nature ecology & evolution

ARTICLES

https://doi.org/10.1038/s41559-017-0413-2

Quantitative species-level ecology of reef fish larvae via metabarcoding

Naama Kimmerling[1,2], Omer Zuqert[3], Gil Amitai[3], Tamara Gurevich[2], Rachel Armoza-Zvuloni[2,8], Irina Kolesnikov[2], Igal Berenshtein[1,2], Sarah Melamed[3], Shlomit Gilad[4], Sima Benjamin[4], Asaph Rivlin[2], Moti Ohavia[2], Claire B. Paris[5], Roi Holzman[2,6*], Moshe Kiflawi[2,7*] and Rotem Sorek[3*]

The larval pool of coral reef fish has a crucial role in the dynamics of adult fish populations. However, large-scale species-level monitoring of species-rich larval pools has been technically impractical. Here, we use high-throughput metabarcoding to study larval ecology in the Gulf of Aqaba, a region that is inhabited by >500 reef fish species. We analysed 9,933 larvae from 383 samples that were stratified over sites, depth and time. Metagenomic DNA extracted from pooled larvae was matched to a mitochondrial cytochrome *c* oxidase subunit I barcode database compiled for 77% of known fish species within this region. This yielded species-level reconstruction of the larval community, allowing robust estimation of larval spatio-temporal distributions. We found significant correlations between species abundance in the larval pool and in local adult assemblages, suggesting a major role for larval supply in determining local adult densities. We documented larval flux of species whose adults were never documented in the region, suggesting environmental filtering as the reason for the absence of these species. Larvae of several deep-sea fishes were found in shallow waters, supporting their dispersal over shallow bathymetries, potentially allowing Lessepsian migration into the Mediterranean Sea. Our method is applicable to any larval community and could assist coral reef conservation and fishery management efforts.

Coral reefs are complex ecosystems with important ecological, cultural and economic values. For over two decades, human activities have been the cause for extensive deterioration of coral reefs worldwide. It has been estimated that 27% of the world's reefs have been substantially degraded through rising sea surface temperatures, fishing, tourism, coastal development, eutrophication and other human-associated activities[1–3]. As a result, considerable efforts have been invested to monitor, preserve and restore coral reefs worldwide[4,5], with special attention directed at the reef-associated fish that form an integral part of this highly diverse ecosystem[6,7].

Coral reef fish are essentially sedentary within their reef system as adults, but their larvae are pelagic[8,9]. The pelagic larval stage extends up to several weeks, during which the larvae drift to the open ocean and can disperse over considerable distances. As such, the larvae of reef fish are largely responsible for the replenishment and stability of adult populations, as well as for the colonization of new habitats. Thus, understanding the ecology of these larvae is key for long-term efforts to conserve and manage coral reefs[10,11]. A crucial component in our ability to understand the dynamics, distribution and structure of coral reef fish communities would be to monitor the composition and structure of their larval pools[9,12,13]. However, although the importance of the larval pool has been long recognized, species-level data on larval pool structure and dynamics are largely lacking, mostly owing to technological limitations[14].

The paucity of species-level information on larval distributions stems mainly from the high degree of morphological similarity between young larvae of different fish species[14], which often limits morphology-based classification of fish larvae to the taxonomic family level only. Moreover, the level of expertise and time investment required for morphology-based classification renders the taxonomic classification of large numbers of larvae both challenging and impractical[14,15]. Hence, to date, the spatio-temporal structure of the larval pool has been studied only at coarse-grained taxonomic resolution: mostly at the family level and rarely below the genus level (that is, identification of the taxonomic family to which the larva belongs, but not the actual species)[14,16,17], with rare exceptions restricted to a few focal species[18,19].

DNA barcoding, using a segment (~650 bp) of the mitochondrial cytochrome *c* oxidase subunit I (COI) as a barcode, has been suggested as a means for species-level identification of individual organisms[20–22]. Successful implementation of the approach requires a pre-compilation of a database that contains reference barcodes of the species in the region[19]. Substantial global effort is underway to map the COI barcodes of all the fish in the world, as part of the Fish Barcode of Life Initiative (FISH-BOL) that, so far, covers ~11,000 of the >32,000 known species[21]. Species identification using COI barcodes typically requires per-individual amplification and sequencing of the COI region (in our case, for each individual larva)[22], which becomes impractical in studies that include large numbers of larvae. Consequently, some studies have adopted a meta-barcoding approach, in which the COI region is PCR-amplified from pooled DNA that is extracted from all individuals within the sample[23]. Such an approach was shown to be prone to strong amplification biases due to differences in primer annealing efficiencies and artifactual

[1]Marine Biology Program, Eilat Campus, Department of Life Sciences, Ben-Gurion University, Eilat, Israel. [2]Interuniversity Institute for Marine Sciences, Eilat, Israel. [3]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel. [4]The Nancy and Stephen Grand Israel National Center for Personalized Medicine (INCPM), Weizmann Institute of Science, Rehovot, Israel. [5]Department of Ocean Sciences, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA. [6]Department of Zoology, Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel. [7]Department of Life Sciences, Ben-Gurion University, Eilat, Israel. Present address: [8]Dead Sea and Arava Science Center, Yotvata, Israel. Naama Kimmerling, Omer Zuqert and Gil Amitai contributed equally to this work. *e-mail: holzman@post.tau.ac.il; mkiflawi@bgu.ac.il; rotem.sorek@weizmann.ac.il

barcode chimeras[24,25], with the level of bias depending on the primers used and the number of PCR cycles[26]. These drawbacks prevent accurate quantitative assessment of species abundance and lead to frequent species misidentification[27,28]. Indeed, universal primer sets for fish COI amplification show a variable degree of alignment accuracy to COI sequences that are recovered from full mitochondrial genomes (Supplementary Table 1 and see Methods).

Here, we introduce a shotgun metagenomic approach, which can be applied feasibly to multiple samples to generate quantitative species-level estimates of larval abundance across space and time. We applied the approach to 383 ichthyoplankton samples, which were obtained over 11 consecutive months in the northern Gulf of Aqaba. This complex reef habitat, which is the northernmost extension of the Red Sea, contains a rich reef-associated ichthyofauna that includes ~1,100 species, primarily of Indo-Pacific origin[29,30]. The high biodiversity and the unique geography of the Gulf of Aqaba have made this region a widely used model for reef ecology and conservation studies, including studies on the connection between global climatic events and coral mortality[31,32] and the processes that drive marine species diversity[29,33,34]. Coupling our approach with a comprehensive DNA barcode reference library, which contains the majority of reef-associated fish species of this species-rich region, we generated the first species-level documentation of the spatio-temporal structure of the regional larval pool.

Results

Intensive larval sampling and genomic barcode collection. To obtain high-resolution insights into the ecology of coral reef fish larvae in the Gulf of Aqaba, we conducted an intensive ichthyoplankton sampling over a period of 11 months between 2010 and 2011 (Fig. 1a,b; Supplementary Table 2). This sampling programme yielded 383 discrete samples, from 10 geographical sites across ~50 km^2 of the northern tip of the Gulf of Aqaba (Fig. 1b). Each site was visited either once or twice per month, and sampled at 1–6 depth layers with 25 m intervals for the upper 100 m (for example, 0–25 or 25–50 m) and 40 m intervals in deeper water (100–180 m). Overall, ~130,000 m^3 of water was filtered during this sampling, yielding 16,695 fish larvae. Larval cohorts that were visually identified as belonging to one of five abundant pelagic taxa were discarded from the set (see Methods), leaving 9,933 fish larvae suspected of belonging to reef-associated and demersal species.

To identify the sampled larvae based on their COI barcode sequence (denoted hereafter as 'COI reads'), we first needed to collect the reference COI sequences of the resident species. We obtained COI barcode sequences for 420 (77.5%) of the 542 demersal and reef-associated fish species that were previously documented in the Gulf of Aqaba, either by sequencing the DNA extracted from tissue samples from locally collected adult fish or from online public repositories (Supplementary Table 3). To account for the possibility that some of the larvae would be of pelagic species, or of Red Sea species not previously documented within the Gulf of Aqaba, we supplemented this database with all available COI sequences of fish that are known to reside in the entire Red Sea, covering ~70% of the known species in this larger basin (Supplementary Table 3, Supplementary Data 1).

