bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sequencing, *de novo* assembly and annotation of the genome of the scleractinian coral, *Pocillopora acuta*

Jeremie Vidal-Dupiol[1*], Cristian Chaparro[2], Marine Pratlong[3,4], Pierre Pontaroti[3,5], Christoph Grunau[2], Guillaume Mitta[2]

1 IHPE, Univ. Montpellier, CNRS, Ifremer, Univ. Perpignan Via Domitia, Montpellier France

2 IHPE, Univ. Montpellier, CNRS, Ifremer, Univ. Perpignan Via Domitia, Perpignan France

3 Aix Marseille Univ,IRD, APHM, Microbe, Evolution, PHylogénie, Infection IHU Méditerranée Infection, Marseille France. Evolutionary Biology team.

4 Aix-Marseille Université, Avignon Université, CNRS, IRD, IMBE, Marseille, France

5 SNC5039 CNRS 19-21 Boulevard Jean Moulin 13005 Marseille

5 Sorbonne Universités, UPMC Univ Paris 06, UMS 2348, Centre de Ressources Biologiques Marines, Observatoire Océanologique, F-66650 Banyuls/Mer, France.

10 Institut de Recherche pour le Développement (IRD), UMR 9220 ENTROPIE & Laboratoire d'Excellence CORAIL, UPVD 52 avenue Paul Alduy, 66860 Perpignan, France.

*Corresponding author: jeremie.vidal.dupiol@ifremer.fr

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Coral reefs are the most divers marine ecosystem. However, under the pressure of global changes and anthropogenic disturbances corals and coral reefs are declining worldwide. In order to better predict and understand the future of these organisms all the tools of modern biology are needed today. However, many NGS based approaches are not feasible in corals because of the lack of reference genomes. Therefore we have sequenced, *de novo* assembled, and annotated, the draft genome of one of the most studied coral species, *Pocillopora acuta* (ex *damicornis*). The sequencing strategy was based on four libraries with complementary insert size and sequencing depth (180pb, 100x; 3Kb, 25x; 8kb, 12x and 20 kb, 12x). The *de novo* assembly was performed with Platanus (352 Mb; 25,553 scaffolds; N50 171,375 bp). 36,140 genes were annotated by RNA-seq data and 64,558 by AUGUSTUS (Hidden-Markov model). Gene functions were predicted through Blast and orthology based approaches. This new genomic resource will enable the development of a large array of genome wide studies but also shows that the *de novo* assembly of a coral genome is now technically feasible and economically realistic.

Introduction

Coral reefs are the most divers marine ecosystem and the second one in terms of diversity after tropical rain forests (Knowlton et al. 2010). In addition to this ecological importance, it provides many ecosystem services, such as a direct access to food and economic-resources through fishing and tourisms (Done et al. 1996, Bryant et al. 1998). The physical and biological support of this ecosystem relies exclusively on one type of organisms, the hermatypic scleractinian coral. Worryingly, these sessile, colonial and symbiotic species are today threatened by an increasing number of various anthropogenic and natural disturbances (Hughes et al. 2003), inducing since the 80's a worldwide decline of this ecosystem (Bellwood et al. 2004). As a consequence the research performed on these biological models has exponentially increased on all aspect of their biology.

Today, most methods studying variations at the transcriptomic, genetic or epigenetic level rely on next generation sequencing. The generalization of these approaches in model organisms such as in the mouse, drosophila or baker's yeast has made it possible to achieve tremendous insights into the biology of these organisms leading to advances in numerous

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

scientific domains (Koboldt et al. 2013). With the cost reduction of the NGS approach, this was also promised but is not yet fulfilled for non-model organisms because of the lack of reference genomes in many phyla or functionnal group. Among the anthozoa, reference genomes are available for the laboratory models *Nematostella vectensis* (Putnam et al. 2007) and *Aiptasia* sp. (Baumgarten et al. 2015). However, there is still few reference genomes for the ecologically important hermatypic corals. To date four species were sequenced and assembled: *Acropora digitifera* (Shinzato et al. 2011), *Pocillopora damicornis* (Cunning et al. 2018), *Stylophora pistillata* (Voolstra et al. 2017) and *Orbicela faveolata* (Prada et al. 2016). However, others are needed to widen the research effort on the future of biodiversity and corals adaptability to global changes.

Among the most studied coral is *Pocillopora acuta* (previously considered as a synonymous of *P. damicornis* until 2014 (Veron and Pichon 1976)). *P. acuta* is encountered on fringing reefs and sheltered areas in the entire Indopacific ocean (Veron 2000). *P. acuta* is also known to be very sensitive to many natural disturbances including coral bleaching (i.e. the symbiosis breakdown between the coral host and its micro-algae endosymbiont) (Loya et al. 2001) and diseases (Ben-Haim and Rosenberg 2002, Luna et al. 2007). It shows relatively good capacities of acclimatization to controlled condition and thus can be maintained in aquaria for decades where it present the advantage to be a fast growing species. These various features have led many scientists to use this coral as a model in a large variety of scientific fields such as population genetics (Adjeroud et al. 2013, Combosch and Vollmer 2015), integrative biology (Vidal-Dupiol et al. 2009, Vidal-Dupiol et al. 2011a), functional genomics (Traylor-Knowles et al. 2011), global changes impact (Vidal-Dupiol et al. 2013, Vidal-Dupiol et al. 2014), transgenerational acclimation (Putnam and Gates 2015), coral diseases (Ben-Haim Rozenblat and Rosenberg 2004, Vidal-Dupiol et al. 2011b), physiology (Richmond 1987, Stimson 1997) and host microbiota interactions (Bourne and Munn 2005) etc. which makes *P. acuta* a major scleractinian coral model.

Although transcriptomic resources are available (Traylor-Knowles et al. 2011, Vidal-Dupiol et al. 2011b) a reference genome for *P. acuta* is still lacking but could enable substantial advances in the understanding of various aspects of coral biology, ecology and evolution. In this context, the aim of the present study is to provide to the scientific community the draft genome sequence and annotation of *P. acuta*. To address this objective, the DNA from a single colony was sequenced using the Illumina technology and *de novo* assembled using Platanus assembler (Kajitani et al. 2014). The annotation of genes and

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

repeats was performed. Gene prediction was conducted; firstly by an experimental annotation using previously published RNA-seq libraries and secondly, by an *ab initio* prediction based on a Hidden-Markov models. Finally, the putative function of each gene and its transcripts were identified by database searches and by an orthology based approach.

Material and methods

Biological material and DNA extraction

The *P. acuta* (Linnaeus, 1758) isolate used in this study was sampled in Lombok, Indonesia (Indonesian CITES Management Authority, CITES number 06832/VI/SATS/LN/2001-E; France Direction de l'Environnement, CITES number 06832/VI/SATS/LN/2001-I) and has been maintained in aquaria since the year 2001. Previously assigned to *Pocillopora damicornis* this isolate was reassigned to *P. acuta*. This assignation was based on the 840 based pair sequence of the ORF marker that enable to separate *P. acuta* and *P. damicornis* (Schmidt-Roach et al. 2014).

In order to avoid contaminations by the zooxanthellae for genome assembly *P. acuta* mini colonies (~7 cm high, ~6 cm diameter) used for DNA extraction were subjected to a menthol treatment inducing bleaching. Briefly, the colonies were placed in a four litter tank filled with seawater. Water motion was created using a submerged water pump (100L/h), temperature was maintained at 27°C and light adjusted to 75µmol/m²/s (PAR). The protocol for the menthol treatment was adapted from previous work (Wang et al. 2012). The first day, the corals were subjected to a concentration of menthol of 0.58mmol/L for 6h. After this exposure they were transferred to the coral nursery for a 18h recovery period. During the second day, the same protocol was applied (menthol treatment and recovery step). The third day, the coral were exposed again to the same treatment but only until the polyps were closed. Once achieved the corals were placed in the coral nursery for recovery while at the same time they lose zooxanthellae. This last step typically takes four to five days.

For DNA extraction, coral tissues of one bleached mini colony (~7 cm high, ~6 cm diameter) were harvested using an airpic in 50 mL of tissue extraction buffer (1M sucrose; 0.05 mM EDTA; 4°C). Then, the extract was centrifuged 10 min at 3000g (4°C) and the pellet was resuspended in the G2 buffer of the QIAGEN Genomic DNA kit. The rest of the protocol was performed according to the manufacturer's instructions with the 100/G Genomic-tip.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

DNA quantity and quality were assessed by spectrophotometry (nanodrop), fluorescence (Qbit) and 0.5% agarose gel electrophoresis.

