9-30-2021

Predictors of marine genetic structure in the Indo-Australian Archipelago

Udhi E. Hernawan
Edith Cowan University

Paul S. Lavery
Edith Cowan University

Gary A. Kendrick

Kor-jent van Dijk

Yaya I. Ulumuddin

See next page for additional authors

10.1016/j.rsma.2021.101919
This is an author's accepted manuscript of: Hernawan, U. E., Lavery, P. S., Kendrick, G. A., van Dijk, K. J., Ulumuddin, Y. I., Triandiza, T., & McMahon, K. M. (2021). Predictors of marine genetic structure in the Indo-Australian Archipelago. *Regional Studies in Marine Science,* 47, Article 101919.
https://doi.org/10.1016/j.rsma.2021.101919

Authors

Udhi E. Hernawan, Paul S. Lavery, Gary A. Kendrick, Kor-jent van Dijk, Yaya I. Ulumuddin, Teddy Triandiza, and Kathryn M. McMahon

Predictors of marine genetic structure in the Indo-Australian Archipelago

Udhi E. Hernawan[a, d], Paul S. Lavery[a], Gary A. Kendrick[b], Kor-jent van Dijk[c], Yaya I. Ulumuddin[d], Teddy Triandiza[d], Kathryn M. McMahon[a]

Affiliation

a. School of Sciences and Centre for Marine Ecosystems Research, Edith Cowan University, 270 Joondalup Dr, Joondalup WA 6027, Australia (udhiehernawan@gmail.com; k.mcmahon@ecu.edu.au; p.lavery@ecu.edu.au).

b. School of Biological Sciences and The Ocean Institute, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia (gary.kendrick@uwa.edu.au).

c. School of Biological Sciences, The University of Adelaide, Adelaide, SA 5005, Australia (korjent.vandijk@adelaide.edu.au).

d. Research Centre for Oceanography (P2O), Indonesian Institute of Sciences (LIPI), Jl. Pasir Putih 1, Ancol Timur, Jakarta 14430, Indonesia (yaya.ulumuddin@gmail.com; teddy.triandiza27@gmail.com).

*Corresponding author: Kathryn M. McMahon

Email: k.mcmahon@ecu.edu.au

HIGHLIGHTS

1. Genetic structure best predicted by pelagic dispersal time and adult mobility
2. Longer pelagic dispersal time promotes higher connectivity among populations
3. Migratory species had less genetic structure compared to sessile species
4. Different genetic markers showed no significant effect on genetic structure
5. Genetic studies should sample sites representatively nested in each ecoregion

ABSTRACT

The spatial genetic structure of marine organisms is related to dispersal and life-history traits, historical processes, current oceanographic connectivity and habitat features. Here, we assessed the relative importance of these factors for the genetic structure of a broad range of marine species in the Indo Australian Archipelago (IAA). We collated published data on 99 marine species from eight taxonomic groups (ascidians, fishes, molluscs, crustaceans, echinoderms, corals, reptiles, and marine plants) and used generalized linear models (GLMs) to estimate the best predictors of genetic structure. Genetic structure was characterized by F_{ST} and the number of genetic clusters over the study area. Predictors tested were: the type of genetic markers; the number of marine ecoregions which are a proxy for habitat variation, historical processes and oceanographic features; species dispersal-related traits (i.e., pelagic larval duration-PLD, adult life habit, reproductive strategy, and egg type); and geographic distance separating populations. The genetic structure of marine species across the IAA was best predicted by traits related to dispersal of larvae or propagules and the mobility of adults; and the number of marine ecoregions sampled not distance was also an important predictor, especially in sedentary and free-swimming species. Our findings highlighted the importance of these key traits to help guide decision-making in spatial management and conservation. There were still many gaps in our understanding of genetic structure, both spatially and within certain taxa, and we recommended future genetic studies focus on habitat-forming taxa and sample sites that are representatively nested in each ecoregion within a marine province or a marine realm, over the spatial extent of the IAA.

KEYWORDS: dispersal, F_{ST}, genetic clusters, conservation, life-history

54 **ABBREVIATIONS**:

55 $F_{ST:}$ Fixation index which is a measure of genetic differentiation between populations.

