MOLECULAR ECOLOGY

Molecular Ecology (2015) 24, 6241–6255 doi: 10.1111/mec.13471

Seascape genetics along environmental gradients in the Arabian Peninsula: insights from ddRAD sequencing of anemonefishes

PABLO SAENZ-AGUDELO,*† JOSEPH D. DIBATTISTA,*‡ MAREK J. PIATEK,§ MICHELLE R. GAITHER,¶** HUGO B. HARRISON,†† GERRIT B. NANNINGA*‡‡ and MICHAEL L. BERUMEN*

**Division of Biological and Environmental Science and Engineering, Red Sea Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia, †Instituto de Ciencias Ambientales y Evolutivas, Universidad Austral de Chile, Valdivia, Chile, ‡Department of Environment and Agriculture, Curtin University, PO Box U1987, Perth, WA 6845, Australia, §Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955, Saudi Arabia, ¶School of Biological and Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, UK, **Section of Ichthyology, California Academy of Sciences, 55 Music Concourse Drive, San Francisco, CA 94118, USA, ††Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, QLD 4811, Australia, ‡‡USR 3278 CRIOBE CNRS-EPHE, CRIOBE BP 1013, Papetoai 98729, Moorea, French Polynesia*

Abstract

Understanding the processes that shape patterns of genetic structure across space is a central aim of landscape genetics. However, it remains unclear how geographical features and environmental variables shape gene flow, particularly for marine species in large complex seascapes. Here, we evaluated the genomic composition of the two-band anemonefish *Amphiprion bicinctus* across its entire geographical range in the Red Sea and Gulf of Aden, as well as its close relative, *Amphiprion omanensis* endemic to the southern coast of Oman. Both the Red Sea and the Arabian Sea are complex and environmentally heterogeneous marine systems that provide an ideal scenario to address these questions. Our findings confirm the presence of two genetic clusters previously reported for *A. bicinctus* in the Red Sea. Genetic structure analyses suggest a complex seascape configuration, with evidence of both isolation by distance (IBD) and isolation by environment (IBE). In addition to IBD and IBE, genetic structure among sites was best explained when two barriers to gene flow were also accounted for. One of these coincides with a strong oligotrophic–eutrophic gradient at around 16–20°N in the Red Sea. The other agrees with a historical bathymetric barrier at the straight of Bab al Mandab. Finally, these data support the presence of interspecific hybrids at an intermediate suture zone at Socotra and indicate complex patterns of genomic admixture in the Gulf of Aden with evidence of introgression between species. Our findings highlight the power of recent genomic approaches to resolve subtle patterns of gene flow in marine seascapes.

Keywords: empirical, fish, hybridization, landscape genetics, population genetics

Received 9 July 2015; revision accepted 9 November 2015

Introduction

Elucidating the processes that shape spatial patterns of genetic variation in natural populations is a central aim of evolutionary biology (Avise 2000). Many species occupy geographical ranges that are heterogeneous in terms of their geography and environment, both of which are fundamental components shaping patterns of gene flow across landscapes. Factors that limit gene flow among populations can lead to divergence via genetic drift (Wright 1943; Slatkin 1987; Riginos &

Correspondence: Pablo Saenz-Agudelo, Fax: +56 9 2221324; E-mail: pablo.saenzagudelo@gmail.com

Liggins 2013). Alternatively, selection in different ecological environments can drive divergence between populations and limit gene flow despite, in some cases, high dispersal (Via 2002; Nosil *et al.* 2009). Overall, disentangling the relative contribution of geography and environment to patterns of genetic diversity in natural populations is important in the identification of drivers of connectivity (Wang & Bradburd 2014).

Discerning the relative contribution of geography and environment in shaping genetic diversity remains a challenging endeavour. This is because in nature, environmental heterogeneity, spatial distance and habitat distribution are part of the same landscape matrix in which populations evolve. One major complication in discerning the role of these two factors in evolution is the fact that geographical distance and environmental differences are often correlated. In addition, identifying and quantifying the underlying processes (selection, drift, dispersion) that shape the patterns of genetic variation in natural populations remains a difficult task (Wang *et al.* 2013; Sexton *et al.* 2014; Wang & Bradburd 2014). However, the development of new statistical methods, the proliferation of protocols that allow for high density SNP discovery and genotyping in large populations, as well as ease of accessing environmental data derived from satellite imagery provide a new opportunity to examine how geography and environment shape genetic structure in natural populations (Balkenhol *et al.* 2009; Wang & Bradburd 2014).

Coral reefs of the Arabian Peninsula are rich and diverse ecosystems characterized by high levels of endemism across many taxa (Klausewitz 1989; DiBattista *et al.* 2015a,b). The geological history of the region has played an important role in generating this unique biodiversity. The Red Sea rift led to the separation of the Arabian plate from the horn of Africa during the Oligocene (~25 Mya), opening the Red Sea to the Indian Ocean via the 30 km wide strait of Bab al Mandab. This narrow and shallow passage was almost completely emerged during Pleistocene glacial cycles (2.5 Mya), when sea level lowered as much as 140 m (Rohling *et al.* 1998) isolating the Red Sea from the Gulf Aden and Indian Ocean. The formation of the Red Sea created a unique set of environmental conditions including minimal freshwater inflow, high evaporation rates and a latitudinal gradient of temperature, salinity and nutrient concentrations (Sofianos 2003; Raitsos *et al.* 2013). These environmental conditions have played an important role in shaping the structure and diversity of shallow-water marine communities in the region (Sheppard & Sheppard 1991; Roberts *et al.* 1992; DiBattista *et al.* 2015b).

Coral reef communities in the Red Sea display heterogeneous composition and structure along the north to south axis (Sheppard & Sheppard 1991; Roberts *et al.* 1992; Khalaf & Kochzius 2002; DiBattista *et al.* 2015b). However, relatively few studies have assessed genetic variation within the region (Berumen *et al.* 2013). These studies include reef fishes (*Larabicus quadrilineatus*, Froukh & Kochzius 2007; *Amphiprion bicinctus*, Nanninga *et al.* 2014), a reef sponge (*Stylissa carteri*, Giles *et al.* 2015) and a coral (*Pocillopora verrucosa*, Robitzch *et al.* 2015), as well as a mussel (*Brachiodontes pharaonis*, Shefer *et al.* 2004). Interestingly, all but one survey (Robitzch *et al.* 2015) found evidence of a genetic break at around 16–20°N latitude. In particular, both Nanninga *et al.* (2014) and Giles *et al.* (2015) found that genetic structure along 1500 km of the Saudi Arabian coast of the Red Sea was best explained by a combination of isolation by distance (IBD) and isolation by environment (IBE). The authors suggest that this internal genetic break coincides with a zone of environmental transition from oligotrophic waters to eutrophic waters associated with a turbid, shallow reef system. It remains unclear, however, if this apparent correlation between genetic and environmental differences is truly the result of adaption to different environmental conditions in the Red Sea. It has been suggested that the sharp environmental gradient around 16–20°N in the Red Sea could be the product of specific topographic and oceanographic conditions in this particular area (Raitsos *et al.* 2013). Under this scenario, it is possible that larval connectivity and thus gene flow across the 16–20°N transition zone are limited by physical (oceanographic and topographic) constraints. If this is the case, then it would still be possible to observe a significant IBE (because both oceanographic/topographic and the environmental transitions occur in the same area) but that would not involve necessarily a causal link to local adaptation to different environments.

Here, we evaluated the genomic composition of the two-band anemonefish *A. bicinctus* across its entire geographical range with a genotyping by sequencing approach using a double-digest restriction-site-associated DNA protocol (ddRAD; Peterson *et al.* 2012). *Amphiprion bicinctus* is a common inhabitant of coral reefs in the Red Sea that is found between 0.5 to 45 m depths in association with five species of anemones (*Entacmaea quadricolor, Stichodactyla haddoni, Heteactis aurora, Heteractis crispa* and *Heteractis magnifica*). While its host anemones occur throughout the Indo-Pacific region, *A. bicinctus* is endemic to the Red Sea and parts of the Gulf of Aden (Nanninga *et al.* 2014; DiBattista *et al.* 2015c). In addition, based on colouration patterns and differentiation at a single nuclear gene (RAG2), a recent study (DiBattista *et al.* 2015c) found that *A. bicinctus* might hybridize with *Amphiprion omanensis* a closely related species endemic to the Southern coast of Oman (Santini & Polacco 2006; Litsios *et al.* 2014).