Library preparation and sequencing

In order to facilitate the *de novo* assembly of the *P. acuta* genome, four different libraries were constructed and sequenced. Sequencing was performed on an Illumina Hiseq 2000 producing 100 bp paired-end reads. The first library was a shotgun (SG) library with an expected average insert size of 180 bp. It was sequenced at an estimated genome coverage of 100X. The second one was a long jumping distance (LJD) library with an average insert size of 3000 bp. It was sequenced for an estimated genome coverage of 25X. The two last one were LJD libraries with an expected average insert size of 8 and 20 kb, respectively. The sequencing coverage was estimated at 12X. Sequencing and library preparation were performed by the Eurofin company.

De novo genome assembly

A stringent cleaning pipeline of the raw reads was applied to the four libraries. First, all reads of the four libraries were filtered in function of their quality. Only reads displaying a Phred quality score > 30 for 95% of its bases were kept for the analysis (FASTX-Toolkit). Secondly, the remaining reads were cleaned from any trace of adaptor using cutadapt program (Martin 2011). Thirdly, reads were interlaced/de-interlaced to separate singleton from paired reads and finally the reads from the SG library were subjected to a correction step using ErrorCorrection program from the SOAPdenovo2 package (Luo et al. 2012).

For the *de novo* assembly of *P. acuta* genome we used the multiple kmer assembler Platanus (Kajitani et al. 2014). For the contiguing step, only the paired read of the SG library were used. For scaffolding, the reads of all libraries were sequentially used from the shortest to the longest insert size (SG 180pb, LJD 3kb, LJD 8kb, LJD 20 kb). In a last step, gaps generated during scaffolding were closed using the paired reads of the SG libraries using the gap closer program included in the Platanus package. Once completed, the assembly quality was assessed by classical metrics (total assembly length, longest scaffold, N50 etc.) compiled in the program QUAST (Gurevich et al. 2013). In order to provide a functional validation of the assembly this quality assessment was also performed using the program CEGMA that looks for the presence of the 248 most conserved core eukaryotic proteins in the assembly (Parra et al. 2007). All scripts and parameters used in command line during the bioinformatics

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

treatments are summarized in supplementary data 1, other tools were run on a local Galaxy instance.

Structural annotation

As a first approach for genome annotation, exon/intron structures of genes were determined using RNA-seq data (experimental approach) previously published (Vidal-Dupiol et al. 2013, Vidal-Dupiol et al. 2014). The cleaned paired reads from our previous work were mapped against the scaffolds with a length above 5 kb using TopHat2 (Trapnell et al. 2009). All parameters were used with the default setting except for the mean and the standard deviation of the inner distance between pairs that were adapted to each specific library. Once mapped, the BAM file output obtained for each RNA-seq library was used to assemble the transcripts defined by TopHat2. This was done with Cufflinks with default parameters (Trapnell et al. 2010). Finally, all the transcript assembly generated (one per mapped RNA-seq library) were merged with Cuffmerge (Trapnell et al. 2010) using the default parameters.

Because our RNA-seq libraries potentially did not cover all putative transcripts in the genome, we performed as a second approach an *ab initio* gene prediction. This step was done using the AUGUSTUS web server (Stanke and Waack 2003) with the genome assembly filtered for scaffolds ≥ 5 kb and the Cuffmerge transcriptome as input file for the training and prediction steps. Genes were predicted on both strands with the option "predict any number of genes".

Functional annotation

In order to attribute a putative function to a maximum of the predicted transcripts but with the lowest probability of wrong annotation we applied two independent methods. The longest ORF of each putative transcripts was found using getorf (Rice et al. 2000), and the ones longer than 100 amino acids were selected for annotation. Then we used orthoMCL (Li et al. 2003) to identify potential orthologous sequences between our predicted ORF and; very well annotated genomes (*Homo sapiens* (Venter et al. 2001), *Mus musculus* (Chinwalla et al. 2002), *Caenorhabditis elegans* (Consortium 1998), *Danio rerio* (Howe et al. 2013), *Drosophila melanogaster* (Adams et al. 2000), *Saccharomyces cerevisiae* (Mewes et al. 1997) and *Strongylocentrotus purpuratus* (Sodergren et al. 2006)); or with a biological and an evolutionary interest with regard to the phylogenetic position of our organism (the cnidarians *Acropora digitifera* (Shinzato et al. 2011), *Hydra magnipapillata* (Chapman et al. 2010) and *Nematostella vectensis* (Putnam et al. 2007), the coral symbiont *Symbiodinium minutum*

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

(Shoguchi et al. 2013) and the sponge *Amphimedon queenslandica* (Srivastava et al. 2010)). Orthology was considered significant when the *e*-value obtained was lower than 10^{-5}. In this case and when available, the annotation of the ortholog(s) was transferred to the *P. acuta* sequence. In parallel to this approach we used Blast2GO (Conesa et al. 2005) version 2.4.2 to perform a semi-automated functional annotation of all putative transcripts using a set of similarity search tools (Conesa et al. 2005) : i) an initial annotation with BLASTX (against the non redundant NCBI database; *e*-value at 1×10^{-3}); ii) a protein domain search using InterProscan; iii) an enzyme annotation using the *Kyoto Encyclopedia of Genes and Genomes* (Kanehisa and Goto 2000) (KEGG) enzyme database; and iv) assignment of a Gene Ontology term (Ashburner et al. 2000).

Repeats annotation

Identification of the transposable elements (TEs) from the SINE family was performed as described previously (Baucom et al. 2009). Briefly, clustering was done by aligning candidates and manually extracting the families which were realigned and borders identified when possible. For the annotation of other TEs, the search using RepeatMasker (Tarailo-Graovac and Chen 2009) and the Repbase (Jurka et al. 2005) library on the assembled scaffolds did not yield any positive results. We therefore opted for another strategy which consists on recuperating all the reads that do not align to the assembled scaffolds using samtools (Li et al. 2009) and assemble them using RepArk (Koch et al. 2014) into contigs which will be searched for signatures of TEs. The RepeatMasker/Repbase combination did not give any results on the assembled contigs, therefore, we used RepeatModeler (Smit and Hubley 2010) to create a specific repeat library for the assembled contigs and used that library with RepeatMasker to search for TEs.

Contamination level and genome heterozygosis

In order to evaluate the level of contamination by *Symbiodinium* sp. DNA that would be co-extracted with the coral DNA, we mapped the reads from the SG library on the genome of *S. kawagutii* (Lin et al. 2015) and *S. minutum* (Shoguchi et al. 2013). This was done with bowtie2 in single-end and sensitive end to end modes (Langmead and Salzberg 2012). In a second step, potential PCR duplicates generated during the amplification step of the gDNA-seq library preparation were removed using RmDup from the Sam tools package (Li et al. 2009) to decrease artefactual redundancy.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to evaluate the level of heterozygosis of the reference genome the raw reads of the SG library were mapped with bowtie2 on the *P. acuta* assembly. This was done in single end and sensitive end to end mode and as previously, PCR duplicates were removed with RmDup. Then, unique hits were filtered and used as input for variant calling using Varscan2 with default parameters (Koboldt et al. 2012).

Results and Discussion

Sequencing and de novo assembly

To enable the development of studies at the genome wide level and to strengthen the position of *Pocillopora acuta* as a coral model we sequenced and *de novo* assembled its genome. The sequencing strategy was based on 4 libraries with complementary insert sizes and sequencing depths (Table 1). The shotgun library (insert size 180 pb) yielded 237,589,976 paired reads of 100 bp each with an average Q score of 36.45 (93.71%>Q30). The corresponding 47,517 Mb produced correspond to an estimated genome coverage of 146X. The 3 kb LJD library has a measured average insert size of 2,479 bp. The sequencing yielded 109,242,128 paired read of 100 bp each with an average Q score of 32.31 (81.79%>Q30) this corresponds to 21,848 Mbp (genome coverage 67 x). The 8 kb LJD library has a measured average insert size of 7,825 bp. The sequencing yielded 44,399,583 paired read of 100 bp each with an average Qvalue of 31.98 (80.56%>Q30) which corresponds to 8,880 Mbp (genome coverage 27). The 20 kb LJD library has a measured average insert size of 32,325 bp. The sequencing yielded 99,813,316 paired read of 100 bp each with an average Qvalue of 30.76 (77.73%>Q30) which corresponds to 19,963 Mbp (genome coverage 61 x). In order to decrease the complexity of the dataset, the raw reads were quality filtered at Q30 for 90% of the read lenght, all remaining traces of adaptor were removed, and finally, sequencing error were corrected. Singleton were filtered and excluded from assembly.

The processed paired-reads were used as input for *de novo* assembly with Platanus assembler (contiging, scaffolding and gap closing). The assembly has resulted in 25,553 scaffolds greater than 1,000 bp and a N50 of 171,375 bp (the main assembly quality statistics are summarized in the Table 2). The longest scaffold reached 1,296,445 bp and the overall assembly is only 8% longer than the predicted genome size (352 Mb *vs* 325 Mb). These metrics are in the same order of magnitude than those classically obtained for other anthozoans (Putnam et al. 2007, Shinzato et al. 2011, Baumgarten et al. 2015). Considering the genome size, our prediction and assembly results stay in the anthozoa range delineated by

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Aiptasia sp. (the smallest; 260 Mb) and *Stylophora pistillata* (the largest; 434 Mb) genome (Baumgarten et al. 2015, Voolstra et al. 2017).

