Received: 3 October 2019 | Revised: 1 February 2020 | Accepted: 10 February 2020

DOI: 10.1111/1755-0998.13150

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

Transcriptome-based target-enrichment baits for stony corals (Cnidaria: Anthozoa: Scleractinia)

Randolph Z. B. Quek[1] | Sudhanshi S. Jain[1] | Mei Lin Neo[1,2] | Greg W. Rouse[3] | Danwei Huang[1,2]

[1]Department of Biological Sciences, National University of Singapore, Singapore, Singapore

[2]Tropical Marine Science Institute, National University of Singapore, Singapore, Singapore

[3]Scripps Institution of Oceanography, University of California San Diego, San Diego, CA, USA

Correspondence
Randolph Z. B. Quek and Danwei Huang, Department of Biological Sciences, National University of Singapore, Singapore, Singapore.
Emails: randolphquek@u.nus.edu; huangdanwei@nus.edu.sg

Funding information
National Research Foundation Singapore, Grant/Award Number: MSRDP-P03

Abstract

Despite the ecological and economic significance of stony corals (Scleractinia), a robust understanding of their phylogeny remains elusive due to patchy taxonomic and genetic sampling, as well as the limited availability of informative markers. To increase the number of genetic loci available for phylogenomic analyses in Scleractinia, we designed 15,919 DNA enrichment baits targeting 605 orthogroups (mean 565 ± *SD* 366 bp) over 1,139 exon regions. A further 236 and 62 barcoding baits were designed for COI and histone H3 genes respectively for quality and contamination checks. Hybrid capture using these baits was performed on 18 coral species spanning the presently understood scleractinian phylogeny, with two corallimorpharians as outgroup. On average, 74% of all loci targeted were successfully captured for each species. Barcoding baits were matched unambiguously to their respective samples and revealed low levels of cross-contamination in accordance with expectation. We put the data through a series of stringent filtering steps to ensure only scleractinian and phylogenetically informative loci were retained, and the final probe set comprised 13,479 baits, targeting 452 loci (mean 531 ± *SD* 307 bp) across 865 exon regions. Maximum likelihood, Bayesian and species tree analyses recovered maximally supported, topologically congruent trees consistent with previous phylogenomic reconstructions. The phylogenomic method presented here allows for consistent capture of orthologous loci among divergent coral taxa, facilitating the pooling of data from different studies and increasing the phylogenetic sampling of scleractinians in the future.

KEYWORDS

coral reef, exon, genome sampling, hybrid capture, multilocus data, phylogenomics

1 | INTRODUCTION

Stony corals (Cnidaria: Anthozoa: Scleractinia), numbering over 1,500 extant species, are of great ecological and economic importance, and can be found from shallow waters to great depths across the world's oceans (Cairns, 2007; Huang, 2012; Kitahara, Fukami, Benzoni, & Huang, 2016; Moberg & Folke, 1999). Despite centuries of research into the systematics of Scleractinia (Fukami et al., 2008; Kitahara et al., 2016; Lamarck, 1816; Linnaeus, 1758; Romano & Palumbi, 1996), and even with recent molecular phylogenetic work (Arrigoni, Berumen, Huang, Terraneo, & Benzoni, 2017; Arrigoni, Berumen, et al., 2014; Arrigoni, Terraneo, Galli, & Benzoni, 2014;

Huang et al., 2016; Huang, Benzoni, Arrigoni, et al., 2014; Huang, Benzoni, Fukami, et al., 2014; Kitano et al., 2014), the classification and phylogeny of this group remain largely unresolved particularly at the species level (Kitahara et al., 2016). The ideal of a well-supported and robust phylogeny of corals remains elusive due to several reasons, including inadequate species sampling and the paucity of informative morphological and molecular markers (Budd, Romano, Smith, & Barbeitos, 2010; Kitahara et al., 2016; but see Quattrini et al., 2018).

Since the advent of DNA sequencing, molecular data have been imperative for phylogenetic reconstructions across the tree of life (Laumer et al., 2019; Pizarro et al., 2018; Wickett et al., 2014), and scleractinian corals are no exception. Molecular phylogenies for Scleractinia were first reconstructed in the mid-1990s, based on single-gene comparisons of either nuclear or mitochondrial ribosomal genes (Chen, Odorico, Tenlohuis, Veron, & Miller, 1995; Romano & Palumbi, 1996, 1997). Since then, multigene analyses with mitochondrial and nuclear markers have been performed (Arrigoni, Berumen, et al., 2014; Arrigoni, Terraneo, et al., 2014; Fukami et al., 2004, 2008; Huang et al., 2016; Huang, Benzoni, Arrigoni, et al., 2014; Huang, Benzoni, Fukami, et al., 2014). Data sets with up to 12 loci are now available, but these have patchy gene coverage among species (Hartmann, Baird, Knowlton, & Huang, 2017; Huang, 2012; Huang & Roy, 2015; Kitahara et al., 2016). Phylogenetic matrices with such limited number of genes are prone to producing uncertain branch length or age estimates and even erroneous species tree topologies (Rokas, Williams, King, & Carroll, 2003; Zhu, Dos Reis, & Yang, 2015).

Today, high-throughput, next-generation sequencing (NGS) platforms have enabled vast amounts of data to be harnessed at relatively low cost (Goodwin, McPherson, & McCombie, 2016; Kulkarni & Frommolt, 2017). Whole genome data, while ideal, remain a distant possibility for Anthozoa due to challenges in whole genome assembly (Sohn & Nam, 2018), especially because of diverse microbial symbionts contaminating the sequencing reads (Artamonova & Mushegian, 2013). Indeed, recent anthozoan whole genome analyses comprise fewer than 15 taxa (Cunning, Bay, Gillette, Baker, & Traylor-Knowles, 2018; Ying et al., 2018). Therefore, alternative NGS methods tailored for phylogenomics are applied more widely among Anthozoa, and include restriction site-associated DNA sequencing (RADseq) and genome skimming for shallow relationships (Forsman et al., 2017; Johnston et al., 2017), as well as phylotranscriptomics (Lin et al., 2016; Quek & Huang, 2019; Richards, Carvajal, Wallace, & Wilson, 2019; Zapata et al., 2015) and target enrichment via hybrid capture (Quattrini et al., 2018) for more inclusive taxon sets.

Capitalizing on recent technological advancements, the hybrid capture approach has been steadily gaining traction over the last decade (Bossert & Danforth, 2018; Bragg, Potter, Bi, & Moritz, 2016; Faircloth et al., 2012; Lemmon, Emme, & Lemmon, 2012). Principally, DNA or RNA probes are designed based on conserved sequences within the genome and used to hybridize to these targeted loci in a DNA library that are subsequently enriched and sequenced. This method has the potential to target hundreds to thousands of homologous loci simultaneously in a cost-effective fashion by pooling a number of samples together in a single hybridization reaction (e.g., n = 96; Liu et al., 2019). Target enrichment has been applied successfully in several marine taxa, including percomorphs (Dornburg et al., 2017), ophiuroids (Hugall, O'Hara, Hunjan, Nilsen, & Moussalli, 2016), mobulids (White et al., 2018) and anthozoans (Quattrini et al., 2018). In particular, the baits designed by Quattrini et al. (2018) were based on both ultraconserved elements and transcriptomes, and were tested in vitro on 33 anthozoan taxa, including just four scleractinians. A further nine genome-enabled taxa were included for phylogenetic analyses which included two scleractinians.

In this study, we designed, screened and tested target-enrichment baits based on 44 scleractinian transcriptomes for the broad purpose of phylogenetic reconstruction among stony corals. Baits designed in this study were tested on 18 species across both the "Robust" and "Complex" clades (sensu Romano & Palumbi, 1996), as well as two corallimorpharian outgroups. With judicious selection of putatively scleractinian loci post-capture, we demonstrate that our baits are able to capture orthologous markers across the scleractinian tree for large-scale phylogenomic analysis.

2 | MATERIALS AND METHODS

2.1 | Bait design and screening

Scleractinian transcriptomes for 43 terminals from 39 species analyzed in Quek and Huang (2019) were translated into their amino acid sequences in Qiagen CLC Genomics Workbench v9.5.4. Clustering of orthologs was performed in ORTHOFINDER v1.1.8 (Emms & Kelly, 2015) under default settings with the *diamond_more_sensitive* flag activated (Buchfink, Xie, & Huson, 2015). Single-copy orthologs were extracted for bait design.

To ensure an even phylogenetic representation for each locus, all single-copy orthologs selected must be represented by a minimum of six (out of 39) scleractinian taxa, of which two must belong to either the "Robust" or "Complex" clade (sensu Romano & Palumbi, 1996). A number of filtering steps were incorporated to retain only putatively scleractinian loci. A BLASTP (*e*-value = 10^{-6}) of all identified sequences was conducted against the gene models of 10 publicly available coral genomes (Table 1). Only orthologs with at least one transcript that had a positive hit to a gene model with bit-score ≥50 were kept. To filter out potential non-scleractinian transcripts originating from the coral holobiont, we conducted a local BLASTN (*e*-value = 10^{-6}) of aforementioned positive hits against GenBank data (downloaded October 2018). A single best hit for each transcript was identified via sorting BLASTN hits by highest bit-score, lowest e-value and highest percentage identity. Transcripts matched to a non-cnidarian with ≥80% sequence similarity across an alignment length ≥100 bp were removed. If the removal of transcript(s) resulted in an ortholog being represented by less than six taxa, the ortholog was removed altogether.