We performed a genomewide survey among 11 sites and compared patterns of genomic divergence in *A. bicinctus* across among two previously described barriers to gene flow in the region: (1) oligotrophic–eutrophic barrier (or possibly a circulation barrier) around 20°N (Roberts *et al.* 1992; Nanninga *et al.* 2014) and (2) the straight of Bab al Mandab at the junction between the Red Sea and Indian Ocean (an ancient bathymetric barrier; DiBattista *et al.* 2013). We also consider a monsoonal upwelling barrier of the monsoonal upwelling system of the Arabian Sea, which may have led to the speciation of *A. bicinctus* and *A. omanensis*. Our first aim was to evaluate genomic divergence across these barriers. We then determined if genomic patterns differed between barriers and whether or not they displayed signals that are consistent with the intrinsic nature of each barrier. If adaptation to environmental conditions is driving patterns of genetic diversity, then we would expect to find a subset of loci under selection and for these to depart from neutral expectations across putative barriers. On the other hand, if drift alone is responsible for the genetic differences, we might expect fewer outlier loci regardless of the degree of divergence among populations on either side of the barrier. Finally, we evaluated the permeability of the Gulf of Aden and Oman upwelling barrier by evaluating patterns of genetic admixture between *A. bicinctus*, *A. omanensis* and their putative hybrids from Socotra Island.

Materials and methods

Study region and sample collection

Samples of *A. bicinctus* (Fig. 1a) were collected from nine sites between the Gulf of Aqaba (28°N, 34°E) in the northern Red Sea and Djibouti (12°N, 43°E) in the Gulf of Aden (Fig. 1d). In addition, we used samples of its congeneric species *A. omanensis* (Fig. 1c) collected by Simpson *et al.* (2014) from the Dhofar region of Oman (16°N, 54°E). Previous reports also indicate that both *A. bicinctus* and *A. omanensis* may co-occur in Socotra, Yemen (12°N, 53°E; Kemp 1998; Zajonz *et al.* 2000). At Socotra, anemonefish are generally rare; only individuals of intermediate colouration were observed and subsequently collected (Fig. 1b). These likely represent putative hybrids of these closely related species (Litsios *et al.* 2014; DiBattista *et al.* 2015c). In total, 144 samples were collected across 12° of latitude and spanning the known geographical distribution for *A. bicinctus* and its putative outposts.

Library preparation and sequencing

Genomic DNA was extracted from fin or gill tissue preserved in 96% ethanol or salt-saturated DMSO buffer using a Nucleospin-96 Tissue kit (Macherey-Nagel, Düren, Germany). Double-digest restriction-associated DNA (ddRAD) libraries were prepared using 500 ng of DNA per sample following the protocol described by Peterson *et al.* (2012) with some modifications. Briefly, genomic DNA was digested at 37 °C for 3 h using the restriction enzymes *SphI* and *MluCI* (NEB) followed by a ligation step, whereby each sample was assigned to one of 16 unique adaptors. Pools of 16 individuals were combined and run on agarose gel, where 300- to 500-bp fragments were manually excised and purified using a Zymoclean Gel DNA recovery kit. Each pool was amplified using 10 PCR cycles in 50 μl reactions containing 25 μl Illumina True Seq Master Mix, 20 μl of library DNA and a unique indexing primer for each pool that corresponds to the standard Illumina multiplexed sequencing protocol. PCRs were carried out in a

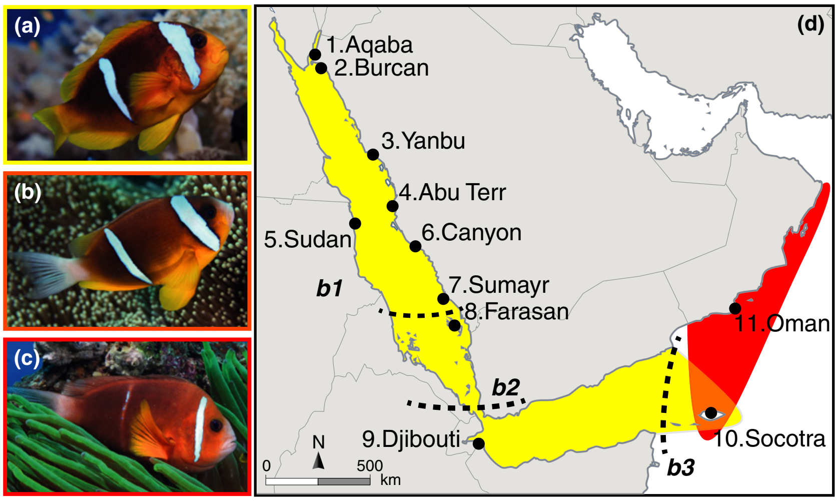

Fig. 1 *Amphiprion bicinctus* (panel a). Suspected hybrid between *A. bicinctus* and *A. omanensis* which exhibits an intermediate morphology (panel b). *A. omanensis* (panel c). Panel d shows the map of sampling sites (numbers in brackets indicate the codes used in Fig. 2) and distribution of *Amphiprion bicinctus* (yellow shading), *A. omanensis* (red shading) and their suspected hybrids, (orange shading). Black dashed lines denote the barriers to dispersal considered in this study: 20°N (*b*1) and the Strait of Bab al Mandab (*b*2), Gulf of Aden and Oman upwelling (*b*3).

Veriti 96-well thermal cycler (Life Technologies) using the following protocol: initial step heating at 98 °C for 30 s, ten cycles (98 °C for 10 s, 60 °C for 30 s and 72 °C for 30 s), followed by a final step at 72 °C for 5 min. DNA libraries were quantified using the high sensitivity DNA analysis kit in a 2100 Bioanalyser (Agilent Technologies). Pools were combined in equimolar concentration to form a single genomic library and sequenced in one lane of a HiSeq 2000 Illumina sequencer (single end, 1 × 101 bp; v3 reagents).

De novo assembly

Sequences were demultiplexed and filtered for quality using the 'process_radtags.pl' pipeline in STACKS version 1.18 (Catchen *et al.* 2011). Individual reads with phred scores below 20 (average on sliding window) or with ambiguous barcodes were discarded. In the absence of a reference genome for this species, RADSeq loci were assembled *de novo* using the 'denovo_map.pl' pipeline in STACKS. Different parameter combinations were evaluated, which resulted in different numbers of loci (Appendix S1, Supporting information) but gave similar results in genetic comparisons (genetic clustering and pairwise F_{ST} among sites). For the main analyses presented here, we used a parameter combination recommended by Mastretta-Yanes *et al.* (2014): minimum read depth to create a stack ($-m$) = 3, number of mismatches allowed between loci within individuals ($-M$) = 2, number of mismatches allowed between loci within catalogue ($-n$) = 2.

Following *de novo* mapping, an initial data-filtering step was performed using the *population* component of STACKS retaining only those loci present in at least 80% of individuals at each site and at least 7 of all 11 sites. A second filtering step was also performed removing loci with minor allele frequencies lower than 0.05 (to reduce the number of false polymorphic loci due to sequencing error) and loci that were absent in more than 20 samples regardless of the site (to remove loci with more than 20% missing data that passed the filtering conditions of STACKS [80% individuals within each site and 7 of 11 sites]). The *write_single_snp* option produced a vcf file with only one single-nucleotide polymorphism (SNP) per stack. The resulting vcf file was converted to other program-specific input files using PGDSPIDER version 2.0.5.1 (Lischer & Excoffier 2012).

F_{ST}, F_{IS} and genetic diversity metrics (percentage of polymorphic loci, average number of alleles and observed and expected heterozygosity) were estimated using GENODIVE version 2.0 (Meirmans & Van Tienderen 2004). The significance of pairwise F_{ST} values was tested using 10 000 permutations..

Population genetic and statistical analyses

To address all aims of this study we performed a principal components analysis (PCA) and a genetic clustering analysis. Both these analyses allowed elucidating the spatial patterns of genetic admixture along the distribution range of *A. bicinctus* and provided information about the levels of admixture between *A. bicinctus*, *A. omanesis* and their putative hybrids from Socotra. To address the second aim we used an information-theoretic approach to elucidate the main factors that shaped genetic differences among sites. In addition, we used a latent factor mixed-effect model to detect potential loci that were strongly associated with the environmental variables. We also performed a test for population splits and mixtures to evaluate for migration events across these barriers. Finally, for the last aim and to complement the Bayesian clustering analysis, we simulated hybrid individuals (F1 and Back crosses) among different sites and compared their distribution in principal component's space relative to all samples.

Principal components analysis

We performed a principal components analysis of the genotype covariance matrix using the 'dudi.pca' function in R and following Jombart *et al.* (2010). We used this to summarize genotypic variation across all sampled individuals.