The GC% was equal to 37.84% and its distribution was unimodal. Since the GC % of anthozoa is around 37-39% (Shinzato et al. 2011) and around 44-46% in *Symbiodinium* sp. (Shoguchi et al. 2013, Lin et al. 2015) our results suggest a very low level of contamination by *Symbiodinium* DNA in our assembly (Sabourault et al. 2009). This very low level of contamination was confirmed through the mapping of the reads constituting the SG library on the *S. kawagutii* and *S. minutum* genome (Shoguchi et al. 2013, Lin et al. 2015). Indeed, only 0.02% and 0.03% of these reads were mapped on the genomes of *S. kawagutii* and *S. minutum* respectively. This results is strengthen by the very low number of contigs (one contig) included in an orthologs group containing Symbiodinium sequences only. These results confirmed that the menthol treatment was very efficient to induce a complete bleaching of *P. acuta* (Wang et al. 2012).

Finally, in order to evaluate the biological significance of this assembly we looked for the presence of 248 ultra-conserved core eukaryotic gene (COG) with the CEGMA package (Parra et al. 2007). In total, 84% of these COG were found in full length, and this number reach 93% when partial length similarities are included, showing the completeness of our assembly. These results are in the upper part of the range of values classically obtained from a *de novo* draft genome assembly performed with Illumina reads. For example, the draft genomes of the arthropods *Sarcoptes scabie* var. *hominis* and var. *suis* contained 98.79% of the 248 COG (Mofiz et al. 2016) while the draft genome of another arthropods, the chinese mitten crab *Eriocheir sinensis* contained 66.9% of these genes (Song et al. 2016).

Structural and functional annotation

We had based this structural and functional annotation on experimental (RNA-seq) and *ab initio* approaches (Hidden-Markov model). Because these two approaches do not reach the same level of confidence we decided to keep the results of these two strategies separated. This will enable the users of these data to choose the experimental and/or *ab initio* gene predictions in function of the scientific question they will investigate.

As a first approach, genes were predicted using experimental data previously obtained. These RNA-seq libraries were chosen in order to cover a large range of physiological condition and include the response to the exposure to normal and high temperatures, virulent and non-virulent bacteria and acidification (Vidal-Dupiol et al. 2013, Vidal-Dupiol et al.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

2014, Vidal-Dupiol et al. submitted). The paired reads of these libraries were used to predict gene and exon/intron structures on scaffolds with a size above or equal to 5 kb. This approaches has predicted 36,140 genes encoding 63,181 alternatively spliced transcripts (Table 3). These predicted transcripts were then used to train AUGUSTUS (Hidden-Markov model) for the *ab initio* gene prediction that had resulted in the identification of 64,558 predicted-genes encoding 79,506 alternatively spliced transcripts. The number of predicted genes using the experimental data is slightly higher to what was found for the two other sequenced symbiotic anthozoan (Shinzato et al. 2011, Baumgarten et al. 2015, Prada et al. 2016, Voolstra et al. 2017, Cunning et al. 2018). However the number of putative transcripts encoded by these genes are in agreement with what was classically obtained in Anthozoan transcriptome assembly (Meyer et al. 2009, Traylor-Knowles et al. 2011, Lehnert et al. 2012, Moya et al. 2012, Vidal-Dupiol et al. 2013, Shinzato et al. 2014, Kitchen et al. 2015). The differences between the number of gene predicted by the experimental and the *ab initio* approaches may result from difficulties encountered by AUGUSTUS rather than by TopHat2. Indeed, the mean exon length is 30% lower in the *ab initio* approach and this may result in an increasing gene number due to gene fragmentation. This hypothesis is confirmed by CEGMA that shows that the completeness of the transcriptome issued from the *ab initio* prediction is low with 56% of the 248 CEGs founds while 93% were found with the transcriptome generated by the experimental approach. In addition *ab initio* approaches are known to annotate some false genes such as pseudogenes and nonfunctional duplicated genes. However, both predictions can be usefull since genes specifically involved in some developmental process, larval stage, etc. may be represented in the *ab initio* annotation only, because genes pertaining to these physiological states were not found with the experimental annotation.

The functional annotation was done using Blast2GO (Conesa et al. 2005) on the transcripts predicted from the RNA-seq and *ab initio* approaches, separately (Table 3). The BlastX performed on the non-redundant NCBI database returned the following results for the transcripts issued from the RNA-seq and *ab initio* approaches, respectively: i) 77.3% and 44.9% presented a significant similarity for a protein with a known function; ii) 4.6 % and 11.3% showed similarities for predicted proteins; iii) 18.1% and 43.8% did not returned any results; iv) GO-terms were then attributed to 52.2% and 22.3% of the transcripts; v) KEGG enzyme code was assigned to 11.3% and 5.5% of the sequences; and vi) conserved protein domains were detected in 52% and 50.6% of the transcript. Analysis of the RNA-seq based approached of GO-term repartition using WEGO (Ye et al. 2006) shows that our predicted transcripts enter in 6, 6 and 20 GO categories (GO level 2, for all level GO repartition, see the

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

supplementary data 2) of the Cellular Component, Molecular Function and Biological Process root, respectively (Fig.1). For the RNA-seq based approach, orthoMCL had generated 16,469 groups of orthologs containing 47,388 transcripts (Table 3). This allowed to successfully annotating 24,024 transcripts with a high level of confidence (50.6% of those in clusters). For the *ab initio* approach, 27,440 groups of orthologs were created; they contain 45,839 transcripts and enable the annotation of 25,335 sequences. In comparison to what is usually obtained in transcriptome assembly of non-model invertebrates the number of transcripts showing significant similarities for protein with a putative function is high. Indeed, this rate fluctuates in general between 20 and 50% in many recent studies (Kitchen et al. 2015, Harney et al. 2016, McGrath et al. 2016). However in draft genome assembly this rates of annotation is higher and comparable to what we obtained (Shinzato et al. 2011, Baumgarten et al. 2015). This is probably due to a better sequencing coverage of the gene set in genome assembly rather than in transcriptome. Firstly, because genome *de novo* assembly need a higher sequencing coverage and secondly because in gDNA-seq the representation of each gene is theoretically equal while it is biased by the expression level in RNA-seq.

Repeat content

The process of repeat annotation shows that at least 15.28% of *P. acuta* genome is composed of repeated sequences. This value is very close to what was found in the genome of *A. digitifera* with 13% of repeat content (Shinzato et al. 2011) and lower than in *Aiptasia* sp. with 26% (Baumgarten et al. 2015). In total, the four main superfamily of TEs (DNA transposons, SINE-, LINE- and LTR-retrotransposons) were found (Table 4) and they represent 18 families of transposons and 19 families of retrotransposons (supplementary data 3). Among them, seven are specific to *P. acuta* and 30 and 20 are shared with *Aiptasia* sp. and *A. digitifera*, respectively (Fig. 2). However, their occurrence is low in comparison to what was found in the two other symbiotic anthozoa (Table 4). Indeed, if 15.28% of the genome is composed of repeated sequences, the identified repeat sequences represent only 3.23% of this 15.36%. The remaining 11.98% correspond essentially to unidentified sequences (11.37%) in addition to satellites (0.03%), simple (0.9%), low complexity repeats (0.16%), and small RNA (0.06%). This high level of unidentified repeated sequences is not an exception. For example, in the *Aiptasia* genome, 63% of the repeat content corresponds to a unique unknown repeat (Baumgarten et al. 2015).

Heterozygosis and SNP library

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In order to evaluate the heterozygosis of the genome the reads from the SG library were re-mapped on the genome (Bowtie2) and variant were called with VarScan2. This analysis reveals the presence of 2,505,660 heterozygote single nucleotide polymorphisms (7.1 SNP/kb) and 321,295 indels. This level of heterozygosis in a single genome can be considered as very high even if we cannot exclude that a small proportion of these SNPs are false positive that can be due to sequencing or mapping error. However, if such a kind of data are absent from the anthozoan literature some comparisons are possible with other invertebrate phyla. In arthropods, SNP density can be as high as 16.5 SNP/kb in the butterflies from the genus *Lycaeides* (Gompert et al. 2010) and as low as 0.062 in the varroato-mite, *Varroa destructor* (Cornman et al. 2010). In these studies these results were obtained from dozen to hundreds of individuals, while in our study we have probably only one genotype per colony. This lead to believe that in the case of a populational study this density can significantly increase and reach very high values of SNP density. Alternatively, some of the polymorphism observed in our study can also be the results of intra colonial genetic variation, a phenomenon more and more observed and quantified in corals (Schweinsberg et al. 2014, Schweinsberg et al. 2015, Barfield et al. 2016) and that can be due to the accumulation of somatic mutations (Van Oppen et al. 2012) or to chimerism (Rinkevich et al. 2016).

