DR. JOSEPH D. DIBATTISTA (Orcid ID : 0000-0002-5696-7574)

Received Date : 25-May-2016

Revised Date : 11-Feb-2017

Accepted Date : 14-Feb-2017

Article type : Resource Article

Molecular Ecology Resources – Permanent Genetic Resources

Using a butterflyfish genome as a general tool for RAD-Seq studies in specialized reef fish

Running title: **RAD-Seq in butterflyfish**

Joseph D. DiBattista[1,2*], Pablo Saenz-Agudelo[1,3], Marek J. Piatek[4], Xin Wang[1], Manuel Aranda[1], and Michael L. Berumen[1]

[1]Red Sea Research Center, Division of Biological and Environmental Science and Engineering, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia, [2]Department of Environment and Agriculture, Curtin University, PO Box U1987, Perth, WA 6845, Australia, [3]Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile, Valdivia 5090000, Chile, [4]Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia

*Correspondence: Joseph DiBattista, Department of Environment and Agriculture, Curtin University, PO Box U1987, Perth, WA 6845, Australia

E-mail: josephdibattista@gmail.com

Keywords: adaptation, Chaetodontidae, ddRAD, ecological genomics, next-generation sequencing, SNP

Abstract

Data from a large-scale restriction site associated DNA (RAD-Seq) study of nine butterflyfish species in the Red Sea and Arabian Sea provided a means to test the utility of a recently published draft genome (*Chaetodon austriacus*) and assess apparent bias in this method of isolating nuclear loci. We here processed double-digest restriction-site (ddRAD) associated DNA sequencing data to identify single nucleotide polymorphism (SNP) markers and their associated function with and without our reference genome to see if it improves the quality of RAD-Seq markers. Our analyses indicate (1) a modest gap between the number of non-annotated versus annotated SNPs across all species, (2) an advantage of using genomic resources for closely related but not distantly related butterflyfish species based on the ability to assign putative gene function to SNPs, and (3) an enrichment of genes among sister butterflyfish taxa related to calcium transmembrane transport and binding. The latter result highlights the potential for this approach to reveal insights into adaptive mechanisms in populations inhabiting challenging coral reef environments such as the Red Sea, Arabian Sea, and Arabian Gulf with further study.

Introduction

Ecologists and evolutionary biologists often mandate genetic data to test hypotheses related to recent population history (i.e. bottlenecks), connectivity among sites, and the spatial distribution of genetic variation for effective species management. Moreover, the capacity to rapidly genotype many individuals' at large numbers of genetic markers improves our ability to estimate demographic parameters (e.g. gene flow, effective population size, admixture), resolve phylogenetic placement, identify genes that may be under selection, and even identify genes that facilitate adaptation to our rapidly changing environment. Now, with the decreasing cost of high-throughput next-generation sequencing (NGS), data can be sampled from across the genome to meet this demand (for review see Kosuri & Church 2014).

Restriction site associated DNA sequencing (RAD-Seq) methods, which use NGS to target sequence data adjacent to restriction enzyme recognition sites (Davey *et al.* 2011), have generated informative single nucleotide polymorphism (SNP) datasets in several fish species. Examples include studies characterizing different trout species and their hybrids (Hand *et al.* 2015), the genetic basis for different forms of stickleback (e.g. oceanic versus freshwater; Hohenlohe *et al.* 2010; Catchen *et al.* 2013a), identifying cryptic lineages of herring in the Baltic Sea (*Clupea harengus*; Corander *et al.* 2013), and resolving relationships among iconic African cichlid species (Wagner *et al.* 2013; Henning *et al.* 2014). Most of these taxa, however, are considered "model organisms", and therefore unique in their ability to map, align, and otherwise trace their SNPs back to annotated genomes.

Reef fish, on the other hand, have received much less attention, despite 5000 species inhabiting tropical seas. We note that RAD-Seq methods have only been applied a few times in reef fish, including surgeonfish (Gaither *et al.* 2015), clownfish (Saenz-Agudelo *et al.* 2015), hamlets (Puebla *et al.* 2014; Picq *et al.* 2016), angelfish (Tariel *et al.* 2016), grunts (Bernal *et al.* 2016), groupers (Jackson *et al.* 2014), and parrotfish (Stockwell *et al.* 2016). Most of these studies, however, are limited in scope (based on small sample sizes, few study species, and minimal sampling sites) or fail to confidently assign gene function to SNPs of interest; all rely exclusively on *de novo* assembly of raw sequence reads. This deficiency in reef fish relative to other aquatic organisms can therefore be, at least in part, attributed to a lack of publicly available genomic resources. All published genomes to date in this group are based on cold water (*Takifugu rubripes*; van de Peer 2004), brackish water (*Tetraodon* *nigroviridis*; van de Peer 2004), pelagic (*Thunnus orientalis*; Nakamura *et al.* 2013), and phylogenetically distinct (*Rhincodon typus*; Read *et al.* 2015) species.

A recently published genome for the Red Sea butterflyfish (*Chaetodon austriacus*; DiBattista *et al.* 2016a), an obligate corallivore often targeted by the ornamental fish trade (Wabnitz 2003; Lawton *et al.* 2013), allows us to explore genes underpinning ecological niche selection (Cole *et al.* 2008), evolutionary distinctness (Bellwood *et al.* 2010), but more importantly biogeographic patterns related to the specialized and speciose Chaetodontidae family (> 130 species). Indeed, the Red Sea has experienced intermittent historical isolation (Bailey 2009), fluctuating environmental conditions (e.g. Raitsos *et al.* 2013), and harbors large numbers of endemic species (DiBattista *et al.* 2016b) whose origins are still subject to debate (DiBattista *et al.* 2016c). The *C. austriacus* genome may therefore enhance RAD-seq analysis in related species.

We therefore have two methodological aims in this study. First, we test whether using the *C. austriacus* genome as a reference improves the quality of RAD-Seq markers for this species, which includes assessing the apparent bias in our method of isolating nuclear loci. Second, to increase the utility of this genome for broader topics of study, we further test the capacity of this genome to serve as a scaffold for RAD-Seq markers in other butterflyfish species. This is the first time comparisons between RAD-Seq analyses, with and without a representative genome, have been conducted across such a large number of related, aquatic organisms.

Material and methods

Sample collection

We collected fin or gill tissue from between 48 and 108 individuals for each of nine species of butterflyfish sampled at sites in the Red Sea and Arabian Sea, including the genomically enabled *C. austriacus* (see Fig.1 and Table 1). We also collected samples from a surgeonfish species (*Ctenochaetus striatus; N* = 120 from 10 populations), which served as an outgroup for our analyses. All samples were preserved in 95% ethanol prior to processing.

Restriction site associated DNA sequencing method

The RAD-Seq method involves digesting genomic DNA with restriction enzymes and sequencing fragments of DNA adjacent to restriction sites. In our case, we used the double-digest RAD tag method (ddRAD; Peterson *et al.* 2012), which uses one common and one rare restriction enzyme cutter but no shearing step, although several other methods exist (Willette *et al.* 2014; Andrews *et al.* 2016).

In brief, genomic DNA was extracted using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA) following the manufacturer's protocol. Total DNA from each extracted aliquot was quantified using a Qubit dsDNA HS Assay Kit (Invitrogen, Carlsbad, CA). Genomic libraries were prepared from 500 ng of DNA per sample by: 1) digesting at 37°C using the restriction enzymes *SphI* and *MluCI* (NEB), 2) ligating to unique combinations of custom adaptors, 3) pooling 16 individuals at a time, 4) size-selecting 300 to 500 bp fragments on agarose gels from each pool, 5) amplifying over 10 PCR cycles to reduce clonality in 50 µl reactions containing 25 µl Illumina True Seq Master Mix, 20µl of library DNA, and unique indexing primers that correspond to the standard Illumina multiplexed sequencing protocol, and 6) combining pools, as appropriate, in equimolar concentration to form nine genomic libraries, which were then run on nine lanes of an Illumina HiSeq2000 (101 bp; v3 reagents) at the King Abdullah University of Science and Technology (KAUST) Bioscience Core Laboratory using the SE option.

RAD-Seq analysis

RAD sequences were first analyzed *de novo* (without a reference genome) using Stacks *vers.* 1.44 (Catchen *et al.* 2011, 2013b). Raw reads for each individual in the RAD tag library were trimmed at the end to a common length of 81 bp in FASTQ format using the “process_radtags” pipeline since read quality was considerably lower from this position onwards; reads with an average phred score < 20 (in a 5 bp sliding window) were discarded. All remaining reads from each species were separately analyzed and SNPs were genotyped using the “denovo_map.pl” pipeline. The parameters used in Stacks were: minimum number of reads required to form a stack (m) set to 3; maximum number of mismatches between loci for individuals (M) set to 4; maximum number of mismatches when aligning secondary reads to primary stacks (N) set to 2; maximum number of mismatches between loci when creating a catalog (n) set to 2. This parameter combination yielded the closest number of SNPs for our focal species (*C. austriacus*) compared to the number of SNPs recovered using the “ref_map.pl” pipeline in Stacks (see below and Appendix S1), and was therefore used as the optimal settings for all the other species. Population filtering parameters used in order to include a locus in the final data set were: 1) minor allele frequency > 0.05 and 2) the locus had to be genotyped in at least 80% of individuals (populations: -r) and present in all (strict quality filter) or all but one (relaxed quality filter) populations (populations: -p)(for more detail see Saenz-Agudelo *et al.* 2015).