TABLE 1 Reference genomes used for putatively scleractinian transcript identification and bait design

Species	Access	Reference
Scleractinia		
Acropora digitifera	http://marinegenomics.oist.jp/	Shinzato et al. (2011)
Acropora tenuis	http://refuge2020.reefgenomics.org/	ReFuGe 2020 Consortium (2015)[a]
Fungia sp.	http://refuge2020.reefgenomics.org/	Ying et al. (2018)[a]
Galaxea fascicularis	http://refuge2020.reefgenomics.org/	Ying et al. (2018)[a]
Coelastrea aspera	http://refuge2020.reefgenomics.org/	Ying et al. (2018)[a]
Montastraea cavernosa	http://matzlab.weebly.com/data--code.html	
Orbicella faveolata	GCF_002042975.1	
Pocillopora damicornis	http://pdam.reefgenomics.org/	Cunning et al. (2018)[a]
Porites lutea	http://refuge2020.reefgenomics.org/	Ying et al. (2018)[a]
Stylophora pistillata	http://spis.reefgenomics.org/	Voolstra et al. (2017)[a]
Symbiodiniaceae		
Breviolum minutum	http//marinegenomics.oist.jp/	Shoguchi et al. (2013)
Cladocopium goreaui	http://symbs.reefgenomics.org/	Liu et al. (2018)[a]
Cladocopium sp.	http://marinegenomics.oist.jp/	Shoguchi et al. (2018)
Fugacium kawagutii	http://symbs.reefgenomics.org/	Liu et al. (2018)[a]
Symbiodinium microadriaticum	http://smic.reefgenomics.org/	Aranda et al. (2016)[a]
Symbiodinium sp.	https://marinegenomics.oist.jp/	Shoguchi et al. (2018)

[a]Reefgenomics.org: Liew, Aranda, and Voolstra (2016).

For the remaining transcripts, we conducted another BLASTP (*e*-value = 10^{-6}) against the same gene models as above. Positive hits were then sorted by highest bit-score, lowest e-value and highest percentage identity, extracting the best hit for each ortholog that was then placed in one of 10 bins, each representing one of the 10 reference genomes (Table 1). Orthologous amino acid sequences were first aligned using MAFFT v7.3.10 with the L-INS-i method (Katoh & Standley, 2013), and then back-translated using PAL2NAL v14 (Suyama, Torrents, & Bork, 2006) into their corresponding nucleotide sequences.

Baits were designed using BAITFISHER v1.2.8 (Mayer et al., 2016) with alignment cutting performed separately for each bin based on the respective reference genome. To maximize taxon and loci recoverability (Schott et al., 2017), we designed multiple 120 bp baits across different lengths of loci located by BAITFISHER in the genome following Bank et al. (2017). We first required that seven baits were designed across a 240 bp window with a 20 bp offset between each bait. Shorter windows were used iteratively following failure to locate a suitable region after alignment cutting: five tiled baits in a 200 bp window; three tiled baits in a 160 bp window; and finally, one bait for a 120 bp window.

To generate a single bait set that contained the best bait locus following alignment cutting, we used BAITFILTER v1.0.6 (Mayer et al., 2016) (*-m fs*) for each of the 10 bait sets designed. As a final check for baits possibly binding to noncoral DNA, we mapped the extracted baits at 70% sequence similarity and 70% length using CLC Genomics Workbench v9.5.4 and searched by blastn (*e*-value = 10^{-4}) against six Symbiodiniaceae genomes (Table 1). All mapped or blastn-matched baits were removed. If there were ≥3 baits removed within a bait region, or only one remaining bait, we removed the entire bait region. Any duplicated baits were removed using SEQKIT v0.7.2 *RMDUP* package (Shen, Le, Li, & Hu, 2016). Finally, we removed baits that had the potential to self-hybridize by running a BLASTN of the baits against one another, searching for baits with regions that were reverse complementary to other baits.

To help flag cross-contamination post-capture and potential sample misidentification, we designed barcoding baits targeting mitochondrial cytochrome c oxidase subunit I (COI) and nuclear histone H3. These genes are suitable for identification to genus (Arrigoni, Berumen, et al., 2014; Huang, Benzoni, Arrigoni, et al., 2014; Huang, Meier, Todd, & Chou, 2008) and can be used to confirm the identities of potentially misidentified or erroneously labelled samples for a broadscale phylogeny. We first downloaded 129 COI and 32 histone H3 sequences from GenBank and obtained 16 additional histone H3 sequences from samples collected in Singapore via Sanger sequencing following Huang, Licuanan, Baird, and Fukami (2011) (Table S1). Taxon coverage spanned the scleractinian phylogeny. BAITFISHER v1.2.7 (Mayer et al., 2016) was used to design 120 bp baits with an offset of 20 bp across the entire length of each gene. Optimal baits were identified using BAITFILTER v1.0.6 (*-m as*) and added to the final bait set, labelled as either "coi" or "h3" and can be removed at the user's discretion.

The full set of biotinylated RNA probes were manufactured by Arbor Biosciences (myBaits Custom Target Capture Kit, USA).

2.2 | Sample collection, target enrichment and sequencing

A total of 18 coral samples, each belonging to a genus analyzed in Quek and Huang (2019), as well as two corallimorpharians, were collected from Singapore reefs (Table S2). Fragments of corals were stored in either 100% ethanol or RNAlater (Invitrogen) until DNA extraction. The remaining skeletal vouchers were treated with a powerful waterjet to remove coral tissue, bleached in dilute sodium hypochlorite, then washed and dried. Voucher specimens were deposited at the Lee Kong Chian Natural History Museum (Table S2).

High quality gDNA was extracted following a modified protocol suggested by Forsman et al. (2017). Briefly, coral samples were crushed and DNA extraction was carried out using E.Z.N.A Mollusc DNA Kit (Omega Bio-tek) with a modified two-elution step. For the first elution, only 35 ul of elution buffer (0.10 nM Tris-HCl) was added to the column, removing a majority of small molecular-weight, fragmented DNA. In the second step, 50 µl of elution buffer was added four times, resulting in a total gDNA volume of 200 µl. High quality, eluted gDNA (200 µl) from the E.Z.N.A. Mollusc DNA Kit extraction was then purified using Zymo Genomic DNA Clean and Concentrator. Four rounds of DNA elution were conducted post-cleanup, with 15 ul of elution buffer added per round to a final volume of 60 µl.

Libraries were prepared by first sonicating purified gDNA using Bioruptor Pico (Diagenode) with a target mode size of 200 bp. Adapters were ligated with KAPA dual-indexed adapters for Illumina platforms (KK8722; KAPA Biosystems) using the KAPA HyperPrep Kit (KK8502; KAPA Biosystems), according to manufacturer's recommendations. An additional double-sided size selection was carried out in the final step to narrow fragment size distribution in final libraries according to KAPA Biosystems protocol using Agencourt AMPure XP beads (Beckman Coulter). Libraries were eluted in 20 µl of Tris-HCl buffer and 2 µl was used for quantification using a Qubit 3 Fluorometer. Five libraries were pooled randomly in equimolar ratios (~100 ng per sample) to a total of 500 ng of DNA per pool, and concentrated to 7 µl per pool (71.42 ng/µl) for hybrid capture.

Hybrid capture was executed following the manufacturer's protocol (myBaits Custom Target Capture Kit; Arbor Biosciences) with 14 cycles of post-capture amplification. Amplified libraries were purified using Agencourt AMPure XP beads (Beckman Coulter) and eluted in 20 µl of nuclease-free water. DNA concentration was quantified using a Qubit 3 Fluorometer and the 20 libraries were pooled in equimolar concentrations. Libraries were sequenced on a single lane of Illumina HiSeq 4000 (150 × 150 bp).

2.3 | Sequence assembly and quality filtering

Raw reads were demultiplexed and processed by trimming low quality bases and adapters using TRIMMOMATIC v0.38 (Bolger, Lohse, & Usadel, 2014) under default settings. In order to identify and assemble orthologous loci targeted by the baits, processed paired reads were parsed into HYBPIPER v1.2 (Johnson et al., 2016) to locate targeted exons. Reads were first mapped to the transcripts that were used for bait design (henceforth referred as 'baitfile') using BWA v0.7.17 (Li, 2013) under default HYBPIPER settings. Mapped reads were assembled into contigs with SPADES v3.12.0 (Bankevich et al., 2012) and exonic regions were identified using EXONERATE v2.2.0 (http://github.com/nathanweeks/exonerate).