Cnidarian core proteome

The cnidarian core proteome was identified through the orthoMCL approach and its annotation. It is composed of 1,781 orthologous group of proteins, shared between the two scleractinian corals *P. acuta* and *A. digitifera,* the anemone *N.vectensis* (actinia) each of them belonging to the anthozoa class, and *H. magnipapillata* from the hydrozoa class (Fig.3). The GO term repartition (GO level 2) between this cnidarian core proteome and the entire proteome content of *P. acuta* is very close (Fig. 4) reflecting the absence of large functional gain or loss.

Conclusion

The draft genome assembly provided by this study constitutes a new coral reference genome and a new resource for the scientific community interested in cnidarians genomics approach (*sensus lato*). This will enable the development of a large array of genome wide studies that will lead to a better understanding of coral physiology, ecology and adaptability.

Availability

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The raw reads of each libraries were submitted to the NCBI Sequence Read Archive (accession numbers are: SRR4254617; SRR4254618; SRR4254619; SRR4254620) The draft genome and related annotation files can be downloaded using the following link http://ihpe.univ-perp.fr/acces-aux-donnees/

Acknowledgements

This study was supported by the Agence Nationale de la Recherche through the Program BIOADAPT (ADACNI ANR-12-ADAP-0016-03) and the French-Israeli High Council for Science and Technology (P2R n u29702YG). The facilities of the Bio-Environnement platform (Perpignan, France) and the ABiMS platform (Roscoff, France) were used for the bioinformatics and molecular biology studies, the Aquarium facilities of the UMS 2348 were used for the sample preparation. The authors are indebted to Rayan Chikhi for his advices in bioinformatic during the assembly processes and to Sebastian Schmidt-Roach for his informations about the differences between *P. acuta* and *P. damicornis*.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, and e. al. 2000. The genome sequence of *Drosophila melanogaster*. Science **287**:2185-2195.

Adjeroud, M., A. Guérécheau, J. Vidal-Dupiol, J. F. Flot, S. Arnaud-Haond, and F. Bonhomme. 2013. Genetic diversity, clonality and connectivity in the scleractinian coral *Pocillopora damicornis*: a multi-scale analysis in the South Pacific. Marine Biology.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene Ontology: tool for the unification of biology. Nat Genet **25**:25-29.

Barfield, S., G. V. Aglyamova, and M. V. Matz. 2016. Evolutionary origins of germline segregation in Metazoa: evidence for a germ stem cell lineage in the coral *Orbicella faveolata* (Cnidaria, Anthozoa). Proceedings of the Royal Society of London B: Biological Sciences **283**:2015-2128.

Baucom, R. S., J. C. Estill, C. Chaparro, N. Upshaw, A. Jogi, J.-M. Deragon, R. P. Westerman, P. J. SanMiguel, and J. L. Bennetzen. 2009. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet **5**:e1000732.

Baumgarten, S., O. Simakov, L. Y. Esherick, Y. J. Liew, E. M. Lehnert, C. T. Michell, Y. Li, E. A. Hambleton, A. Guse, M. E. Oates, J. Gough, V. M. Weis, M. Aranda, J. R. Pringle, and C. R. Voolstra. 2015. The genome of *Aiptasia*, a sea anemone model for coral symbiosis. Proceedings of the National Academy of Sciences of The United States Of America **112**:11893-11898.

Bellwood, D. R., T. P. Hughes, C. Folke, and M. Nyström. 2004. Confronting the coral reef crisis. Nature **429**:827-833.

Ben-Haim Rozenblat, Y. and E. Rosenberg 2004. Temperature-regulated bleaching and tissue lysis of *Pocillopora damicornis* by the novel pathogen *Vibrio coralliilyticus*. Pages 301-324 *in* E. Rosenberg and Y. Loya editors. Coral health and disease. Spinger-Verlag, New-York.

Ben-Haim, Y. and E. Rosenberg. 2002. A novel *Vibrio* sp. pathogen of the coral *Pocillopora damicornis*. Marine Biology **141**:47-55.

Bourne, D. G. and C. B. Munn. 2005. Diversity of bacteria associated with the coral Pocillopora damicornis from the Great Barrier Reef. Environmental Microbiology **7**:1162-1174.

Bryant, D., L. Burke, J. W. McManus, and M. D. Spalding. 1998. Reefs at risk: a map based indicator of threats to the world's coral reefs., World Resources Institute, Washington, D.C.

Chapman, J. A., E. F. Kirkness, O. Simakov, S. E. Hampson, T. Mitros, T. Weinmaier, T. Rattei, P. G. Balasubramanian, J. Borman, D. Busam, K. Disbennett, C. Pfannkoch, N. Sumin, G. G. Sutton, L. D. Viswanathan, B. Walenz, D. M. Goodstein, U. Hellsten, T. Kawashima, S. E. Prochnik, N. H. Putnam, S. Shu, B. Blumberg, C. E. Dana, L. Gee, D. F. Kibler, L. Law, D. Lindgens, D. E. Martinez, J. Peng, P. A. Wigge, B. Bertulat, C. Guder, Y. Nakamura, S. Ozbek, H. Watanabe, K. Khalturin, G. Hemmrich, A. Franke, R. Augustin, S. Fraune, E. Hayakawa, S. Hayakawa, M. Hirose, J. S. Hwang, K. Ikeo, C. Nishimiya-Fujisawa, A. Ogura, T. Takahashi, P. R. H. Steinmetz, X. Zhang, R. Aufschnaiter, M.-K. Eder, A.-K. Gorny, W. Salvenmoser, A. M. Heimberg, B. M. Wheeler, K. J. Peterson, A. Bottger, P. Tischler, A. Wolf, T. Gojobori, K. A.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Remington, R. L. Strausberg, J. C. Venter, U. Technau, B. Hobmayer, T. C. G. Bosch, T. W. Holstein, T. Fujisawa, H. R. Bode, C. N. David, D. S. Rokhsar, and R. E. Steele. 2010. The dynamic genome of Hydra. Nature **464**:592-596.

Chinwalla, A. T., L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. W. Hillier, E. R. Mardis, J. D. McPherson, and e. al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature **420**:520-562.

Combosch, D. J. and S. V. Vollmer. 2015. Trans-Pacific RAD-Seq population genomics confirms introgressive hybridization in Eastern Pacific Pocillopora corals. Molecular Phylogenetics and Evolution **88**:154-162.

Conesa, A., S. Götz, J. M. Garcia-Gomez, J. Terol, M. Talon, and M. Robles. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21**:3674-3676.

Consortium, S. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. Science **282**:2012-2018.

Cornman, R. S., M. C. Schatz, J. S. Johnston, Y.-P. Chen, J. Pettis, G. Hunt, L. Bourgeois, C. Elsik, D. Anderson, C. M. Grozinger, and J. D. Evans. 2010. Genomic survey of the ectoparasitic mite *Varroa destructor*, a major pest of the honey bee Apis mellifera. BMC Genomics **11**:1-15.

Cunning, R., R. A. Bay, P. Gillette, A. C. Baker, and N. Traylor-Knowles. 2018. Comparative analysis of the Pocillopora damicornis genome highlights role of immune system in coral evolution. Scientific Reports **8**:16134.

Done, T. J., J. C. Ogden, and W. J. Wiebe 1996. Biodiversity and ecosystem function of coral reefs. Pages 393-429 *in* H. A. Mooney, E. Cushman, O. E. S. Medina, and E. D. Schulze, editors. Functional role of biodiversity: A global perspective. John Wiley and Sons, Chichester.

Gompert, Z., M. L. Forister, J. A. Fordyce, C. C. Nice, R. J. Williamson, and C. Alex Buerkle. 2010. Bayesian analysis of molecular variance in pyrosequences quantifies population genetic structure across the genome of *Lycaeides* butterflies. Molecular Ecology **19**:2455-2473.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics **29**:1072-1075.

Harney, E., B. Dubief, P. Boudry, O. Basuyaux, M. B. Schilhabel, S. Huchette, C. Paillard, and F. L. Nunes. 2016. De novo assembly and annotation of the European abalone *Haliotis tuberculata* transcriptome. Marine genomics.

Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, and L. Matthews. 2013. The zebrafish reference genome sequence and its relationship to the human genome. Nature **496**:498-503.

Hughes, T., A. Baird, D. Bellwood, M. Card, S. Connolly, C. Folke, R. Grosberg, H. Guldberg, J. Jackson, J. Kleypas, J. Lough, P. Marshall, M. Nyström, S. Palumbi, J. Pandolfi, B. Rosen, and J. Roughgarden. 2003. Climate change, human impacts, and the resilience of coral reefs. Science **301**:929-933.