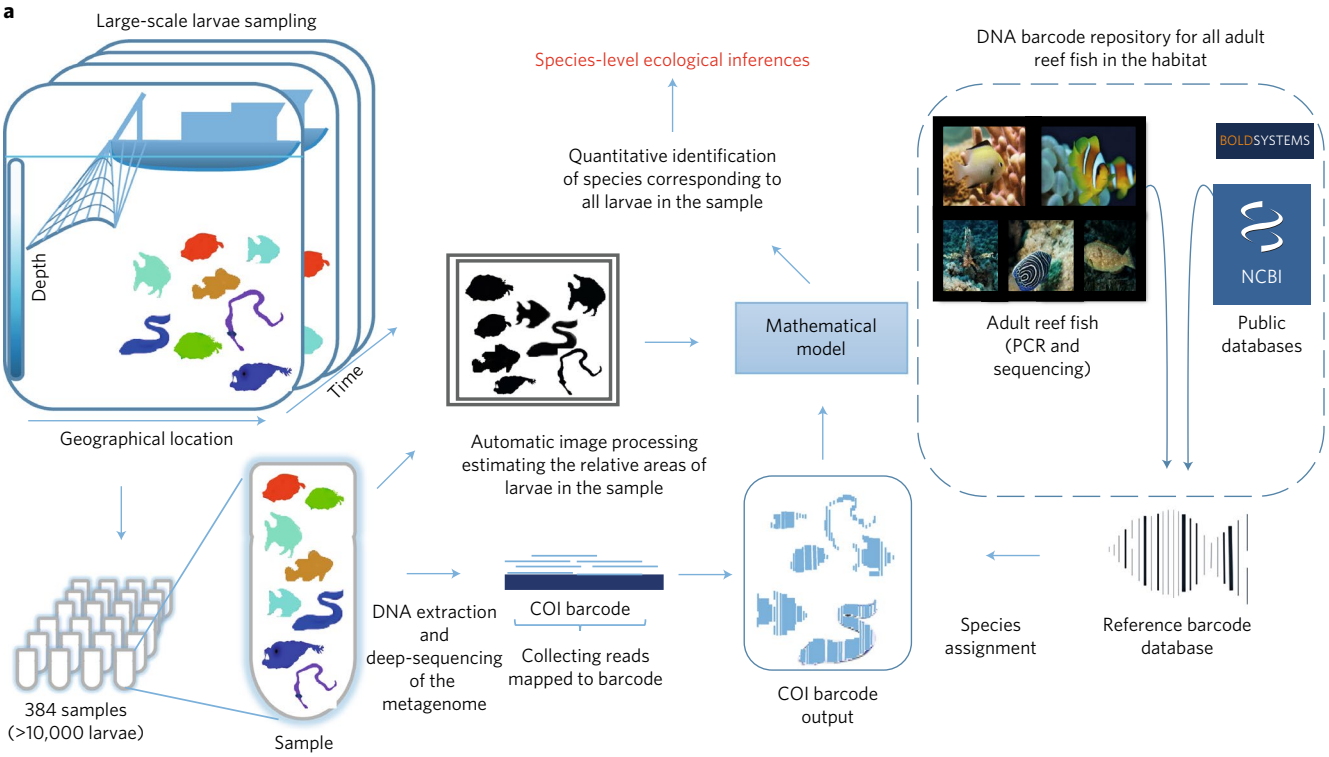
To verify that the COI repository of the Red Sea fish species can provide unique species-level identification, we quantified the distance (for example, the number of mismatches within the COI barcode) between all possible pairing of species of the same genus in our collection. In the overwhelming majority of cases (97.4%), a COI sequence was >20 base pairs (bp) from its closest match (average distance: 83.5 bp; Fig. 2). We compared these results to sequence distances observed between COIs that were sampled from pairs of individuals belonging to the same species, and found an average intra-species distance of 2.7 bp, with distances rarely exceeding 5 bp (Fig. 2). The high inter-species diversity in COI sequences, combined with the low intra-species diversity, confirms that fish species in the Gulf of Aqaba can be unequivocally identified based on their COI barcode (as in other locations[19]).

Unbiased metabarcoding of larval samples. We set out to use the COI barcode approach to accurately identify all sampled larvae. Because the large number of larvae in this study rendered amplification and sequencing of the COI from each individual larva unfeasible, we used a high-coverage metagenomic approach. The entire pooled genomic DNA from each of the 383 samples (Supplementary Table 2) was sequenced using the Illumina HiSeq technology. The sequence coverage for each sample was adjusted to yield ~20 COI-derived reads per larva, so that samples containing more larvae were proportionally sequenced more deeply (see Methods). In total, 3.54×10^9 paired-end reads (two paired reads, each of a 101 bp) were generated in this study, 201,145 of which were mapped to the COI barcode sequence (Supplementary Fig. 1, Supplementary Data 2; see Methods).

We then compared the COI reads in each sample with each of the COI barcodes in our set (Supplementary Fig. 1). Only reads that provided high-confidence identification of species, namely, those showing unique best mapping with up to two mismatches to one of the barcodes, were used for species assignment (see Methods). Surprisingly, only 71.5% of the COI reads showed such high-confidence mapping to one of the COIs in our set, which was in contradiction to our expectation based on the 77.5% coverage of the reef-associated species space in our barcode database. We reasoned that some of the larvae may belong to abundant pelagic species, the COIs of which were not targeted in our adult sampling effort. To account for these species, we took advantage of the high-coverage metagenomic data, and new COI barcodes were de novo assembled based on a strict, iterative assembly procedure that was verified independently for a subset of assemblies (Supplementary Fig. 1, Supplementary Data 1, Supplementary Table 2; see Methods). Approximately 60% of the de novo-assembled barcodes (95 out of 158) were taxonomically assigned based on sequence phylogeny and morphological inspection (see Methods), corroborating that many of these mapped to pelagic taxa (Supplementary Table 3, Supplementary Data 2). Overall, 184,826 of the 201,145 COI reads (91.8%) mapped to the extended COI barcode set (including the de novo-assembled COI barcodes), providing a framework for high-resolution, high-accuracy species identification.

To verify that our strategy produced accurate taxonomic identifications, we selected 96 samples, each comprising between 1 and 15 larvae (subset 1 in Supplementary Table 2). For 474 of the larvae in these samples, it was possible to visually determine the taxonomic family based on meristic, morphological and pigmentation criteria, overall identifying larvae belonging to 33 families (see Methods). In comparison, our metagenomic approach mapped 92% of the COI reads onto barcodes of known species in these 96 samples, identifying 138 individual species. Furthermore, 412 (86%) of the 474 morphologically identified larvae were accurately accounted for in the COI identifications. For an additional 21 larvae, we recorded discrepancies between the morphology-based and barcode-based taxonomic assignments, but upon re-examination of the morphology data, these cases were resolved as clear morphological misclassifications, verifying the barcode-based classification. Thus, the consistency of our barcode-based approach with the morphology-based taxonomic assignment was at least 91.4%, and with a much higher taxonomic resolution. Only 41 (8.6%) morphologically identified larvae were not supported in the sequencing data, which was probably owing to their COI being missing from our barcode database.

Bias-free metabarcoding reveals species abundance. A major limitation of PCR-based metabarcoding approaches is the loss of quantitative data due to amplification biases, which means that it



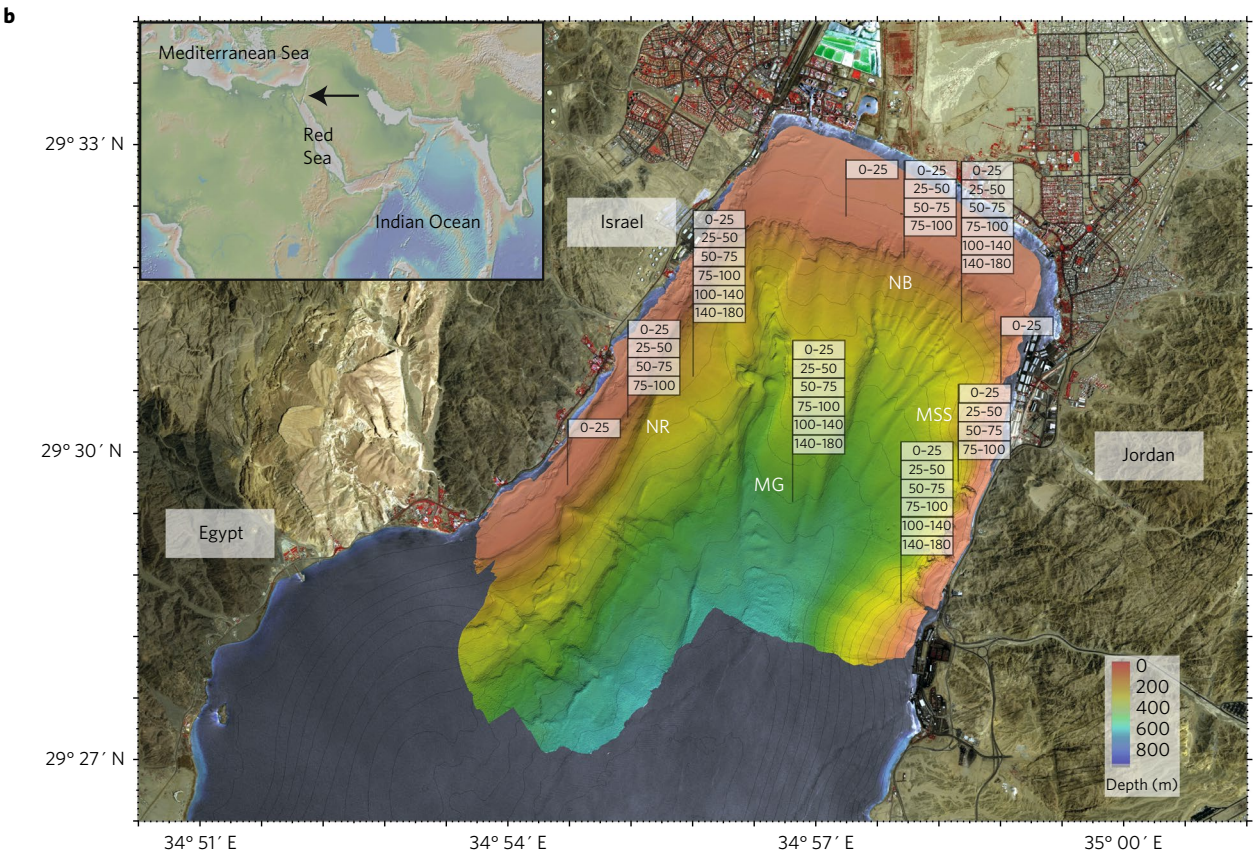


Fig. 1 | Larval sampling. a, Flowchart of the study and methods. **b**, Map showing the study sites in the Gulf of Aqaba (marked by an arrow in the inset), which were sampled extensively in 2010–2011. Within each site, sampling was conducted over three bottom depths (and therefore three sub-sites per site), except for the mid-Gulf deep water (MG) site. Multiple Opening and Closing Net and Environmental Sensing System (MOCNESS) sampling was conducted for each sub-site over the depth stratum indicated in the figure. MSS, Marine Science Station; NB, North Beach; NR, nature reserve. Credits: **a**, NCBI logo, https://www.ncbi.nlm.nih.gov/; BOLDSYSTEMS logo, www.boldsystems.org. The map in **b** was created by the authors of ref. [84], based on data from that article, and is reproduced with their permission.