Clustering analysis

To explore genetic structure across sampling sites, a clustering analysis was performed in STRUCTURE version 2.3.4 (Pritchard *et al.* 2000) without a priori information of the geographical origin of each sample. The analyses were run under the admixture model with correlated allele frequencies (Falush *et al.* 2003), with a burn-in period of 200 000 MCMC iterations, followed by 300 000 iterations for each run. The number of K (putative populations) ranged from 1 to 11 when all loci and all sites were included and ranged from one to six for analyses that involved subsets of the main data set. Five replicate analyses were run for each value of K. The number of clusters was inferred by comparing the ln Pr ($X|K$) among different values of K. The value of K for which ln Pr ($X|K$) was highest or reached a plateau was selected as the most parsimonious number of populations in our sample. The *ad hoc* statistic ΔK (Evanno *et al.* 2005) was also considered.

Determinants of genetic differentiation within the Red Sea

We used an information-theoretic approach (Anderson 2008) to determine the factors that may have shaped the observed patterns of population structure among sites. This approach is widely used in ecology, especially in situations where model uncertainty prevails (Johnson & Omland 2004), and has the advantage of ranking alternative models according to empirical evidence rather than simply accepting or rejecting each putative model (Correa & Hendry 2012). Environmental and geographical variables were included in the model selection and each model was ranked based on their evidence ratio and posterior probability. Environmental data were gathered from the NASA Giovanni website using the MODIS-Aqua 4 km database (http://oceancolor.gsfc.nasa.gov) with standard NASA estimate algorithms. Colour radiometry measurements of daytime sea surface temperature (day DSST°C), chlorophyll-a (CHLA mg/m^3), coloured dissolved organic matter (CDOM), particulate organic carbon (POC mg/m^3) and sea surface salinity (‰) were used. For all variables, 9-year annual averages (January 2003 to December 2012) were downloaded and used. Values of these variables for each site are included as Appendix S1 (Supporting information). Environmental distances among sites (*env*) were estimated by standardizing all variables, performing a principal components analysis using the 'prcomp' function in R and then calculating the distances between sites plotted on the resulting first two PC axes (Variance explained by PC1 + PC2 = 97.12%). Geographical factors included the Euclidean distance over water between sites (*geo*) and the presence of three putative barriers to gene flow: i) oligotrophic–eutrophic barrier around 16–20°N (*b*1; Nanninga *et al.* 2014); ii) the Strait of Bab Al Mandab that connects the Red Sea and Gulf of Aiden (bathymetric barrier) (*b*2); and iii) the monsoonal upwelling system in the Arabian Sea (*b*3) that may limit the distribution of *A. bicinctus*. All barriers were considered independently and in combination (*b*1_2, *b*1_3, *b*2_3 and *b*1_2_3). Each barrier was modelled as a factor with two levels (same side vs different side of barrier).

Overall, 49 models were fitted using the linear model 'lm' function in R. For each model, the sample size-corrected Akaike information criterion (AIC) was computed as AICc = AIC + $2K(K + 1)/(n-K-1)$, where AIC = −2log-likelihood + $2K$ (K = number of parameters in model, n = number of observations). Models were then ranked based on increasing AICc and further interpretation based on model probabilities (w) and evidence ratios (Anderson 2008). To compensate for the fact that pairwise F_{ST} values are not independent among sites, tests of significance were performed using a randomized permutation ($n = 10^4$) procedure as implemented in the MMRR function (Wang *et al.* 2013).

Testing for association between loci and the environmental gradient

We tested for the presence of loci that exhibited high correlation with the environmental gradient for *A. bicinctus* (sites 1–9). To do this, we used a latent factor mixed-effect model (LFMM) implemented in the package LEA in R (Frichot & François 2015). The algorithm implemented in LFMM detects correlations between environmental and genetic variation while simultaneously inferring background levels of population structure. Background population structure is introduced into these models using unobserved variables (latent factors). First, we used the function 'snmf' from the package LEA to estimate the most likely number of ancestral populations that best explains the genotypic data by evaluating the cross-entropy criterion for different values of K. Second, we performed an ecological association test using the LEA function 'lfmm' with numbers of latent factors ranging from $K = 2$ to $K = 3$ to account for background population structure. $K = 2$ corresponded to the most likely number of clusters based on STRUCTURE as well as the value of K with the minimal cross-entropy according to 'snmf' results (Appendix S2, Supporting information). As environmental variable, we used the first PC estimated in the previous. We ran the Gibbs sampler algorithm for a period of 10 000 cycles following a burn-in period of 5000 cycles. This was performed five times for each value of K. The genomic inflation factor (λ) was 1.75 for $K = 2$ and 1.63 for $K = 3$. While this value should ideally be close to 1, we found that this calibration produced a correct distribution of adjusted P-values (drawn from a uniform distribution) (Appendix S2, Supporting information). To control for false discoveries, we applied the Benjamin–Hochberg algorithm using standard R scripts as described in the LEA package documentation and using a false discovery rate (FDR) $q = 10\%$.

Testing for population splits and mixtures

We built a population tree to infer patterns of genetic admixture among the 11 sampled sites using the program TREEMIX version 1.12 (Pickrell & Pritchard 2012). The model used was based on the construction of population trees to infer their relationships and allowed for the inclusion of both the effects of population splits and gene flow. In particular, we tested if a simple bifurcating population tree could fully explain the genetic variation in the data (reflecting the amount of genetic drift among populations), or if additional migration events were needed to explain the data. We first built a maximum-likelihood tree for all 11 sites under the assumption that all SNPs were independent. Given the fact that no reference genome, genetic map or linkage was available for this species, *Amphiprion omanensis* was set as the root of the tree. To assess the confidence of the topology of the tree, ten replicate trees were generated using the bootstrap option (resampling blocks of 250 SNPs). We then built trees allowing for different numbers of migration events (1 to 10) and assessed the percentage of variance explained for each iteration. Three and four population tests were used to explore the robustness of the inferred migration events (Reich *et al.* 2009).

Simulation of hybrids

We used the function 'hybridize' from the package adegenet (Jombart *et al.* 2010) in R to simulate hybrid individuals. This function performs hybridization between two sets of genotypes (populations). The function estimates allelic frequencies for each population and gametes are sampled following a multinomial distribution. For each simulation a set of 12 hybrid genotypes were produced from two parental populations (sites). Hybrid individuals were then included in a second PCA as described previously.

Results

Raw sequence filtering and assembly

A total of 107 186 385 reads of 101 bp each were obtained for 144 individual samples from 11 locations in the Red Sea and Arabian Sea (Fig. 1d). As a conservative measure, 29 samples were discarded due to low read recovery (<250 000), which comprised only 2.7% of all reads. Of the remaining reads, 86.7% were successfully built into 'STACKS' with 180 776 to 2 566 557 reads per individual. On average, ~800 000 reads and ~200 000 STACKS per individual were recovered and used to build loci. The minimum average depth of coverage per individual ranged between 6× and 20× and averaged 10.5× across all samples. A total of 60 142 loci with at least one SNP were recovered. Additional filtering was performed to eliminate loci with more than 20% missing data and individual samples for which more than 25% of loci were missing. Overall, 4559 loci were retained for 115 samples from 11 sites. Depth of coverage per sample after filtering ranged from 7× to 75× and averaged 25× coverage per locus per sample.

Principal components analysis

The amount of genotypic variation explained by the first three principal components (PC) was low (11%). Yet, PC1 (accounting for 5.77% of the variation) discriminated well simples from sites 10 and 11 (Socotra and Oman) from the rest. PC1 also discriminated to a lesser extent samples from Sumayr, Farasan and Djibouti in the southern Red Sea (sites 7–9) from samples from the central and northern Red Sea (sites 1–6). PC2 (3.03% of variation) further distinguished samples from the southern Red Sea (sites 7–9) from the central and northern Red Sea region. Interestingly, PC3 (2.13% of variation) separated site 10 from site 11 indicating a small but evident difference between these sites (Fig. 2). Six different sets of simulated hybrids were then added to the PCA (Appendix S2, Supporting information). Interestingly, some simulated hybrids were in proximity to Socotra when plotting PC1 and PC2. In particular, simulated F1 hybrids between sites Sumayr and Oman as well as backcrossed hybrids (F1 Oman-Farasan X Oman) were plotted in close proximity to Socotra individuals. However, PC3 clearly separated Socotra samples from all other samples including all simulated hybrids.