RAD sequences were also analyzed using the *C. austriacus* genome as a reference (DiBattista *et al.* 2016a). For this approach, trimmed RAD reads from each species (and thus SNPs) were first aligned to the *C. austriacus* scaffolds using Bowtie *vers.* 1.0.1 (Langmead *et al.* 2009). We configured Bowtie to report only one (i.e. the best) alignment per read and allowed up to 3 mismatches in the default seed length. Aligned reads were then analyzed with ref_map.pl using the same minimum number of reads to build a stack as outlined above for denovo_map.pl; population filtering parameters also remained the same (see Appendix S1). This step additionally allowed us to estimate homology levels for all studied species against the *C. austriacus* genome.

Complimentary Maximum Likelihood (ML) phylogenetic trees were built with RAD sequences using an unpartitioned GTRCAT approximation model in RAxML *vers.* 2.0 (Stamatakis 2006) as implemented in Geneious *vers.* 10.0.2 (Drummond *et al.* 2009)(also see Ree & Hipp 2015). The input for these analyses consisted of a full sequence (.phylip) alignment generated in STACKS (denovo_map: -m 3 -M 4 -N 2 -n 6) by compiling four individuals per species, resulting in 38,070 total bp representing 2846 shared SNP loci (2170 fixed SNPs and 676 variable SNPs within species). Only loci present in all 10 species and at least 3 individuals per species were included. Branch support was based on 100,000 rapid-bootstrap replicates with *Ct. striatus* used as an outgroup. Trees were visualized using FigTree *vers.* 1.4.2 (http://tree.bio.ed.ac.uk/software/figtree/).

We next explored the form and function of SNPs that mapped to the *C. austriacus* scaffolds in several ways. First, we identified SNPs from *C. austriacus* and its closest relative, *C. melapterus*, which mapped to regions containing gene models. These represent the only two species where using the *C. austriacus* reference genome increased the number of SNPs identified (see below). The associated genes were then used to look for potential enrichment of specific gene functions using GO enrichment analyses. We identified 3775 SNPs from the *C. austriacus* dataset mapping to 3056 different genes in the reference genome of *C. austriacus*, and 2341 SNPs from the *C. melapterus* dataset mapping to 2011 different genes of the reference genome. GO enrichment analyses were performed using TopGo *vers.* 2.22.0 from the R Bioconducter package (Alexa & Rahnenfuhrer 2010), the weight01 model, a cut-off of $P < 0.01$, and specific background gene sets covered by stacks from the ref_map.pl analysis. These analyses were used to find which GO terms were over-represented (or under-represented) using annotations for that gene set. Second, we estimated the number of SNPs from *C. austriacus,* with corresponding tolerance intervals (95% confidence in 99% of the sample), that mapped to the assembled genome as a function of scaffold size. This particular analysis highlights scaffolds that host a statistically higher number of SNPs for their length. Frequency distributions for the number of SNPs and scaffold lengths were also calculated.

Results

We used Stacks to assemble and genotype SNPs in two different ways. First, we used denovo_map.pl to assemble and genotype SNPs *de novo* for ten different species, including *C. austriacus*, the only butterflyfish with a fully sequenced genome. We also used the ref_map.pl pipeline to assemble and map reads from each of these species to the *C. austriacus* scaffolds, which are now publicly available (NCBI Bioproject PRJNA292048; http://caus.reefgenomics.org; see Liew *et al.* 2016). Using an optimal parameter combination (see Appendix S1) and denovo_map.pl, we found 289,504 loci, with

8,842 (strict quality filter) or 10,711 (relaxed quality filter) variable loci containing at least one SNP site within or between individuals for *C. austriacus* (Table 1). On average, there were 262,247 loci across all butterflyfish species (range: 189,554 to 363,221), and between 476 (*Chaetodon trifascialis*) and 10,257 (*Chaetodon paucifasciatus*) loci per species passing our filters based on the strict criteria, and between 1,271 (*Chaetodon trifascialis*) and 13,539 (*Chaetodon paucifasciatus*) loci per species passing our filters based on the relaxed criteria (Table 1). Average depth of coverage ranged from 10.2X (*Chaetodon fasciatus*) to 16.2X (*Chaetodon paucifasciatus*) for denovo_map.pl (average depth of coverage = 12.3X), and from 12.1X (*Chaetodon fasciatus*) to 16.1X (*Chaetodon paucifasciatus*) for ref_map.pl (average depth of coverage = 13.5X).

We note a 38% (strict quality filter) or 56% (relaxed quality filter) increase in the number SNPs identified using ref_map.pl versus denovo_map.pl for *C. melapterus* (Table 1), the closest relative to *C. austriacus*. In contrast, there was a 2- to 6-fold decrease (strict quality filter), or 4- to 12-fold decrease (relaxed quality filter) in the number of SNPs identified for the remaining butterflyfish species using ref_map.pl (excluding *C. austriacus*; Table 1). This downward trend was more pronounced when we considered fish outside of the Chaetodontidae family (i.e. *Ct. striatus*, family Acanthuridae), with a 52-fold (strict quality filter) or a 56-fold (relaxed quality filter) decrease in the number of SNPs identified. Indeed, our surgeonfish outgroup, *Ct. striatus*, had only 8 (strict quality filter) or 27 (relaxed quality filter) useable SNP loci after filtering when using ref_map.pl.

Gene function under putative selection

In order to determine if variable SNPs in *C. austriacus* and its closest relative, *C. melapterus*, were associated with specific biological processes, we performed GO enrichment analyses on SNPs lying within gene models. To this end, we first analysed if SNPs were randomly distributed with respect to gene function. Analysis of the GO terms covered by the species-specific SNPs showed a strong bias towards particular gene functions, with the strongest enrichment for genes involved in the positive regulation of GTPase activity, axon guidance, and chloride transmembrane transport, among others.

To determine if this observed bias in gene function was a consequence of our RAD-seq method and thus chosen restriction enzyme, we analysed the gene models covered by our loci in comparison to all proteins. Briefly, the *SphI* recognition site (GCATGC) contains an ATG followed by a C, which in coding sequences translates to a methionine (M) followed by a leucine (L). Consequently, using this restriction enzyme could enrich for genomic regions encoding proteins starting with or harboring internal Methionine-Leucine (ML) amino acid stretches. We therefore analyzed the frequency of ML amino acid stretches in all gene models for *C. austriacus* and compared them to the frequency within the set of gene models associated with our loci. This analysis revealed a highly significant enrichment of such genes within our loci, thus identifying this as a source of the previously observed bias (Chi-square test: $P < 0.0001$). Based on this result we generated species-specific GO term backgrounds for our subsequent enrichment analyses to account for the observed bias, which may be an important consideration for other RAD-seq approaches.

SNP specific GO enrichment analyses identified 29 and 37 biological processes in *C. austriacus* and *C. melapterus*, respectively, that were significantly enriched ($P < 0.01$), of which six were shared between the species. These shared terms included calcium ion transmembrane transport, axon guidance, neuron cell-cell adhesion, positive regulation of excitatory postsynaptic potential, and positive regulation of cardiac muscle cell proliferation. Variable SNPs exclusively found in *C. austriacus* were predominantly associated with genes involved in response to regulation of cytosolic calcium concentrations, morphogenesis, and locomotory behavior (Figure 2a), whereas SNPs exclusively found in *C. melapterus* were enriched for functions involved in ion transmembrane transport, synaptic transmission, and social behavior (Figure 2b). Moreover, we found that only 385 SNPs (i.e. 4%) identified with the denovo_map.pl for *C. austriacus* approach failed to map to the reference genome, indicating that both approaches produced representative SNPs.

We next mapped SNP loci identified from *C. austriacus* with ref_map.pl to the gap-closed scaffolds in order to track the distribution of SNP density across its own genome. We found that (as expected) the number of SNPs mapping to the reference genome increased with scaffold size (R^2 = 0.71961; Fig. 3c). That said, 3055 scaffolds had < 10 SNPs that mapped back to them versus 115 scaffolds with > 10 SNPs that mapped back to them (Fig. 3a) despite a large mean scaffold size (~ 150 kb; Fig. 3b). These two categories (i.e. > 10 SNPs per scaffold versus < 10 SNPs per scaffold) represent sets of SNPs above and below the tolerance threshold, respectively, with a clear difference in GO enrichment between them (see Appendix S2).