Verification of cross-contamination was first conducted by running the following barcoding check. Trimmed reads were first mapped to COI and histone H3 sequences from 12 samples—six "Robust" and five "Complex" corals spanning eight families and one corallimorpharian *Rhodactis indosinensis* (Table 2)—using BWA MEM under default settings (Li, 2013). Mapped reads were extracted in FASTQ format using SAMTOOLS (Li et al., 2009) and assembled using SPADES v3.12.0 under default settings with the *--careful* flag activated (Bankevich et al., 2012). Assembled contigs with a minimum length of 200 bp were searched by BLASTN against the assembled Sanger sequences. Positive hits were then filtered for the single best hit with the highest sequence similarity (≥98%) over ≥200 bp to check sample identity.

Having verified that there were low levels of cross-contamination between samples (see 3.1 Bait design and screening), we reconstructed gene trees for loci located by HYBPIPER using FASTTREE v2.1.9 (Price, Dehal, & Arkin, 2010). Gene trees were based on coding sequences and separated into two sets, either potentially paralogous loci or single-copy loci as identified by HYBPIPER. The visualization of gene trees served several functions: (a) to remove potentially paralogous sequences; (b) to check for contamination, and (c) to check for locus capture efficiency. We treated paralogous loci following Johnson et al. (2016) and kept the ".main" paralog; or the "0.0" paralog if no ".main" was present for Type I paralogs (recent duplicates or alleles). For genes indicative of type II paralogy (deep divergences or early gene/genome duplications), we conservatively removed the loci from further downstream analyses.

Contamination was checked first by visually inspecting gene trees and using the following criteria to remove contaminant sequences: (a) if a paralogous locus had two sequences with one in the expected major clade and the other not (e.g., a "Complex" coral with a paralogous locus having one sequence in the "Complex" clade and the other in the "Robust" clade), the contaminant sequence was removed; (b) if a taxon for a single-copy locus was placed in the wrong major clade (e.g., "Robust" coral in the "Complex" clade), the taxon was removed, and (c) if a taxon/clade exhibited an unusually long branch within a tree, the taxon/clade was removed from the locus. Finally, based on the gene trees, we removed loci which captured less than three scleractinian taxa, gave spurious topologies, or were phylogenetically uninformative (i.e., poor or no phylogenetic signal due to highly conserved gene sequences).

For all loci and barcodes retained, read coverage was determined with the following pipeline. Trimmed reads were first mapped to assembled contigs (n = 18 samples) and barcode(s) (n = 12 samples for COI; n = 8 samples for histone H3) using BWA MEM under default settings (Li, 2013). Mapped reads were extracted using SAMTOOLS in BAM format (Li et al., 2009) and funneled into QUALIMAP v2.2.1 (Okonechnikov, Conesa & García-Alcalde, 2016) to compute the mean coverage per locus.

Finally, to determine if mitochondrial contigs could be recovered at mitochondrial loci other than COI, thereby allowing for the safe removal of barcoding baits, we conducted a similar analysis to that of Quek, Chang, Ip, and Huang (2019). Briefly, trimmed reads for all scleractinian samples (n = 18) were assembled by SPADES v3.1.2 under default settings. Mitochondrial contigs assembled were then identified by executing a blastn of all contigs to either the mitogenome or mitochondrial genes of a closely-related taxon; all contigs

TABLE 2 Summary statistics of loci assembled per sample for both exons-only and exons + supercontigs data sets

Species	Number of loci (#/%)	Locus length range (bp) (exons-only/supercontigs + exons)	Mean locus length (±*SD* bp) (exons-only/supercontigs + exons)
Robust corals			
Cyphastrea serailia[a]	386/85.40	93–1,923/156–4,356	443 ± 223/847 ± 597
Diploastrea heliopora	401/88.71	87–1,920/117–4,414	444 ± 228/802 ± 555
Dipsastraea maxima[a]	381/84.29	93–1,611/126–4,694	432 ± 194/853 ± 594
Goniastrea retiformis	393/86.95	93–1,395/99–4,479	439 ± 199/860 ± 608
Herpolitha limax[a]	349/77.21	93–1,209/126–3,690	425 ± 194/834 ± 560
Lobophyllia radians	389/86.06	93–1,572/117–3,619	435 ± 203/794 ± 520
Oulastrea crispata[a]	323/71.46	81–1,878/135–4,247	425 ± 236/798 ± 569
Platygyra sinensis[a]	383/84.73	90–1,674/192–3,950	443 ± 215/834 ± 583
Plesiastrea versipora[a]	354/78.32	93–1,383/147–5,192	437 ± 209/823 ± 564
Pocillopora acuta	258/57.08	123–1,587/153–3,531	473 ± 234/833 ± 523
Complex corals			
Acropora aspera	263/58.19	111–1,938/144–5,286	490 ± 238/1,005 ± 714
Astreopora expansa[a]	219/48.45	72–1,413/189–2,818	424 ± 209/774 ± 507
Fimbriaphyllia ancora[a]	343/75.88	90–2,895/105–6,837	513 ± 322/998 ± 765
Galaxea astreata[a]	343/75.88	99–2,697/105–8,843	521 ± 310/1,053 ± 871
Goniopora lobata	277/61.28	54–1,521/114–5,123	430 ± 214/693 ± 498
Pachyseris speciosa[a]	325/71.90	66–2,889/147–4,947	475 ± 288/899 ± 659
Porites lobata[a]	311/68.81	54–2,208/63–6,595	461 ± 259/875 ± 676
Turbinaria mesenterina	329/72.79	96–3,798/111–5,537	463 ± 312/934 ± 693
Corallimorpharia			
Rhodactis inchoata	43/9.51	183–717/183–1,773	375 ± 144/541 ± 352
Rhodactis indosinensis[a]	47/10.40	138–759/138–1,291	356 ± 123/448 ± 226

Note: Percentage of loci is based on total number of loci (*n* = 452). A supercontig includes both exon and intron regions in a sequence.
[a]Samples used in identity checks with COI and histone H3 barcodes.

with sequence similarity of 90% with an overlap of 200 bp were extracted. Contigs were then assembled using CAP3 (Huang & Madan, 1999) and annotated using MITOS2 (Bernt et al., 2013).

2.4 | Phylogenetic inference

Two separate data matrices were prepared for phylogenetic analysis. In the first data set, we combined only filtered coding sequences for both paralogy-filtered loci and single-copy loci (exons-only data set). In the second, we included filtered sequences with both introns and exons (supercontigs in HYBPIPER) for nonparalogous sequences and combined it with the paralogy-filtered coding sequences (exons + supercontigs data set). Sequences were first aligned for each locus in MAFFT v7.427 with the L-INS-i method (Katoh & Standley, 2013), and poorly aligned regions were trimmed using TRIMAL v1.4 under the heuristic setting (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009). Trimmed alignments were concatenated into a single matrix and partitioned by loci (*n* = 452) for the respective data sets.

For each data set, the maximum likelihood (ML) phylogeny was reconstructed using RAXML v8.2.11 (Stamatakis, 2014) under the rapid hill climbing mode and GTRGAMMA substitution model (100 random tree searches and 500 bootstrap pseudoreplicates). Bayesian analysis was performed using EXABAYES v1.5 (Aberer, Kobert, & Stamatakis, 2014), generating four coupled Markov chain Monte Carlo chains in four independent runs, each with 3 million generations and sampling every 500 generations. Convergence was checked based on average standard deviation of split frequencies (ASDSF < 0.001%). A consensus tree across all four runs was generated after the first 25% of generations had been discarded as burnin.

For the exons-only data set, we further reconstructed a ML phylogeny for each loci with at least four taxa (*n* = 438) using RAXML v8.2.11 (Stamatakis, 2014) under the rapid hill climbing mode and GTRGAMMA substitution model (10 random tree searches and 100 bootstrap pseudoreplicates). The gene trees were input for species tree analysis using ASTRAL-III v5.6.3 (Zhang, Rabiee, Sayyari, & Mirarab, 2018). Low support branches (<10% bootstrap support) were removed from all gene trees using NEWICK UTILITIES (Junier & Zdobnov, 2010) as per developers' recommendation. Finally, gene tree incongruence relative to the species tree (Figure 1) was assessed with DISCOVISTA (Sayyari, Whitfield, & Mirarab, 2018) based on family-level splits (bootstrap support ≥75 as recommended).