Population genetic statistics

A summary of the principal statistics (number of individuals per site, percentage of polymorphic loci, average number of alleles, observed and expected heterozygosity and F_{IS}) is presented in Table 1. Global F_{ST} among all sites was 0.054. The distribution of F_{ST} per SNP is presented as Appendix S2 (Supporting information). Briefly, 90% of all F_{ST} values were below 0.127. The 90th to 100th percentile comprised values ≥ 0.127. Site pairwise F_{ST} values varied between 0.0 (Yanbu – Abu Terr) and 0.178 (Sudan–Socotra) (Table 2). When populations were grouped according to Bayesian clustering results (see below), global F_{CT} among $K = 3$ groups (Fig. 3) was 0.064 and F_{ST} between sites but within groups was 0.015.

Bayesian clustering analysis

With the complete data set of 4559 SNPs from 113 samples across 11 sites, the Bayesian clustering analysis suggested $K = 2$ populations as the most parsimonious partitioning of individuals based on the metric ΔK (Evanno *et al.* 2005). This division separates sites from the north and central Red Sea (sites 1–6) from Socotra and Oman (sites 10–11), with sites in the southern Red Sea (7–9) displaying different levels of admixture between these two clusters. When plotting ln $\Pr(X|K)$ against K to determine the most likely number of clusters, we found $K = 3$ as the most likely partition of the samples (Fig. 3), with a curve-like trend as K increased. This result is indicative of scenarios where one contact

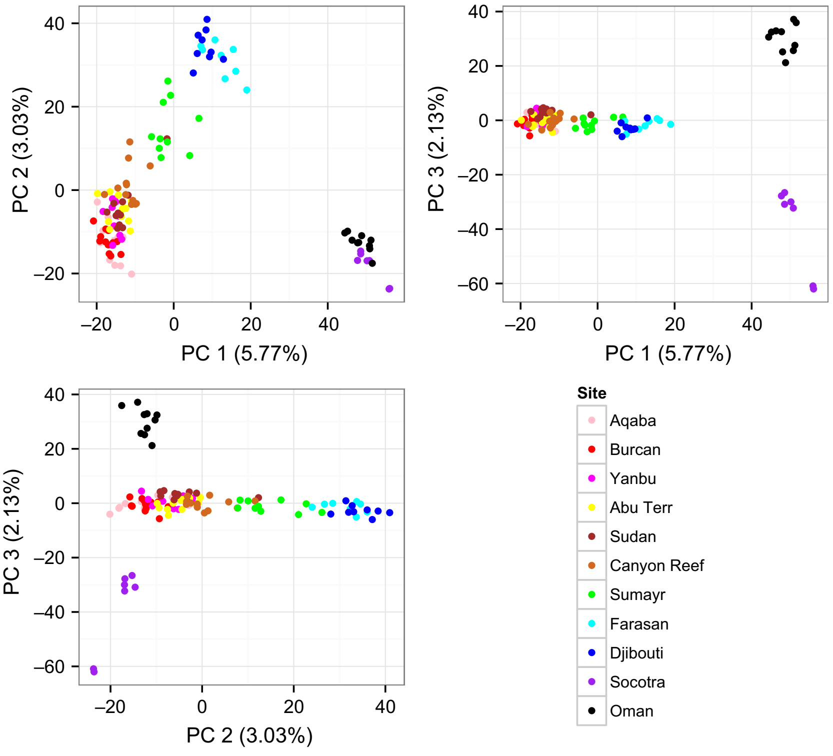

Fig. 2 Principal components analysis of multilocus genotypes for *Amphiprion bicinctus, Amphiprion omanensis* and putative hybrids from Socotra. Each point represents one individual fish. Colour codes correspond to sampling sites.

Table 1 Sites, sample size and molecular metrics for *Amphiprion bicintus* and *A. omanensis* populations based on 4559 RADSeq loci (SNPs)

Site	SC	Lat	Lon	N	Na	Ho	He	F_{IS}
Aqaba	Aqa	28.404	34.741	12	1.874	0.237	0.262	0.093
Burcan	Bur	27.910	35.065	12	1.863	0.234	0.26	0.099
Yanbu	Yan	24.150	37.675	10	1.792	0.203	0.246	0.173
Abu Terr	At	21.676	38.841	12	1.862	0.226	0.256	0.119
Sudan	Sud	21.207	37.232	11	1.758	0.187	0.239	0.219
Canyon Reef	Can	19.890	39.961	12	1.879	0.235	0.257	0.085
Sumayr	Sum	17.787	41.442	10	1.83	0.225	0.255	0.12
Farasan	Far	16.618	41.938	8	1.774	0.245	0.26	0.057
Djibouti	Dji	11.721	43.165	9	1.705	0.221	0.237	0.067
Socotra	Soc	12.659	53.907	7	1.632	0.2	0.216	0.078
Oman	Oma	16.969	55.085	10	1.649	0.18	0.214	0.16

SC, site code; Na, average number of alleles; Ho, observed heterozygosity; He, expected heterozygosity; F_{IS}, inbreeding coefficient.

zone has significant levels of substructure within zones and is likely driven by the differences among species rather than among sites (Evanno *et al.* 2005). Both summary statistics (ΔK and ln $\Pr(X|K)$) are presented as Appendix S3 (Supporting information). In an attempt to identify population substructure in *A. bicinctus* only, Socotra and Oman were excluded from further Bayesian clustering analyses. In this case, $K = 2$ was the most likely solution, providing further support for the general $K = 3$ as the most likely overall scenario (Fig. 2). While $K = 2$ suggests introgression of *A. omanensis* DNA into *A. bicinctus* in the southern Red Sea, a $K = 3$ solution would suggest a genetic substructure of *A. bicinctus* between northern (1–6) and southern (8–9) Red Sea sites, with Farasan Islands (site 8) showing introgression with *A. omanensis* and Sumayr (site 7) showing intermediate levels of admixture between the northern and southern Red Sea genetic clusters.

To evaluate if patterns of genetic clustering changed among different loci, clustering analyses were also

Table 2 Pairwise F_{ST} values for *Amphiprion bicintus* and *A. omanensis* populations based on 4559 RADSeq loci (SNPs)

	Aqa	Bur	Yan	At	Sud	Can	Sum	Far	Dji	Soc	Oma
Aqa	—										
Bur	0.002	—									
Yan	**0.006**	0.002	—								
AT	**0.006**	0.003	−0.001	—							
Sud	**0.013**	**0.009**	0.004	0.005	—						
Can	**0.012**	**0.009**	**0.004**	0.002	**0.008**	—					
Sum	**0.034**	**0.029**	**0.020**	**0.017**	**0.018**	0.008	—				
Far	**0.075**	**0.077**	**0.065**	**0.06**	**0.062**	**0.047**	0.014	—			
Dji	**0.088**	**0.083**	**0.074**	**0.066**	**0.071**	**0.057**	**0.032**	**0.028**	—		
Soc	**0.165**	**0.172**	**0.177**	**0.169**	**0.178**	**0.165**	**0.148**	0.137	**0.172**	—	
Oma	**0.156**	**0.163**	**0.165**	**0.158**	**0.161**	**0.153**	**0.132**	**0.118**	**0.152**	**0.099**	—

Values in bold were significant at 0.05 after Bonferroni's correction (based on 10 000 permutations). Abbreviations are listed in Table 1.

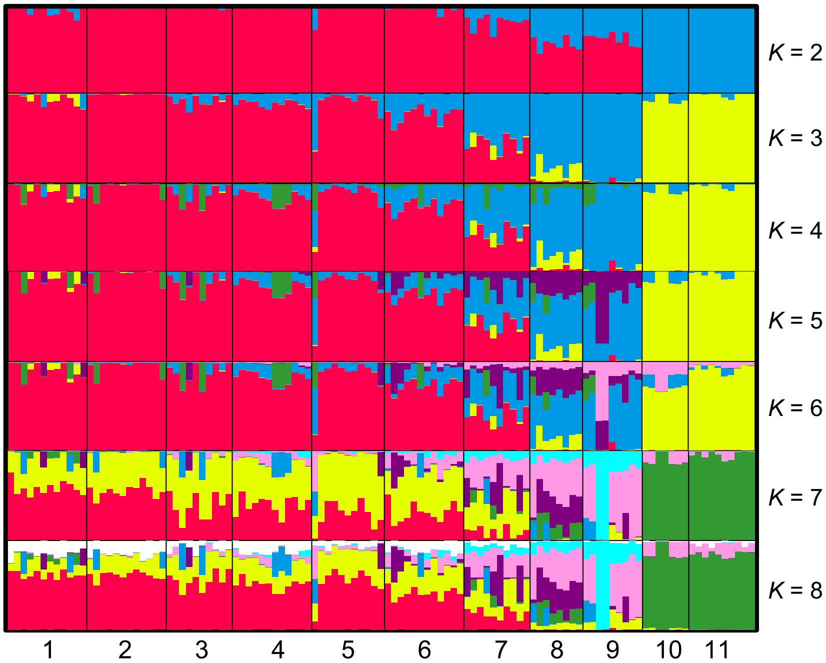

Fig. 3 Results of Bayesian clustering for different values of K (2–8) using 4559 SNPs. Each bar corresponds to an individual fish and colours in each bar correspond to estimates of admixture proportion. Numbers at the bottom correspond to sampling sites in fig. 1.

performed for different subsets of SNPs based on the distribution of global F_{ST} (Appendix S2, Supporting information). While no evidence of genetic clustering was identified with the SNP subset in the 60th percentile based on global F_{ST} (see ln $\Pr(X|K)$ plot in Appendix S2, Supporting information), all other SNP subsets indicated similar groupings between sites as when the analysis was performed with all loci, with $K = 3$ being the most likely model. No remarkable differences in admixture patterns were observed for different subsets of loci in sites 1–9 compared to results when all loci were included. However, site 10 (Socotra) did show variation in admixture proportions depending on the subset of loci analysed. In particular, loci in the 60th to 80th percentile showed evidence of admixture in Socotra between the southern Red Sea cluster and the Oman cluster (Appendix S2, Supporting information).