Phylogenetic overview

In order to test the phylogenetic generality of these analyses, we assigned SNPs identified within each butterflyfish species, and the outgroup surgeonfish, to annotated versus non-annotated regions of the *C. austriacus* scaffolds. We found that approximately 16% more SNPs were assigned to non-annotated versus annotated regions of the genome, and that the absolute number of SNPs assigned varied among species (Fig. 4a). An estimate of the percent homology to the *C. austriacus* reference genome for all RAD data shows that *C. austriacus* and its closest relative, *C. melapterus*, mapped at a level of 86% and 87% respectively, with the rest of the butterflyfish mapping between 19% and 28% similarity; the outgroup surgeonfish had the lowest level of homology (~2%) (Fig. 4b). This phylogenetic decay is consistent with the inferred relationships among butterflyfish species revealed by a Maximum Likelihood consensus tree built with shared SNP loci (Fig. 4c).

Discussion

We compared analysis of RAD-Seq data using a *de novo* approach (i.e. without a genome) to analysis using a reference genome for a series of Red Sea and Arabian Sea butterflyfish. Our analyses indicate (1) a modest gap between the number of non-annotated versus annotated SNPs across all species, (2) an advantage of using genomic resources for closely related but not distantly related butterflyfish species based on the ability to assign putative gene function to SNPs, and (3) an enrichment of genes among sister butterflyfish taxa related to calcium transmembrane transport and binding. The RAD-Seq dataset we present here thus provides the most comprehensive estimate of genetic polymorphism to date for any reef fish, and further creates the potential to elevate the Chaetodontidae family to a 'model group' for future genomic studies. Overall, our results suggest that RAD-Seq approaches inherently provide valuable insight to population genomic questions for non-model organisms (i.e. those lacking reference genomes), but that detailed analysis of RAD-Seq data is improved when a reference genome is available from a closely related species (as per Pecoraro *et al.* 2015) or when apparent bias in the method of SNP isolation can be identified and accounted for.

Most RAD tags are thought to be randomly distributed across the genome based on the stochastic placement of restriction cut sites (Davey *et al.* 2013), and indeed we found no evidence for what we here refer to as SNP "hotspots" (e.g. Myers *et al.* 2005; also see Fig. 3). The genome assembly of *C. austriacus*, however, consists of gap-closed scaffolds and does not include a linkage map (DiBattista *et al.* 2016a); the exact location of each SNP relative to autosomal (or sex-linked) chromosomes is therefore not yet known. Moreover, we enabled the "select one SNP per read" filtering option in Stacks in order to comply with population genetic assumptions of independent loci, our intended downstream application for these data, which further reduced our ability to detect multiple SNPs sitting within 81 bp of each other. Despite this uncertainty, a large portion of the identified SNPs could be assigned to annotated regions of the genome for most of the butterflyfish species we considered (average number of annotated SNPs = 1199; Fig. 4a). We did, however, detect a strong bias with respect to covered gene functions (i.e. proteins containing ML stretches), which apparently stems from our choice of restriction enzyme. Restriction enzymes that target gene rich regions (e.g. Roda *et al.* 2013), or even specific primers flanking variable SNPs (Campbell *et al.* 2014), have previously been used when more specificity is required. This refinement provides an advantage for researchers who may wish to select a smaller number of SNP sites for genotyping with known spacing, location or function, but also a direct application to projects estimating connectivity and adaptability of fishes. However, such approaches are likely to introduce their own biases that need to be considered in subsequent analyses as we have shown here. Moreover, if a mutation has occurred in the cut-site of some (but not all) individuals, allelic dropout can occur, which has been shown to further bias the inference of genetic variability (Gautier *et al.* 2013). Without an available reference genome, these types of methodological biases are often missed.

We observed a clear advantage to using genomic resources for closely related but not distantly related butterflyfish species in terms of the ability to assign SNPs a putative function. This advantage would likely extend to other sister taxa of reef fish, and as such it would be ideal for studies focused on between-species comparisons *within* monophyletic lineages. Similar advantages were recently shown for the Pacific bluefin tuna (*Thunnus orientalis;* Pecoraro *et al.* 2015). For example, in our case, the availability of the *C. austriacus* genome sequence allowed for functional interrogation of the data using GO annotations within sister taxa, which revealed that variable SNPs were significantly enriched for certain gene functions (Fig. 2). These functions included genes associated with neurological processes, which might indicate potential differences in behavior or sensory perception for these specialized reef fish *within* or *between* the different species, and therefore represent plausible candidate genes for future functional interrogation. This is consistent with the hypothesis that detecting and interpreting visual and olfactory cues are highly important for butterflyfish given that most of these fish are monogamous and need to maintain pair bonds, recognize mates, and defend territories from conspecifics (e.g. Boyle & Tricas 2014). More importantly we found that genes associated with calcium transmembrane transport and binding were significantly overrepresented in both *C. austriacus* and *C. melapterus*, suggesting potential convergent adaptations to the prevailing Red Sea and Arabian Sea environments, respectively. A plausible alternative is that these SNPs are not themselves adaptive but linked to other regions of the genome that were not sequenced, resulting in a consistent association and statistically significant enrichment of sequence variation and specific gene functions. If we assume that these functions are under selection, calcium/chlorine ratios are remarkably similar in the Red Sea and adjacent Indian Ocean (Krumgalz 1982), although the Gulf of Aqaba is much lower, suggesting that adaptation to these conditions may dictate which gene functions are conserved. The assimilation of calcium from seawater is an important process for the formation of structural skeleton in corals, protective shells for many of the marine invertebrates, as well as promotes optimal larval development in fish via ossification and gill formation (e.g. Malvezzi *et al.* 2015). Further study of additional fish from the Red Sea and Arabian Sea region could test the hypothesis that genetic adaptations allow species to handle the unique chemical challenges present in regional waters, and particularly could provide some insight to the mechanisms underlying high rates of endemism in the region.

We note that the percentage of reads mapping to the reference genome was actually higher for *C. melapterus* (87.5%) versus the genomically enabled *C. austriacus* (86.4%)(Fig. 4b). Although this result may appear unusual, these species are closely related to each other (Fig. 4c), have only recently diverged (~50 Kya), share mitochondrial DNA haplotypes (see Waldrop *et al.* 2016), and likely hybridize in regions of overlap based on similar behavior observed in nearby seas (DiBattista *et al.* 2015). Recent and/or on-going genetic exchange suggests that introgression across the nuclear genome must be high for these two species (see Waldrop *et al.* 2016). Moreover, despite phylogenetic similarity, *C. melapterus* had lower levels of polymorphism (0.78% versus 2.3%, respectively) and fewer restriction sites recovered (5913 versus 8812 variable loci passing strict criteria, respectively) compared to *C. austriacus*, and yet the average depth of coverage for the former increased from 12.2X to 14.9X when mapping to the reference genome (Table 1). We may therefore have achieved marginally better mapping results for *C. melapterus* simply because we had a lower probably of discarding reads with ref_map.pl based on increased homogeneity and higher per locus coverage. We additionally attribute the large range of pre- and post-filtered loci among species to differences in sample size, in addition to stochastic factors such as variable DNA quality for each sample and library preparation protocol.

Conclusion

The rapid identification of a large number of SNPs within populations likewise holds promise for a number of evolutionary-oriented studies. While the number of publicly available genomes is steadily increasing, our results suggest that increasing phylogenetic distance decreases the utility of genomes, but that this remains an important means to identify apparent bias in SNP isolation approaches as well as assess conserved biological processes. Substantial advances in our understanding of evolutionary mechanisms may therefore be possible without the substantial investment of resources required for full comparative genomic studies.

Acknowledgements. This research was supported by the KAUST Office of Competitive Research Funds (OCRF) under Award No. CRG-1-2012-BER-002 and baseline research funds to M.L.B., as well as a National Geographic Society Grant 9024-11 to J.D.D. For support in Socotra, we kindly thank the Ministry of Water and Environment of Yemen, staff at the Environment Protection Authority (EPA) Socotra, and especially Salah Saeed Ahmed, Fouad Naseeb and Thabet Abdullah Khamis, as well as Ahmed Issa Ali Affrar from Socotra Specialist Tour for handling general logistics. For logistic support elsewhere, we thank Eric Mason at Dream Divers in Saudi Arabia; the Red Sea State Government and The Red Sea University in Sudan, as well as Equipe Cousteau including N. Hussey, S. Kessel, C. Scarpellini, and M. Younis; Nicolas Prévot at Dolphin Divers and the crew of the M/V Deli in Djibouti; the KAUST Coastal and Marine Resources Core Lab and Amr Gusti; and the Ministry of Agriculture and Fisheries in Oman including Abdul Karim. For specimen collections, we thank Tilman Alpermann, Giacomo Bernardi, Brian Bowen, Camrin Braun, Richard Coleman, Michelle Gaither, Jean-Paul Hobbs, Jennifer McIlwain, Gerrit Nanninga, Mark Priest, Luiz Rocha, Tane Sinclair-Taylor, and members of the Reef Ecology Lab at KAUST. For assistance with bench work at KAUST we thank Craig Michell. We also acknowledge important contributions from Luiz Rocha and David Catania for assistance with specimen archiving at the California Academy of Sciences, Tane Sinclair-Taylor for formatting figures at later stages, and the KAUST Bioscience Core Laboratory with Sivakumar Neelamegam and Hicham Mansour for their assistance with Illumina sequencing.