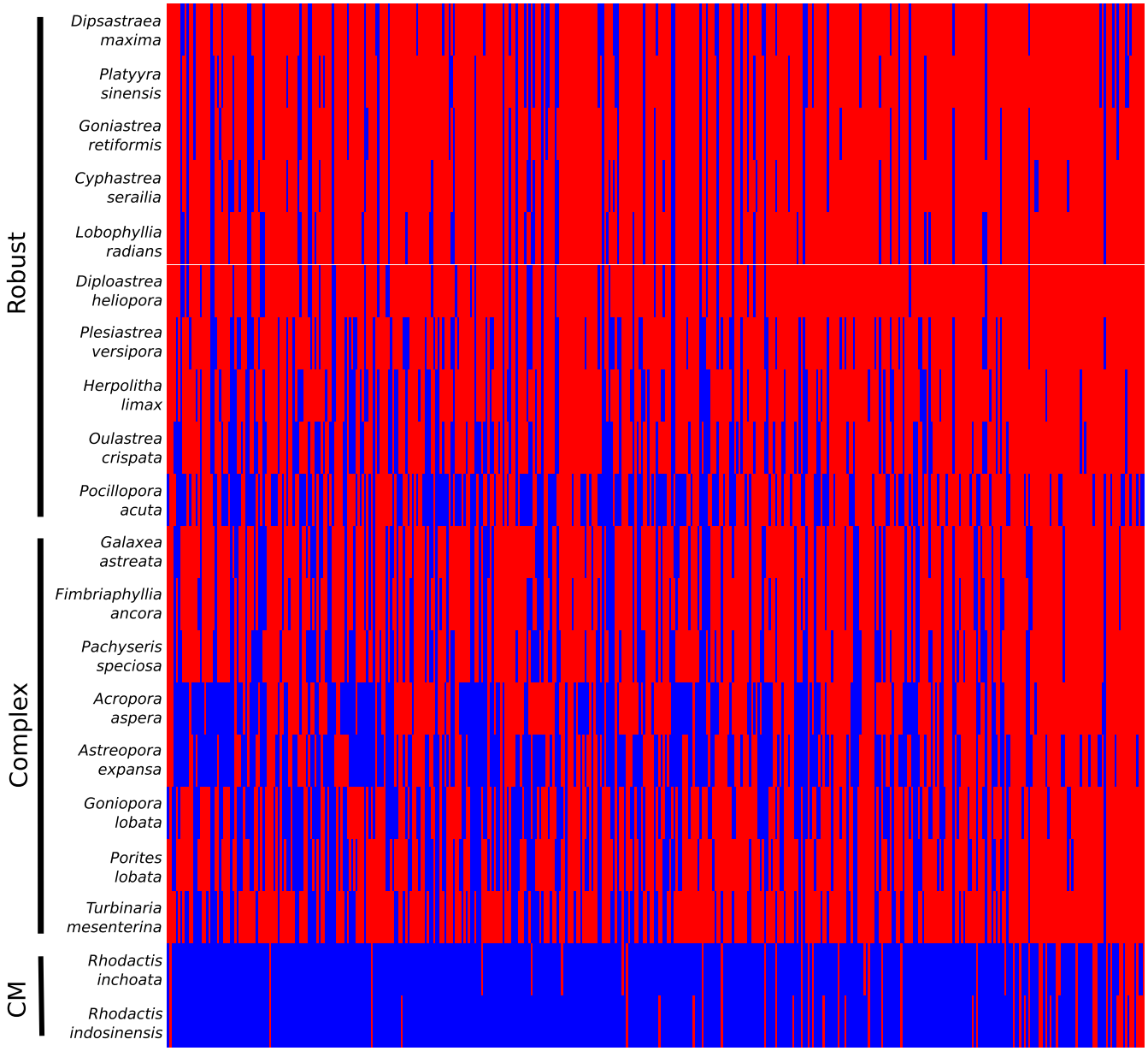

FIGURE 1 Coverage of loci captured by target-enrichment baits as determined by HYBPIPER post-filtering (blue = absent, red = present) [Colour figure can be viewed at wileyonlinelibrary.com]

3 | RESULTS

3.1 | Bait design and screening

From an initial 3,567 potential loci identified by ORTHOFINDER v1.1.8 (Emms & Kelly, 2015), our blastp returned 853 positive hits (bit-score ≥50). Following the BLASTN filter, we retained a total of 842 alignments for bait design. With alignment to genomes and bait tiling, a total of 665 loci had baits designed over 1,267 exon regions. After the final mapping to Symbiodiniaceae, removal of self-hybridizing baits and duplicates, we retained a total of 15,919 baits for 605 orthogroups over 1,139 exon regions. A further 236 and 62 baits were designed for COI and histone H3 genes respectively. The reference genomes used with the respective number of alignments, features and baits designed are detailed in Table S3.

All raw sequencing reads are available as NCBI Sequence Read Archive under BioProject accession number PRJNA602211. Following demultiplexing, the total number of reads per sample was between 6,220,293 and 16,119,962 (mean 10,027,194 ± *SD* 2,511,802; Table S4). The total number of trimmed reads remaining that were mapped to the baitfile in HYBPIPER was between 4,941,626 and 14,368,535 per sample (mean 8,617,176 ± *SD* 2,216,674; Table S4). The proportion of reads that mapped to sequenced barcodes ranged from 18.45% to 51.06% (mean 32.30% ± *SD* 10.44%; Table S4). In accordance with expectations of cross-sample contamination (Bank et al., 2017; Hugall et al., 2016), the assembly of barcoding baits revealed that a sample may have several sequences passing the filtering criteria with the best BLASTN hit to a nontarget taxon. However, these matches could be recognized and disregarded as the k-mer coverage was at least 100 times lower than that

TABLE 3 Concatenated matrix statistics for both exons-only and exons + supercontigs data sets

Data set	Missing data (%)	Concatenated matrix length (bp)	Mean locus length (±*SD* bp)	Locus length range (bp)	Parsimony informative sites (#/%)
Exons-only	30.86	201,137	456 ± 233	6–2133	68,997/34.30
Exons + supercontigs	32.43	287,749	636 ± 351	54–2491	119,095/41.39

Note: Missing data percentages as defined in Quek and Huang (2019).
A supercontig includes both exon and intron regions in a sequence.

of the correct hit, so we could recover the correct samples based on one or both barcoding genes. In total, we recovered accurate COI barcodes for all 11 tested scleractinians and a corallimorpharian, as well as histone H3 barcodes for eight scleractinians and a corallimorpharian. A summary of the barcoding results is available at Zenodo (http://dx.doi.org/10.5281/zenodo.3590246; barcoding_baits_summary.csv).

Read coverage was high across all loci captured, ranging from a mean coverage read depth of 749 (±*SD* 4,826) in *Goniopora lobata* to 3,424 (±*SD* 3,223) in *Acropora aspera*, with an overall mean coverage of 2,201 (±*SD* 676) across all samples (n = 18). This coverage was ~150× lower than that of the barcoding baits, with a mean read depth of 366,183 (±*SD* 194,946) for COI (n = 11) and 323,547 (±*SD* 261,479) for histone H3 (n = 8). We were able to recover a majority of mitochondrial genes from each sample (Figure S2), with the exception of species without suitably close relatives represented among GenBank's pool of mitogenomes (*Diploastrea heliopora*, *Herpolitha limax*, *Lobophyllia radians* and *Oulastrea crispata*).

Following stringent quality filtering of loci to remove paralogs, contaminant sequences and uninformative loci, we obtained a final bait set targeting 452 putatively scleractinian loci (mean 531 ± *SD* 307 bp) over 865 exon regions. The length of contigs recovered ranged from 54 bp for *Goniopora lobata* and *Porites lobata*, to 3,798 bp for *Turbinaria mesenterina* based on the exons-only data set (mean 453 ± *SD* 242 bp); and from 63 bp for *Porites lobata* to 8,843 bp for *Galaxea astreata* based on the exons + supercontigs data set (mean 857 ± *SD* 625 bp) (Table 2). The original (15,919 baits) and filtered (13,479 baits) bait sets, as well as baitfile comprising the final 452 loci targeted are available at Zenodo (http://dx.doi.org/10.5281/zenodo.3590246; baits_designed.fa, baits_designed_filtered.fa and baitfile_452.fa respectively).

3.2 | Phylogenetic inference

Of the 605 loci targeted, HYBPIPER was able to generate contigs for 581 loci. Following filtering of loci to remove paralogs (43 loci), contaminant sequences (154 sequences) and uninformative loci (129 loci, including those with <3 taxa captured), a total of 452 loci spanning 865 exon regions were retained for phylogenomic inference. Loci recovered across scleractinian samples were fairly evenly distributed, with completeness ranging from 48.45% to 88.71% (mean 74.08% ± *SD* 11.65%; Table 2; Figure 1). The average taxon occupancy of the final matrix was 13.33 (±*SD* 4.14) scleractinians per locus, out of 18 scleractinians tested. Contigs assembled for each locus and concatenated alignments are available at Zenodo (http://dx.doi.org/10.5281/zenodo.3590246). Post-trimming, the concatenated matrices contained a total of 201,137 sites with 30.86% missing data (following Quek & Huang, 2019) for the exons-only data set (69.14% complete; mean alignment length = 456 ± *SD* 233 bp), and 287,749 sites with 32.43% missing data for the exons + supercontigs data set (67.57% complete; mean alignment length = 636 ± *SD* 352 bp). The latter showed greater sequence variability with 41.39% parsimony-informative sites, compared to 34.30% in the former data set (Table 3).