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

Kajitani, R., K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, and T. Itoh. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Research.

Kanehisa, M. and S. Goto. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic acids research **28**:27-30.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Kitchen, S. A., C. M. Crowder, A. Z. Poole, V. M. Weis, and E. Meyer. 2015. De novo assembly and characterization of four anthozoan (phylum Cnidaria) transcriptomes. G3: Genes| Genomes| Genetics **5**:2441-2452.

Knowlton, N., R. E. Brainard, R. Fisher, M. Moews, L. Plaisance, and M. J. Caley. 2010. Coral reef biodiversity. Pages 65-74 *in* A. D. McLntyre, editor. Life in the World's Oceans: Diversity Distribution and Abundance. Wiley-Blackwell, Singapore.

Koboldt, Daniel C., Karyn M. Steinberg, David E. Larson, Richard K. Wilson, and E. R. Mardis. 2013. The next-generation sequencing revolution and ts impact on genomics. Cell **155**:27-38.

Koboldt, D. C., Q. Zhang, D. E. Larson, D. Shen, M. D. McLellan, L. Lin, C. A. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research **22**:568-576.

Koch, P., M. Platzer, and B. R. Downie. 2014. RepARK—de novo creation of repeat libraries from whole-genome NGS reads. Nucleic acids research.

Langmead, B. and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods **9**:357-359.

Lehnert, E., M. Burriesci, and J. Pringle. 2012. Developing the anemone *Aiptasia* as a tractable model for cnidarian-dinoflagellate symbiosis: the transcriptome of aposymbiotic A. pallida. BMC Genomics **13**:271.

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

Li, L., C. J. Stoeckert, and D. S. Roos. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research **13**:2178-2189.

Lin, S., S. Cheng, B. Song, X. Zhong, X. Lin, W. Li, L. Li, Y. Zhang, H. Zhang, J. Zhiliang, M. Cai, Y. Zhuang, X. Shi, L. Lin, L. Wang, Z. Wang, X. Liu, S. Yu, P. Zeng, H. Hao, Q. Zou, C. Chen, Y. Li, Y. Wang, C. Xu, S. Meng, X. Xu, J. Wang, H. Yang, D. A. Campbell, N. R. Sturm, S. Dagenais-Bellefeuille, and D. Morse. 2015. The *Symbiodinium kawagutii* genome illuminates dinoflagellate gene expression and coral symbiosis. Science **350**:691-694.

Loya, Y., K. Sakai, K. Yamazato, Y. Nakano, R. Sambali, and R. V. Van Woesik. 2001. Coral bleaching: the winners and the losers. Ecology Letters **4**:122-131.

Luna, G. M., F. Biavasco, and R. Danovaro. 2007. Bacteria associated with the rapid tissue necrosis of stony corals. Environmental Microbiology **9**:1851-1857.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**:18.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal **17**:10-12.

McGrath, L. L., S. V. Vollmer, S. T. Kaluziak, and J. Ayers. 2016. De novo transcriptome assembly for the lobster *Homarus americanus* and characterization of differential gene expression across nervous system tissues. BMC Genomics **17**:1-16.

Mewes, H., K. Albermann, M. Bähr, D. Frishman, A. Gleissner, J. Hani, K. Heumann, K. Kleine, A. Maierl, and S. Oliver. 1997. Overview of the yeast genome. Nature **387**:7-8.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Meyer, E., G. Aglyamova, S. Wang, J. Buchanan-Carter, D. Abrego, J. Colbourne, B. Willis, and M. Matz. 2009. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics **10**:219.

Mofiz, E., D. C. Holt, T. Seemann, B. J. Currie, K. Fischer, and A. T. Papenfuss. 2016. Genomic resources and draft assemblies of the human and porcine varieties of scabies mites, *Sarcoptes scabiei* var. *hominis* and var. *suis*. GigaScience **5**:1.

Moya, A., L. Huisman, E. E. Ball, D. C. Hayward, L. C. Grasso, C. M. Chua, H. N. Woo, J. P. Gattuso, S. Forêt, and D. J. Miller. 2012. Whole transcriptome analysis of the coral *Acropora millepora* reveals complex responses to CO_2-driven acidification during the initiation of calcification. Molecular Ecology **21**:2440-2454.

Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics **23**:1061-1067.

Prada, C., B. Hanna, A. F. Budd, C. M. Woodley, J. Schmutz, J. Grimwood, R. Iglesias-Prieto, J. M. Pandolfi, D. Levitan, K. G. Johnson, N. Knowlton, H. Kitano, M. DeGiorgio, and M. Medina. 2016. Empty Niches after Extinctions Increase Population Sizes of Modern Corals. Current Biology **26**:3190-3194.

Putnam, H. M. and R. D. Gates. 2015. Preconditioning in the reef-building coral *Pocillopora damicornis* and the potential for trans-generational acclimatization in coral larvae under future climate change conditions. Journal of Experimental Biology **218**:2365-2372.

Putnam, N. H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro, E. Lindquist, V. V. Kapitonov, J. Jurka, G. Genikhovich, I. V. Grigoriev, S. M. Lucas, R. E. Steele, J. R. Finnerty, U. Technau, M. Q. Martindale, and D. S. Rokhsar. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317**:86-94.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: the European molecular biology open software suite. Trends in genetics **16**:276-277.

Richmond, R. H. 1987. Energetics, competency, and long-distance dispersal of planula larvae of the coral *Pocillopora damicornis*. Marine Biology **93**:527-533.

Rinkevich, B., L. Shaish, J. Douek, and R. Ben-Shlomo. 2016. Venturing in coral larval chimerism: a compact functional domain with fostered genotypic diversity. Scientific Reports **6**:19493.

Sabourault, C., P. Ganot, E. Deleury, D. Allemand, and P. Furla. 2009. Comprehensive EST analysis of the symbiotic sea anemone *Anemonia viridis*. BMC Genomics **10**:333.

Schmidt-Roach, S., K. J. Miller, P. Lundgren, and N. Andreakis. 2014. With eyes wide open: a revision of species within and closely related to the Pocillopora damicornis species complex (Scleractinia; Pocilloporidae) using morphology and genetics. Zoological Journal of the Linnean Society **170**:1-33.

Schweinsberg, M., R. G. Pech, R. Tollrian, and K. Lampert. 2014. Transfer of intracolonial genetic variability through gametes in *Acropora hyacinthus* corals. Coral reefs **33**:77-87.

Schweinsberg, M., L. C. Weiss, S. Striewski, R. Tollrian, and K. P. Lampert. 2015. More than one genotype: how common is intracolonial genetic variability in scleractinian corals? Molecular Ecology **24**:2673-2685.

Shinzato, C., M. Inoue, and M. Kusakabe. 2014. A snapshot of a coral "holobiont": a transcriptome assembly of the scleractinian coral, *Porites*, captures a wide variety of genes from both the host and symbiotic zooxanthellae. PLoS ONE **9**:e85182.

Shinzato, C., E. Shoguchi, T. Kawashima, M. Hamada, K. Hisata, M. Tanaka, M. Fujie, M. Fujiwaea, R. Koyanagi, T. Ikuta, A. Fujiyama, D. Miller, and N. Satoh. 2011. Using

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the *Acropora digitifera* genome to understand coral responses to environmental change. Nature **476**:320-323.

Shoguchi, E., C. Shinzato, T. Kawashima, F. Gyoja, S. Mungpakdee, R. Koyanagi, T. Takeuchi, K. Hisata, M. Tanaka, M. Fujiwara, M. Hamada, A. Seidi, M. Fujie, T. Usami, H. Goto, S. Yamasaki, N. Arakaki, Y. Suzuki, S. Sugano, A. Toyoda, Y. Kuroki, A. Fujiyama, M. n. Medina, MaryÂ A. Coffroth, D. Bhattacharya, and N. Satoh. 2013. Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. Current Biology **23**:1399-1408.

Smit, A. and R. Hubley. 2010. RepeatModeler Open-1.0. Repeat Masker Website.