References

Alexa A, Rahnenfuhrer J (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2.22.0.

Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA (2016) Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, **17,** 81-92.

Bailey R (2009) *Ecosystem Geography: From Ecoregions to Sites*. 2nd edition, New York, Springer Publishing.

Bellwood DR, Klanten S, Cowman PF, Pratchett MS, Konow N, van Herwerden L (2010) Evolutionary history of the butterflyfishes (f: Chaetodontidae) and the rise of coral feeding fishes. *Journal of Evolutionary Biology*, **23,** 335-349.

Bernal MA, Gaither MR, Simison WB, Rocha LA (2016) Introgression and selection shaped the evolutionary history of sympatric sister-species of coral reef fishes (genus: *Haemulon*). *Molecular Ecology*. DOI: 10.1111/mec.13937.

Boyle KS, Tricas TC (2014) Discrimination of mates and intruders: visual and olfactory cues for a monogamous territorial coral reef butterflyfish. *Animal Behaviour*, **92,** 33-43.

Campbell NR, Harmon SA, Narum SR (2014) Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular Ecology Resources*, **15,** 855-867.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, **1,** 171-182.

Catchen J, Bassham S, Wilson T, Currey M, O'Brien C, Yeates Q, Cresko WA (2013a) The population structure and recent colonization history of Oregon threespine stickleback determined using restriction-site associated DNA-sequencing. *Molecular Ecology*, **22,** 2864-2883.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013b) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, **22,** 3124-3140.

Cole AJ, Pratchett MS, Jones GP (2008) Diversity and functional importance of coral-feeding fishes on tropical coral reefs. *Fish and Fisheries*, **9,** 286-307.

Corander J, Majander KK, Cheng L, Merilä J (2013) High degree of cryptic population differentiation in the Baltic Sea herring *Clupea harengus*. *Molecular Ecology*, **22,** 2931-2940.

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12,** 499-510.

Davey JW, Cezard T, Fuentes-Utrilla P, Eland C, Gharbi K, Blaxter ML (2013) Special features of RAD Sequencing data: implications for genotyping. *Molecular Ecology*, **22,** 3151-3164.

DiBattista JD, Wang X, Saenz-Agudelo P, Piatek M, Aranda M, Berumen ML (2016a) Draft genome of an iconic Red Sea reef fish, the blacktail butterflyfish (*Chaetodon austriacus*): current status and its characteristics. *Molecular Ecology Resources*. Online Early, doi:10.1111/1755-0998.12588.

DiBattista JD, Roberts M, Bouwmeester J, Bowen BW, Coker DF, Lozano-Cortés DF, Choat JH, Gaither MR, Hobbs JP, Kahil M, Kochzius M, Myers R, Paulay G, Robitzch V, Saenz-Agudelo P, Salas E, Sinclair-Taylor TH, Toonen RJ, Westneat M, Williams S, Berumen ML (2016b) A review of contemporary patterns of endemism for shallow water reef fauna in the Red Sea. *Journal of Biogeography*, **43,** 423-439.

DiBattista JD, Choat JH, Gaither MR, Hobbs JP, Lozano-Cortés DF, Myers R, Paulay G, Rocha LA, Toonen RJ, Westneat M, Berumen ML (2016c) On the origin of endemic species in the Red Sea. *Journal of Biogeography*, **43,** 13-30.

DiBattista JD, Rocha LA, Hobbs JP, He S, Priest MA, Sinclair-Taylor TH, Bowen BW, Berumen ML (2015) When biogeographical provinces collide: hybridization of reef fishes at the crossroads of marine biogeographical provinces in the Arabian Sea. *Journal of Biogeography*, **42,** 1601-1614.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2009) Geneious v4.8. Available at: http://www.geneious.com/

Gaither MR, Bernal MA, Coleman RR, Bowen BW, Jones SA, Simison WB, Rocha LA (2015) Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Molecular Ecology*, **24,** 1543-1557.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet JM, Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22,** 3165-3178.

Hand BK, Hether TD, Kovach RP, Muhlfeld CC, Amish SJ, Boyer MC, O'Rourke SM, Miller MR, Lowe WH, Hohenlohe PA, Luikart G (2015) Genomics and introgression: Discovery and mapping of thousands of species-diagnostic SNPs using RAD sequencing. *Current Zoology*, **61,** 146-154.

Henning F, Lee HJ, Franchini P, Meyer A (2014) Genetic mapping of horizontal stripes in Lake Victoria cichlid fishes: benefits and pitfalls of using RAD markers for dense linkage mapping. *Molecular Ecology*, **23,** 5224-5240.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6,** e1000862.

Jackson AM, Semmens BX, de Mitcheson YS, Nemeth RS, Heppell SA, Bush PG, Aguilar-Perera A, Claydon JAB, Calosso MC, Sealey KS, Schärer MT, Bernardi G (2014) Population structure and phylogeography in Nassau grouper (*Epinephelus striatus*), a mass-aggregating marine fish. *PLoS ONE*, **9,** e97508.

Krumgalz BS (1982) Calcium distribution in the world ocean waters. *Oceanologica Acta*, **5,** 121-128.

Kosuri S, Church, GM (2014) Large-scale *de novo* DNA synthesis: technologies and applications. *Nature Methods*, **11,** 499-507.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, **10,** R25.

Lawton RJ, Pratchett MS, Delbeek JC (2013) Harvesting of butterflyfishes for aquarium and artisanal fisheries. In: *Biology of Butterflyfishes* (eds: Pratchett MS, Berumen ML, Kapoor BG), 269-291.

Liew YJ, Aranda M, Voolstra CR (2016) Reefgenomics.Org – a repository for marine genomics data. *Database*, 1-4.

Malvezzi AJ, Murray CS, Feldheim KA, DiBattista JD, Garant D, Gobler CJ, Chapman DD, Baumann H (2015) A quantitative genetic approach to assess the evolutionary potential of a coastal marine fish to ocean acidification. *Evolutionary Applications*, **8,** 352-362.

Nakamura Y, Mori K, Saitoh K, Oshima K, Mekuchi M, Sugaya T, et al. (2013) Evolutionary changes of multiple visual pigment genes in the complete genome of Pacific bluefin tuna. *Proceedings of the National Academy of Sciences USA*, **110,** 11061-11066.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. *Science*, **310,** 321-324.

Pecoraro C, Babbucci M, Villamor A, Franch R, Papetti C, Leroy B, Ortega-Garcia S, Muir J, Rooker J, Arcoha F, Murua H, Zudaire I, Chassot E, Bodin N, Tinti F, Bargelloni L, Cariani A (2015) Methodological assessment of 2b-RAD genotyping technique for population structure inferences in yellowfin tuna (*Thunnus albacares*). *Marine Genomics*, Online Early.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS ONE*, **7,** e37135.

Picq S, McMillan WO, Puebla O (2016) Population genomics of local adaptation versus speciation in coral reef fishes (*Hypoplectrus* spp, Serranidae). *Ecology and Evolution*, **6,** 2109-2124.

Puebla O, Bermingham E, McMillan WO (2014) Genomic atolls of differentiation in coral reef fishes (*Hypoplectrus* spp., Serranidae). *Molecular Ecology*, **23,** 5291-5303.

Raitsos DE, Pradhan Y, Brewin RJW, Stenchikov G, Hoteit I (2013) Remote sensing the phytoplankton seasonal succession of the Red Sea. PloS ONE, e64909.

Read TD, Petit III RA, Joseph SJ, Alam MT, Weil R, Ahmad M, Bhimani R, Vuong JS, Haase CP, Webb DH, Dove AD (2015) Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: *Rhincodon typus* Smith 1828. *PeerJ PrePrints*, e1036.

Ree RH, Hipp AL (2015) Inferring phylogenetic history from restriction site associated DNA (RADseq). *Next-generation Sequencing in Plant Systematics*, 181-204.

Roda F, Ambrose L, Walter GM, Liu HL, Schaul A, Lowe A, Pelser PB, Prentis P, Rieseberg LH, Ortiz-Barrientos D (2013) Genomic evidence for the parallel evolution of coastal forms in the *Senecio lautus* complex. *Molecular Ecology*, **22,** 2941-2952.

Saenz-Agudelo P, DiBattista JD, Piatek MJ, Gaither MR, Harrison HB, Nanninga GB, Berumen ML (2015) Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes. *Molecular Ecology*, **24,** 6241-6255.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22,** 2688-2690.