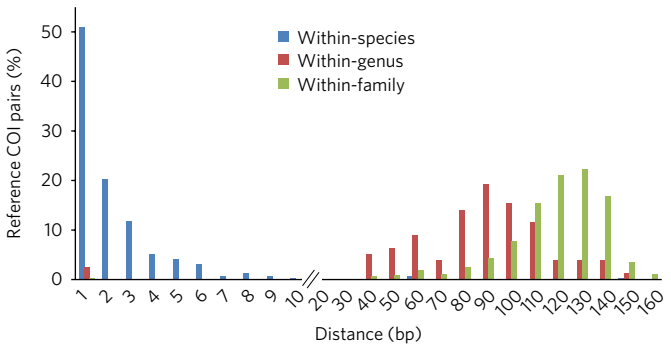


Fig. 2 | COI sequences in the reference barcode set provide distinctive species identification. Histogram of the pairwise distances (number of mismatches) separating individuals within species ($n=384$ bp), different species within genera ($n=79$ bp) and different genera within families ($n=580$ bp). The y axis represents the percentage of pairs falling within each distance bin.

is possible to detect the presence of a specific species in a given sample, but not its abundance (the number of individuals)[27,35]. We reasoned that as our shotgun metagenome-based metabarcoding approach does not involve PCR amplification, the fraction of COI reads derived from each individual larva should be proportional to the relative tissue mass of that larva in the sample. As measuring tissue mass for each larva is impractical, we used a two-dimensional projection as a proxy to tissue mass. Hence, prior to DNA extraction, we recorded silhouette pictures of all the larvae in each sample and, using image-processing software, calculated the proportion of the area occupied by each larva out of the total larval area in the image (Fig. 3a,b). Using a subset of samples in which all larvae were taxonomically pre-identified to the taxonomic family level based on morphology (subset 2 in Supplementary Table 2), we found a high correlation ($R^2=0.74$) between the relative size of the larvae and the fraction of the COI reads mapping to the barcode of that larva in the sample (Fig. 3c). Further correction based on family-specific biases (for example, eel larvae consistently produce less reads than expected by their relative projected area due to the thinness of their body) improved the correlation ($R^2=0.82$; Fig. 3c; see Methods). These results show that the relative size of each larva in the sample is an excellent predictor of the amount of COI reads derived from that larva.

Drawing on the above observation, we developed a statistical model that uses the relative sizes of larvae per sample and the corresponding number of COI reads per species within the sample, to estimate the most likely abundance of these species (see Methods). We evaluated the accuracy of this model by examining the resulting abundance profiles in samples in which taxonomic identity, at the family level, was inferred morphologically. Our results indicate high accuracy of the abundance estimates, with errors rarely exceeding 1–2 larvae per species per sample when compared with the actual abundances (Fig. 3d; see Methods).

High-resolution spatio-temporal larval distribution. Applying the above method to the 383 samples we collected resulted in taxonomic classification of 9,262 larvae, 5,388 of which were classified to the species level (Supplementary Table 4). Overall, larvae belonging to 278 species were identified, 255 of which were categorized as reef-associated or demersal. These species belong to 186 genera from 79 different taxonomic families, accounting for almost all known families in the region.

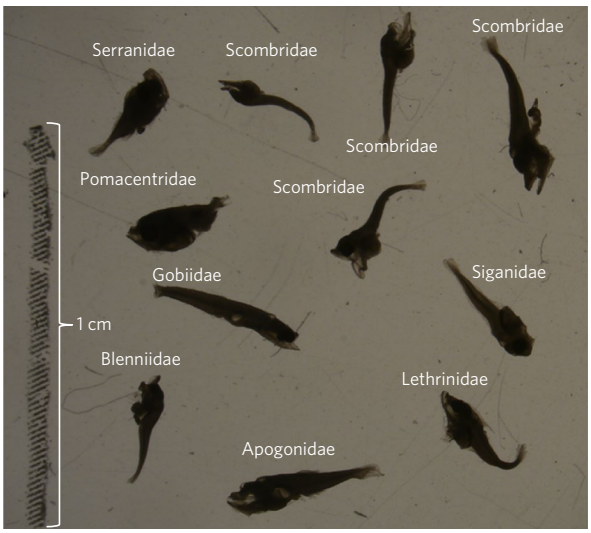
Fish larvae can actively control their vertical position in the water column from early ontogenetic stages[17]. Although it is well established that vertical position can affect larval dispersal trajectories (for example, when current velocity is vertically non-uniform)[36,37], little is known of the extent to which species differ in their preferred depths. We examined the spatio-temporal distribution of larvae from the five most abundant families of the reef-associated fish in our set, encompassing ~60% of the corresponding larvae we identified. Importantly, our data show species-specific depth distributions in all families examined (Fig. 4). For example, a clear bi-modal depth distribution is observed for larvae of the family Gobiidae (Fig. 4). Whereas the larvae of about half of the species belonging to this family in our samples tend to dwell in surface water (0–25 m) and appear during July, the other half dwell in deeper water (50–75 m) and have a higher tendency to appear between September and November (Fig. 4). Similarly, the larvae of *Plectranthias winniensis* (family: Serranidae) frequently reside in depth ranges of 50–100 m, whereas the larvae of its confamilial species *Pseudanthias squamipinnis* and *Pseudanthias taeniatus* are rarely found deeper than 50 m (Fig. 4).

The water column in the Gulf of Aqaba is stratified in the summer, but in the winter, it is mixed down to a depth of up to 700 m. Consequently, vertical gradients of temperature, chlorophyll and zooplankton densities disappear during the mixed period[31,32]. If larvae are responding to these environmental variables, we would expect seasonal differences in their vertical distribution. Nevertheless, we found little indication of a pronounced effect of the time of year on larval vertical distribution (Mantel test for the correlation between the spatial (depth) and temporal (month) dissimilarity matrices: Labridae: $r=0.10$; Serranidae: $r=0.02$; Pomacentridae: $r=0.23$; Apogonidae: $r=0.11$; all $P<0.05$); with the Gobiidae as the only possible exception ($r=0.25$, $P=0.02$). These results suggest that larvae actively position themselves in the water column according to species-specific depth preferences.

Larval supply and local adult population size. The size and stability of populations at small spatial scales depend on a set of hierarchical processes involving larval supply (that is, which species arrive to the site), settlement and post-settlement processes (that is, which larvae survive after arrival), which ultimately determine the number of recruits replenishing the population[38,39]. The relative importance of larval supply is a central question in reef fish ecology, but, to date, could not be directly assessed, certainly not for a wide range of species. In three of the six families examined, we found a positive cross-species relationship between larval abundance and the corresponding abundance of adults within local assemblages (Fig. 5), despite a >4-year separation between the sampling of the adult and larval data sets (see Methods). As the larval pool probably represents the reproductive output of a meta-community spanning an area of at least an order of magnitude larger than the local adult assemblage, we view these relationships as indicative of the major part played by larval supply in determining local adult densities (Fig. 5). Coincidently, however, several species seem to be consistently absent from the local assemblage despite ample larval supply (black circles in Fig. 5), presumably because of local environmental filtering.

Documentation of non-native species. One of the advantages of shotgun metagenomic sequencing is the ability to discover, in the larval pool, species whose presence as adults have never been documented in the region of interest. The fish fauna of the Gulf of Aqaba consists of a sub-sample of the species known from the Red Sea, and a long-standing question is whether the missing species have failed to arrive or failed to become established[29,40]. Using strict criteria (see Methods), we searched the larval pool for Red Sea species that have never been documented as adults in the Gulf of Aqaba. Surprisingly, we found 17 such non-native species (Supplementary Table 4). Notably, the actual number is probably higher, as not all of the known local species were recovered as larvae in our samples,

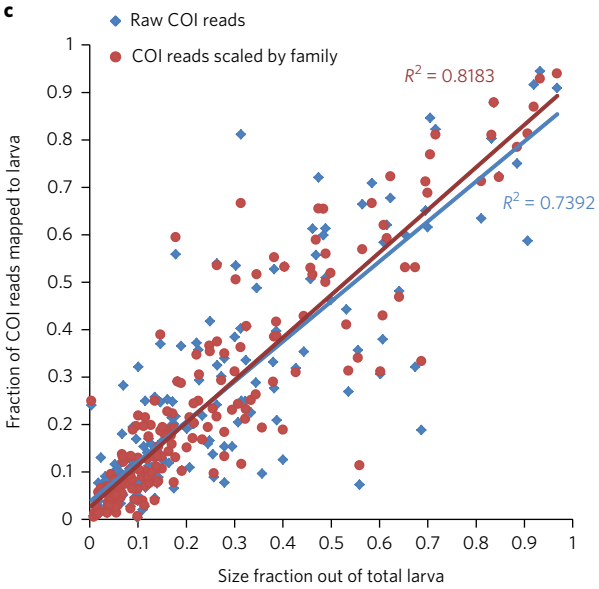
a



b

Family	Species	Size (pixels)	COI reads	Size fraction %	Fraction of COI reads %
Apogonidae	*Pristiapogon fraenatus*	97,129	75	12.66	15.99
Blenniidae	*Cirripectes stigmaticus*	46,559	34	6.07	7.25
Gobiidae	Unknown (assembled COI)	85,908	72	11.20	15.35
Lethrinidae	*Lethrinus lentjan*	65,137	13	8.49	2.77
Pomacentridae	*Dascyllus trimaculatus*	91,610	66	11.94	14.07
Scombridae	*Katsuwonus pelamis*	247,311	117	32.24	24.95
Serranidae	Unknown (assembled COI)	70,285	51	9.16	10.87
Siganidae	*Siganus sutor*	63,202	41	8.24	8.74

c



d

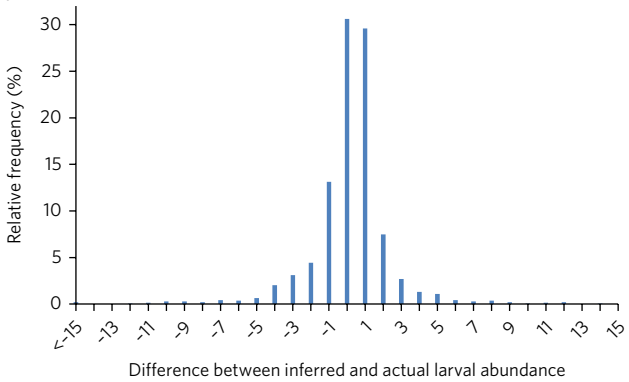


Fig. 3 | Size-based model for quantitative inference of species abundance. a, Silhouette picture of larvae from sample 97, collected on 19 October 2010 from a depth layer of 25–50 m of the mid-Gulf site. Family identity per larva was inferred based on morphological, meristic and pigmentation criteria. **b**, Species identified in sample 97, based on COI mapping. Larval sizes were inferred from the silhouette picture (see Methods). **c**, Correlation between relative larval size and the fraction of mapped COI reads in 47 samples in which all larvae ($n = 303$; subset 2 in Supplementary Table 2) were taxonomically assigned by both morphology (to the family level) and sequencing (to the species level). **d**, Differences between the number of larvae estimated using morphological assignments and the number estimated using COI sequencing and our quantitative model. Data were based on a larger set of 3,736 larvae from 234 samples (subset 3 in Supplementary Table 2) that were taxonomically assigned to the family level using the morphological criteria (see Methods).

and, in addition, our COI reference database does not contain all the species of the Red Sea.