56 IAA: Indo Australian Archipelago

57 PLD: Pelagic larval duration

58 GLM: Generalized linear models

59 AIC : Akaike's Information Criterion

60 MEOW: Marine Ecoregion of the World

61 IUCN: International Union for Conservation of Nature

62 SNP: single nucleotide polymorphisms

63 ISSR: Inter Simple Sequence Repeat

64 SSR: Simple Sequence Repeat

65 EPIC: Exon-primed intron-crossing.

66

1. INTRODUCTION

Most species are composed of spatially separated populations that are connected by dispersal. Successful dispersal, that is when migrants settle and interbreed with members of a recipient population, results in exchange of genetic material or gene flow. The level of gene flow, together with mutation, selection and genetic drift, can influence the spatial distribution of genetic variation within and among populations known as genetic structure. High gene flow homogenizes genetic variation by counteracting the effect of mutation, selection and genetic drift, while low gene flow can lead to isolation and increased genetic differentiation. Barriers to gene flow among populations will result in populations drifting apart and become more distinct within a species distribution, such that they no longer behave as a single, randomly mating (panmictic) population (Slatkin, 1987; Charlesworth et al., 2003).

Spatial genetic structuring is a consequence of the interaction between intrinsic (e.g., life-history traits) and extrinsic factors (e.g., habitat heterogeneity and dispersal barriers) over time (Lowe et al., 2004; Cowen and Spongaule, 2009). Among the intrinsic factors, the duration of early life stages (pelagic larval duration–PLD) has been highlighted as a key factor in determining genetic structure. A longer PLD increases the species' dispersal potential as larvae or propagules are transported by currents for a greater period of time (Shanks et al., 2003; Shanks, 2009; Treml et al., 2015). As dispersal facilitates gene flow (Wright, 1931; Slatkin, 1987), PLD should be inversely correlated with genetic structure (Palumbi, 1992; Doherty et al., 1995; Siegel et al., 2003). However, some analyses have found weak or no correlation between PLD and genetic structure (Weersing and Toonen, 2009; Liggins et al., 2016; Costantini et al., 2018), while other dispersal-related traits (e.g. reproductive strategy, phenology and adult mobility) have been identified as important in influencing spatial genetic structure (Bradbury et al., 2008; Galarza et al., 2009; Riginos et al., 2011; Selkoe et al., 2014; Treml et al., 2015).

Extrinsic factors influencing genetic structure include geological history, past and/or contemporary oceanography, and habitat heterogeneity. For example, historical geological processes generate biogeographic barriers that restrict gene flow in many marine species (Avise, 1992; Jacobs et al., 2004; Ayre et al., 2009; Pelc et al., 2009; Evans et al., 2016; Crandall et al., 2019). Barriers to gene flow also could emerge from contemporary geological features (e.g. islands, deep-sea trenches and continental shelves) (Palumbi, 1994; Galarza et al., 2009; Riginos and Liggins, 2013) and oceanographic processes (e.g. ocean currents and upwelling) but oceanographic currents may also act as the dispersal vector, facilitating gene flow among populations (Hu et al., 2013; Simpson et al., 2014; Treml et al., 2015). Habitat heterogeneity acts as a driver of local selection and adaptation, thus contributing to patterns of genetic structure (Riginos and Liggins, 2013; Wang and Bradburd, 2014; Donati et al., 2019).

The Indo-Australian Archipelago–IAA (Fig. 1) comprises more than 20,000 islands situated in the Central Indo-Pacific and is one of the most geologically dynamic and complex regions on Earth (Lohman et al., 2011). Although it occupies only about 4% of the planet's land surface (Lohman et al., 2011) the IAA is the epicentre of biodiversity; not only of corals, but also fishes, echinoderms, molluscs, crustaceans and seagrasses (Hoeksema, 2007; Short et al., 2007; Evans et al., 2016). Despite its importance, many habitats and species in this region are threatened with extinction under current and predicted future anthropogenic pressures (Hoegh-Guldberg, 2010; McLeod et al., 2010). A meta-analysis by Selig et al. (2014) highlighted this region as one of the global priorities for marine biodiversity conservation (Fisher et al., 2011).