Stockwell BL, Larson WA, Waples RK, Abesamis RA, Seeb LW, Carpenter KE (2016) The application of genomics to inform conservation of a functionally important reef fish (*Scarus niger*) in the Philippines. *Conservation Genetics*, **17,** 239-249.

Tariel J, Longo GC, Bernardi G (2016) Tempo and mode of speciation in *Holacanthus* angelfishes based on RADseq markers. *Molecular Phylogenetics and Evolution*, **98,** 84-88.

van de Peer Y (2004) Tetraodon genome confirms *Takifugu* findings: most fish are ancient polyploids. *Genome Biology*, **5,** 250.

Wabnitz C (2003) *From Ocean to Aquarium: The Global Trade in Marine Ornamental Species* (No. 17) Cambridge, UNEP/Earthprint.

Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O (2013) Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Molecular Ecology*, **22,** 787-798.

Waldrop E, Hobbs JP, Randall JE, DiBattista JD, Rocha LA, Kosaki RK, Berumen ML, Bowen BW (2016) Phylogeography, population structure and evolution of coral-feeding butterflyfishes (Subgenus *Corallochaetodon*). *Journal of Biogeography*, **43,** 1116-1129.

Willette DA, Allendorf FW, Barber PH, Barshis DJ, Carpenter KE, Crandall ED, Cresko, WA, Fernandez-Silva, I, Matz, MV, Meyer, E, Santos, MD, Seeb LW, Seeb JE (2014) So, you want to use next-generation sequencing in marine systems? Insight from the Pan-Pacific Advanced Studies Institute. *Bulletin of Marine Science*, **90,** 79-122.

Data Accessibility

Raw RAD reads (FASTQ format) are available in the NCBI repository, BioProject: PRJNA292048.

Genome browser URL: http://caus.reefgenomics.org

Filtered RAD reads in .vcf format are available from Dryad: doi:10.5061/dryad.f09rh.

Author contributions

J.D.D., P.S.-S., and M.L.B. conceived of and designed the RAD-Seq study. J.D.D., P.S.-S., and M.J.P. produced and analyzed the RAD libraries. X.W. and M.A. performed GO enrichment analyses. All authors developed the manuscript and approve of the final paper.

Fig. 1 Map of the Red Sea and Arabian Sea, including study species and collection sites.

Fig. 2 Top ten highest enriched ($P < 0.01$) gene ontologies for (a) *Chaetodon austriacus* and (b) *C. melapterus* based on annotated SNPs produced using a ddRAD protocol that have protein and GO information available. Gene ontology categories include molecular function, biological process, and cellular component.

Fig. 3 (a) Frequency of SNPs per scaffold for *Chaetodon austriacus* mapping to the assembled genome, (b) frequency distribution of scaffold size (in million base pairs), and (c) number of SNPs as a function of scaffold size using a ddRAD protocol and "ref_map.pl" option in Stacks. Mean = black dashed line; + 2 standard deviations = green dashed lines; tolerance interval = red dashed lines.

Fig. 4 (a) Number of annotated versus non-annotated SNPs (total number of SNPs, i.e. annotated plus non-annotated, is in brackets), (b) percentage of reads mapping to the *Chaetodon austriacus* reference genome using Bowtie for Red Sea and Arabian Sea resident butterflyfish (and an outgroup surgeonfish), and (c) Maximum Likelihood (ML) phylogenetic relationship among species. Mapping parameters used for Bowtie were -n = 3 and –k = 1. A SNP was considered annotated if located in any region of the scaffold identified as protein coding; a non-annotated SNP was any SNP that did not meet this criteria.

Fig. 1

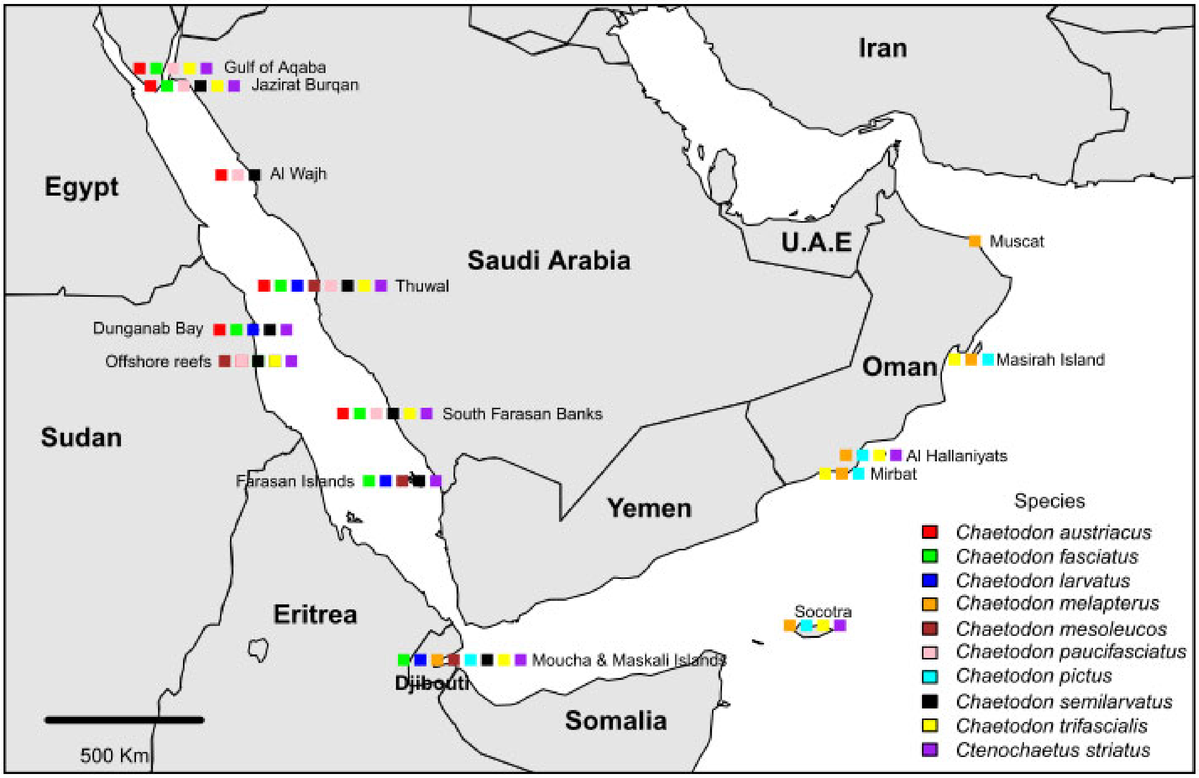
Iran
Gulf of Aqaba
Jazirat Burqan
Al Wajh
Egypt
Saudi Arabia
U.A.E
Muscat
Thuwal
Dunganab Bay
Offshore reefs
Oman
Masirah Island
South Farasan Banks
Sudan
Al Hallaniyats
Mirbat
Farasan Islands
Yemen
Species
Chaetodon austriacus
Chaetodon fasciatus
Chaetodon larvatus
Chaetodon melapterus
Chaetodon mesoleucos
Chaetodon paucifasciatus
Chaetodon pictus
Chaetodon semilarvatus
Chaetodon trifascialis
Ctenochaetus striatus
Eritrea
Socotra
Moucha & Maskali Islands
Djibouti
Somalia
500 Km

