolution. *Sci. Rep.* **8**, 16134 (2018).

26. Ying, H. *et al.* Comparative genomics reveals the distinct evolutionary trajectories of the robust and complex coral lineages. *Genome Biol.* **19**, 175 (2018).
27. Celis, J. S. *et al.* Binning enables efficient host genome reconstruction in cnidarian holobionts. *GigaScience* **7**, (2018).
28. Shumaker, A. *et al.* Genome analysis of the rice coral Montipora capitata. *Sci. Rep.* **9**, 2571 (2019).
29. Robbins, S. J. *et al.* A genomic view of the reef-building coral Porites lutea and its microbial symbionts. *Nat. Microbiol.* **4**, 2090–2100 (2019).
30. Ying, H. *et al.* The Whole-Genome Sequence of the Coral Acropora millepora. *Genome Biol. Evol.* **11**, 1374–1379 (2019).
31. Buitrago-López, C., Mariappan, K. G., Cárdenas, A., Gegner, H. M. & Voolstra, C. R. The Genome of the Cauliflower Coral Pocillopora verrucosa. *Genome Biol. Evol.* **12**, 1911–1917 (2020).
32. Romano, S. L. & Palumbi, S. R. Evolution of Scleractinian Corals Inferred from Molecular Systematics. *Science* (1996) doi:10.1126/science.271.5249.640.
33. Shinzato, C. *et al.* Eighteen Coral Genomes Reveal the Evolutionary Origin of Acropora Strategies to Accommodate Environmental Changes. *Mol. Biol. Evol.* **38**, 16–30 (2021).
34. Helmkampf, M., Bellinger, M. R., Geib, S. M., Sim, S. B. & Takabayashi, M. Draft Genome of the Rice Coral Montipora capitata Obtained from Linked-Read Sequencing. *Genome Biol. Evol.* **11**, 2045–2054 (2019).
35. Tørresen, O. K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res.* **47**, 10994–11006 (2019).
36. Belser, C. *et al.* Telomere-to-telomere gapless chromosomes of banana using nanopore sequencing. *Commun. Biol.* **4**, 1–12 (2021).
37. Boulay, J. N., Hellberg, M. E., Cortés, J. & Baums, I. B. Unrecognized coral species

diversity masks differences in functional ecology. *Proc. R. Soc. B Biol. Sci.* **281**, 20131580 (2014).

38. Traut, W. *et al.* The telomere repeat motif of basal Metazoa. *Chromosome Res.* **15**, 371–382 (2007).
39. Tan, K.-T., Slevin, M. K., Meyerson, M. & Li, H. *Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres*. 2022.01.11.475254 https://www.biorxiv.org/content/10.1101/2022.01.11.475254v1 (2022) doi:10.1101/2022.01.11.475254.
40. Fuller, Z. L. *et al.* Population genetics of the coral Acropora millepora: Toward genomic prediction of bleaching. *Science* **369**, eaba4674 (2020).
41. Cooke, I. *et al.* Genomic signatures in the coral holobiont reveal host adaptations driven by Holocene climate change and reef specific symbionts. *Sci. Adv.* **6**, eabc6318.
42. Park, E. *et al.* Estimation of divergence times in cnidarian evolution based on mitochondrial protein-coding genes and the fossil record. *Mol. Phylogenet. Evol.* **62**, 329–345 (2012).
43. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
44. Shoja, V. & Zhang, L. A Roadmap of Tandemly Arrayed Genes in the Genomes of Human, Mouse, and Rat. *Mol. Biol. Evol.* **23**, 2134–2141 (2006).
45. Leister, D. Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. *Trends Genet.* **20**, 116–122 (2004).
46. Plomion, C. *et al.* Oak genome reveals facets of long lifespan. *Nat. Plants* **4**, 440–452 (2018).
47. Liu, C. *et al.* Genome-wide analysis of tandem duplicated genes and their contribution to stress resistance in pigeonpea (Cajanus cajan). *Genomics* **113**, 728–735 (2021).
48. Palmer, C. V. Immunity and the coral crisis. *Commun. Biol.* **1**, 1–7 (2018).
49. Hamada, M. *et al.* The Complex NOD-Like Receptor Repertoire of the Coral Acropora

digitifera Includes Novel Domain Combinations. *Mol. Biol. Evol.* **30**, 167–176 (2013).