Total larval densities, our proxy for propagule pressure, did not differ significantly between species documented in the Gulf of Aqaba (native species) and those that have not (non-native). Specifically, both shared the same relationship between total density and incidence (ANCOVA, $P > 0.84$ and $P > 0.14$ for slopes and intercepts, respectively), with mostly overlapping distributions of incidence (that is, the number of samples in which a species was present; Fig. 6). In addition, cluster analysis found little distinction between native and non-native species, based on the location and time of samples in which they were found (adjusted Rand index = 0.023). Together, these findings indicate that for multiple Red Sea species, their absence from the northern Gulf of Aqaba represents ecological barriers to colonization rather than limited larval supply.

An interesting case of a deep-sea species found in our samples was the occurrence of *Tylerius spinosissimus* (Spiny blaasop)—a small pufferfish native to the Indian Ocean, known to dwell at depths of 250–435 m (ref. [41]). Recent observation of this fish in the Aegean Sea suggest that the species is a Lessepsian migrant, that is, a species that invaded the Mediterranean Sea via the Suez Canal[42]. The depths in which this species dwells are at odds with the shallow depth of the Suez Canal (24 m), leaving its mode of migration via the Suez Canal enigmatic. We detected larvae of *T. spinosissimus* in three samples collected from depths of 50–100 m (Supplementary Table 4), which indicates that its larvae make it to waters that are much shallower than the adults and provides a possible mechanism of its migration through the Suez Canal. For additional deep-sea species—*Diaphus coeruleus*, *Ophichthus echeloides* and *Acropoma japonicum* (all of which were documented as adults only below the depths of 100 m)—we recovered larvae from water shallower (as shallow as 25 m) than the upper depth documented for their adults. These results suggest that for deep-sea fish, pelagic larvae can be an efficient mode of dispersal between habitats separated by shallow-water barriers.

Discussion

We have developed an approach that allows unbiased, quantitative analysis of large planktonic communities, at species-level resolution.

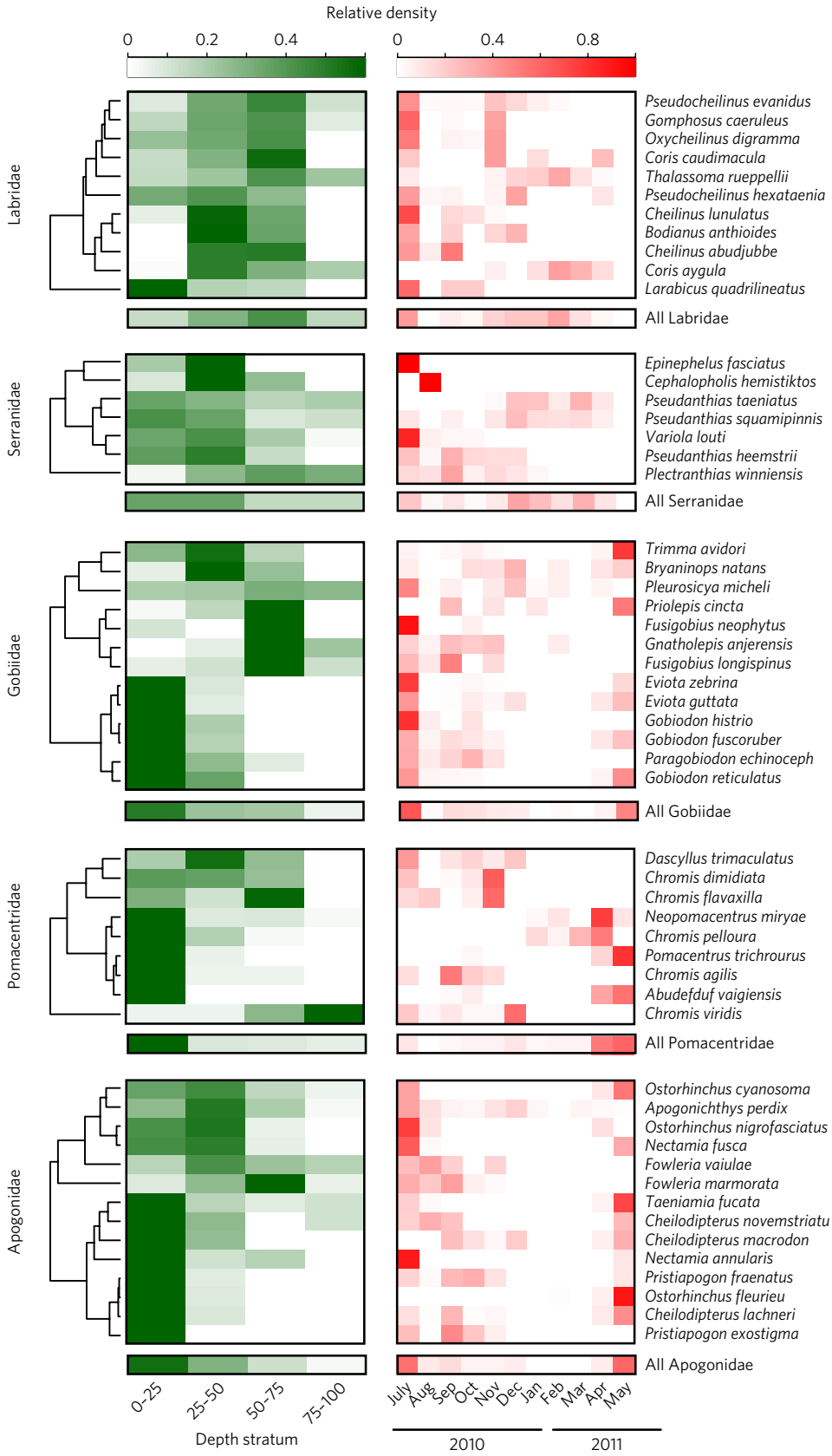


Fig. 4 | Spatio-temporal distribution of species from the five most abundant families of reef-associated fish in our study. The relative density of each species in each category (depth stratum or month) is given as the proportion from the total number of individuals of that species, after correcting for sampling effort (volume of water filtered and the number of hauls in each depth stratum or month; see Methods). Clustering is based on average linkage and Bray–Curtis dissimilarity between samples. Only species represented in at least five hauls were included in the analysis (see Methods; Supplementary Table 2, Supplementary Table 4). Note the difference between species-level distribution and family-level profiles (displayed below each panel).

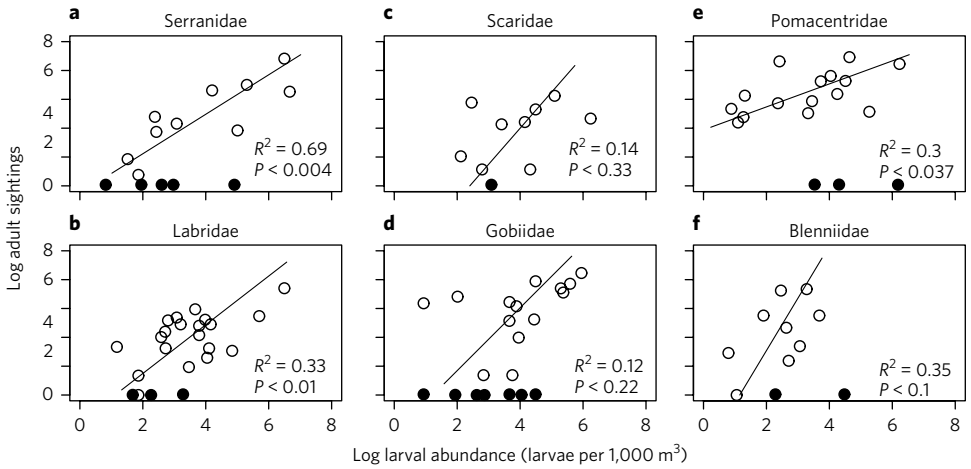


Fig. 5 | The relationship between larval and adult abundance in six common families within our data set. a–f, Adult abundances were estimated from sightings of adults from multiple benthic surveys across Eilat (Israel) reefs (see Methods). Families for which adult-sighting and larval-density data were available for at least nine species are included in this analysis. Larval abundances were derived from the metabarcoding data in the current study and calculated as summed densities across all Eilat sampling sites (Methods). A major axis regression model is presented for each family. Black circles depict species that are present in the larval community but not in the local assemblage; white circles depict species present in both larval community and local adult assemblages.

Although metabarcoding of metazoans has been previously suggested[23,27], our approach allows rigorous quantitative estimation of the number of individuals per sample, which is a basic parameter needed in ecological studies[14,22]. We verified this approach using a large sample set of coral reef fish larvae. Our comprehensive COI reference library, which covered much of the region's rich species pool, allowed the surprising discovery of the high influx of 'non-native' species into the sampled area. This capacity to detect potentially invasive species at their dispersal stage, before establishing viable populations, could be a strong asset in the conservation and management of reef and other marine ecosystems.

Our method requires shotgun metagenomic sequencing of the total genomic DNA derived from the sample. As the COI sequence is encoded on the mitochondrial genome, and as an average eukaryotic cell is estimated to contain ~600–1,000 mitochondria[43,44], the COI sequence is naturally enriched in the sample by 2–3 orders of magnitude as compared to nuclear genes. Moreover, usage of mitochondria enrichment[45] can further increase the effective COI coverage and reduce sequencing costs, although in our case, mitochondrial enrichment did not work on ethanol-preserved samples.