Fig. 2a

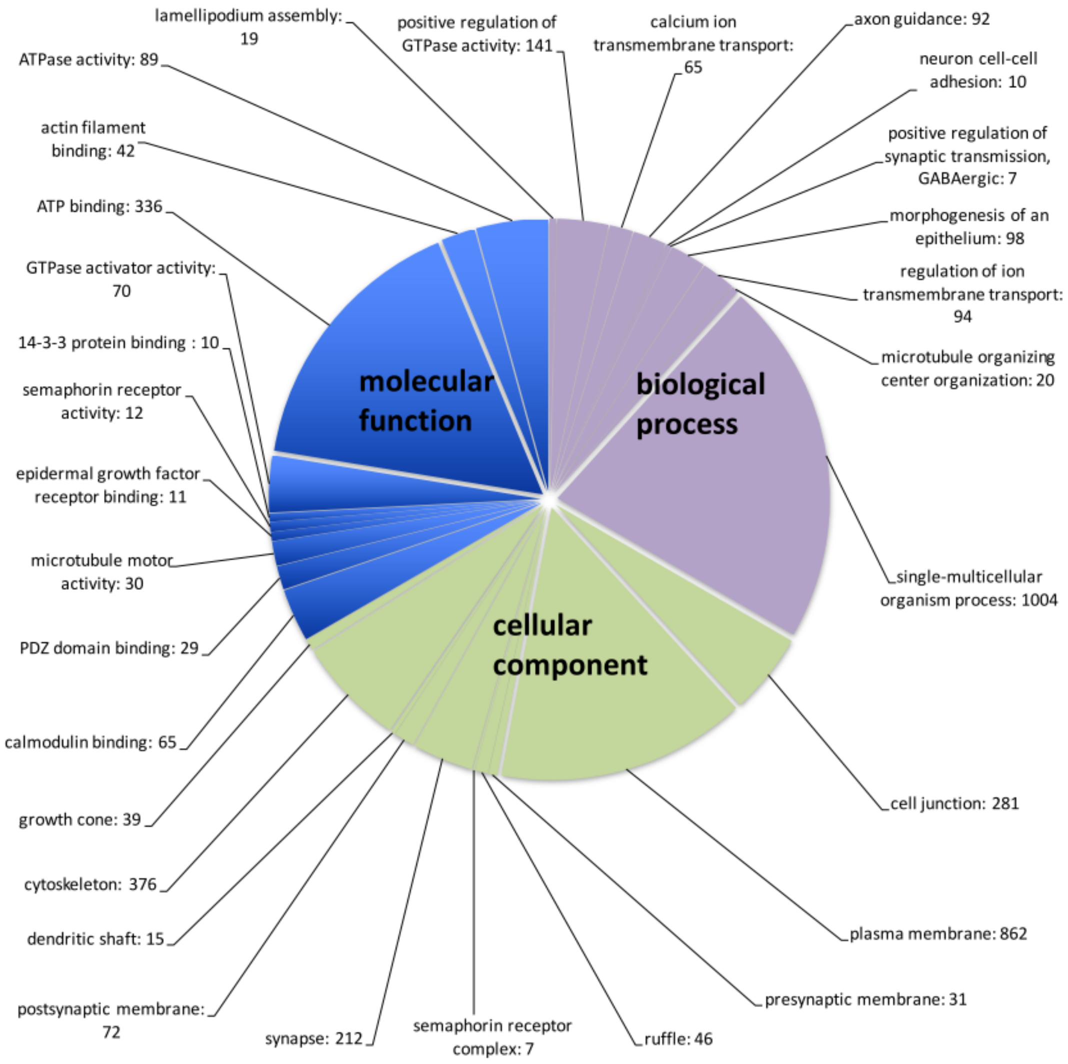
lamellipodium assembly: 19
positive regulation of GTPase activity: 141
calcium ion transmembrane transport: 65
axon guidance: 92
ATPase activity: 89
neuron cell-cell adhesion: 10
actin filament binding: 42
positive regulation of synaptic transmission, GABAergic: 7
ATP binding: 336
morphogenesis of an epithelium: 98
GTPase activator activity: 70
regulation of ion transmembrane transport: 94
14-3-3 protein binding : 10
microtubule organizing center organization: 20
semaphorin receptor activity: 12
molecular function
biological process
epidermal growth factor receptor binding: 11
microtubule motor activity: 30
single-multicellular organism process: 1004
PDZ domain binding: 29
cellular component
calmodulin binding: 65
cell junction: 281
growth cone: 39
cytoskeleton: 376
plasma membrane: 862
dendritic shaft: 15
presynaptic membrane: 31
postsynaptic membrane: 72
synapse: 212
semaphorin receptor complex: 7
ruffle: 46

Fig. 2b

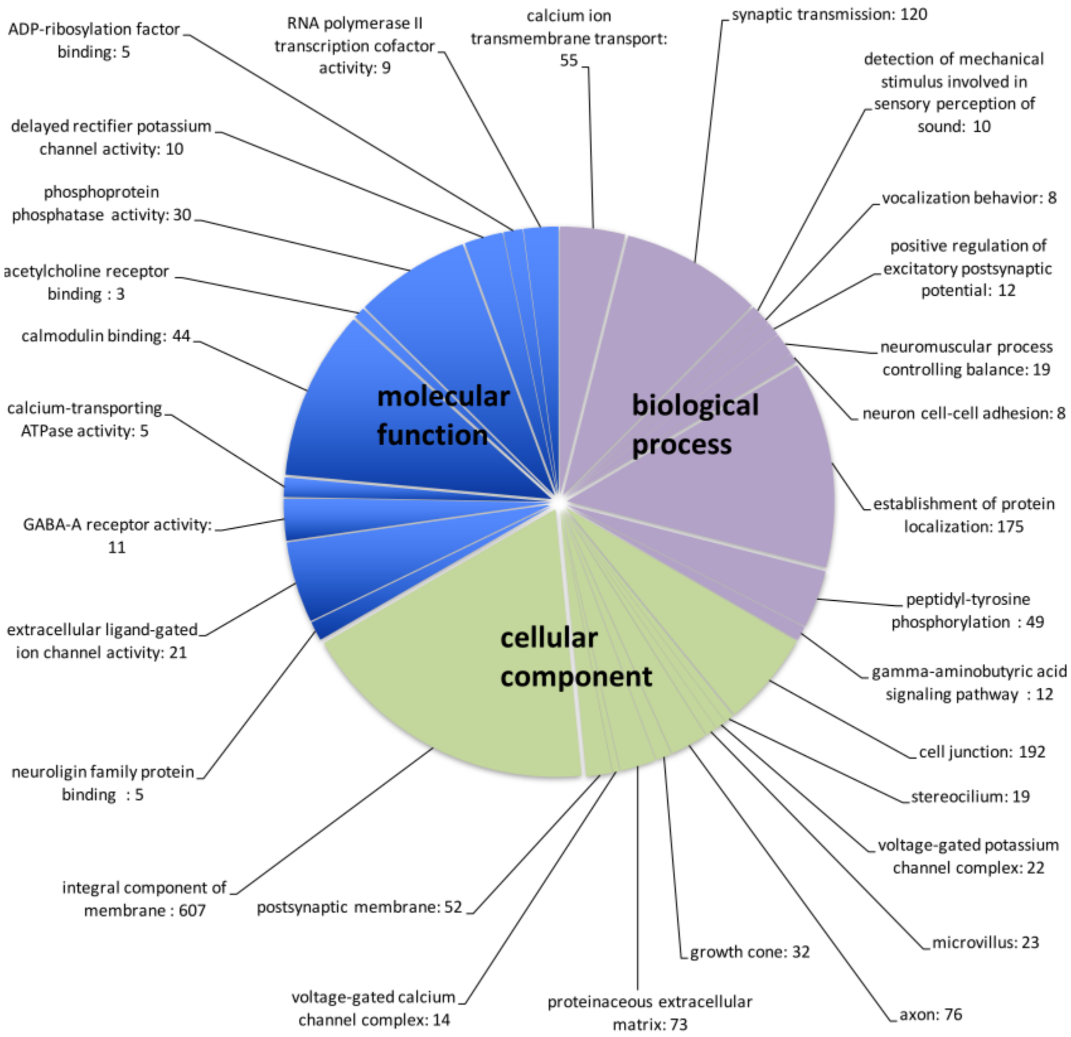
ADP-ribosylation factor binding: 5
RNA polymerase II transcription cofactor activity: 9
calcium ion transmembrane transport: 55
synaptic transmission: 120
detection of mechanical stimulus involved in sensory perception of sound: 10
delayed rectifier potassium channel activity: 10
phosphoprotein phosphatase activity: 30
vocalization behavior: 8
positive regulation of excitatory postsynaptic potential: 12
acetylcholine receptor binding : 3
calmodulin binding: 44
neuromuscular process controlling balance: 19
molecular function
biological process
neuron cell-cell adhesion: 8
calcium-transporting ATPase activity: 5
establishment of protein localization: 175
GABA-A receptor activity: 11
peptidyl-tyrosine phosphorylation : 49
extracellular ligand-gated ion channel activity: 21
cellular component
gamma-aminobutyric acid signaling pathway : 12
cell junction: 192
neuroligin family protein binding : 5
stereocilium: 19
voltage-gated potassium channel complex: 22
integral component of membrane : 607
postsynaptic membrane: 52
microvillus: 23
growth cone: 32
voltage-gated calcium channel complex: 14
proteinaceous extracellular matrix: 73
axon: 76