50. Loya, Y. *et al.* Coral bleaching: the winners and the losers. *Ecol. Lett.* **4**, 122–131 (2001).
51. McClanahan, T. R. The relationship between bleaching and mortality of common corals. *Mar. Biol.* **144**, 1239–1245 (2004).
52. Razak, T. B., Roff, G., Lough, J. M. & Mumby, P. J. Growth responses of branching versus massive corals to ocean warming on the Great Barrier Reef, Australia. *Sci. Total Environ.* **705**, 135908 (2020).
53. Wang, X. *et al.* The Evolution of Calcification in Reef-Building Corals. *Mol. Biol. Evol.* **38**, 3543–3555 (2021).
54. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).
55. Nei, M. & Rooney, A. P. Concerted and Birth-and-Death Evolution of Multigene Families. *Annu. Rev. Genet.* **39**, 121–152 (2005).
56. Necsulea, A. & Kaessmann, H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat. Rev. Genet.* **15**, 734–748 (2014).
57. Lan, X. & Pritchard, J. K. Coregulation of tandem duplicate genes slows evolution of subfunctionalization in mammals. *Science* **352**, 1009–1013 (2016).
58. Kvennefors, E. C. E. *et al.* Analysis of evolutionarily conserved innate immune components in coral links immunity and symbiosis. *Dev. Comp. Immunol.* **34**, 1219–1229 (2010).
59. Parisi, M. G., Parrinello, D., Stabili, L. & Cammarata, M. Cnidarian Immunity and the Repertoire of Defense Mechanisms in Anthozoans. *Biology* **9**, 283 (2020).
60. Miller, D. J. *et al.* The innate immune repertoire in Cnidaria - ancestral complexity and stochastic gene loss. *Genome Biol.* **8**, R59 (2007).
61. Poole, A. Z. & Weis, V. M. TIR-domain-containing protein repertoire of nine anthozoan species reveals coral-specific expansions and uncharacterized proteins. *Dev. Comp. Immunol.* **46**, 480–488 (2014).

62. O'Neill, L. A. J., Golenbock, D. & Bowie, A. G. The history of Toll-like receptors — redefining innate immunity. *Nat. Rev. Immunol.* **13**, 453–460 (2013).

63. Hayes, M. L., Eytan, R. I. & Hellberg, M. E. High amino acid diversity and positive selection at a putative coral immunity gene (tachylectin-2). *BMC Evol. Biol.* **10**, 150 (2010).

64. Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V. & Aravind, L. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol. Direct* **8**, 15 (2013).

65. Bertucci, A. *et al.* Carbonic anhydrases in anthozoan corals—A review. *Bioorg. Med. Chem.* **21**, 1437–1450 (2013).

66. Bhattacharya, D. *et al.* Comparative genomics explains the evolutionary success of reef-forming corals. *eLife* **5**, e13288 (2016).

67. Zoccola, D. *et al.* Bicarbonate transporters in corals point towards a key step in the evolution of cnidarian calcification. *Sci. Rep.* **5**, 9983 (2015).

68. Moya, A. *et al.* Carbonic anhydrase in the scleractinian coral Stylophora pistillata: characterization, localization, and role in biomineralization. *J. Biol. Chem.* **283**, 25475–25484 (2008).

69. Le Goff, C. *et al.* Carbonic Anhydrases in Cnidarians: Novel Perspectives from the Octocorallian Corallium rubrum. *PloS One* **11**, e0160368 (2016).

70. Drake, J. L. *et al.* Proteomic analysis of skeletal organic matrix from the stony coral Stylophora pistillata. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3788–3793 (2013).

71. Mass, T. *et al.* Cloning and characterization of four novel coral acid-rich proteins that precipitate carbonates in vitro. *Curr. Biol. CB* **23**, 1126–1131 (2013).

72. Addadi, L. & Weiner, S. Interactions between acidic proteins and crystals: stereochemical requirements in biomineralization. *Proc. Natl. Acad. Sci. U. S. A.* **82**, 4110–4114 (1985).

73. Takeuchi, T., Yamada, L., Shinzato, C., Sawada, H. & Satoh, N. Stepwise Evolution of Coral Biomineralization Revealed with Genome-Wide Proteomics and Transcriptomics.

PLOS ONE **11**, e0156424 (2016).