Explaining patterns of genetic structure in the Arabian Sea

Results from Bayesian clustering showed a gradient of admixture in sites 6 to 8 providing alternative putative locations to place the *b*1 barrier) between sites 5 and 6, 6 and 7 or 7 and 8). We compared three independent models to evaluate which placement of *b*1 explained best pairwise genetic variation. We found that pairwise genetic differentiation was best explained when *b*1 was placed between sites 7 and 8 (Fig. 1d) (Appendix S1, Supporting information). This model had a probability of 99% compared to the second best model, which placed *b*1 between sites 6 and 7 and had a probability of less than 1%. Therefore, *b*1 was placed between sites 7 and 8 for all subsequent analyses.

When all sites were included in the models, genetic differentiation (F_{ST}) was best explained by models that included the presence of at least one barrier. That is, the top 30 ranked models (of 67) included at least one barrier (Appendix S4, Supporting information). One model outperformed all the other models (model probability of 0.86 compared to 0.06 of the next model in the rank). This model included *env*, *geo*, the presence of the *b*3 barrier (Gulf of Aden and Oman upwelling) as well as an interaction between *env* and *b*3 as well as *geo* and *b*3 (Fig. 4 panels a–e). This model confirmed that: (i) The average pairwise F_{ST} among sites on *different* sides of the Gulf of Aden and Oman upwelling barrier was 2.9 times higher than the average F_{ST} among sites on the same side of the barrier (F_{ST} same = 0.057 ± 0.0040 SE, $P < 0.001$; F_{ST} different = 0.149 ± 0.0100 SE, $P < 0.001$). (ii) Pairwise F_{ST} values were positively correlated with *geo* within the same side of *b*3 but not across it (same side: 0.038 ± 0.0062 SE, $P < 0.001$; different sides: 0.0104 ± 0.0045, $P = 0.076$) (Fig. 4b). (iii) Pairwise F_{ST} values were positively correlated with *env* for comparisons within the same side of the barrier (slope: 0.0104 ± 0.0041 SE, $P < 0.008$), but not across different sides (slope: −0.006 ± 0.006 SE; $P = 0.31$) (Fig. 4c). Overall, F_{ST} values increased as a function of the combined effect of *env* plus *geo* within the same side of the barrier (combined slope: 0.049), but was close to zero among sites in different sides of the barrier (combined slope: 0.007) (Fig. 4d). There was a small but significant correlation between *env* and *geo* ($R^2 = 0.165$, $F = 10.46$, $P = 0.002$) (Fig. 4e).

We repeated this analysis excluding sites 10 and 11 (*A. omanensis* and putative hybrid samples) and thus *b*3 (Gulf of Aden and Oman upwelling) (Fig. 4, panels f–j). Genetic differentiation within sites 1–9 was best explained by models that included the presence of at least one barrier. That is, the top 25 ranked models (of 31 evaluated) included one or two barriers (*b*1: the oligotrophic–eutrophic barrier between sites 7 and 8; *b*2: the straight of Bab al Mandeb barrier) as effects (Appendix S5, Supporting information). One model outperformed all other models (model probability of 52% compared to 22% of the next model in the rank). This model included *geo*, *env*, *b*1_2 (a variable including both *b*1 and *b*2) and an interaction between *env* and *b*1_2. This model suggested that: (i) the average pairwise F_{ST} among sites within the same side of the *b*1 barrier was 2.3 times lower than the average F_{ST} among sites across *b*1 and 4.1 times lower than the average F_{ST} among sites

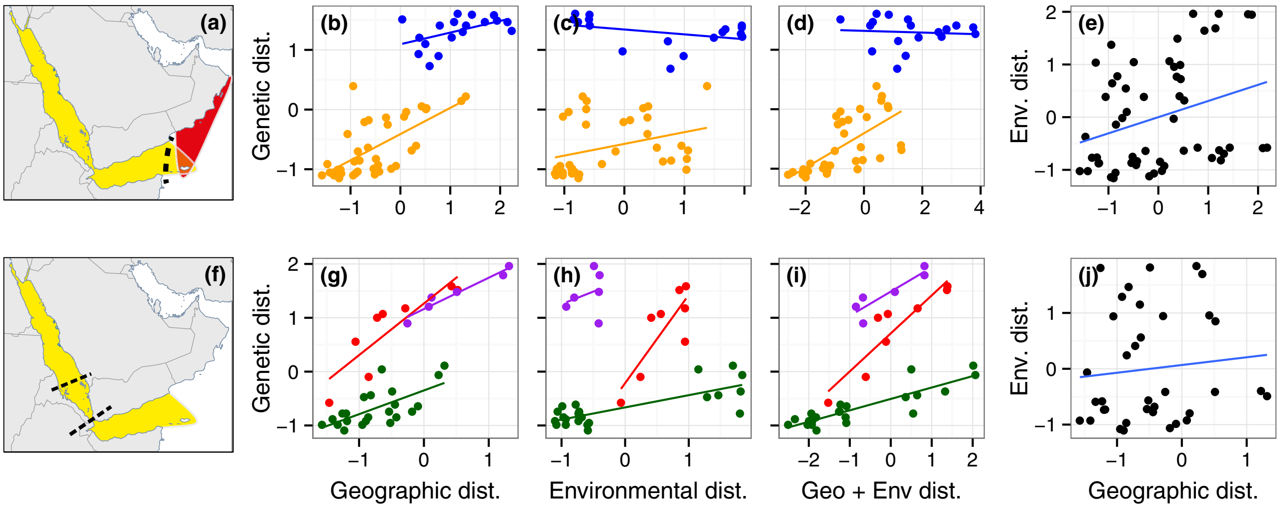

Fig. 4 Correlation between pairwise genetic distance, geographical and environmental distances for *Amphiprion bicinctus* and *A. omanensis* around the Arabian Peninsula (a–e) and *A. bicinctus* in the Red Sea (f–j). In each case, genetic distance (F_{ST}) was compared to geographical distance (b, g), environmental distance (c, h) and the combined geographical–environmental distance (d, i). Correlation between geographical and environmental distance are shown in (e, j). For panels b–d, orange dots correspond to the pairwise comparison of sites on the same side of the *b*3 barrier (Gulf of Aden and Oman upwelling) and blue dots indicate pairwise comparisons among sites on opposite sides of this barrier. For panels g–i, green dots correspond to pairwise comparisons on the same side of the *b*1 barrier (oligotrophic–eutrophic barrier), red dots indicate pairwise comparisons among sites on opposite sides of *b*1, and purple dots indicate pairwise comparisons among sites on opposite sides of *b*2 (bathymetric barrier).

across *b*2 (F_{ST} same = 0.015 ± 0.0015 SE; $P < 0.001$; F_{ST} across *b*1 = 0.034 ± 0.0012 SE; $P < 0.001$; F_{ST} across *b*2 = 0.059 ± 0.0089 SE; $P < 0.001$). (ii) Pairwise F_{ST} was positively correlated with *geo* both within and among barriers with no differences among slopes among factors (barriers) (0.0092 ± 0.002 SE, $P < 0.001$) (Fig. 4g). (iii) F_{ST} increased with *env* but the rate of change (slope) of F_{ST} with respect *to env* was similar within same side of barriers and across barrier *b*2 (slope: 0.004 ± 0.001 SE, $P < 0.001$), but was higher across *b*1 (slope different side of *b*1: 0.032, $P < 0.001$) (Fig. 3h). Overall, with both *geo* and *env* combined, the rate of increase in F_{ST} was 2.9 times higher for pairwise comparisons across the *b*1 barrier than for comparisons in the same side of barriers (combined slope same side = 0.014 ± 0.0015 SE; combined slope different side of *b*1: 0.041 ± 0.006 SE). The combined slope for comparisons across the *b*2 barrier was similar to that of comparisons within same side of barriers (combined slope different side of *b*2: 0.007 ± 0.012 SE) (Fig. 4i). There was no significant correlation between *env* and *geo* ($R^2 = 0.01$, $F = 0.34$, $P = 0.556$) (Fig. 4j).