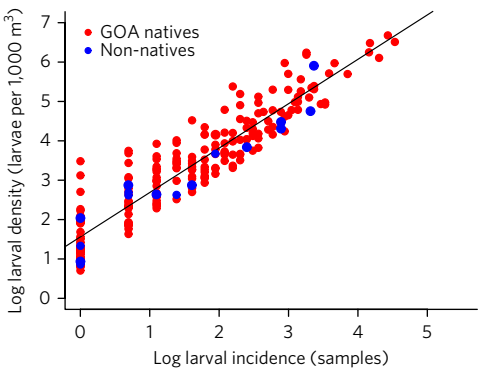


Fig. 6 | The relationship between larval incidence and density for demersal and reef-associated species. The *y* axis represents natural log-transformed densities (larvae per 1,000 m^3) of species that are known to occur as adults in the Gulf of Aqaba (GOA; $n = 230$) and those that are not ($n = 17$). Larval abundances were derived from the metabarcoding data. Incidence is defined as the number of samples with non-zero abundance.

Over the past decade, metagenomics has revolutionized the field of microbial ecology and microbiome research[46]. With the extension of this technique to metazoa, it now has the potential to have similar revolutionary effects on ecological studies of higher-order organisms. The method that we introduce is applicable to study species composition in any complex ecosystem. Specifically, the distribution and dispersal of zooplankton is crucially important for the entire food web in the ocean[47] and we envision that bias-free metabarcoding studies using our approach, combined with additional oceanographic models, could reveal important insights into this poorly resolved community.

Methods

Intensive larval sampling. An extensive ichtyoplankton sampling effort has been carried out along the western and eastern coasts of the northern tip of the Gulf of Aqaba (29° 26′–29° 32′ N, 34° 54′–34° 59′ E). The sampling commenced on July 2010 in a 2-day sampling bimonthly format, such that the first day of sampling was carried out on the western side of the coast and the following day on the eastern side (Fig. 1b). In December 2010, the sampling format was altered into a 2-day sampling monthly layout, and continued until May 2011. The sampling efforts were reduced towards the end of autumn, along with the decline of the major reproductive season. Plankton samples were collected using a 1 m² MOCNESS[48], mounted with 600 µm nets. MOCNESS nets were thoroughly cleaned between uses.

The sampling transects run parallel to the north-western shore of the gulf, over a bottom depth of ~70, ~170, ~250 and ~500 m. The nets were first lowered to the maximum depth and then opened sequentially, such that for the nearest-to-shore transect (above a bottom depth of ~70 m), a single net sampled the depth range of 25–0 m; for the next transect (above a bottom depth of ~170 m), four nets sampled the depth ranges of 100–75, 75–50, 50–25 and 25–0 m; and for the two farthest-from-shore transects (above bottom depths of ~250 and ~500 m), the nets sampled the depths of 180–140, 140–100, 100–75, 75–50, 50–25 and 25–0 m. The MOCNESS was towed at a speed of ~2 knots, for 5 min per net, and the flux of water through the net was measured. The number of larvae per net was normalized to the volume of water filtered to yield larval concentrations (individuals per 1,000 m^3).

The ichthyoplankton samples were instantly preserved on board in 80% ethanol, which was replaced by fresh absolute ethanol the following day. Ethanol preservation allows subsequent morphological and molecular identification. The samples were stored in a 4 °C cold room. The ichthyoplankton were then manually separated from the rest of the plankton. Larval cohorts visually identified as belonging to one of five abundant pelagic taxa—Myctophidae, Phosichthyidae, Paralepididae, Trichiuridae and Sternoptychidae—were discarded from the set. In addition, one visually identified morphotype of larvae, encompassing the genus *Cirrhilabrus* and the species *Paracheilinus octotaenia* ($n = 285$ larvae), were

manually excluded from the larvae samples for the purpose of a separate study that will be published elsewhere.

Silhouette images of larvae samples. The larvae from each of the 383 nets (samples) were photographed using a Panasonic DMC-G5 camera mounted on a Nikon SMZ1500 dissecting microscope. Photos were taken under two sets of conditions: the first entailed full illumination to provide high-quality pictures in case re-examination of larval morphological characteristics was needed at a later time. The second set of conditions involved the bottom illumination source alone to create a silhouette image of the larvae next to a 10 mm scale. The later photographs were used for measuring larval area that was used in the automatic quantification procedure.

Morphology-based taxonomic assignment of larvae families. Classic taxonomic identification for a subset of larvae ($n=234$ samples; subset 3 in Supplementary Table 2) was done in most cases down to the family level under a Nikon SMZ1500 dissecting microscope, using published identification guides[49,50]. Larval meristic, morphological and pigmentation criteria were examined for the purpose of their identification. Morphological identification was necessary for the following purposes: (1) validation of species identified by sequencing and taxonomic ascription of COI sequences that were missing from the local and public databases; (2) calibration and assessment of the quantification method; and (3) removal of the most abundant non-reef-associated taxa, to reduce sequencing costs and to prevent the masking of the less common reef-associated sequences.

Collection of adult fish and COI barcoding. Fin clips and muscle tissue samples were collected from adult fish belonging to 203 species. Fish were caught during scuba dives in Eilat (Israel), using barrier nets and clove oil. At least one specimen from each species was transferred to the shore, identified and euthanized using an overdose of MS-222. A muscle tissue sample was obtained from this individual, which was then deposited as a voucher specimen in The Steinhardt Museum of Natural History, Israel National Center for Biodiversity Studies (SMNHTAU). In addition, more specimens (1–8 individuals) were caught for each species, for which identification and sampling of fin clips (usually a 1 cm^2 clip from the caudal or dorsal fin) were sampled underwater. Tissue and fin samples were preserved in analytical grade absolute ethanol. From each sample, DNA was extracted, and the COI gene was amplified following PCR with universal primer cocktails and sequenced using SANGER sequencing following refs [19,51].

List of fish known to dwell in the Gulf of Aqaba and the Red Sea. The list of fish known to dwell in the Gulf of Aqaba and the Red Sea was compiled from the primary literature[34,52–66]. For this purpose, only published, peer-reviewed articles and species reports that included the physical collection and identification of species in the area were used. In addition, three guide books on fish and one graduate thesis published by fish taxonomists were also used[63,64,67,68]. Finally, fish captured as adults during our adult sampling between 2011 and 2013 were also included in the list of known fish. The list is provided in Supplementary Table 3. Species names were verified with fishbase.org.

Compilation of COI reference barcode database. In addition to the set of COI reads sequenced as part of this study (see 'Collection of adult fish and COI barcoding'), publicly available COI barcodes of Red Sea fish were obtained from the National Center for Biotechnology Information (NCBI) and Barcode of Life Database (BOLD)[21]). COIs were collected from public records in two main steps: (1) COIs of Red Sea fish were collected by searching the organism name, and a representative COI was chosen. (2) A BLAST search with the 'nt' database (downloaded on June 2014) was performed for reads that could not be associated with any COI in the database. COIs that yielded hits with at least 98% identity over >90 bp were added to the database.

DNA extraction, library preparation and sequencing. For DNA extraction from mixed larvae samples, each sample was gently centrifuged (1,000 g) and washed 3 times with 1 ml of PBS buffer (pH 7.2, 50 mM potassium phosphate and 150 mM NaCl) to discard residual ethanol. The DNA was then extracted from each sample separately by using the DNeasy Blood & Tissue kit (Qiagen) according to the manufacturer protocol for purification of total DNA from animal tissues using spin columns. In a few cases, when the amount of larvae exceeded 100 larvae, samples were split during the morphological identification step and reunited after sequencing. Genomic DNA was measured using the Qubit dsDNA fluorometric assay (ThermoFisher Scientific). The integrity of the genomic DNA was assessed by the 2200 TapeStation instrument (Agilent Technologies).

Extracted DNA samples were prepared for sequencing as previously described[69] with the following modifications: ~200–1,000 ng of DNA was sheared using the Covaris S220 sonicator to receive a peak of ~250 bp. End repair was performed in 80 µl reaction at 20 °C for 30 min. Following Agencourt AmPURE XP beads cleanup (Beckman Coulter) in a ratio of 0.75× beads/DNA volume, an adenine base was added, and indexed adapters were ligated in a final concentration of 0.125 µM. Following another AmPURE XP beads cleanup procedure in a ratio of 0.75× beads/DNA volume, 8 PCR cycles for enrichment of adaptor-ligated fragments, using primers on the library adaptors, were performed using 2X HiFi HotStart ReadyMix (Kappa Biosystems) in a total volume of 25 µl with the following programme: 2 min at 98 °C, 8 cycles of 20 s at 98 °C, 30 s at 55 °C, 60 s at 72 °C, followed by 72 °C at 10 min. The libraries were pooled and sequenced at paired-end 101-bp lanes of Illumina HiSeq 2500 instrument, overall sequencing 19 HiSeq lanes.

Species identification from metagenomic COI reads based on available COI barcodes. All metagenomic reads were aligned, using BLAST search (blastn parameters: -F F -e 1e–10), to all barcodes in our COI database, and reads passing the threshold of evalue <1e–10 were considered COI reads. Reads having a single best hit to one of the COIs in the reference barcode set with an alignment length >90 bp and percent identity >98% were assigned to that barcode COI ('species'). Reads mapping with alignment length of >50 bp to one of the ends of the barcode (where an end was defined as overlapping the first or last 10 bp of the COI barcode) were defined as 'edge reads'. Mapping of edge reads was only maintained if the paired-end read mate was fully aligned to the COI, or if there was at least one full-length read mapping to that barcode in the sample. In case that a read had multiple best hits, its COI identity was resolved by its paired-end read mate. In case paired-end resolution was not possible for a multiple-hit read, but only one of the COI barcodes in the database was supported by other reads from the same sample uniquely mapped to it, that COI barcode was assigned to the read. Reads that had multiple best hits that could not be resolved were counted as 'ambiguous reads' (overall representing 0.3% of the data). COI barcode (and hence species identification) were reported as present in a given sample only if at least 50% of the barcode was covered by COI reads.