What factors affect spatial distribution of genetic variation is one of the primary questions related to marine conservation in the IAA (Palumbi, 2004; Barber, 2009; Barber et al., 2011; Carpenter et al., 2011). Many single-species phylogeographic studies have addressed this question but the conclusions vary depending on the focal taxon and methodology. Carpenter et al. (2011) attempted to reveal commonalities in the patterns of genetic structure in the IAA

across a broad range of marine taxa using a qualitative approach from the published phylogeographic genetic data from invertebrate and fish species. This was based on genetic data and did not consider dispersal-related traits. Crandall et al. (2019) identified a single consistent barrier west of the Sunda Shelf for many species in their multi-species (56 species) analysis based on mitochondrial sequence data, but further barriers within the IAA were not identified. However, Treml et al. (2015) did identify common barriers to dispersal within the IAA using biophysical larval dispersal models under a range of scenarios, representing species with a range of reproductive strategies. In this case genetic data were not used to validate their findings, nor were variations in habitat that could limit recruitment or geological history considered. These studies have provided great leaps forward in the understanding of the patterns and processes influencing these patterns in the IAA.

Here, we used a robust but simplified approach (i.e. generalized liner modelling) to interrogate a range of potential drivers of genetic structure in the IAA. We tested the hypothesis that habitat heterogeneity, oceanographic-geologic features and dispersal-related traits would best predict genetic structure in a range of marine species. Understanding of the patterns and drivers of genetic structure can help guide decision-making in spatial management and conservation in the IAA.

2. MATERIALS AND METHODS

2.1. Literature survey

Peer-reviewed publications reporting population genetic structure of marine species in the IAA were searched using the Web of Science and Google Scholar databases (August 2020). The search terms included: “gene flow”; “genetic structure”; “phylogeography”; and “population

genetics". As this study was spatially limited to the IAA, we refined the search results using these following terms: "Indo-Australian Archipelago"; "East Indies"; "Coral Triangle"; "Indo-Malay"; "Indonesia"; "Malaysia"; "Philippines"; or "Australia". This yielded 285 publications. We verified if each publication contained all of the following: 1) marine species; 2) more than three sampling locations within the IAA; and 3) data from which spatially explicit genetic structure could be determined. This filtering resulted in 101 publications for further analysis (http://dx.doi.org/10.25958/4sj2-sw67).

2.2. Data extraction

We collected data on the variables listed in Table 1 from each publication. Based on our hypothesis, we extracted four sets of variables: measures of genetic structure, habitat heterogeneity and oceanographic-geologic features, geographic distance and species dispersal-related traits. We used two measures of genetic structure: (i) global F_{ST} and (ii) the number of genetic clusters (*cluster*) derived from a range of different approaches depending on the specific paper. F_{ST} does not provide spatial information about genetic breaks, while genetic clustering does, so these are complementary approaches. F_{ST} is a common measure of genetic structure but it can be influenced by the type of marker and the spatial extent and resolution of sampling (Meirmans and Hedrick, 2011). Global F_{ST} of all sites sampled in the study was more commonly presented than the pairwise F_{ST} between sites, therefore this format was selected. Negative values of F_{ST} were changed to "0". Using pairwise F_{ST} as a measure of genetic structure would enable a more spatially explicit interrogation of the significance of habitat heterogeneity, oceanographic-geologic features and dispersal-related traits and remove biases that could arise from genetic drift amongst populations sampled over different distances with different dispersal potential (Crandall et al., 2019). However, to maximise the data available across multiple taxa with different life-histories we selected global F_{ST}. To explore another measure of genetic structure, the number of genetic clusters or panmictic populations in the

study area was extracted. Spatial groupings of sites with similar genetic structure is very relevant in conservation management (Reiss et al., 2009; von der Heyden et al., 2014) and cannot be fully addressed using only F_{ST}. The use of genetic clusters enables an assessment without the inherent biases of F_{ST} described above. The most supported number of genetic clusters was based on either the *K*-value of STRUCTURE analyses, the number of significant clusters in a principal coordinate analysis, or the number of distinct clades in a phylogenetic tree and/or haplotype network provided in each study. When multiple clustering techniques were used for a particular species in a paper, the author's interpretation of the most strongly supported number of clusters was used. While many reviews or meta-analyses have used F_{ST} for examining the patterns of genetic structure (Kelly and Palumbi, 2010; Nanninga and Manica 2018), the use of genetic clustering could be an alternative approach and more observations were able to be extracted with this variable, 142 compared to 117 for F_{ST}.