All phylogenetic trees inferred—from maximum likelihood, Bayesian and species tree analyses—were congruent with maximum bootstrap values and posterior probabilities at all nodes (Figure 2 and S3; available at Zenodo, http://dx.doi.org/10.5281/zenodo.3590246). The "Robust" and "Complex" clades, as well as monophyletic families Acroporidae, Euphylliidae, Poritidae and Merulinidae were unambiguously recovered. Furthermore, the vast majority of gene trees analyzed supported the phylogeny in both the concatenated and species tree reconstructions (Figure S1).

4 | DISCUSSION

In this study, we have designed hybrid-capture baits to target 605 putatively scleractinian loci over 1,139 exon regions. Laboratory testing of the baits shows that they are highly accurate and specific, enriching 452 loci across 865 exons that map to coral genomes following rigorous post-sequencing filtering. Our test on 18 species spanning 18 genera and 12 families also demonstrate that our baits are able to capture loci effectively—despite using just a minimum of six scleractinian taxa per locus for bait design—with minimal taxonomic bias across the "Robust" and "Complex" clades (Figure 1), which account for >98% of all species in Scleractinia (Huang, 2012; Huang & Roy, 2015; Kitahara et al., 2016). A slight bias towards the capture of "Robust" (mean 80% ± *SD* 9.7%) relative to "Complex" corals (mean 67% ± 9.8%) might be due to the larger number of baits designed based on "Robust" coral genomes (n = 9,149) compared to "Complex" coral genomes (n = 6,770). Future studies need to verify the efficacy of these baits in recovering sequences from the "Basal" clade (Stolarski et al., 2011). Nevertheless, our analyses have recovered a maximally-supported phylogeny, with a topology congruent with recent broad-scale phylogenies (Figure 2; Kitahara et al., 2016; Quek & Huang, 2019). We noted a slightly elevated level of gene tree incongruence among members of the "Robust" clade

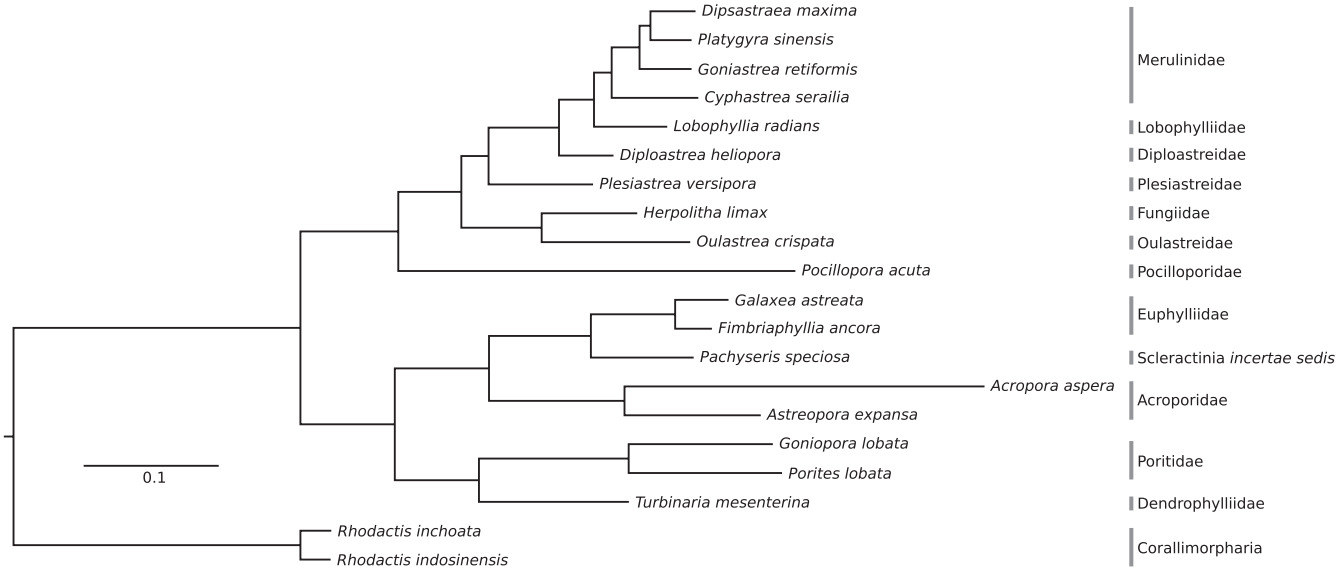

FIGURE 2 Maximum likelihood phylogeny of Scleractinia for exons-only data set (minimum taxon occupancy of three scleractinian taxa per locus; 30.86% missing data; 452 loci over 865 exon regions; 201,137 bp) with *Rhodactis* as outgroup. All nodes have maximum bootstrap values and posterior probabilities

(Figure S1), which could be attributed to factors such as incomplete lineage sorting (Woodhams, Lockhart, & Holland, 2016) and varying phylogenetic signal between clades (Gonçalves, Simpson, Ortiz, Shimizu, & Jansen, 2019). Analyses similar to that of Ying et al. (2018) performed at a broader scale would reveal factors driving these differences.

The method we employed in designing target-enrichment baits involves multiple filtering steps to capture putatively scleractinian loci. Considering the diversity of the coral holobiont (Stat et al., 2012; Thompson, Rivera, Closek, & Medina, 2014; Wainwright, Afiq-Rosli, Zahn, & Huang, 2019), a number of transcripts assembled would inevitably be of noncoral origins despite preliminary filters for Symbiodiniaceae transcripts (Quek & Huang, 2019). We circumvented symbiont contamination by leveraging a number of published coral genomes, using both blastp to locate putatively coral loci and alignment cutting used for bait design. In other words, our protocol ensured that the targeted loci were present in at least one of the reference genomes. Genome-based bait design is useful for locating intron-exon boundaries for optimal bait design (Bank et al., 2017; Hugall et al., 2016), and specifically in this study, the approach further aids in the accurate identification of coral loci.

In our post-sequencing analysis, we took advantage of a unique phylogenetic signature of Scleractinia: the deep split between the "Robust" and "Complex" clades (Huang, 2012; Huang & Roy, 2015; Kitahara, Cairns, Stolarski, Blair, & Miller, 2010; Kitahara et al., 2016; Romano & Palumbi, 1996, 1997; Stolarski et al., 2011; Ying et al., 2018). By inspecting individual gene trees for this pattern, we ensured that contaminant sequences were appropriately removed, and only phylogenetically informative, putatively coral and orthologous loci were retained for phylogenomic reconstruction. While laborious for a large number of loci, we recommend using well-substantiated prior information to help select loci for future target enrichment studies. We note that the identification of paralogs can now be expedited by newly-developed tools such as Clan_Check (Siu-Ting et al., 2019), which detects potential instances of hidden paralogy from a large number of trees, highlighting genes that may warrant further investigation (e.g., visual inspection of gene trees and sequence alignments, etc). Ultimately, meticulous data curation is of utmost importance in any phylogenetic reconstruction; molecular data sets are susceptible to cross-contamination (Bank et al., 2017; Hugall et al., 2016) and paralogy (Siu-Ting et al., 2019), both of which compound inaccuracies in phylogenomic analyses and therefore ought to be carefully checked.

While only exon sequences have been used for all phylogenomic analyses due to the divergent taxa sampled, we also provide an alternative pipeline to include intron sequences. Contigs comprising both exon and intron regions (supercontigs) can readily be extracted from HYBPIPER (Johnson et al., 2016). Since introns are more variable than exons (Table 3; Thomson, Wang, & Johnson, 2010), they may be useful for clarifying cryptic species complexes and resolving shallow divergences (Concepcion, Crepeau, Wagner, Kahng, & Toonen, 2008; Oppen, Willis, Vugt, & Miller, 2000; Pinzón & LaJeunesse, 2011).

Beyond the recovery of phylogenetic relationships, a handy aspect of the bait set designed here lies in the inclusion of baits targeting barcodes that are highly enriched in our assemblies. Considering that coral taxa are notoriously difficult to tell apart and taxonomic misidentifications even among families are not uncommon (e.g., *Turbinaria* sp. identified as *Astreopora* sp., noted in Quek & Huang, 2019), we provide an additional safeguard in the form of baits designed to assign samples to their genera. Furthermore, preliminary quality checks estimating the levels of cross-contamination can be assessed based on the number of contigs and depth of sequencing recovered per barcode. Nearly one-third of our reads

mapped to both barcoding baits, which is high when compared to previous target-enrichment studies (e.g., 5% of all reads originating from COI in Hugall et al., 2016). However, the two barcoding loci we used are essential for reliable verification of sample identities, and confer redundancy in case a single locus is not recovered for any one sample. Indeed, nine samples in Hugall et al. (2016) did not retain COI barcodes, while one and three of our samples did not recover COI and histone H3, respectively, but every sample had at least one barcoding locus. Ultimately, we have been able to obtain numerous loci per sample that contribute to robust, consistent inferences despite the large proportion of barcode reads produced.