De novo assembly of COIs. De novo COI barcode sequences were constructed from the pool of unmapped metagenomics COI reads (that is, reads that were not assigned to any COI barcode in our database) by an iterative procedure. In the first step, clusters of overlapping COI reads were collected using BLASTCLUST[70] by clustering reads that showed 100% identity over 80% of the sequence (blastclust parameters: -p F -L. 8 -b T -S 100). Next, COI contigs were constructed from clusters with at least 10 reads using the Velvet assembler[71]. The resulting contigs were used as a database for BLASTn search to re-map the unmapped COI reads. Each contig defined a cluster of reads that exhibited 100% identity over at least 80 bp to the contig. Next, the clusters of reads were used to re-construct the contigs using Velvet. At this point, very short COI contigs (<450 bp) were removed from the set and the remaining contigs were exposed to the following extension procedure: the unmapped reads were blasted (blastn parameters: -e 1e–8) against the contigs; reads that were mapped to the start or the end position of the contig and showed 100% identity with overlap of 90–100 bp were used to extend the COI contig. At each iteration, the hits with the largest overlap were used to extend the COI. The process was terminated either by the construction of a full-length COI (655 bp) or at the point in which no reads were found to further extend the contig.

Taxonomic classification of de novo-assembled COIs. To taxonomically classify the 158 COI sequences that were constructed using the de novo assembly method, this set was first compared to all COI sequences present in the BOLD database[21]. In case a sequence showed >99% identity to a barcode sequence of a known species, and for which at least two barcodes were deposited in the BOLD databases, the species name was assigned based on the BOLD annotation. In cases in which only one sequence was found or multiple species exceeded the 99% threshold, the common taxonomic unit (genus) was used. In case multiple species were identified as exceeding the 96% threshold by BOLD, the common taxonomic assignment (usually genus) was given to the COI. In the same manner, family was assigned in cases in which the sequence showed >92% identity or multiple genera passed the 96% threshold.

For taxonomic assignment of the remaining COIs, a phylogenetic tree was first constructed from all COIs in the database (including the identified ones) using MAFFT (version 7.13)[72] for the construction of the multiple sequence alignment. FastTree (version 2.1.7)[73] was later used for the tree construction (gtr method), and the resulting tree was visualized by FigTree software (version 1.4.2; http://tree.bio.ed.ac.uk/software/figtree/). Assembled COIs falling in a clear clade, that is, where all other COIs in the clade belonged to a defined taxonomic group (genus or family), were taxonomically assigned to that group. For additional taxonomic assignments, we used the pplacer software tool[74].

The third approach for taxonomical inference was based on the independent family-level morphological identifications performed on a subset of 234 samples (subset 3 in Supplementary Table 2). Assembled COIs that were consistently present in multiple samples in which a morphologically observed family was absent from the sequence identification, and that no other explanation was possible, were assigned to that family.

Inference of larval area from silhouette images. For each sample with a silhouette picture, the Fiji software[75] was used to measure the area occupied by each larvae in the following manner: the larval boundaries were detected using the 'Otsu' threshold method, and the default 'Measure particle' option was used to extract the area. The relative area of each larva was estimated as

$$\text{area}_i = \frac{\text{area}(\text{larva}_i)}{\sum_i^{n \text{ larvae in sample}} \text{area}(\text{larva}_i)}$$

The relative area of a given species was calculated as the fraction of the sum of all larvae that belongs to that species in the sample out of the sum of all area.

Normalization of expected COI reads per larval area by taxonomic families. To refine the model for quantitative estimation of the larvae abundance in the sample, a set of 47 samples (subset 2 in Supplementary Table 2) in which all larvae were taxonomically assigned based on morphology (before taking the silhouette image), was used for learning a linear model that aimed to normalize the amount of COI reads expected per area unit for each separate taxonomical classification (family or order). To this end, for each family that occurred in at least five annotated samples a factor w was inferred, such that w minimized the L_2 norm (Euclidean distance) between the relative areas of the larvae of this family in the sample and the relative amount of COI that were observed. This factor was later used to scale the reads associated with species of the particular family. A similar factor was calculated for the taxonomical order, for families appearing in less than five annotated samples.

A statistical method for inferring the species abundance in a sample. Based on the observation that the relative sizes of larvae are highly correlated with the fraction of COI reads derived from that larvae in the sample, a statistical method was developed for inferring the relative abundance of the different species in the sample given the larvae size estimates (based on silhouette pictures) and the COI reads derived from metagenomic sequencing of the sample, as follows:

Given a sample with (1) n larvae, (2) s species $(s_1, \ldots, s_s)$ (identified by the sequencing data), (3) the observed COI reads vector $R = (r_1, \ldots, r_s)$ normalized for taxonomic family, and (4) unassigned areas vector $a = (a_1, \ldots, a_n)$:

(1) Derive 1,000 initial random assignments $Z = (Z_1, \ldots, Z_{1,000})$ of the n larvae to the s species such that Z_i maps the areas vector a to the species vector s.
(2) For each assignment, Z_i, derive the vector of the relative area of each species in the sample $A = (A_1, \ldots, A_s)$. Calculate the log likelihood score of the assignment, defined as $L(Z_i \mid r_1 \ldots r_s) = \sum_{i=1}^{s} r_i \log(A_i)$. This likelihood score assumes that the observed reads are distributed according to a multinomial distribution, where the probability of a read belonging to a given larva is proportional to its relative area.
(3) For score maximization of each initial assignment, iteratively use the following algorithm until convergence:
   I. Define the vector $v = A - R$.
   II. For each species s_i, define the assignment vector $p_{\text{assign}_s} = 1 + v_s$ and the replacement vector $p_{\text{replace}_s} = 1 - v_s$
   III. Normalize the vectors to represent probabilities:
   $p_{\text{change}_s} = \frac{p_{\text{change}_s}}{\sum_{s=1}^{S} p_{\text{change}_s}}, p_{\text{replace}_s} = \frac{p_{\text{replace}_s}}{\sum_{s=1}^{S} p_{\text{replace}_s}}$
   IV. For each larvae l in the sample:
      a. Check its current assignment to species s (from Z_i)
      b. With probability p_{replace_s} choose a new random assignment to this larvae according to the probabilities p_{assign_s}. This step defines a new assignment Z_i'
   V. Calculate the log likelihood score of the new assignment $L(Z_i')|r_1, \ldots, r_s)$. Compare the score to the score of the previous assignment, and use the assignment of the better score for the next iteration.
(4) Repeat steps I–V until convergence (defined as 10,000 steps with no score improvement).
(5) From the 1,000 final assignments, select the one with the best score.

The performance of this algorithm was assessed by comparing its results with the results derived from morphology assignments (Fig. 3).

The necessity for a stochastic heuristic algorithm is apparent. For samples with a small number of larvae, it is possible to calculate the likelihood of each possible assignment. However, this quickly becomes computationally complex as the number of larvae and species increases. The rationale behind this heuristic algorithm is that we take a step in a direction that probably improves the assignment score. If the relative size of the species is over-represented with respect to its observed reads, we have a higher chance of picking one of its larvae and changing its species assignment.

Spatio-temporal patterns in larval distribution. Heat maps and dendrograms for the spatio-temporal distribution of larvae were generated with the heatmap.2 command of the R package 'gplots'[76]. The analyses were limited to species represented in at least five sampled nets, with clustering based on Bray–Curtis dissimilarity and average linkage. A Mantel test (R package 'vegan'[76]) on the distance matrices was used to look for an association between depth preference and time of year. Differences in sampling effort (that is, in the number of hauls per depth stratum or month) were controlled for by considering average per-haul densities. For our spatial analysis, the total density per depth stratum for each month was divided by the number of hauls in that stratum; averages were then summed across months and expressed as a proportion of the grand-total (that is, the sum, or the monthly sums, across the four depth strata of the upper 100 m). For our temporal analysis, the total density for each month was divided by the number of hauls for that month and expressed as a proportion of the sum of the monthly averages.

Species representation in a local assemblage and the larval pool. The total number of adult sightings during replicated censuses along the north-western Gulf of Aqaba was used as a quantitative proxy for population size and stability. The data were collected during two periods: September 1999 to September 2000 and December 2003 to April 2006. For the first period, we took the mean of the total counts from 42 belt-transects (2×50 m), with the mean calculated across four seasonal replicates[68]. For the second period, we took the total counts from 42 non-replicated belt-transects (2×25 m), conducted at depths down to 65 m (ref. [64]). The counts from both periods were summed (by species). Log-transformed counts were used along with log-transformed total larval-density estimates in six separate major axis regression models, one for each of the six families for which we had adult-sighting and larval-density data for at least nine species. The analysis was run with the 'lmodel2' command of the R package 'lmodel2'[76]. All statistical tests reported throughout the manuscript are two-sided.

For the identification of non-native species in the larval pool, we demanded that there would be more than one deposited COI in the BOLD database (to avoid possible species misidentification), and that the read coverage in our samples would be >90%.

Summing larval densities across samples (that is, across space and time), we treated total larval density, per species, as a proxy for propagule pressure, which is a reliable factor in predicting colonization success[77]. As a means of comparing the propagule pressure of species native to the Gulf of Aqaba to non-native species, an ANCOVA model was used to test the relationship between total larval density, per species, and species' incidence. Equality of slopes was tested first, using an interaction term for the covariate and status, followed by a test for the equality of intercepts.

To further evaluate the similarity between Gulf of Aqaba natives and non-natives, model-based cluster-analysis was performed with the 'Mclust' function in the R library 'mclust'[1]. Sampling site and sampling month were entered as descriptor variables; the Bayesian Information Criterion was used as the criteria for choosing among competing mixture models. The quality of the selected clustering was evaluated using the adjusted Rand index, which takes a value of zero when the agreement between the observed and modelled classification is no better than chance. The analysis was limited to 35 species from 11 genera, which included at least one species from each category (21 Gulf of Aqaba natives and 14 non-natives).

Analysis of universal primer set alignment to fish mitochondrial genomes. For the analysis presented in Supplementary Table 1, we first downloaded 2,151 fish mitochondrial genomes available at the MitoFish database (http://mitofish.aori.u-tokyo.ac.jp/)[78]. Primer sets were retrieved from refs [51,79–83]. Degenerate primers were parsed to include all possible sequence options. BLASTn was then used to align each primer to all 2,151 genomes using flags -W 7 -F F.

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. The code used in this study was deposited in github: https://github.com/omerzu/larvae_project.

Data availability. COI barcodes sequenced in this study were deposited in BOLD (http://www.boldsystems.org/). BOLD reference numbers appear in Supplementary Table 3.

Raw metagenomics data were deposited in the ENA repository, under the accession number PRJEB20625.