Association between loci and the environmental gradient

Using 2 latent factors, and after controlling the FDR at $q = 10\%$, we obtained a list of 98 candidate SNPs, representing 2.1% of the total number of loci. That is, 98 loci presented significant correlations with the environmental gradient after controlling for population structure in the data set, adjusting the *P*-values of these correlations and correcting for false discoveries. We note that the use of the genomic inflation factor to adjust *P*-values is a conservative procedure (Frichot *et al.* 2013). Results when using three latent factors, and after controlling the FDR at $q = 10\%$, were quite similar. The number of candidate loci with significant correlations with the environmental gradient was 103. Of these, 95 were the same as when two latent factors were used. Putative loci under selection had associated *P*-values that ranged from 0.0021 to $9.17e^{-12}$. Overall, after controlling for population structure there was evidence of putative adaptive loci for *A. bicinctus* along the Red Sea.

Population trees

The most likely population tree in TREEMIX explained 98.04% of the variance in relatedness between populations (Fig. 5). The tree was rooted using Oman and placed Socotra as the closest site to Oman, but also the most divergent one (longest branch of the tree). Following Socotra are Farasan and Djibouti, with the latter having diverged for twice as long as Farasan. Of the remaining sites in the Red Sea, Sudan had the longest branch. We examined the residuals of the model's fit to identify aspects of ancestry not captured by the tree (Fig. 5). Overall only five sites showed evidence of admixture that was not captured by the original tree (Oman, Socotra, Djibouti, Farasan and Burcan). We then sequentially added migration events to the tree. By adding five migration events ($P < 1 \times 10^{-3}$ in all cases), the amount of the variance in relatedness between sites that could be explained increased to 99.6% (Fig. 5). Interestingly, these events included one admixture event from Oman to Farasan ($w = 0.12$), one event from Socotra to the northern Red Sea ($w = 0.049$) and three events from southern Red Sea (Djibouti or Farasan) to the northern Red Sea (Abu Terr, Aqaba, Canyon Reef and Sumayr) ($w = 0.09$–0.48). We tested the significance of these admixture events with three population tests (Reich *et al.* 2009), which provided support for admixture of sites in the northern Red Sea (Abu Terr, Canyon Reef and Sumayr) with sites from the southern Red Sea (Farasan and Djibouti) (significance was set at $Z < -3.3$, $P < 0.0005$). The admixture event between Socotra and

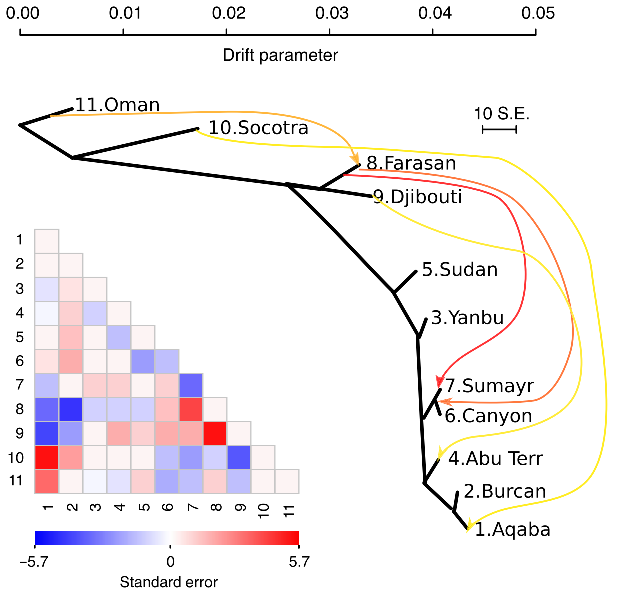

Fig. 5 Maximum-likelihood tree showing the relationship among the 11 sampled sites inferred by TREEMIX. The scale bar shows ten times the average standard error of the covariance matrix. The pairwise matrix below the tree shows the residual fit from the maximum-likelihood tree before migration events were added. Residuals above zero (in red) represent sites that are more closely related to each other than in the best-fit tree and thus are candidates for admixture events. Coloured arrows in the tree indicate the five migration events that were added to the tree (coloured according to their weight: yellow < orange < red). Admixture weight values for these migration events are as follows: Oma→Far = 0.12; Soc→Aqa = 0.05, Far→Sum = 0.48; Far→Sum: Can = 0.21; Dji→At = 0.09.

the northern Red Sea sites was not significant for this test (all tests with $Z > -0.40$, $P > 0.34$). The admixture event between Oman and the southern Red Sea (Farasan) was not significant for this test either (all tests with $Z > 0$).

Discussion

Understanding the processes that shape genetic structure and lead to population differentiation across landscapes is one of the primary objectives of population genetics and speciation studies. Here, we shed light about the main processes that shaped the genetic structure of a coral reef fish, *A. bicinctus*, in the Red Sea and its genetic relationship to one of its closest relatives *A. omanensis* in the Arabian Sea. Our analysis of genomic data from ddRAD sequencing identified one prominent genetic discontinuity in the southern Red Sea. This pattern of genetic structure in *A. bicinctus* coincides with a putative oligotrophic–eutrophic environmental transition that divides the basin into its central-northern and southern bioregions. While our results support the idea that isolation by distance (IBD) and isolation by environment (IBE) may be important mechanisms that define the observed genetic structure, the presence of a physical (circulation) barrier cannot be ruled out. Here, we discuss how these new results add to our understanding of seascape genetics, identify caveats to our genomic approach and provide guidelines for future research in the discipline.

Bayesian clustering analyses supported the presence of three genetic clusters. The limits of these clusters coincide with the position of two of the three putative barriers evaluated (*b*1 and *b*3). One clearly delimited the distribution of *A. bicinctus* outside the Red Sea (*b*3). The second (*b*1) gave rise to a genetic break within *A. bicinctus* somewhere between Canyon and Farasan (sites 5 and 8). The analysis, however, did not support the presence of a barrier at the Strait of Bab al Mandab (*b*2). Principal components analysis was mostly congruent with Bayesian clustering but provided evidence of further substructure. Perhaps the only remarkable difference was that the Bayesian clustering analysis placed sites 10 and 11 together, while they were clearly separated by one of the principal components (PC3) in the PCA. Overall, our data support the presence of the *b*1 barrier, consistent with previous reports from microsatellite studies in *A. bicinctus* (Nanninga *et al.* 2014) and the reef sponge *Stylissa carteri* (Giles *et al.* 2015). Interestingly, these studies clearly differentiate the Farasan Islands from the rest of the Red Sea and suggest that this southern Red Sea genetic cluster displays an introgression that diminishes gradually towards the north. Our results are in accordance with these previous reports. In addition, Bayesian clustering and TREEMIX analyses indicated unidirectional gene flow from the southern Red Sea to the northern Red Sea. While we cannot confirm the directionality of introgression using this data set, estimated gene flow does coincide with the dominant direction of water flow through the Strait of Bab el Mandab (intrusion of Gulf of Aden Intermediate Water; Sofianos 2003; Sofianos & Johns 2007). However, high seasonal variability in both the direction and magnitude of surface flows are common in the southern Red Sea (P. Zhan, personal communication). A coupled biophysical modelling approach (e.g. Paris *et al.* 2013) of ocean circulation would be necessary to determine how oceanographic patterns affect the dispersal potential of *A. bicinctus* larvae across this boundary.