Sodergren, E. and G. M. Weinstock and E. H. Davidson and R. A. Cameron and R. A. Gibbs and R. C. Angerer and L. M. Angerer and M. I. Arnone and D. R. Burgess and R. D. Burke and J. A. Coffman and M. Dean and M. R. Elphick and C. A. Ettensohn and K. R. Foltz and A. Hamdoun and R. O. Hynes and W. H. Klein and W. Marzluff and D. R. McClay and R. L. Morris and A. Mushegian and J. P. Rast and L. C. Smith and M. C. Thorndyke and V. D. Vacquier and G. M. Wessel and G. Wray and L. Zhang and C. G. Elsik and O. Ermolaeva and W. Hlavina and G. Hofmann and P. Kitts and M. J. Landrum and A. J. Mackey and D. Maglott and G. Panopoulou and A. J. Poustka and K. Pruitt and V. Sapojnikov and X. Song and A. Souvorov and V. Solovyev and Z. Wei and C. A. Whittaker and K. Worley and K. J. Durbin and Y. Shen and O. Fedrigo and D. Garfield and R. Haygood and A. Primus and R. Satija and T. Severson and M. L. Gonzalez-Garay and A. R. Jackson and A. Milosavljevic and M. Tong and C. E. Killian and B. T. Livingston and F. H. Wilt and N. Adams and R. Bellé and S. Carbonneau and R. Cheung and P. Cormier and B. Cosson and J. Croce and A. Fernandez-Guerra and A.-M. Genevière and M. Goel and H. Kelkar and J. Morales and O. Mulner-Lorillon and A. J. Robertson and J. V. Goldstone and B. Cole and D. Epel and B. Gold and M. E. Hahn and M. Howard-Ashby and M. Scally and J. J. Stegeman and E. L. Allgood and J. Cool and K. M. Judkins and S. S. McCafferty and A. M. Musante and R. A. Obar and A. P. Rawson and B. J. Rossetti and I. R. Gibbons and M. P. Hoffman and A. Leone and S. Istrail and S. C. Materna and M. P. Samanta and V. Stolc and W. Tongprasit and Q. Tu and K.-F. Bergeron and B. P. Brandhorst and J. Whittle and K. Berney and D. J. Bottjer and C. Calestani and K. Peterson and E. Chow and Q. A. Yuan and E. Elhaik and D. Graur and J. T. Reese and I. Bosdet and S. Heesun and M. A. Marra and J. Schein and M. K. Anderson and V. Brockton and K. M. Buckley and A. H. Cohen and S. D. Fugmann and T. Hibino and M. Loza-Coll and A. J. Majeske and C. Messier and S. V. Nair and Z. Pancer and D. P. Terwilliger and C. Agca and E. Arboleda and N. Chen and A. M. Churcher and F. Hallböök and G. W. Humphrey and M. M. Idris and T. Kiyama and S. Liang and D. Mellott and X. Mu and G. Murray and R. P. Olinski and F. Raible and M. Rowe and J. S. Taylor and K. Tessmar-Raible and D. Wang and K. H. Wilson and S. Yaguchi and T. Gaasterland and B. E. Galindo and H. J. Gunaratne and C. Juliano and M. Kinukawa and G. W. Moy and A. T. Neill and M. Nomura and M. Raisch and A. Reade and M. M. Roux and J. L. Song and Y.-H. Su and I. K. Townley and E. Voronina and J. L. Wong and G. Amore and M. Branno and E. R. Brown and V. Cavalieri and V. Duboc and L. Duloquin and C. Flytzanis and C. Gache and F. Lapraz and T. Lepage and A. Locascio and P. Martinez and G. Matassi and V. Matranga and R. Range and F. Rizzo and E. Röttinger and W. Beane and C. Bradham and C. Byrum and T. Glenn and S. Hussain and G. Manning and E. Miranda and R. Thomason and K. Walton and A. Wikramanayke and S.-Y. Wu and R. Xu and C. T. Brown and L. Chen and R. F. Gray and P. Y. Lee and J. Nam and P. Oliveri and J. Smith and D. Muzny and S. Bell and J. Chacko and A. Cree and S. Curry and C. Davis and H. Dinh and S. Dugan-Rocha and

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

J. Fowler and R. Gill and C. Hamilton and J. Hernandez and S. Hines and J. Hume and L. Jackson and A. Jolivet and C. Kovar and S. Lee and L. Lewis and G. Miner and M. Morgan and L. V. Nazareth and G. Okwuonu and D. Parker and L.-L. Pu and R. Thorn and R. Wright. 2006. The genome of the sea urchin *Strongylocentrotus purpuratus*. Science **314**:941-952.

Song, L., C. Bian, Y. Luo, L. Wang, X. You, J. Li, Y. Qiu, X. Ma, Z. Zhu, and L. Ma. 2016. Draft genome of the Chinese mitten crab, *Eriocheir sinensis*. GigaScience **5**:1.

Srivastava, M., O. Simakov, J. Chapman, B. Fahey, M. E. A. Gauthier, T. Mitros, G. S. Richards, C. Conaco, M. Dacre, U. Hellsten, C. Larroux, N. H. Putnam, M. Stanke, M. Adamska, A. Darling, S. M. Degnan, T. H. Oakley, D. C. Plachetzki, Y. Zhai, M. Adamski, A. Calcino, S. F. Cummins, D. M. Goodstein, C. Harris, D. J. Jackson, S. P. Leys, S. Shu, B. J. Woodcroft, M. Vervoort, K. S. Kosik, G. Manning, B. M. Degnan, and D. S. Rokhsar. 2010. The *Amphimedon queenslandica* genome and the evolution of animal complexity. Nature **466**:720-726.

Stanke, M. and S. Waack. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics **19**:ii215-ii225.

Stimson, J. 1997. The annual cycle of density of zooxanthellae in the tissues of field and laboratory-held *Pocillopora damicornis* (Linnaeus). Journal of Experimental Marine Biology and Ecology **214**:35-48.

Tarailo-Graovac, M. and N. Chen. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics **25**:1-14.

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

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, S. L. Salzberg, B. J. Wold, and L. Pachter. 2010. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology **28**:511-515.

Traylor-Knowles, N., B. Granger, T. Lubinski, J. Parikh, S. Garamszegi, Y. Xia, J. Marto, L. Kaufman, and J. Finnerty. 2011. Production of a reference transcriptome and a transcriptomic database (PocilloporaBase) for the cauliflower coral, *Pocillopora damicornis*. BMC Genomics **12**:585.

Van Oppen, M. J. H., P. Souter, E. J. Howells, A. Heyward, and R. Berkelmans. 2012. Novel genetic diversity through somatic mutations: fuel for adaptation of reef corals? Diversity **3**:405-423.

Venter, J. C. and M. D. Adams and E. W. Myers and P. W. Li and R. J. Mural and G. G. Sutton and H. O. Smith and M. Yandell and C. A. Evans and R. A. Holt and J. D. Gocayne and P. Amanatides and R. M. Ballew and D. H. Huson and J. R. Wortman and Q. Zhang and C. D. Kodira and X. H. Zheng and L. Chen and M. Skupski and G. Subramanian and P. D. Thomas and J. Zhang and G. L. Gabor Miklos and C. Nelson and S. Broder and A. G. Clark and J. Nadeau and V. A. McKusick and N. Zinder and A. J. Levine and R. J. Roberts and M. Simon and C. Slayman and M. Hunkapiller and R. Bolanos and A. Delcher and I. Dew and D. Fasulo and M. Flanigan and L. Florea and A. Halpern and S. Hannenhalli and S. Kravitz and S. Levy and C. Mobarry and K. Reinert and K. Remington and J. Abu-Threideh and E. Beasley and K. Biddick and V. Bonazzi and R. Brandon and M. Cargill and I. Chandramouliswaran and R. Charlab and K. Chaturvedi and Z. Deng and V. D. Francesco and P. Dunn and K. Eilbeck and C. Evangelista and A. E. Gabrielian and W. Gan and W. Ge and F. Gong and Z. Gu and P. Guan and T. J. Heiman and M. E. Higgins and R.-R. Ji and Z. Ke and K. A. Ketchum and Z. Lai and Y. Lei and Z. Li and J. Li and Y. Liang and X. Lin and F. Lu and G. V. Merkulov and N. Milshina and H. M. Moore and A. K. Naik and V. A.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Narayan and B. Neelam and D. Nusskern and D. B. Rusch and S. Salzberg and W. Shao and B. Shue and J. Sun and Z. Y. Wang and A. Wang and X. Wang and J. Wang and M.-H. Wei and R. Wides and C. Xiao and C. Yan and A. Yao and J. Ye and M. Zhan and W. Zhang and H. Zhang and Q. Zhao and L. Zheng and F. Zhong and W. Zhong and S. C. Zhu and S. Zhao and D. Gilbert and S. Baumhueter and G. Spier and C. Carter and A. Cravchik and T. Woodage and F. Ali and H. An and A. Awe and D. Baldwin and H. Baden and M. Barnstead and I. Barrow and K. Beeson and D. Busam and A. Carver and A. Center and M. L. Cheng and L. Curry and S. Danaher and L. Davenport and R. Desilets and S. Dietz and K. Dodson and L. Doup and S. Ferriera and N. Garg and A. Gluecksmann and B. Hart and J. Haynes and C. Haynes and C. Heiner and S. Hladun and D. Hostin and J. Houck and T. Howland and C. Ibegwam and J. Johnson and F. Kalush and L. Kline and S. Koduru and A. Love and F. Mann and D. May and S. McCawley and T. McIntosh and I. McMullen and M. Moy and L. Moy and B. Murphy and K. Nelson and C. Pfannkoch and E. Pratts and V. Puri and H. Qureshi and M. Reardon and R. Rodriguez and Y.-H. Rogers and D. Romblad and B. Ruhfel and R. Scott and C. Sitter and M. Smallwood and E. Stewart and R. Strong and E. Suh and R. Thomas and N. N. Tint and S. Tse and C. Vech and G. Wang and J. Wetter and S. Williams and M. Williams and S. Windsor and E. Winn-Deen and K. Wolfe and J. Zaveri and K. Zaveri and J. F. Abril and R. Guigó and M. J. Campbell and K. V. Sjolander and B. Karlak and A. Kejariwal and H. Mi and B. Lazareva and T. Hatton and A. Narechania and K. Diemer and A. Muruganujan and N. Guo and S. Sato and V. Bafna and S. Istrail and R. Lippert and R. Schwartz and B. Walenz and S. Yooseph and D. Allen and A. Basu and J. Baxendale and L. Blick and M. Caminha and J. Carnes-Stine and P. Caulk and Y.-H. Chiang and M. Coyne and C. Dahlke and A. D. Mays and M. Dombroski and M. Donnelly and D. Ely and S. Esparham and C. Fosler and H. Gire and S. Glanowski and K. Glasser and A. Glodek and M. Gorokhov and K. Graham and B. Gropman and M. Harris and J. Heil and S. Henderson and J. Hoover and D. Jennings and C. Jordan and J. Jordan and J. Kasha and L. Kagan and C. Kraft and A. Levitsky and M. Lewis and X. Liu and J. Lopez and D. Ma and W. Majoros and J. McDaniel and S. Murphy and M. Newman and T. Nguyen and N. Nguyen and M. Nodell and S. Pan and J. Peck and M. Peterson and W. Rowe and R. Sanders and J. Scott and M. Simpson and T. Smith and A. Sprague and T. Stockwell and R. Turner and E. Venter and M. Wang and M. Wen and D. Wu and M. Wu and A. Xia and A. Zandieh and X. Zhu. 2001. The sequence of the human genome. Science **291**:1304-1351.