As mitochondrial DNA is naturally enriched, sequencing reads of target-enriched libraries typically include background mitochondrial sequences as byproducts. For example, Allio et al. (2019) were able to extract COI barcodes and other mitochondrial genes from 501 hybrid-capture libraries in ants (Formicidae), and Taucce et al. (2018) assembled mitogenomes for five frog species (Anura) from hybrid-capture libraries. Here, we show that mitochondrial genes other than COI could be recovered for all samples in this study (Figure S2). Despite having on average >2,000× read depth across all 452 loci targeted, both COI and histone H3 have higher read depth (~150×) in comparison. The lower number of reads for the targeted loci could raise sequencing cost, particularly with lower throughput sequencers (e.g., Illumina MiSeq), but clearly, without targeted capture of these barcodes, mitochondrial loci may not be consistently captured even if they are naturally enriched (Figure S2).

We stress that the barcodes have been designed as a safeguard against sample misidentification, and it is at researchers' discretion to remove these barcodes. In particular, their removal is recommended if: (a) there is little to no risk of sample misidentification or mix-ups; (b) a low-throughput sequencing strategy is employed, or (c) a lower read depth is an acceptable tradeoff to maximize the number of samples sequenced. Furthermore, the depth of sequencing reflected in this study suggests that a few hundred samples can be safely combined into a single sequencing run with still sufficient read depth, especially with the removal of barcoding baits. Finally, it must be emphasized that these baits are useful for identification up to genus level. Where necessary, we strongly recommend the design of more specific baits, such as *Pax-C* 46/47 intron for *Acropora* spp. (Márquez, Van Oppen, Willis, Reyes & Miller, 2002; Van Oppen, Willis, Van Vugt & Miller, 2003) and open reading frame region for *Pocillopora* spp. (Flot & Tillier, 2006; Schmidt-Roach, Miller, Lundgren & Andreakis, 2014) following the pipeline outlined above.

The target-enrichment baits and sequence processing method presented here—leveraging recent developments in molecular techniques, sequencing and bioinformatics—represent another major step towards building large, gene-rich scleractinian trees. In particular, Quattrini et al. (2018) had designed baits targeting the more inclusive clade of Anthozoa. The baits designed were tested in vitro on 33 anthozoan taxa (four scleractinians), with a further nine genome-enabled taxa (two scleractinians) included for phylogenetic reconstruction. The data set thus incorporated a total of 42 anthozoans comprising 22 hexacorals (six scleractinians) and 20 octocorals. When comparing our phylogenetic data matrix (452 loci of the exons-only data set) with theirs containing a similar number of loci (438 loci of the 50% taxon occupancy matrix for Hexacorallia), there are 68,997 parsimony-informative sites in the former, more than twice the 34,390 parsimony-informative sites in the latter. Only when the total amount of missing data in Quattrini et al. (2018) are increased by lowering the taxon occupancy to 25% does the number of parsimony-informative sites increase to 63,968. The bait set designed in our study is clearly highly specific and targeted towards scleractinians, with much lower coverage for the sister-group corallimorpharians (Figure 1). Not surprisingly then, a blastn of the baits in Quattrini et al. (2018) against our final set of 13,479 baits reveals no overlap between the loci targeted, highlighting immense bait dissimilarities as a result of the distinct taxonomic level targeted (Shaffer, McCartney-Melstad, Near, Mount, & Spinks, 2017). Taken together, we suggest combining the two bait sets in future studies of Anthozoa to maximize the loci captured for scleractinian corals.

Resolving the phylogeny of scleractinian corals is critical for elucidating processes related to their evolutionary success and trajectories based on comparative genomics (Bhattacharya et al., 2016; Ying et al., 2018), and for reconstructing their origin and trait evolution (Hartmann et al., 2017; Madin et al., 2016; Stolarski et al., 2011). To date, the largest coral phylogeny reconstructed using NGS data includes only 39 scleractinian species represented by 43 samples (Quek & Huang, 2019). Over the next few years, we aim to increase the number of taxa placed on the phylogenomic tree by several fold using the method developed here to advance our understanding of coral evolution.

ACKNOWLEDGEMENTS

Phylogenomic analyses were performed partly with computational resources from the National Supercomputing Centre, Singapore (http://www.nscc.sg). This study was funded by the National Research Foundation, Prime Minister's Office, Singapore under its Marine Science R&D Programme (MSRDP-P03).

AUTHOR CONTRIBUTIONS

R. Z. B. Q., G. W. R., and D. H. conceptualized the study. R. Z. B. Q., S. S. J., and M. L. N. collected the samples. R. Z. B. Q. conducted the laboratory work, designed the pipeline, and analyzed the data. The manuscript was written by R. Z. B. Q., and D. H., with input from all authors. Publication of this manuscript was approved by all authors.

DATA AVAILABILITY STATEMENT

Baits designed, bait file, contigs assembled and scripts used in this study are available at Zenodo (http://dx.doi.org/10.5281/zenodo.3590246). Raw sequences and barcodes are available at NCBI SRA database (PRJNA602211) and GenBank respectively (Table S1). Voucher specimens have been deposited at Lee Kong Chian Natural History Museum, Singapore (Table S2).

ORCID

Randolph Z. B. Quek https://orcid.org/0000-0001-7998-7052
Danwei Huang https://orcid.org/0000-0003-3365-5583

REFERENCES

Aberer, A. J., Kobert, K., & Stamatakis, A. (2014). EXABAYES: Massively parallel Bayesian tree inference for the whole-genome era. *Molecular Biology and Evolution*, *31*(10), 2553–2556.

Allio, R., Schomaker-Bastos, A., Romiguier, J., Prosdocimi, F., Nabholz, B., & Delsuc, F. (2019). MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *bioRxiv*. https://doi.org/10.1101/685412

Aranda, M., Li, Y., Liew, Y. J., Baumgarten, S., Simakov, O., Wilson, M. C., … Voolstra, C. R. (2016). Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. *Scientific Reports*, *6*, 39734.

Arrigoni, R., Berumen, M. L., Huang, D., Terraneo, T. I., & Benzoni, F. (2017). *Cyphastrea* (Cnidaria : Scleractinia : Merulinidae) in the Red Sea: Phylogeny and a new reef coral species. *Invertebrate Systematics*, *31*(2), 141–156.

Arrigoni, R., Berumen, M. L., Terraneo, T. I., Caragnano, A., Bouwmeester, J., & Benzoni, F. (2014). Forgotten in the taxonomic literature: Resurrection of the scleractinian coral genus *Sclerophyllia* (Scleractinia, Lobophylliidae) from the Arabian Peninsula and its phylogenetic relationships. *Systematics and Biodiversity*, *13*(2), 40–63.

Arrigoni, R., Terraneo, T. I., Galli, P., & Benzoni, F. (2014). Lobophylliidae (Cnidaria, Scleractinia) reshuffled: Pervasive non-monophyly at genus level. *Molecular Phylogenetics and Evolution*, *73*, 60–64.

Artamonova, I. I., & Mushegian, A. R. (2013). Genome sequence analysis indicates that the model eukaryote *Nematostella vectensis* harbors bacterial consorts. *Applied and Environmental Microbiology*, *79*(22), 6868–6873.

Bank, S., Sann, M., Mayer, C., Meusemann, K., Donath, A., Podsiadlowski, L., … Niehuis, O. (2017). Transcriptome and target DNA enrichment sequence data provide new insights into the phylogeny of vespid wasps (Hymenoptera: Aculeata: Vespidae). *Molecular Phylogenetics and Evolution*, *116*, 213–226.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477.

Bhattacharya, D., Agrawal, S., Aranda, M., Baumgarten, S., Belcaid, M., Drake, J. L., … Falkowski, P. G. (2016). Comparative genomics explains the evolutionary success of reef-forming corals. *eLife*, *5*. https://doi.org/10.7554/eLife.13288

Bolger, A. M., Lohse, M., & Usadel, B. (2014). TRIMMOMATIC: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bossert, S., & Danforth, B. N. (2018). On the universality of target-enrichment baits for phylogenomic research. *Methods in Ecology and Evolution*, *9*(6), 1453–1460.

Bragg, J. G., Potter, S., Bi, K., & Moritz, C. (2016). Exon capture phylogenomics: Efficacy across scales of divergence. *Molecular Ecology Resources*, *16*(5), 1059–1068.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*, 59–60.

Budd, A. F., Romano, S. L., Smith, N. D., & Barbeitos, M. S. (2010). Rethinking the phylogeny of scleractinian corals: A review of morphological and molecular data. *Integrative and Comparative Biology*, *50*(3), 411–427.

Cairns, S. D. (2007). Deep-water corals: An overview with species reference to diversity and distribution of deep-water scleractinian corals. *Bulletin of Marine Science*, *81*(3), 311–322.

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973.

Chen, C. A., Odorico, D. M., Tenlohuis, M., Veron, J. E. N., & Miller, D. J. (1995). Systematic relationships within the Anthozoa (Cnidaria: Anthozoa) using the 5′-end of the 28S rDNA. *Molecular Phylogenetics and Evolution*, *4*(2), 175–183.

Concepcion, G. T., Crepeau, M. W., Wagner, D., Kahng, S. E., & Toonen, R. J. (2008). An alternative to ITS, a hypervariable, single-copy nuclear intron in corals, and its use in detecting cryptic species within the octocoral genus *Carijoa*. *Coral Reefs*, *27*(2), 323–336.