Fig. 3a

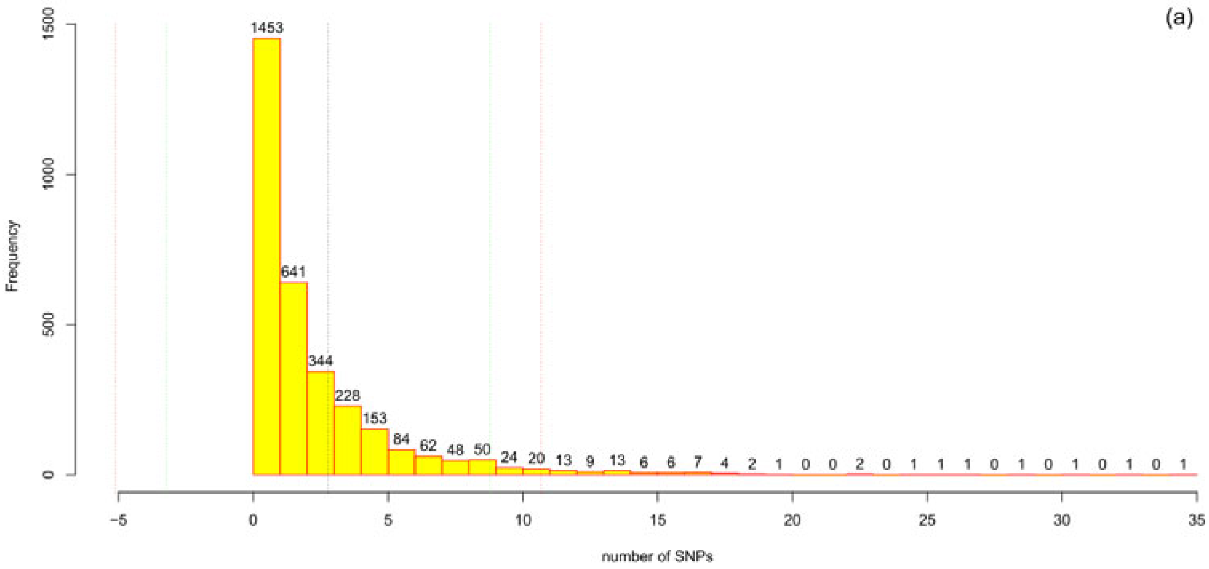
(a)
Frequency
1500
1000
500
0
1453
641
344
228
153
84
62
48
50
24
20
13
9
13
6
6
7
4
2
1
0
0
2
0
1
1
1
0
1
0
1
0
1
0
1
-5
0
5
10
15
20
25
30
35
number of SNPs

Fig. 3b

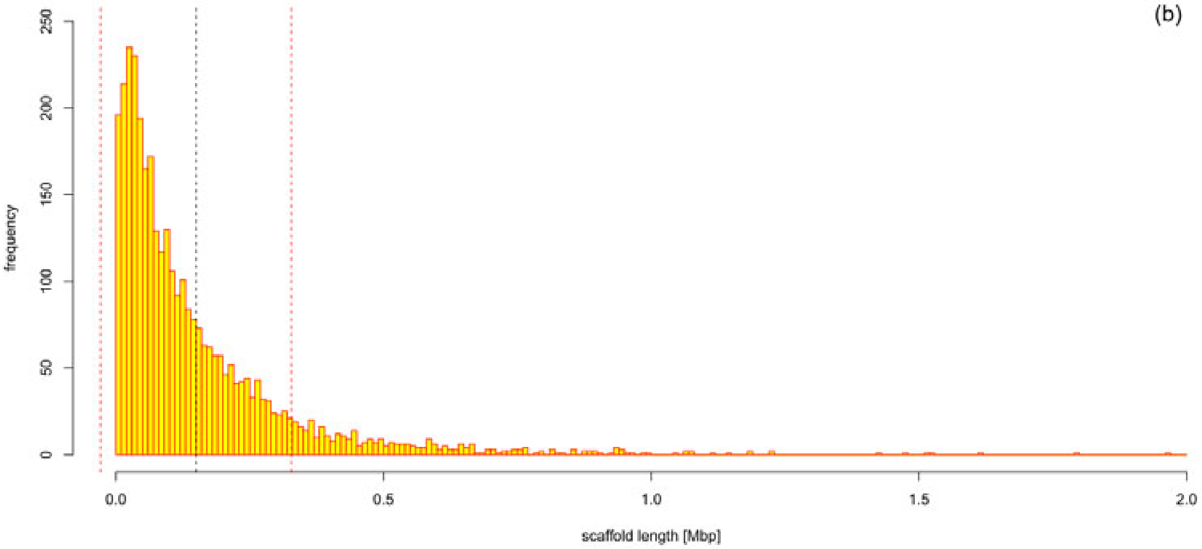
(b)
frequency
0
50
100
150
200
250
0.0
0.5
1.0
1.5
2.0
scaffold length [Mbp]

Fig. 3c

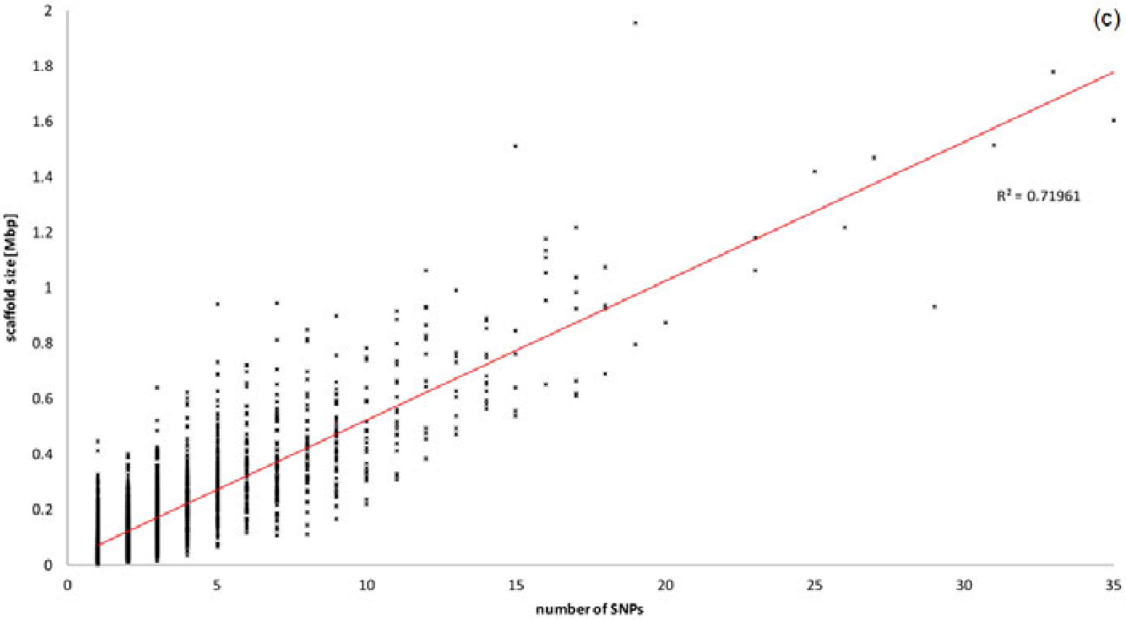
(c)
R² = 0.71961
scaffold size [Mbp]
number of SNPs
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
0
5
10
15
20
25
30
35

Fig. 4a

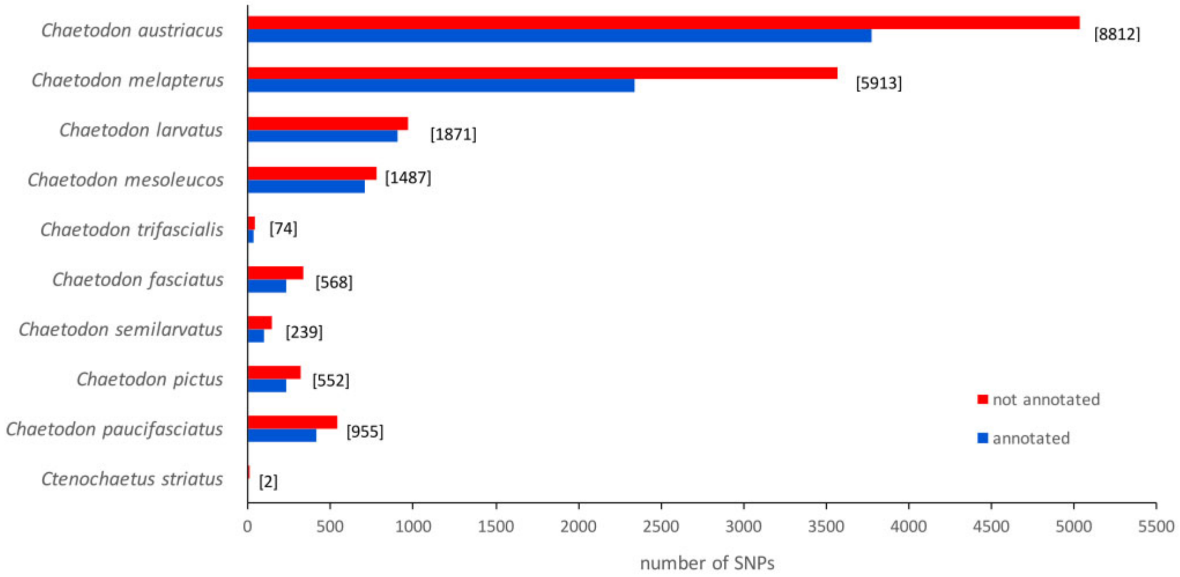
Chaetodon austriacus
Chaetodon melapterus
Chaetodon larvatus
Chaetodon mesoleucos
Chaetodon trifascialis
Chaetodon fasciatus
Chaetodon semilarvatus
Chaetodon pictus
Chaetodon paucifasciatus
Ctenochaetus striatus
[8812]
[5913]
[1871]
[1487]
[74]
[568]
[239]
[552]
[955]
[2]
🟥 not annotated
🟦 annotated
0
500
1000
1500
2000
2500
3000
3500
4000
4500
5000
5500
number of SNPs