Habitat heterogeneity and oceanographic-geological features were represented by the number of marine ecoregions covered by each study (*ecoregion*) as defined from the Marine Ecoregions of the World system (MEOW, Spalding et al., 2007) (Fig. 1). There are many classifications of ecoregions that cover the IAA (for example Crandall et al., 2019). This particular classification included a subset of ecoregions (28) from the Central Indo-Pacific Province. The hierarchy with the highest number of divisions was selected because we wanted to have the maximum resolution of habitat, oceanographic and geological features to examine drivers of genetic structure. The dataset had observations for all but 1 ecoregion with a median of 17 observations per ecoregion and a maximum of 70 (Figure 1, Table 1). The variable geographic distance (*distance*) was calculated by measuring the pairwise minimum distance by sea (without crossing any landmass) among all sampling sites in each study in Google Earth v7.1.2.2041. Then, the largest pairwise minimum geographical distance was included in the analysis as the maximum geographic distance to represent the spatial scale of the study.

190 The dispersal-related traits examined were obtained from peer-reviewed publications,
191 International Union for Conservatio of Nature (IUCN) Redlist (iucnredlist.org), FishBase
192 (fishbase.org) and LarvalBase (larvalbase.org); and included pelagic larval duration (PLD),
193 adult life habit (with respect to adult mobility), reproductive strategy (with respect to how
194 sperm and eggs are released) and egg type (related to how fertilized eggs are dispersed). Treml
195 et al. (2015) identified that reproductive output, spawning time and frequency were also
196 important predictors of connectivity based on larval dispersal modelling in this region but this
197 information was not available across the 99 species so could not be included in the analysis.
198 The variable *PLD* was defined as the maximum recorded pelagic larval duration in hours for
199 each species and for marine plants (seagrasses and mangroves) the *PLD* was determined based
200 on the maximum viability of the reproductive propagule before settlement. The maximum PLD
201 was used as Weersing and Toonen (2009) identified this as the better predictor of genetic
202 structure and this was verified by Treml et al. (2015) specifically in this region. The PLD was
203 not available for 9 species. Adult life habit (*adult*) represents the species motility in the adult
204 phase, which has the potential to influence dispersal and genetic structure. This variable was
205 classified into sessile (e.g. corals), sedentary (restricted movement, e.g. sea urchin), motile
206 (freely moving/swimming e.g. fishes) and migratory (e.g. the skipjack tuna *Katsuwonus*
207 *pelamis*) (Maguire et al., 2006; de Juan et al., 2009). The reproductive strategy (*rep. strategy*)
208 pertaining to the mode that sperm and eggs are released was classified into broadcaster and
209 brooder. Brooders potentially exhibit greater genetic structure than the broadcast-spawning
210 species due to the lack of a planktonic dispersive stage (Foggo et al., 2007; Bradbury et al.,
211 2008). The variable egg type (*egg*) related to the mode that fertilized eggs are dispersed, either
212 in the pelagic or benthic zone or as direct development (e.g. some sharks) and we predicted
213 that pelagic eggs have a greater dispersal potential (Bradbury et al., 2008; Riginos et al., 2011).
214 Species that mouth-/pouch brood (e.g. seahorses) or guards their eggs (e.g. *Amphiprion*

215 *ocellaris*) were classified as benthic eggs (Table 1). For marine plants the variable *egg* was
216 defined by the potential for dispersal of the reproductive propagule based on its buoyancy
217 (buoyant phase=pelagic or no buoyant phase=benthic). Additionally, we recorded the genetic
218 markers in the study as "Seq"-sequence, "Allo"-allozymes, "SNP"-single nucleotide
219 polymorphisms, "MSat"-microsatellite, including Inter Simple Sequence Repeat (ISSR) and
220 Simple Sequence Repeat (SSR), and "EPIC" for Exon-Primed Intron-Crossing.

221 **2.3. Statistical analysis**

222 ***2.3.1. Response and predictor variables***

223 We used generalized linear models (GLMs) to investigate which variables best predicted the
224 genetic structure with separate analyses run for the two response variables, 1) F_{ST}, and 2)
225 genetic cluster. The basic model formulation used in GLMs for each response variable included
226 the predictor variables *ecoregion*, *distance*, *PLD*, *adult*, *rep. strategy* and *egg*. The variable
227 *marker* was treated as a fixed factor in the model due to differences in attributes and sensitivity
228 of the genetic markers to detect genetic variation (Parker et al., 1998; Schlötterer, 2004).