Veron, J. E. N. 2000. Corals of the World. Australian Institute of Marine Science, Townsville.

Veron, J. E. N. and M. Pichon. 1976. Scleractinia of eastern Australia, Part I. Families Thamnasteriidae, Astrocoeniidae, Pocilloporidae.

Vidal-Dupiol, J., M. Adjeroud, E. Roger, L. Foure, D. Duval, Y. Mone, C. Ferrier-Pages, E. Tambutte, S. Tambutte, D. Zoccola, D. Allemand, and G. Mitta. 2009. Coral bleaching under thermal stress: putative involvement of host/symbiont recognition mechanisms. BMC Physiology **9**:14.

Vidal-Dupiol, J., N. M. Dheilly, R. Rondon, C. Grunau, C. Cosseau, K. M. Smith, M. Freitag, M. Adjeroud, and G. Mitta. 2014. Thermal stress triggers broad Pocillopora damicornis transcriptomic remodeling, while Vibrio coralliilyticus infection induces a more targeted immuno-suppression response. PLoS ONE **9**:e107672.

Vidal-Dupiol, J., O. Ladrière, D. Destoumieux-Garzon, P.-E. Sautière, A. L. Meistertzheim, E. Tambutté, S. Tambutté, D. Duval, L. Fouré, M. Adjeroud, and G. Mitta. 2011a. Innate immune responses of a scleractinian coral to vibriosis. The Journal of Biological Chemistry **286**:22688-22698.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Vidal-Dupiol, J., O. Ladrière, A. L. Meistertzheim, L. Fouré, M. Adjeroud, and G. Mitta. 2011b. Physiological responses of the scleractinian coral *Pocillopora damicornis* to bacterial stress from *Vibrio coralliilyticus*. The Journal of Experimental Biology **214**:1533-1545.

Vidal-Dupiol, J., E. Toulza, O. Rey, D. Roquis, C. Chaparro, C. Cosseau, A. Picart-Picolo, P. Romans, M. Pratlong, K. Brener-Raffalli, P. Pontaroti, M. Adjeroud, G. Mitta, and C. Grunau. submitted. Genetic and epigenetic changes mediate rapid adaptation to global warming in a tropical coral

Vidal-Dupiol, J., D. Zoccola, E. Tambutté, C. Grunau, C. Cosseau, K. M. Smith, M. Freitag, N. M. Dheilly, D. Allemand, and S. Tambutté. 2013. Genes related to ion-transport and energy production are upregulated in response to CO_2-driven pH decrease in corals: New insights from transcriptome analysis. PLoS ONE **8**:e58652.

Voolstra, C. R., Y. Li, Y. J. Liew, S. Baumgarten, D. Zoccola, J.-F. Flot, S. Tambutté, D. Allemand, and M. Aranda. 2017. Comparative analysis of the genomes of Stylophora pistillata and Acropora digitifera provides evidence for extensive differences between species of corals. Scientific Reports **7**:17583-17583.

Wang, J.-T., Y.-Y. Chen, K. S. Tew, P.-J. Meng, and C. A. Chen. 2012. Physiological and biochemical performances of menthol-induced aposymbiotic corals. PLoS ONE **7**:e46406.

Ye, J., L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, and J. Wang. 2006. WEGO: a web tool for plotting GO annotations. Nucleic acids research **34**:W293-W297.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Data used for assembly

Library	Average insert size (measured; bp)	Paired-reads (million)	Yield (Mb)	Genome coverage*	Average Q	Q>30(%)
Shot-gun	178	238	47,517	146	36.45	93.71
LJD 3Kb	2,479	109	21,848	67	32.31	81.79
LJD 8Kb	7,825	44	8,880	27	31.98	80.56
LJD 20 Kb	32, 325	100	19,963	61	30.78	77.73

*For an estimated (kmer distribution) genome size of 325 Mb

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Assembly statistics; all statistics are based on sequence of size ≥ 500 bp, unless otherwise noted

Statistics	Contigs	Scaffolds
Number (seq ≥ 0 bp)	196,891	168,465
Number (seq ≥ 500 bp)	81,286	58,326
Number (seq ≥1000 bp)	47,580	25,553
Assembly length bp (seq ≥ 0bp)	336,691,489	352,019,984
Assembly length bp (seq ≥ 500bp)	309,335452	325,576,138
Assembly length bp (seq ≥ 1000bp)	285,593,703	302,677350
Largest	112,653	1,296,445
N50	11,125	171,375
NG50*	10,244	171,375
N75	3,847	16,625
NG75*	2,912	16,732
N90	1,120	696
NG90*	818	521
Number of N's per 100 kbp	1.69	4697.98
% Core eukaryotic genes (full lenght)	Not done	84.3
% Core eukaryotic genes (partial)	Not done	93.15
GC%	37.84%	

*based on the predicted genome size: 325 Mb

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: Structural and functional annotation summary results

	Experimental approach	*Ab initio* approach*
Number of predicted gene	36,140	64,558
Number of predicted transcript	63,181	79,506
Mean exon length (pb)	394	184
Mean intron length (pb)	675	522
Transcript with functional annotation	48,987	35,726
Matching to predicted or hypothetical protein	2,986	9,004
No results	11,508	34,776
With GO terms	32,986	17,754
With Enzyme code	7,147	4,389
With a conserved protein domain	33,002	40,240
Transcript in an orthologs group	47,402	45,839
Number of orthologs group	16,469	27,440
Ortholog group with a annotation	9,916	12,132
Transcript in an annotated orthologs group	24,024	25,335

*Include the sequences predicted by the experimental approach

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 4: *Pocillopora acuta*, repeat content

Element (super family)	Number of occurrence	Number of bp covered	Per cent of genome coverage
DNA transposon	27,081	3,708,267	1.05%
SINE retrotransposon	8,504	11,051,79	0.31%
LINE retrotransposon	45,381	4,352,179	1.24%
LTR retrotransposon	3,457	838,736	0.24%
Unclassified	364,626	40,009,286	11.37%
Small RNA	2,751	198,347	0.06%
Satellites	463	105,608	0.03%
Simple repeats	62,511	3,182,003	0.90%
Low complexity	11,794	576,390	0.16%