Cunning, R., Bay, R. A., Gillette, P., Baker, A. C., & Traylor-Knowles, N. (2018). Comparative analysis of the *Pocillopora damicornis* genome highlights role of immune system in coral evolution. *Scientific Reports*, *8*(1), 16134.

Dornburg, A., Townsend, J. P., Brooks, W., Spriggs, E., Eytan, R. I., Moore, J. A., … Near, T. J. (2017). New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. *Molecular Phylogenetics and Evolution*, *110*, 27–38.

Emms, D. M., & Kelly, S. (2015). ORTHOFINDER: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*, 157.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726.

Flot, J.-F., & Tillier, S. (2006). Molecular phylogeny and systematics of the scleractinian coral genus *Pocillopora* in Hawai'i. In *Proceedings of the 10th International Coral Reef Symposium* (pp. 24–29).

Forsman, Z. H., Knapp, I. S. S., Tisthammer, K., Eaton, D. A. R., Belcaid, M., & Toonen, R. J. (2017). Coral hybridization or phenotypic variation? Genomic data reveal gene flow between *Porites lobata* and *P. compressa*. *Molecular Phylogenetics and Evolution*, *111*, 132–148.

Fukami, H., Budd, A. F., Paulay, G., Sole-Cava, A., Chen, C. A., Iwao, K., & Knowlton, N. (2004). Conventional taxonomy obscures deep divergence between Pacific and Atlantic corals. *Nature*, *427*, 832–835.

Fukami, H., Chen, C. A., Budd, A. F., Collins, A., Wallace, C., Chuang, Y.-Y., … Knowlton, N. (2008). Mitochondrial and nuclear genes suggest that stony corals are monophyletic but most families of stony corals are not (Order Scleractinia, Class Anthozoa, Phylum Cnidaria). *PLoS ONE*, *3*(9), e3222.

Gonçalves, D. J. P., Simpson, B. B., Ortiz, E. M., Shimizu, G. H., & Jansen, R. K. (2019). Incongruence between gene trees and species trees and phylogenetic signal variation in plastid genes. *Molecular Phylogenetics and Evolution*, *138*, 219–232.

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351.

Hartmann, A. C., Baird, A. H., Knowlton, N., & Huang, D. (2017). The paradox of environmental symbiont acquisition in obligate mutualisms. *Current Biology*, *27*(23), 3711–3716.e3.

Huang, D. (2012). Threatened reef corals of the world. *PLoS ONE*, *7*(3), e34459.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Research*, *9*, 868–877.

Huang, D., Arrigoni, R., Benzoni, F., Fukami, H., Knowlton, N., Smith, N. D., … Budd, A. F. (2016). Taxonomic classification of the reef coral family Lobophylliidae (Cnidaria: Anthozoa: Scleractinia). *Zoological Journal of the Linnean Society*, *178*(3), 436–481.

Huang, D., Benzoni, F., Arrigoni, R., Baird, A. H., Berumen, M. L., Bouwmeester, J., … Budd, A. F. (2014). Towards a phylogenetic classification of reef corals: The Indo-Pacific genera *Merulina*, *Goniastrea* and *Scapophyllia* (Scleractinia, Merulinidae). *Zoologica Scripta*, *43*(5), 531–548.

Huang, D., Benzoni, F., Fukami, H., Knowlton, N., Smith, N. D., & Budd, A. F. (2014). Taxonomic classification of the reef coral families Merulinidae, Montastraeidae, and Diploastraeidae (Cnidaria: Anthozoa: Scleractinia). *Zoological Journal of the Linnean Society*, *171*(2), 277–355.

Huang, D., Licuanan, W. Y., Baird, A. H., & Fukami, H. (2011). Cleaning up the "Bigmessidae": Molecular phylogeny of scleractinian corals from Faviidae, Merulinidae. *Pectiniidae and Trachyphylliidae*. *BMC Evolutionary Biology*, *11*, 37.

Huang, D., Meier, R., Todd, P. A., & Chou, L. M. (2008). Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. *Journal of Molecular Evolution*, *66*(2), 167–174.

Huang, D., & Roy, K. (2015). The future of evolutionary diversity in reef corals. *Philosophical Transactions of the Royal Society B*, *370*, 20140010.

Hugall, A. F., O'Hara, T. D. O., Hunjan, S., Nilsen, R., & Moussalli, A. (2016). An exon-capture system for the entire Class Ophiuroidea. *Molecular Biology and Evolution*, *33*(1), 281–294.

Johnson, M. G., Gardner, E. M., Liu, Y., Medina, R., Goffinet, B., Shaw, A. J., … Wickett, N. J. (2016). HYBPIPER: Extracting coding sequence and introns for phylogenetics from high-throughput sequencing reads using target enrichment. *Applications in Plant Sciences*, *4*(7), 1600016.

Johnston, E. C., Forsman, Z. H., Flot, J.-F., Schmidt-Roach, S., Pinzón, J. H., Knapp, I. S. S., & Toonen, R. J. (2017). A genomic glance through the fog of plasticity and diversification in *Pocillopora*. *Scientific Reports*, *7*(1), 5991.

Junier, T., & Zdobnov, E. M. (2010). The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*, *26*(13), 1669–1770.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kitahara, M. V., Cairns, S. D., Stolarski, J., Blair, D., & Miller, D. J. (2010). A comprehensive phylogenetic analysis of the Scleractinia (Cnidaria, Anthozoa) based on mitochondrial CO1 sequence data. *PLoS ONE*, *5*(7), e11490.

Kitahara, M. V., Fukami, H., Benzoni, F., & Huang, D. (2016). The new systematics of Scleractinia: Integrating molecular and morphological evidence. In S. Goffredo, & Z. Dubinsky (Eds.), *The Cnidaria, past, present and future* (pp. 41–59). Basel, Switzerland: Springer International Publishing.

Kitano, Y. F., Benzoni, F., Arrigoni, R., Shirayama, Y., Wallace, C. C., & Fukami, H. (2014). A phylogeny of the family Poritidae (Cnidaria, Scleractinia) based on molecular and morphological analyses. *PLoS ONE*, *9*(5), e98406.

Kulkarni, P., & Frommolt, P. (2017). Challenges in the setup of large-scale next-generation sequencing analysis workflows. *Computational and Structural Biotechnology Journal*, *15*, 471–477.

Lamarck, J.-B.-P. (1816). *Histoire Naturelle des Animaux sans Vertèbres*. Paris, France: Verdière.

Laumer, C. E., Fernández, R., Lemer, S., Combosch, D., Kocot, K. M., Riesgo, A., … Giribet, G. (2019). Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1906), 20190831.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, *61*(5), 727–744.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 1303.3997.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., & Homer, N. … 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079.

Liew, Y. J., Aranda, M., & Voolstra, C. R. (2016). Reefgenomics. Org - a repository for marine genomics data. *Database*, *2016*(2016), baw152.

Lin, M. F., Chou, W. H., Kitahara, M. V., Chen, C. L. A., Miller, D. J., & Forêt, S. (2016). Corallimorpharians are not "naked corals": Insights into relationships between Scleractinia and Corallimorpharia from phylogenomic analyses. *PeerJ*, *4*, e2463.

Linnaeus, C. (1758). *Systema Naturæ per Regna Tria Naturæ: Secundum Classes, Ordines, Genera, Species, Cum Characteribus, Differentiis, Synonymis, Locis*. Stockholm, Sweden: Laurentii Salvii.

Liu, H., Stephens, T. G., González-Pech, R. A., Beltran, V. H., Lapeyre, B., Bongaerts, P., … Chan, C. X. (2018). *Symbiodinium* genomes reveal adaptive evolution of functions related to coral-dinoflagellate symbiosis. *Communications Biology*, *1*, 95.

Liu, Y., Johnson, M. G., Cox, C. J., Medina, R., Devos, N., Vanderpoorten, A., … Goffinet, B. (2019). Resolution of the ordinal phylogeny of mosses using targeted exons from organellar and nuclear genomes. *Nature Communications*, *10*(1), 1485.

Madin, J. S., Anderson, K. D., Andreasen, M. H., Bridge, T. C. L., Cairns, S. D., Connolly, S. R., … Baird, A. H. (2016). The Coral Trait Database, a curated database of trait information for coral species from the global oceans. *Scientific Data*, *3*, 160017.

Márquez, L. M., Van Oppen, M. J., Willis, B. L., Reyes, A., & Miller, D. J. (2002). The highly cross-fertile coral species, Acropora hyacinthus and Acropora cytherea, constitute statistically distinguishable lineages. *Molecular Ecology*, *11*(8), 1339–1349.

Mayer, C., Sann, M., Donath, A., Meixner, M., Podsiadlowski, L., Peters, R. S., … Niehuis, O. (2016). BAITFISHER: A software package for multispecies target DNA enrichment probe design. *Molecular Biology and Evolution*, *33*(7), 1875–1886.

Moberg, F., & Folke, C. (1999). Ecological goods and services of coral reef ecosystems. *Ecological Economics*, *29*(2), 215–233.