74. Denoeud, F. *et al.* The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* (2014) doi:10.1126/science.1255274.
75. Ganley, A. R. D. Concerted Evolution☆. in *Reference Module in Life Sciences* (Elsevier, 2017). doi:10.1016/B978-0-12-809633-8.06264-6.
76. Eirín-López, J. M., Rebordinos, L., Rooney, A. P. & Rozas, J. The Birth-and-Death Evolution of Multigene Families Revisited. *Repetitive DNA* **7**, 170–196 (2012).
77. NEI, M. Balanced polymorphism and evolution by the birth- and -death process in the MHC loci. *11th Histocompat. Workshop Conf. 1992* (1992).
78. Mansfield, K. M. & Gilmore, T. D. Innate immunity and cnidarian-Symbiodiniaceae mutualism. *Dev. Comp. Immunol.* **90**, 199–209 (2019).
79. Kolora, S. R. R. *et al.* Origins and evolution of extreme life span in Pacific Ocean rockfishes. *Science* (2021) doi:10.1126/science.abg5332.
80. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).
81. Bythell, J. C., Brown, B. E. & Kirkwood, T. B. L. Do reef corals age? *Biol. Rev.* **93**, 1192–1202 (2018).
82. Putnam, H. M. & Gates, R. D. Preconditioning in the reef-building coral Pocillopora damicornis and the potential for trans-generational acclimatization in coral larvae under future climate change conditions. *J. Exp. Biol.* **218**, 2365–2372 (2015).
83. McClanahan, T. Changes in coral sensitivity to thermal anomalies. (2017) doi:10.3354/MEPS12150.
84. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* **4**, 170093 (2017).
85. Engelen S, Aury JM. fastxtend. https://www.genoscope.cns.fr/fastxtend/.
86. Li, R., Li, Y., Kristiansen, K. & Wang, J. SOAP: short oligonucleotide alignment program. *Bioinformatics* **24**, 713–714 (2008).

87. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

88. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).

89. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1**, 2047-217X-1–18 (2012).

90. rrwick/Filtlong: quality filtering tool for long reads. https://github.com/rrwick/Filtlong.

91. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* **2021**, 1–9 (2021).

92. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **17**, 155–158 (2020).

93. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nat. Biotechnol.* **37**, 540–546 (2019).

94. Vaser, R., Sović, I., Nagarajan, N. & Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).

95. Aury, J.-M. & Istace, B. Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genomics Bioinforma.* **3**, (2021).

96. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **35**, 543–548 (2018).

97. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).

98. Zerbino, D. R. Using the Velvet de novo assembler for short-read sequencing technologies. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **CHAPTER**, Unit-11.5 (2010).

99. Schulz, M. H., Zerbino, D. R., Vingron, M. & Birney, E. Oases: robust de novo RNA-seq

assembly across the dynamic range of expression levels. *Bioinformatics* **28**, 1086–1092 (2012).

100. Smit, AFA, Hubley, R & Green, P. RepeatMasker. http://repeatmasker.org/.
101. Kent, W. J. BLAT—The BLAST-Like Alignment Tool. *Genome Res.* **12**, 656–664 (2002).
102. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988 (2004).
103. Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci. CABIOS* **13**, 477–478 (1997).
104. Dubarry, M. *et al.* Gmove a tool for eukaryotic gene predictions using various evidences. *F1000Research* **5**, (2016).
105. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
106. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
107. RepeatModeler2: automated genomic discovery of transposable element families | bioRxiv. https://www.biorxiv.org/content/10.1101/856591v1.
108. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).
109. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
110. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
111. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
112. Krzywinski, M. I. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* (2009) doi:10.1101/gr.092759.109.

113. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

114. Jehl, P., Sievers, F. & Higgins, D. G. OD-seq: outlier detection in multiple sequence alignments. *BMC Bioinformatics* **16**, 269 (2015).

115. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).

116. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE* **5**, e9490 (2010).

117. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).

118. Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr. Protoc. Bioinforma.* **69**, e96 (2020).

119. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34**, W609–W612 (2006).

120. Yang, Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

121. Storey JD, Bass AJ, Dabney A, Robinson D. *qvalue: Q-value estimation for false discovery rate control*. (Storey Lab, 2021).

122. Kolde, R. *pheatmap: Pretty Heatmaps*. (2019).

123. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

124. Chen, S., He, C., Li, Y., Li, Z. & Melançon, C. E., III. A computational toolset for rapid identification of SARS-CoV-2, other viruses and microorganisms from sequencing data. *Brief. Bioinform.* **22**, 924–935 (2021).