Compared to the previous study by Nanninga *et al.* (2014) using microsatellite markers, the results of this study provide further support regarding the complex nature of the oligotrophic–eutrophic barrier located somewhere between the Farasan and Sumayr sites. Our results also extend the geographical range of the previous study by including samples from Djibouti and providing support for the straight of Bab Al Mandab as a putative barrier. Our present approach suggested that models that explicitly included both of these barriers were better than a simple IBD or IBE model, or even a model combining both IBD and IBE. In particular, we found significant signals of both IBD and IBE for sites on the same side of *b*1 as well as when comparing sites across the *b*1 and *b*2 barriers. Interestingly, even if the average pairwise F_{ST} changed considerably among comparisons within and between the barriers, the IBD slope of this correlation was similar for all comparisons regardless of their position with respect to the *b*1 and *b*2 barriers. The IBE slope on the other hand, was significantly steeper for comparisons across the *b*1 barrier than for comparisons within the same side of the barrier (sites 1–7) or for comparisons across barrier *b*2. Assuming that *b*1 was a nearly impermeable physical barrier, we could expect to find significant IBD among sites on the same sides of *b*1 and relatively high pairwise F_{ST} values among sites across *b*1 but without significant IBD (or at least weaker correlations). This is because in the presence of a strong physical barrier, allele frequencies would be allowed to drift independently on either side of it, erasing the relationship between genetic and geographical distances among sites on different sides of the barrier (Hutchison & Templeton 1999). As the rate of change (slope) in pairwise F_{ST} is maintained despite these barriers, our results essentially rule out the possibility of both *b*1 and *b*2 as being impermeable barriers. In contrast, we did not detect signals of IBD or IBE when comparing sites across the

Gulf of Aden and Oman upwelling barrier (*b*3) (the slopes for comparisons across the barrier were not different from zero). We note, however, that the Gulf of Aden and Oman upwelling barrier (*b*3), which serves as our reference for what should be expected under an impermeable physical barrier scenario, is an old barrier that delimits the *A. bicinctus* distribution. As such, it is possible that the lack of IBD and IBE among sites on different sides of the barrier is a consequence of the age of the barrier rather than its nature (strong physical barrier). Whether or not a more recent physical barrier or an old semi-permeable physical barrier (DiBattista *et al.* 2013) would produce the same pattern remains to be tested.

With this caveat in mind, these results support the idea of a permeable barrier at *b*1 and *b*2. In addition, the results of the correlations between pairwise F_{ST} and environmental distances also advocate for the presence of a significant ecological component in *b*1 but not *b*2. Evidence for a strong ecological component was further supported by the results of the latent factor mixed-effect models. This analysis revealed the presence of almost 100 loci that showed significant association of their allele frequencies with the environmental gradient after accounting for population structure. Assuming that the nature of this barrier is indeed ecological, our results suggest that multigenic and/or weak selection against immigrants that cross this barrier might explain to some extent the observed patterns of genomic divergence observed here (De Villemereuil *et al.* 2014). We note, however, that we cannot rule out the possibility that barrier *b*1 has instead a strong oceanographic physical component (Sofianos 2003; Sofianos & Johns 2007) that simply coincides with the environmental gradient. Such a scenario could produce similar signals of IBD and IBE without necessarily involving divergent selection in different environments. Unfortunately, testing between these two alternative hypotheses is beyond the limits of our current approach and data set. A promising avenue of research to more thoroughly address these questions includes the development of high resolution biophysical modelling tools for the southern Red Sea as well as fine scale habitat maps and fine spatial scale genomic surveys of this region. Further experiments, involving reciprocal transplantations, coupled with high density genomic surveys and comparisons of genetic differentiation of quantitative traits (Orsini *et al.* 2013) may also shed further light onto the nature of this barrier and the potential for selection to act against dispersers moving across it..

Finally, based on SNPs, our results suggest that divergence between *A. bicinctus* and *A. omanensis* is strong, despite the lack of differences in mtDNA previously reported (Litsios & Salamin 2014; DiBattista *et al.* 2015c). These results add to the growing literature that highlights the use of RAD sequencing in delimiting closely related species and identifying cryptic species (Puebla *et al.* 2014; Gaither *et al.* 2015; Pante *et al.* 2015). We found different patterns of admixture depending on the percentile of loci used for clustering analyses. It seems that while loci in the 90th percentile showed introgression from *A. omanensis* to one of the sites in the south of the Red Sea (Farasan), lower percentiles (60th–80th) indicated some level of introgression from *A. bicinctus* to samples of the putative hybrids at site 10 (Socotra). A finding of introgression at RADSeq loci confirms previous work by DiBattista *et al.* (2015c) that identified seven anemonefish individuals collected from Socotra as putative hybrids based on intermediate colouration between the parental species and differentiation at a single nuclear gene (RAG2). Even though anemonefishes are very rare at Socotra, both *A. bicinctus* and *A. omanensis* have been recorded in previous surveys (Kemp 1998; Zajonz *et al.* 2000), setting the scene for hybridization between the two species at the edge of their distributional range. Bayesian clustering results were partially supported by the hybrid simulation analysis. However, the inclusion of simulated hybrids in PCA revealed that despite the fact that Socotra seems similar to some simulated hybrids when plotting the first two PC axes, the third axis still clearly separates Socotra from all samples, suggesting that if hybridization is indeed occurring here, it might be more complex than the simplistic simulations we have evaluated. Admixture of RADSeq loci between the species further highlights the importance of the Socotra Archipelago as a hybrid hotspot in the tropical Indo-Pacific. Still, further genomic surveys of clown fishes in this region seem necessary to fully resolve the complex history of anemonefishes in this suture zone. Overall, our results reinforce the notion that coastal seascapes are extremely complex systems, highlighting the value of powerful new genomic tools, such as RAD seq, to resolve subtle patterns of gene flow and connectivity in marine organisms.

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 logistic support elsewhere, we thank Eric Mason at Dream Divers, Nicolas Prévot at Dolphin Divers and the crew of the M/V Deli in Djibouti, Equipe Cousteau support for the Sudan Shark and Ray Conservation and Management Program including N. Hussey, S. Kessel, C. Scarpellini and M. Younis, the KAUST Coastal and

Marine Resources Core Lab and Amr Gusti, Oman logistics and the Wildlife Conservation General Administration, Red Sea State Government and Red Sea University of Sudan. For specimen collections, we thank B. Bowen, J.P. Hobbs, M. Priest, L. Rocha, T. Sinclair-Taylor, Steve Simpson and members of the Coral Reef Ecology lab at KAUST. For assistance with bench work at KAUST, we thank C. Michell. We also acknowledge important contributions from David Catania and Luiz Rocha for assistance with specimen archiving, the KAUST Bioscience Core Laboratory with Sivakumar Neelamegam and Hicham Mansour for their assistance with Illumina sequencing and the UC Davis Genome Center – DNA Technologies and Expression Analysis Core with Lutz Froenicke. We thank Eric Frichot for his assistance with analyses with the LEA package. Finally, we thank Eric Crandall and two anonymous reviewers for constructive comments on an earlier version of the manuscript. This research was carried out under the general auspices of KAUST's arrangements for marine research with the Saudi Arabian Coast Guard and the Saudi Arabian Presidency of Meteorology and Environment, which are the relevant Saudi Arabian authorities governing all sea-going and marine research activities. Fish from Oman (Simpson *et al.* 2014) were collected with specific approval from both the University of Edinburg Ethical Review Committee, the University of Edinburgh Expedition Committee, and with permission from the Ministry of Regional Municipalities, Environment and Water Resources, Sultanate of Oman and Department of Nature Conservation and Wildlife, Governate of Dhofar. Samples from Oman were collected with permission of the Ministry of Water and Environment, Environment Protection Authority of the Republic of Yemen.

References

Anderson D (2008) *Model Based Inference in the Life Sciences: A Primer on Evidence*. Springer Science + Business Media, New York.

Avise JC (2000) *Phylogeography: The History and Formation of Species*. Harvard University Press, Cambridge, Massachusetts.

Balkenhol N, Gugerli F, Cushman SA *et al.* (2009) Identifying future research needs in landscape genetics: where to from here? *Landscape Ecology*, **24**, 455–463.

Berumen ML, Hoey AS, Bass WH *et al.* (2013) The status of coral reef ecology research in the Red Sea. *Coral Reefs*, **32**, 737–748.

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

Correa C, Hendry AP (2012) Invasive salmonids and lake order interact in the decline of puye grande *Galaxias platei* in western Patagonia lakes. *Ecological Applications*, **22**, 828–842.

De Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

DiBattista JD, Berumen ML, Gaither MR *et al.* (2013) After continents divide: comparative phylogeography of reef fishes from the Red Sea and Indian Ocean. *Journal of Biogeography*, **40**, 1170–1181.

DiBattista JD, Choat JH, Gaither MR *et al.* (2015a) On the origin of endemic species in the Red Sea. *Journal of Biogeography*, doi: 10.1111/jbi.12631, in press.

DiBattista JD, Roberts M, Bouwmeester J *et al.* (2015b) A review of contemporary patterns of endemism for shallow water reef fauna in the Red Sea. *Journal of Biogeography*, doi: 10.1111/jbi.12649, in press.

DiBattista JD, Rocha LA, Hobbs J-PA *et al.* (2015c) 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.

Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies (ed O'Meara B). *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Froukh T, Kochzius M (2007) Genetic population structure of the endemic fourline wrasse (*Larabicus quadrilineatus*) suggests limited larval dispersal distances in the Red Sea. *Molecular Ecology*, **16**, 1359–1367.

Gaither MR, Bernal MA, Coleman RR *et al.* (2015) Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Molecular Ecology*, **24**, 1543–1557.

Giles EC, Saenz-Agudelo P, Hussey NE, Ravasi T, Berumen ML (2015) Exploring seascape genetics and kinship in the reef sponge *Stylissa carteri* in the Red Sea. *Ecology and Evolution*, **5**, 2487–2502.

Hutchison DW, Templeton AR (1999) Correlation of pairwise genetic and geographic distance measures: inferring the relative influences of geneflow and drift on the distribution of genetic variability. *Evolution*, **53**, 1898–1914.

Johnson JB, Omland KS (2004) Model selection in ecology and evolution. *Trends in Ecology & Evolution*, **19**, 101–108.

Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, **11**, 94.

Kemp J (1998) Zoogeography of the coral reef fishes of the Socotra Archipelago. *Journal of Biogeography*, **25**, 919–933.

Khalaf MA, Kochzius M (2002) Community structure and biogeography of shore fishes in the Gulf of Aqaba, Red Sea. *Helgoland Marine Research*, **55**, 252–284.

Klausewitz W (1989) Evolutionary history and zoography of the Red Sea ichthyofauna. *Fauna of Saudi Arabia*, **10**, 310–337.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–299.

Litsios G, Salamin N (2014) Hybridisation and diversification in the adaptive radiation of clownfishes. *BMC Evolutionary Biology*, **14**, 245.

Litsios G, Pearman PB, Lanterbecq D, Tolou N, Salamin N (2014) The radiation of the clownfishes has two geographical replicates. *Journal of Biogeography*, **41**, 2140–2149.

Mastretta-Yanes A, Arrigo N, Alvarez N *et al.* (2014) Restriction site-associated DNA sequencing, genotyping error estimation and de novo assembly optimization for population genetic inference. *Molecular Ecology Resources*, **15**, 28–41.

Meirmans PG, Van Tienderen PH (2004) GENOTYPE and GENODIVE: two programs for the analysis of genetic diversity of asexual organisms. *Molecular Ecology Notes*, **4**, 792–794.

Nanninga GB, Saenz-Agudelo P, Manica A, Berumen ML (2014) Environmental gradients predict the genetic population structure of a coral reef fish in the Red Sea. *Molecular Ecology*, **23**, 591–602.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Orsini L, Vanoverbeke J, Swillen I, Mergeay J, De Meester L (2013) Drivers of population genetic differentiation in the wild: isolation by dispersal limitation, isolation by adaptation and isolation by colonization. *Molecular Ecology*, **22**, 5983–5999.

Pante E, Abdelkrim J, Viricel A *et al.* (2015) Use of RAD sequencing for delimiting species. *Heredity*, **114**, 450–459.

Paris CB, Helgers J, van Sebille E, Srinivasan A (2013) Connectivity modeling system: a probabilistic modeling tool for the multi-scale tracking of biotic and abiotic variability in the ocean. *Environmental Modelling & Software*, **42**, 47–54.

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.

Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics*, **8**, e1002967.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

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*, **8**, e64909.

Reich D, Thangaraj K, Patterson N, Price AL, Singh L (2009) Reconstructing Indian population history. *Nature*, **461**, 489–494.

Riginos C, Liggins L (2013) Seascape genetics: populations, individuals, and genes marooned and adrift. *Geography Compass*, **7**, 197–216.

Roberts CM, Shepherd ARD, Ormond RFG (1992) Large-scale variation in assemblage structure of Red Sea butterflyfishes and angelfishes. *Journal of Biogeography*, **19**, 239.

Robitzch V, Banguera-Hinestroza E, Sawall Y, Al-Sofyani A, Voolstra CR (2015) Absence of genetic differentiation in the coral *Pocillopora verrucosa* along environmental gradients of the Saudi Arabian Red Sea. *Frontiers in Marine Science*, **2**, doi:10.3389/fmars.2015.00005.

Rohling EJ, Fenton M, Jorissen FJ *et al.* (1998) Magnitudes of sea-level lowstands of the past 500,000 years. *Nature*, **394**, 162–165.

Santini S, Polacco G (2006) Finding Nemo: molecular phylogeny and evolution of the unusual life style of anemonefish. *Gene*, **385**, 19–27.

Sexton JP, Hangartner SB, Hoffmann AA (2014) Genetic isolation by environment or distance: which pattern of gene flow is most common? *Evolution*, **68**, 1–15.

Shefer S, Abelson A, Mokady O, Geffen E (2004) Red to Mediterranean Sea bioinvasion: natural drift through the Suez Canal, or anthropogenic transport? *Molecular Ecology*, **13**, 2333–2343.

Sheppard CR, Sheppard AL (1991) Corals and coral communities of Arabia. *Fauna of Saudi Arabia*, **12**, 3–170.

Simpson SD, Harrison HB, Claereboudt MR, Planes S (2014) Long-distance dispersal via ocean currents connects Omani clownfish populations throughout entire species range. *PLoS ONE*, **9**, e107610.

Slatkin M (1987) Gene flow and the geographic structure of natural populations. *Science*, **236**, 787–792.

Sofianos SS (2003) An Oceanic General Circulation Model (OGCM) investigation of the Red Sea circulation: 2. Three-dimensional circulation in the Red Sea. *Journal of Geophysical Research*, **108**, 3066.

Sofianos SS, Johns WE (2007) Observations of the summer Red Sea circulation. *Journal of Geophysical Research*, **112**, C06025.

Via S (2002) The ecological genetics of speciation. *The American Naturalist*, **159**, S1–S7.

Wang I, Bradburd G (2014) Isolation by environment. *Molecular Ecology*, **23**, 5649–5662.

Wang IJ, Glor RE, Losos JB (2013) Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology Letters*, **16**, 175–182.

Wright S (1943) Isolation by distance. *Genetics*, **28**, 114–138.

Zajonz U, Khalaf M, Krupp F (2000) *Coastal Fish Assemblages of the Socotra Archipelago*. Conservation and Sustainable Use of Biodiversity of the Socotra Archipelago. Marine Habitat, Biodiversity and Fisheries Surveys and Management. Progress Report of Phase III – Senckenberg Research Institute, Frankfurt aM, pp. 127–170.

P.S.A., J.D. and M.B. designed the study with input from G.N., M.P. and H.H. The majority of the sampling was conducted by G.N., H.H. and J.D. Laboratory analyses were conducted by M.G., P.S.A. and J.D. M.P. and P.S.A. conducted the majority of data and bioinformatics analyses. All authors contributed to the writing of the manuscript.

Data accessibility

Illumina RAD-tag sequences are accessible at NCBI SRA accession no. SRP064543 (Biosample accession nos: SAMN04151387–SAMN04151499). The following data are accessible at Dryad (doi:10.5061/dryad.n1432): STACKS *de novo* assembly files, Individual SNP genotypes for 4559 loci and 113, environmental and geographical distance matrix used in the information-theoretic approach and R code used in this analysis. ML tree output files from TREEMIX and three population test result files.

Supporting information

Additional supporting information may be found in the online version of this article.

Appendix S1 This file contains Tables S1-S3.

Table S1 The following table describes the different parameter combinations that were evaluated in Stacks.

Table S2 Values of environmental variables for each site that were used to estimate environmental distances (*env*) among pairs of sites.

Table S3 Summary statistics and ranking of three linear models that differed in the relative position of the *b*1 barrier.

Appendix S2 This file contains Figures S1-S4.

Figure S1 Values of the cross-entropy criterion as a function of the number of factors in snmf runs including sites 1–9.

Figure S2 Histogram of adjusted *P*-values obtained with lfmm using $K = 2$ and $K = 3$ latent factors (5 runs each).

Figure S3 Principal components analysis of genotype variation for all samples in this study (filled circles) as well as simulated hybrid individuals (crosses).

Figure S4 Results of Bayesian clustering for $K = 3$ for (a) all SNPs ($n = 4559$) and for the subsets of SNPs (b) above the 90th percentile, (c) between the 90th and 80th percentile, (d) between the 70th and 80th percentile, (e) between the 60th and 80th percentile, and (f) the distribution of global F_{ST} estimates for all 4559 SNPs.

Appendix S3 Excel file including all STRUCTURE output summary statistics for all runs.

Appendix S4 Linear model ranking including all sites.

Appendix S5 Linear model ranking including sites 1 to 9.