Received: 29 August 2016; Accepted: 13 November 2017;
Published online: 18 December 2017

References

1. Bellwood, D. R., Hoey, A. S. & Hughes, T. P. Human activity selectively impacts the ecosystem roles of parrotfishes on coral reefs. *Proc. R. Soc. B* **279**, 1621–1629 (2012).
2. McClanahan, T. R. et al. Critical thresholds and tangible targets for ecosystem-based management of coral reef fisheries. *Proc. Natl Acad. Sci. USA* **108**, 17230–17233 (2011).
3. Hoegh-Guldberg, O. et al. Coral reefs under rapid climate change and ocean acidification. *Science* **318**, 1737–1742 (2007).
4. Bellwood, D. R., Hughes, T. P., Folke, C. & Nyström, M. Confronting the coral reef crisis. *Nature* **429**, 827–833 (2004).
5. Gardner, T. A., Cote, I. M., Gill, J. A., Grant, A. & Watkinson, A. R. Long-term region-wide declines in Caribbean corals. *Science* **301**, 958–960 (2003).
6. Garpe, K. C., Yahya, S. A. S., Lindahl, U. & Öhman, M. C. Long-term effects of the 1998 coral bleaching event on reef fish assemblages. *Mar. Ecol. Prog. Ser.* **315**, 237–247 (2006).

7. Campbell, L. M., Gray, N. J., Hazen, E. L. & Shackeroff, J. M. Beyond baselines: rethinking priorities for ocean -conservation. *Ecol. Soc.* **14**, 14 (2009).
8. Cowen, R. K. & Sponaugle, S. Larval dispersal and marine population connectivity. *Annu. Rev. Mar. Sci.* **1**, 443–466 (2009).
9. Cowen, R. K. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 149–170 (Academic, London, 2002).
10. Doherty, P. J., Fowlert, T. & Fowler, T. An empirical test of recruitment limitation in a coral reef fish. *Science* **263**, 935–939 (1994).
11. Armsworth, P. R. Recruitment limitation, population regulation, and larval connectivity in reef fish metapopulations. *Ecology* **83**, 1092–1104 (2002).
12. Werner, F. E., Cowen, R. C. & Paris, C. B. Coupled biological and physical models: present capabilities and necessary developments for future studies of population connectivity. *Oceanography* **20**, 54–69 (2007).
13. Llopiz, J. K. & Cowen, R. K. Variability in the trophic role of coral reef fish larvae in the oceanic plankton. *Mar. Ecol. Prog. Ser.* **381**, 259–272 (2009).
14. Leis, J. M. Taxonomy and systematics of larval Indo-Pacific fishes: a review of progress since 1981. *Ichthyol. Res.* **62**, 9–28 (2014).
15. Ko, H. L. et al. Evaluating the accuracy of morphological identification of larval fishes by applying DNA barcoding. *PLoS ONE* **8**, e53451 (2013).
16. Limouzyparis, C., Mcgowan, M. F., Richards, W. J., Umaran, J. P. & Cha, S. S. Diversity of fish larvae in the Florida-Keys—results from SEFCAR. *Bull. Mar. Sci.* **54**, 857–870 (1994).
17. Irisson, J., Paris, C., Gulgand, C. & Planes, S. Vertical distribution and ontogenetic 'migration' in coral reef fish larvae. *Limnol. Oceanogr.* **55**, 909–919 (2009).
18. Evans, N. T. et al. Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Mol. Ecol. Resour.* **16**, 29–41 (2016).
19. Hubert, N., Delrieu-trottin, E., Irisson, J., Meyer, C. & Planes, S. Identifying coral reef fish larvae through DNA barcoding : a test case with the families Acanthuridae and Holocentridae. *Mol. Phylogenet. Evol.* **55**, 1195–1203 (2010).
20. Hebert, P. D. N., Ratnasingham, S. & Waard, J. Barcoding animal life: cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proc. R. Soc. Lond. B* **270**, S96–S99 (2003).
21. Ratnasingham, S. & Hebert, P. D. N. BOLD : the Barcode of Life Data System (www.barcodinglife.org). *Mol. Ecol. Notes* **7**, 355–364 (2007).
22. Hubert, N., Espiau, B., Meyer, C. & Planes, S. Identifying the ichthyoplankton of a coral reef using DNA barcodes. *Mol. Ecol. Resour.* **15**, 57–67 (2015).
23. Leray, M. & Knowlton, N. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. *Proc. Natl Acad. Sci. USA* **112**, 2076–2081 (2015).
24. Qiu, X. et al. Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning. *Appl. Environ. Microbiol.* **67**, 880–887 (2001).
25. Galan, M., Pagés, M. & Cosson, J. F. Next-generation sequencing for rodent barcoding: species identification from fresh, degraded and environmental samples. *PLoS ONE* **7**, e48374 (2012).
26. Bucklin, A., Steinke, D. & Blanco-Bercial, L. DNA barcoding of marine metazoa. *Ann. Rev. Mar. Sci.* **3**, 471–508 (2011).
27. Zhou, X. et al Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. *Gigascience* **2**, 4 (2013).
28. Deagle, B. E. et al. DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. *Biol. Lett.* **10**, 1789–1793 (2014).
29. Kiflawi, M., Belmaker, J., Brokovich, E., Einbinder, S. & Holzman, R. The determinants of species richness of a relatively young coral-reef ichthyofauna. *J. Biogeogr.* **33**, 1289–1294 (2006).
30. Golani, D. & Bogorodsky, S. V. The fishes of the Red Sea—reappraisal and updated checklist. *Zootaxa* **2463**, 1–135 (2010).
31. Genin, A., Lazar, B. & Brenner, S. Vertical mixing and coral death in the Red Sea following the eruption of Mount Pinatubo. *Nature* **377**, 507–510 (1995).
32. Fine, M., Gildor, H. & Genin, A. A coral reef refuge in the Red Sea. *Glob. Change Biol.* **19**, 3640–3647 (2013).
33. Hughes, T. P., Bellwood, D. R. & Connolly, S. R. Biodiversity hotspots, centers of endemicity, and the conservation of coral reefs. *Ecol. Lett.* **5**, 775–784 (2002).
34. Brokovich, E., Einbinder, S., Shashar, N., Kiflawi, M. & Kark, S. Descending to the twilight-zone: changes in coral reef fish assemblages along a depth gradient down to 65 m. *Mar. Ecol. Prog. Ser.* **371**, 253–262 (2008).
35. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol. Ecol.* **21**, 2045–2050 (2012).
36. Leis, J. M. & Mccormick, M. I. in *Coral Reef Fishes: Dynamics and Diversity in a Complex Ecosystem* (ed. Sale, P. F.) 171–200 (Academic, London, 2002).
37. Paris, C. B. & Cowen, R. K. Direct evidence of a biophysical retention mechanism for coral reef fish larvae. *Limnol. Oceanogr.* **49**, 1964–1979 (2004).
38. Ottosson, U., Sandberg, R. & Pettersson, J. Orientation cage and release experiments with migratory Wheatears (*Oenanthe oenanthe*) in Scandinavia and Greenland: the importance of visual cues. *Ethology* **86**, 57–70 (1990).
39. Pineda, J., Porri, F., Starczak, V. & Blythe, J. Causes of decoupling between larval supply and settlement and consequences for understanding recruitment and population connectivity. *J. Exp. Mar. Biol. Ecol.* **392**, 9–21 (2010).
40. Dibattista, J. D. et al. A review of contemporary patterns of endemism for shallow water reef fauna in the Red Sea. *J. Biogeogr.* **43**, 423–439 (2016).
41. Matasuura, K. & Tyler, J. C. *Resultats DES Campagnes Musorstom* Vol. 17 (ed. Séret, B.) 173–208 (Museum National d'Histoire Naturelle, Paris, 1997).
42. Turan, C. & Yaglioglu, D. First record of the spiny blaasop *Tylerius spinosissimus* (Regan, 1908) (Tetraodontidae) from the Turkish coasts. *Mediterr. Mar. Sci.* **12**, 247–252 (2011).
43. Clayton, D. Replication of animal mitochondrial DNA. *Cell* **28**, 693–705 (1982).
44. Chu, H. T. et al. Quantitative assessment of mitochondrial DNA copies from whole genome sequencing. *BMC Genomics* **13**, S5 (2012).
45. Munwes, I., Geffen, E., Friedmann, A., Tikochinski, Y. & Gafny, S. Variation in repeat length and heteroplasmy of the mitochondrial DNA control region along a core-edge gradient in the eastern spadefoot toad (*Pelobates syriacus*). *Mol. Ecol.* **20**, 2878–2887 (2011).
46. Blaser, M., Bork, P., Fraser, C., Knight, R. & Wang, J. The microbiome explored: recent insights and future challenges. *Nat. Rev. Microbiol.* **11**, 213–217 (2013).
47. Cowles, T. in *Handbook of Scaling Methods in Aquatic Ecology: Measurement, Analysis, Simulation* (eds Seuront, L. & Strutton, P. G.) 31–49 (CRC Press, Boca Raton, 2003).
48. Wiebe, P. H. et al. New development in the MOCNESS, an apparatus for sampling zooplankton and micronekton. *Mar. Biol.* **87**, 313–323 (1985).
49. Leis, J. & Carson-Ewart, B. M. (eds) *The Larvae of Indo-Pacific Coastal Fishes: An Identification Guide to Marine Fish Larvae* (Brill, Leiden, 2000).
50. Richards, W. J. *Early Stages of Atlantic Fishes: An Identification Guide for the Western Central North Atlantic* (CRC Press, Boca Raton, 2005).
51. Ivanova, N. V., Zemlak, T. S., Hanner, R. H. & Hebert, P. D. N. Universal primer cocktails for fish DNA barcoding. *Mol. Ecol. Notes* **7**, 544–548 (2007).
52. Klöppel, A., Brümmer, F., Schwabe, D. & Morlock, G. Detection of bioactive compounds in the mucus nets of *Dendropoma maxima*, Sowerby 1825 (Prosobranch Gastropod Vermetidae, Mollusca). *J. Mar. Biol.* **2013**, 283506 (2013).
53. Khalaf, M. Fish fauna of the Jordanian coast, Gulf of Aqaba, Red Sea. *J. King Abdulaziz Univ. Mar. Sci.* **15**, 23–50 (2004).
54. Hensley, D. A. Two new flatfish records from the Red Sea, an Indopacific samarid (*Samariscus inornatus*) and the European plaice (*Pleuronectes platessa*). *Isr. J. Zool.* **39**, 371–379 (1993).
55. Russell, B. C. & Golani, D. A review of the fish genus *Parascolopsis* (Nemipteridae) of the western Indian Ocean, with description of a new species from the northern Red Sea. *Isr. J. Zool.* **39**, 337–347 (1993).
56. Ben-Tuvia, A. A review of the Indo-West Pacific congrid fishes of genera *Rhynchoconger* and *Bathycongrus* with the description of three new species. *Isr. J. Zool.* **39**, 349–370 (1993).
57. Randall, J. E. & Golani, D. Review of the moray eels (Anguilliformes: Muraenidae) of the Red Sea. *Bull. Mar. Sci.* **56**, 849–880 (1995).
58. Kimura, S., Golani, D., Iwatsuki, Y., Tabuchi, M. & Yoshino, T. Redescriptions of the Indo-Pacific atherinid fishes *Atherinomorus forskalii*, *Atherinomorus lacunosus*, and *Atherinomorus pinguis*. *Ichthyol. Res.* **54**, 145–159 (2007).
59. Golani, D. *Upeneus Davidaromi*, a new deep water goatfish (Osteichthyes, Mullidae) from the Red Sea. *Isr. J. Zool.* **47**, 111–121 (2001).
60. Golani, G. & Lerner, A. A long-term study of the sandy shore icthyofauna in the northern Red Sea (Gulf of Aqaba) with reference to adjacent mariculture activity. *Raffles Bull. Zool.* **14**, 255–264 (2007).
61. Baranes, A. & Golani, D. An annotated list of deep-sea fishes collected in the northern Red-Sea, Gulf-of-Aqaba. *Isr. J. Zool.* **39**, 299–336 (1993).
62. Herler, J., Bogorodsky, S. V. & Suzuki, T. Four new species of coral gobies (Teleostei: Gobiidae: *Gobiodon*), with comments on their relationships within the genus. *Zootaxa* **3709**, 301–329 (2013).
63. Khalaf, M. A. & Kochzius, M. Community structure and biogeography of shore fishes in the Gulf of Aqaba, Red Sea. *Helgol. Mar. Res.* **55**, 252–284 (2002).
64. Randall, J. E. & van Egmond, J. in *Results of the 'Oceanic Reefs' Expedition to the Seychelles (1992–1993)* Vol. 1 (ed. van der Land, J.) 1–70 (Nationaal Natuurhistorisch Museum, Leiden, 1994).
65. Freinschlag, M. & Patzner, R. A. Shrimp-gobies in the southern Gulf of Aqaba (Red Sea). *Zool. Middle East* **55**, 41–46 (2012).
66. Fricke, R., Golani, D., Appelbaum-Golani, B. & Zajonz, U. New record of the spiny pufferfish, *Tylerius spinossissimis* (Regan, 1908), from Israel, Gulf of Aqaba, Red Sea (Actinopterygii : Tetraodontiformes: Tetraodontidae). *Acta Ichthyol. Piscat.* **46**, 115–118 (2016).
67. Pietsch, T. W. & Grobecker, D. B. *Frogfishes of the World: Systematics, Zoogeography, and Behavioral Ecology* (Stanford Univ. Press, Stanford, 1987).
68. Brokovich, E. *The Community Structure and Biodiversity of Reef Fishes at the Northern Gulf of Aqaba (Red Sea) with Relation to their Habitat*. MSc thesis, Tel Aviv Univ. (2001).