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure legends

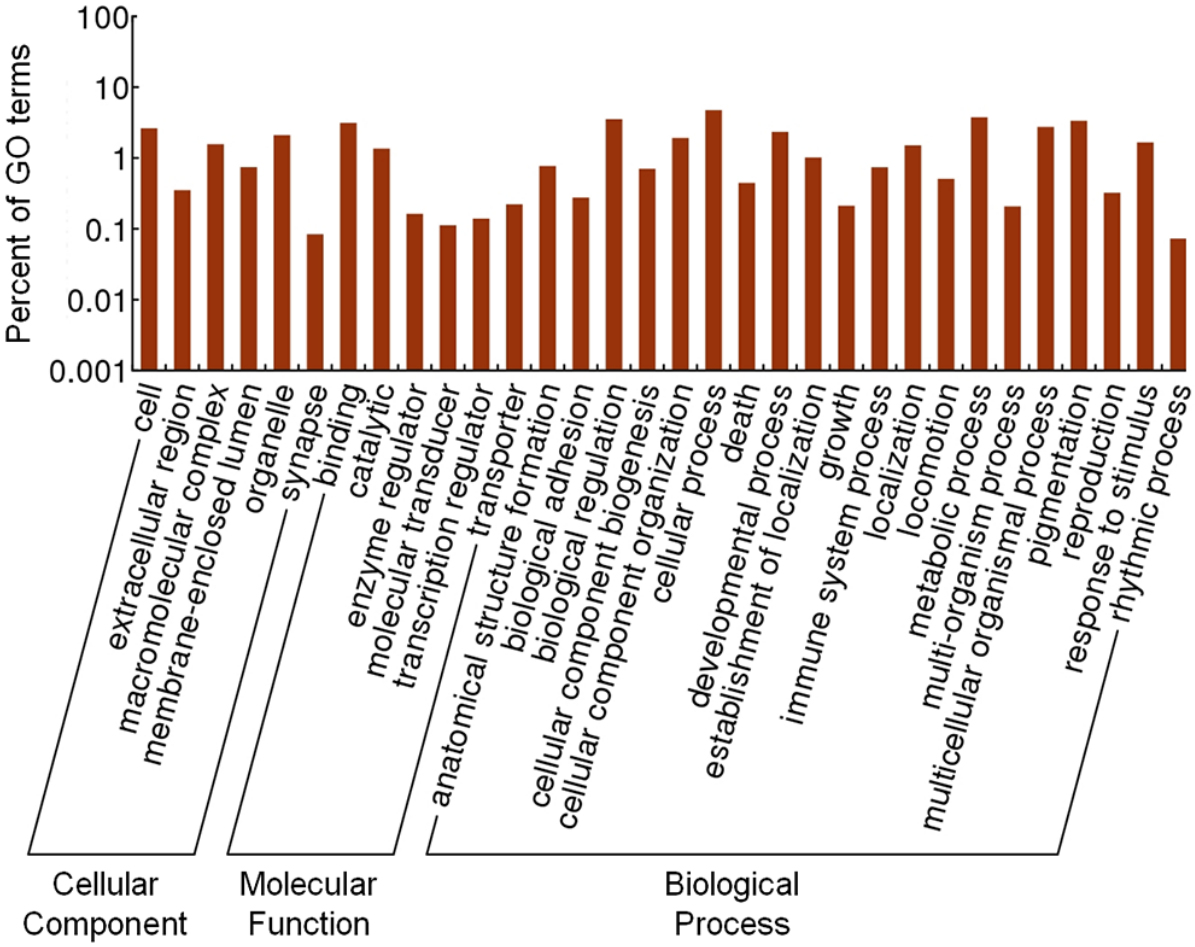

Figure 1: Distribution of GO terms belonging to the level 2 of the GO arborescence in the three main GO categories: Cellular Component, Molecular Function and Biological Process. Transcripts from the experimental annotation.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

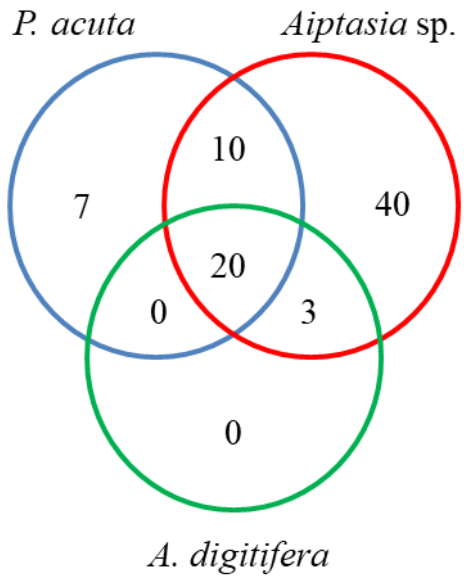

Figure 2: Venn diagram representing the number of transposable elements specific or shared between the three sequenced symbiotic anthozoa, the scleractinian corals *Acropora digitifera* and *Pocillopora acuta* and the anemonia *Aiptasia* sp. Transposable elements for *A. digitifera* and *Aiptasia* sp. were obtained from [8, 9].

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3: The cnidarian core proteome. This diagram represent the comparison between the protein content encoded in the genome of *P. acuta* and those of three other cnidarians: another scleractinia *A digitifera*; an actinia *N. vectensis*; and a hydrozoa *H. magnipapillata*. OrthoMCL was used to identify orthologous sequences between the predicted ORF (longer than 100AA) translated from the RNA-seq annotation and those annotated in these other genomes. The values in the diagram represent the number of group of orthologs created by orthoMCL (a group contains at least 2 sequences where at least one belongs to *P. acuta*). The value corresponding to the proteome restricted to *P. acuta* is composed by the number of ortholog groups containing *P. acuta* sequences only, plus the number of orphan sequences (sequences that are not included in a group). In total, this analysis include the 63,181 (47,402 are included in a group of ortholog, 15,779 are orphans) translated protein sequences annotated by RNA-seq in *P. acuta*, and 11,369, 10,636 and 7,278 sequences for *A digitifera*, *N. vectensis* and *H. magnipapillata*, respectively. NA signify not applicable: these values were not determined by this approach since the clustering was done to classify and annotate the proteome of *P. acuta*. Therefore, the proteins belonging to the other species and that do not has an ortholog in the genome of *P. acuta* could not be included in a ortholog group and so could not be taken into account by the analysis.

bioRxiv preprint doi: https://doi.org/10.1101/698688; this version posted July 11, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

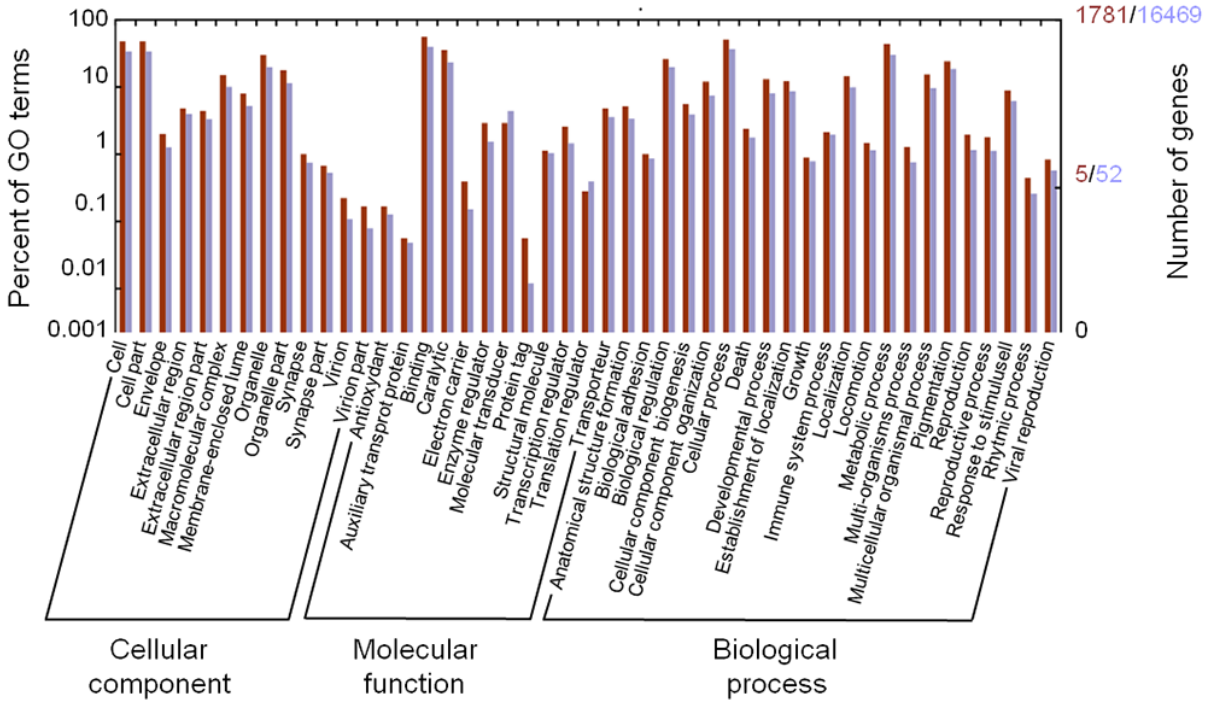

Figure 4: Distribution and comparison of GO terms included in the cnidarian core proteome (red) and the *P. acuta* proteome (blue). This GO terms belong to the level 2 of the GO arborescence and represent the three main GO categories: Cellular Component, Molecular Function and Biological Process. Transcripts from the experimental annotation.