Oppen, M. J., Willis, B. L., Vugt, H. W., & Miller, D. J. (2000). Examination of species boundaries in the *Acropora cervicornis* group (Scleractinia, Cnidaria) using nuclear DNA sequence analyses. *Molecular Ecology*, *9*(9), 1363–1373.

Pinzón, J. H., & LaJeunesse, T. C. (2011). Species delimitation of common reef corals in the genus *Pocillopora* using nucleotide sequence phylogenies, population genetics and symbiosis ecology. *Molecular Ecology*, *20*(2), 311–325.

Pizarro, D., Divakar, P. K., Grewe, F., Leavitt, S. D., Huang, J.-P., Dal Grande, F., … Lumbsch, H. T. (2018). Phylogenomic analysis of 2556 single-copy protein-coding genes resolves most evolutionary relationships for the major clades in the most diverse group of lichen-forming fungi. *Fungal Diversity*, *92*(1), 31–41.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FASTTREE 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, *5*(3), e9490.

Quattrini, A. M., Faircloth, B. C., Dueñas, L. F., Bridge, T. C. L., Brugler, M. R., Calixto-Botía, I. F., … McFadden, C. S. (2018). Universal target-enrichment baits for anthozoan (Cnidaria) phylogenomics: New approaches to long-standing problems. *Molecular Ecology Resources*, *18*(2), 281–295.

Quek, Z. B. R., Chang, J. J. M., & Ip, Y. C. A., & Huang, D. (2019). Complete mitochondrial genome of the sea star *Archaster typicus* (Asteroidea: Archasteridae). *Mitochondrial DNA Part B*, *4*(2), 3130–3132.

Quek, Z. B. R., & Huang, D. (2019). Effects of missing data and data type on phylotranscriptomic analysis of stony corals (Cnidaria: Anthozoa: Scleractinia). *Molecular Phylogenetics and Evolution*, *134*, 12–23.

ReFuGe 2020 Consortium. (2015). The ReFuGe 2020 Consortium—using "omics" approaches to explore the adaptability and resilience of coral holobionts to environmental change. *Frontiers in Marine Science*, *2*, 68.

Richards, Z. T., Carvajal, J. I., Wallace, C. C., & Wilson, N. G. (2019). Phylotranscriptomics confirms *Alveopora* is sister to *Montipora* within the family Acroporidae. *Marine Genomics*, https://doi.org/10.1016/j.margen.2019.100703

Rokas, A., Williams, B. L., King, N., & Carroll, S. B. (2003). Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, *425*(6960), 798–804.

Romano, S. L., & Palumbi, S. R. (1996). Evolution of scleractinian corals inferred from molecular systematics. *Science*, *271*(5249), 640–642.

Romano, S. L., & Palumbi, S. R. (1997). Molecular evolution of a portion of the mitochondrial 16S ribosomal gene region in scleractinian corals. *Journal of Molecular Evolution*, *45*, 397–411.

Sayyari, E., Whitfield, J. B., & Mirarab, S. (2018). DiscoVista: Interpretable visualizations of gene tree discordance. *Molecular Phylogenetics and Evolution*, *122*, 110–115.

Schott, R. K., Panesar, B., Card, D. C., Preston, M., Castoe, T. A., & Chang, B. S. W. (2017). Targeted capture of complete coding regions across divergent species. *Genome Biology and Evolution*, *9*(2), 398–414.

Schmidt-Roach, S., Miller, K. J., Lundgren, P., & Andreakis, N. (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), 1–33.

Shaffer, H. B., McCartney-Melstad, E., Near, T. J., Mount, G. G., & Spinks, P. Q. (2017). Phylogenomic analyses of 539 highly informative loci dates a fully resolved time tree for the major clades of living turtles (Testudines). *Molecular Phylogenetics and Evolution*, *115*, 7–15.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS ONE*, *11*(10), e0163962.

Shinzato, C., Shoguchi, E., Kawashima, T., Hamada, M., Hisata, K., Tanaka, M., … Satoh, N. (2011). Using the *Acropora digitifera* genome to understand coral responses to environmental change. *Nature*, *476*, 320–323.

Shoguchi, E., Beedessee, G., Tada, I., Hisata, K., Kawashima, T., Takeuchi, T., … Shinzato, C. (2018). Two divergent *Symbiodinium* genomes reveal conservation of a gene cluster for sunscreen biosynthesis and recently lost genes. *BMC Genomics*, *19*, 458.

Shoguchi, E., Shinzato, C., Kawashima, T., Gyoja, F., Mungpakdee, S., Koyanagi, R., … Satoh, N. (2013). Draft assembly of the *Symbiodinium minutum* nuclear genome reveals dinoflagellate gene structure. *Current Biology*, *23*(15), 1399–1408.

Siu-Ting, K., Torres-Sánchez, M., Mauro, D. A., Wilcockson, D., Wilkinson, M., Pisani, D., … Creevey, C. J. (2019). Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. *Molecular Biology and Evolution*, *36*(6), 1344–1356.

Sohn, J.-I., & Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23–40.

Stamatakis, A. (2014). RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Stat, M., Baker, A. C., Bourne, D. G., Correa, A. M. S., Forsman, Z., Huggett, M. J., … Gates, R. D. (2012). Chapter one – molecular delineation of species in the coral holobiont. In M. Lesser (Ed.), *Advances in marine biology*, Vol. *63* (pp. 1–65). London, UK: Academic Press.

Stolarski, J., Kitahara, M. V., Miller, D. J., Cairns, S. D., Mazur, M., & Meibom, A. (2011). The ancient evolutionary origins of Scleractinia revealed by azooxanthellate corals. *BMC Evolutionary Biology*, *11*, 316.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34(Web Server, Issue)*, W609–W612.

Taucce, P. P. G., Canedo, C., Haddad, C. F. B., Lemmon, A. R., Lemmon, E. M., Vences, M., & Lyra, M. (2018). The mitochondrial genomes of five frog species of the Neotropical genus Ischnocnema (Anura: Brachycephaloidea: Brachycephalidae). *Mitochondrial DNA Part B*, *3*(2), 915–917.

Thompson, J. R., Rivera, H. E., Closek, C. J., & Medina, M. (2014). Microbes in the coral holobiont: Partners through evolution, development, and ecological interactions. *Frontiers in Cellular and Infection Microbiology*, *4*, 176.

Thomson, R. C., Wang, I. J., & Johnson, J. R. (2010). Genome-enabled development of DNA markers for ecology, evolution and conservation. *Molecular Ecology*, *19*(11), 2184–2195.

Van Oppen, M. J. H., Willis, B. L., Van Vugt, H. W. J. A., & Miller, D. J. (2003). Examination of species boundaries in the Acropora cervicornis group (Scleractinia, Cnidaria) using nuclear DNA sequence analyses. *Molecular Ecology*, *9*(9), 1363–1373.

Voolstra, C. R., Li, Y., Liew, Y. J., Baumgarten, S., Zoccola, D., Flot, J.-F., … Aranda, M. (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.

Wainwright, B. J., Afiq-Rosli, L., Zahn, G. L., & Huang, D. (2019). Characterisation of coral-associated bacterial communities in an urbanised marine environment shows strong divergence over small geographic scales. *Coral Reefs*, *38*(6), 1097–1106. https://doi.org/10.1007/s00338-019-01837-1

White, W. T., Corrigan, S., Yang, L., Henderson, A. C., Bazinet, A. L., Swofford, D. L., & Naylor, G. J. P. (2018). Phylogeny of the manta and devilrays (Chondrichthyes: Mobulidae), with an updated taxonomic arrangement for the family. *Zoological Journal of the Linnean Society*, *182*(1), 50–75.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., … Leebens-Mack, J. (2014). Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences*, *111*(45), E4859–E4868.

Woodhams, M. D., Lockhart, P. J., & Holland, B. R. (2016). Simulating and Summarizing Sources of Gene Tree Incongruence. *Genome Biology and Evolution*, *8*(5), 1299–1315.

Ying, H., Cooke, I., Sprungala, S., Wang, W., Hayward, D. C., Tang, Y., … Miller, D. J. (2018). Comparative genomics reveals the distinct evolutionary trajectories of the robust and complex coral lineages. *Genome Biology*, *19*(1), 175.

Zapata, F., Goetz, F. E., Smith, S. A., Howison, M., Siebert, S., Church, S. H., … Cartwright, P. (2015). Phylogenomic analyses support traditional relationships within Cnidaria. *PLoS ONE*, *10*(10), e0139068.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(Suppl 6), 153.

Zhu, T., Dos Reis, M., & Yang, Z. (2015). Characterization of the uncertainty of divergence time estimation under relaxed molecular clock models using multiple loci. *Systematic Biology*, *64*(2), 267–280.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Quek RZB, Jain SS, Neo ML, Rouse GW, Huang D. Transcriptome-based target-enrichment baits for stony corals (Cnidaria: Anthozoa: Scleractinia). *Mol Ecol Resour*. 2020;20:807–818. https://doi.org/10.1111/1755-0998.13150