69. Blecher-Gonen, R. et al. High-throughput chromatin immunoprecipitation for genome-wide mapping of in vivo protein–DNA interactions and epigenomic states. *Nat. Protoc.* **8**, 539–554 (2013).
70. Dondoshansky, I. & Wolf, Y. *Blastclust (NCBI Software Development Toolkit)* (NCBI, Bethesda, 2002).
71. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
72. Standley, K. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
73. Price, M. N., Dehal, P. S. & Arkin, A. P. Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
74. Matsen, F. A., Kodner, R. B. & Armbrust, E. V. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11**, 538 (2010).
75. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
76. R Development Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, 2011).
77. Hufbauer, R. A., Rutschmann, A., Serrate, B., Vermeil de Conchard, H. & Facon, B. Role of propagule pressure in colonization success: disentangling the relative importance of demographic, genetic and habitat effects. *J. Evol. Biol.* **26**, 1691–1699 (2013).
78. Iwasaki, W. et al. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol. Biol. Evol.* **30**, 2531–2540 (2013).
79. Palumbi, S. in *Molecular Systematics* 2nd edn (eds Hillis, D.M., Moritz, C. & Mable, B.K.) 205–247 (Sinauer Associates, Sunderland, 1996).
80. Baldwin, C. C., Mounts, J. H., Smith, D. G. & Weigt, L. A. Genetic identification and color descriptions of early life-history stages of Belizean *Phaeoptyx* and *Astrapogon* (Teleostei: Apogonidae) with comments on identification of adult *Phaeoptyx*. *Zootaxa* **2008**, 1–22 (2009).
81. Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H. & Hallwachs, W. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proc. Natl Acad. Sci. USA* **101**, 14812–14817 (2004).
82. Ivanova, N. V., Dewaard, J. R. & Hebert, P. D. N. An inexpensive, automation-friendly protocol for recovering high-quality DNA. *Mol. Ecol. Notes* **6**, 998–1002 (2006).
83. Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. & Hebert, P. D. N. DNA barcoding Australia's fish species. *Phil. Trans. R. Soc. B* **360**, 1847–1857 (2005).
84. Sade, A. R. et al. The Israel National Bathymetric Survey: northern Gulf of Aqaba/Eilat poster. *Isr. J. Earth Sci.* **57**, 139–144 (2008).

Acknowledgements

We thank M. McGrouther from the Australian Museum, P. L. Munday from James Cook University, J. Herler from the University of Vienna and P. Borsa from Universitas Udayana for providing tissue samples for this study, the staff of the Inter-University Institute for Marine Sciences in Eilat, Israel, and the Marine Science Station of The University of Jordan and Yarmouk University for their help in conducting the research. This study was supported by the United States–Israel Binational Science Foundation (BSF grant 2008/144 to M.K. and C.B.P.), the Israeli Ministry of the Environment (grant 111-51-6 to M.K. and R.H.), the Angel Faivovich Foundation (to R.S.) and by the Nancy & Stephen Grand Israel National Center for Personalized Medicine. Field sampling was supported in part by the World Bank, as part of the Red Sea–Dead Sea Water Conveyance Study Program.

Author contributions

M.K., C.B.P., R.S. and R.H. designed the study. N.K., I.K., I.B., A.R. and M.O. performed the field sampling of larvae. N.K., O.Z., G.A., T.G., R.A.-Z., I.K., S.M., C.B.P., M.K. and R.H. processed the field samples and collected data for the COI database. O.Z., G.A., S.G., S.B. and R.S. performed and analysed the high-throughput sequencing. O.Z., N.K., M.K., R.H. and R.S. analysed the data. M.K., R.H., R.S., O.Z. and N.K. wrote the paper. All authors contributed to writing the manuscript through comments and discussions.

Competing interests

The authors declare no competing financial interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-017-0413-2.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to R.H. or M.K. or R.S.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature research

Corresponding author(s): Kiflawi, Holzman, Sorek

☐ Initial submission ☐ Revised version ☒ Final submission

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

Experimental design

1. Sample size

   Describe how sample size was determined.

   Sampling was determined based on the availability of ship time. Available cruise days were spread to allow monthly sampling of multiple sites, and to provide the best possible spatial resolution (sampling sites and depths).

2. Data exclusions

   Describe any data exclusions.

   Larval cohorts visually identified as belonging to one of five abundant pelagic taxa (Myctophidae, Phosichthyidae, Paralepididae, Trichiuridae, and Sternoptychidae) were discarded from the study. In addition, one visually identified morphotype of larvae, encompassing the genus Cirrhilabrus and the species Paracheilinus octotaenia (n=285 larvae), were manually excluded from the larvae samples for the purpose of a separate study that will be published elsewhere.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   All the computational stages that are non-deterministic were repeated to verify reliable reproduction.

4. Randomization

   Describe how samples/organisms/participants were allocated into experimental groups.

   This is a field study, and no allocation is possible. All the organisms contained in the samples were analyzed (but see 2)

5. Blinding

   Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

   During all laboratory stages (DNA extraction, silhouette imaging, and sequencing), investigators were blinded to the allocation of groups to sites/depth/sampling date. The computer code used for bioinformatic analysis was identical for all samples regardless of sites/depth/sampling date

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
☐	☒ A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ A statement indicating how many times each experiment was replicated
☐	☒ The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
☐	☒ A description of any assumptions or corrections, such as an adjustment for multiple comparisons
☐	☒ The test results (e.g. *P* values) given as exact values whenever possible and with confidence intervals noted
☐	☒ A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
☐	☒ Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	Code used in the bioinformatic analysis was deposited in github: https://github.com/omerzu/larvae_project Statistical analysis was done using the software R statistics, using libraries 'vegan', 'mclust', and 'lmodel2'

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* guidance for providing algorithms and software for publication provides further information on this topic.

Materials and reagents

Policy information about availability of materials

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.	COI barcodes sequenced in this study were deposited in Barcode Of Life Database (BOLD; http://www.boldsystems.org/). BOLD Reference numbers appear in Supplementary Table 3. Raw metagenomics data was deposited in the ENA repository, under accession number PRJEB20625.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).	No antibodies were used in this study

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	No eukaryotic cell lines were used in this study
b. Describe the method of cell line authentication used.	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used have been authenticated OR state that no eukaryotic cell lines were used.*
c. Report whether the cell lines were tested for mycoplasma contamination.	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination OR state that no eukaryotic cell lines were used.*
d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	*Provide a rationale for the use of commonly misidentified cell lines OR state that no commonly misidentified cell lines were used.*

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.	The study used environmental samples of ichthyoplankton, and no laboratory animals were used. Sampling of larvae and adult reef fish was done under the approval of the Authority of Nature Protection in Israel.

Policy information about studies involving human research participants

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.	The study did not involve human research participants