Fig. 4b

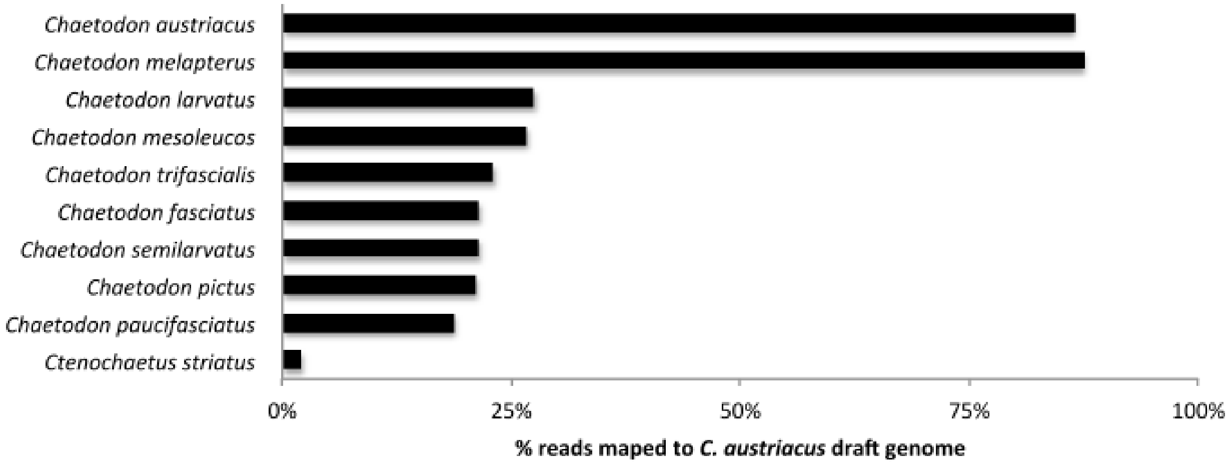
Chaetodon austriacus
Chaetodon melapterus
Chaetodon larvatus
Chaetodon mesoleucos
Chaetodon trifascialis
Chaetodon fasciatus
Chaetodon semilarvatus
Chaetodon pictus
Chaetodon paucifasciatus
Ctenochaetus striatus
0%
25%
50%
75%
100%
% reads maped to C. austriacus draft genome

Fig. 4c

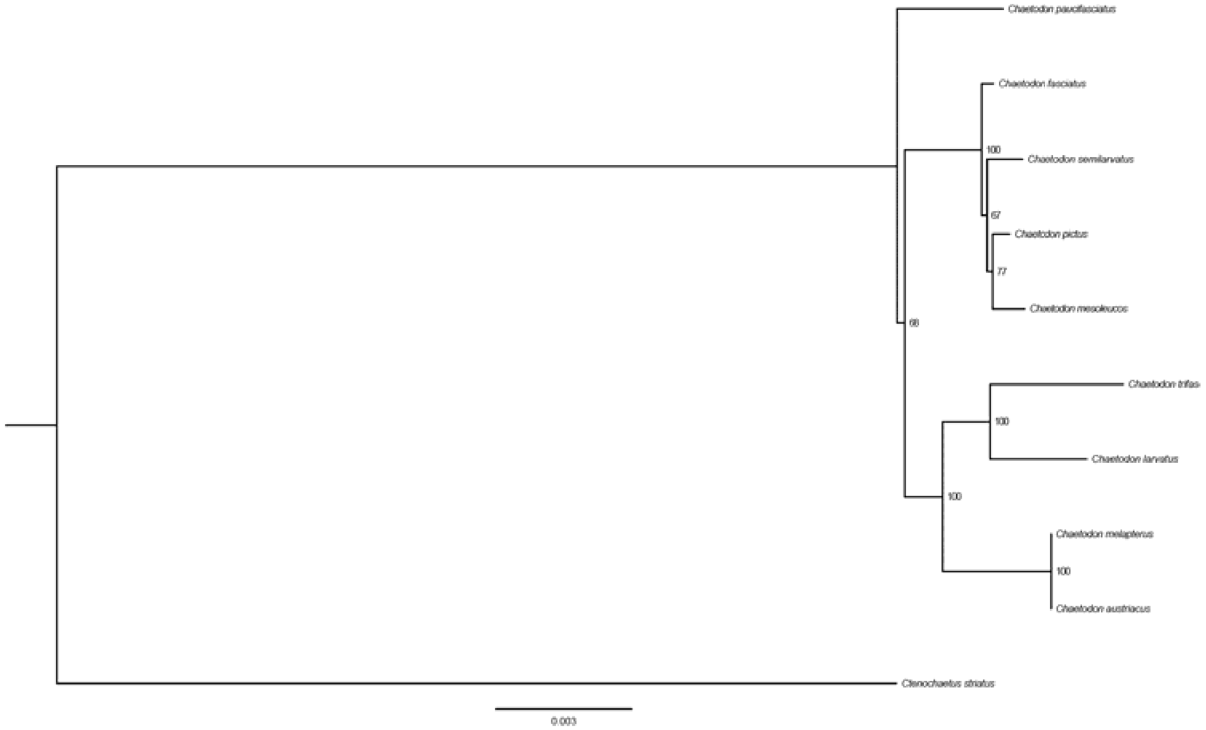
Chaetodon paucifasciatus
Chaetodon fasciatus
100
Chaetodon semilarvatus
67
Chaetodon pictus
77
Chaetodon mesoleucos
68
Chaetodon trifas-
100
Chaetodon larvatus
100
Chaetodon melapterus
100
Chaetodon austriacus
Ctenochaetus striatus
0.003

Table 1. Stacks results of ddRAD data for nine species of Red Sea and Arabian Sea resident butterflyfish (and an outgroup surgeonfish) before and after quality filtering using denovo_map.pl and ref_map.pl options in Stacks. In all cases, 12 individuals were sampled per population. Population filtering parameters included: 1) minor allele frequency > 0.05, 2) the locus had to be genotyped in at least 80% of individuals, and 3) the locus was present in all (or all but one) populations, which are represented by numbers outside and inside the parentheses, respectively, for “# of variable loci passing filter”.

			denovo_map.pl				ref_map.pl			
Species	Sample size	Number of populations (geographic range of sampling)	# of reads used	av. depth of coverage	# of loci	# of variable loci passing	# of reads used	av. depth of coverage	# of loci	# of variable loci passing
Chaetodon austriacus (exquisite butterflyfish)	84	7 (Gulf of Aqaba to South Farasan Banks)	106,207,151	13.7	289,504	8,842 (10,711)	91,785,299	13.4	194,235	8,812 (10,780)
Chaetodon fasciatus (Red Sea racoon butterflyfish)	96	8 (Gulf of Aqaba to Djibouti)	93,749,728	10.2	262,590	1,343 (2,650)	20,137,784	12.1	42,029	568 (697)
Chaetodon larvatus (hooded butterflyfish)	60	5 (Thuwal to Djibouti)	85,686,503	12.6	298,347	10,028 (12,393)	23,585,269	13.1	64,641	1,871 (2,179)
Chaetodon melapterus (Arabian butterflyfish)	72	6 (Djibouti to Muscat)	97,774,030	12.2	273,706	2,275 (4,384)	85,591,171	14.9	184,187	5,913 (7,761)
Chaetodon mesoleucos (white-face butterflyfish)	48	4 (Thuwal to Djibouti)	54,334,800	11.9	218,077	7,608 (11,151)	14,493,353	13.3	50,966	1,487 (1,806)
Chaetodon paucifasciatus (Eritrean butterflyfish)	72	6 (Gulf of Aqaba to South Farasan Banks)	100,273,743	16.2	363,221	10,257 (13,539)	18,726,021	16.1	44,073	955 (1,145)

Chaetodon pictus (horseshoe butterflyfish)	60	5 (Djibouti to Masirah Island)	46,939,818	10.9	225,976	2,073 (4,131)	9,907,802	12.8	40,368	552 (759)
Chaetodon semilarvatus (bluecheek butterflyfish)	96	8 (Jazirat Burqan to Djibouti)	89,373,543	12.3	189,554	1,270 (2,053)	19,177,775	13.9	37,766	239 (338)
Chaetodon trifascialis (chevron butterflyfish)	108	9 (Gulf of Aqaba to Masirah Island)	110,578,849	10.7	239,244	476 (1,271)	25,527,940	12.2	45,670	74 (331)
Outgroup										
Ctenochaetus striatus (striated surgeonfish)	120	10 (Gulf of Aqaba to Al Hallaniyats)	126,536,786	9.8	516,163	415 (1,508)	2,573,428	12.4	8,516	8 (27)

Abbreviation: av = average; SNP, single nucleotide polymorphism.