229 ***2.3.2. Data set for GLMs***

230 Five different sets of models were run, the first on the full dataset to examine general patterns
231 across all species (1) and four sets using subsets of the data. In subset 2, all records with no
232 pelagic life-history phase (n=103 or 119) were removed to negate the potential bias from the
233 absence of a larval duration in some observations in our analysis. As GLMs results from the
234 full dataset analysis identified adult life habit as a significant predictor of genetic structure and
235 migratory species were different to all other types, the remaining subsets allowed us to examine
236 if the drivers of genetic structure were consistent across each adult life habit type. Subset 3
237 focused on free swimming species (n=35 or 50), subset 4 on sedentary species (n=35 or 40)

238 and subset 5 on sessile species (n=34 or 37). As there were only 15 records for migratory
239 species, there were not enough observations for this group to run GLMs.

240 ***2.3.3. Test for independence and multicollinearity***

241 A key assumption for GLMs is independence among continuous predictor variables (Fox and
242 Weisberg, 2011). This was tested for *ecoregion*, *distance*, and *PLD* using Hoeffding's D test in
243 function *hoeffd* of **Hmisc** 3.15-0 package (Harrell Jr, 2015), confirming low dependency for
244 PLD with ecoregion and PLD with maximum distance (max. Hoeffding's D value of pair-wise
245 comparison 0.4 and 0.6) (Appendix Table S1). As the pairwise comparison was less than 0.8,
246 where 1.0 = total dependency, we incorporated the three continuous variables into the analysis.
247 Multicollinearity was also not detected from the variable inflation factor-VIF for both F_{ST} and
248 genetic clusters (*PLD*= 1.01, 1.07; *ecoregion* = 1.84, 1.87; *distance* = 1.84, 1.79) calculated
249 using **car** 2.0-25 package (Fox et al., 2015) (Appendix Table S2).

250 ***2.3.4. Model generation and selection***

251 We used **glmulti** 1.0.7 to calculate the GLMs by generating all possible model formulas and
252 fits them with a GLM (Calcagno and Mazancourt, 2010). This approach does not require 'a
253 priori' selection of candidate models, which is needed in other packages (e.g. **MuMIn**). In the
254 case of missing data (PLD values for 9 observations), **glmulti** excluded the corresponding
255 variable from the calculation. For F_{ST}, we used $\log((F_{ST}+0.001)/(1-(F_{ST}+0.001))$ to improve
256 the approximation of linearity and the GLMs was run using the Gaussian distribution family
257 with an identity link function. For the response variable *cluster,* we ran the GLMs using the
258 Poisson distribution family with a log link function because *cluster* is count data. Model
259 selection was based on Akaike's Information Criterion (AIC). Models within the two lowest
260 AIC units are considered best at explaining the response variable (Burnham and Anderson,
261 2002). To examine the contribution of each predictor in determining genetic structure, relative

evidence weight of the predictor was calculated as the sum of the relative evidence weights of all models in which the predictor appears where a value of >0.8 indicates a significant contribution (Calcagno and Mazancourt, 2010).

2.3.5. Effect of predictor variables

The influence of the important predictors resulting from **glmulti** (if any) was examined using the best models. Multiple comparison of means in the package **multcomp** was used to test the effect of categorical predictors (Hothorn et al., 2008). All statistical analysis was done in the statistical computing environment, R version 3.2.2 (R Development Core Team, 2015) and RStudio version 0.98.1103.

3. RESULTS

3.1. Literature survey

From 101 publications, we collated data on 99 marine species from eight taxonomic groups (numbers are unique species per group; ascidians: 1; fishes: 54; molluscs: 11, crustaceans: 10, echinoderms: 4; corals: 11 marine plants: 6 and reptiles: 2) (doi: currently being generated, can supply as supplementary file for review process). For most there was one record per species (76%), but in some cases there were two records (15%) with the remainder (8%) having more than two records per species. The species that had multiple records were in different locations so were considered as independent observations. The maximum number of records was 8, for the clam *Tridacna crocea*. The full dataset comprised 150 records, with fishes contributing to 45% of the records followed by molluscs at 17%. The remainder of the groups accounted for 10% of the records or less. After filtering the full dataset, a subset of 116 records of species

284 with pelagic larval state (PLD>0) was generated, with 42 records of sessile, 43 of sedentary
285 species, 50 free-swimming and 15 migratory species (Table 1).

286 For the dependent variables there was a large range in the global F_{ST} (0 to 0.905) and the number
287 of genetic clusters identified (1-13) across all observations (Table 1). It was a similar case for
288 the predictor variables e.g. the maximum overwater distance ranged from 23 to 9728 km and
289 PLD from 0 to 5640 hours. For the dispersal related traits, pelagic egg types were best
290 represented (102) compared to benthic egg developers (32) and direct developers (16). There
291 were more records for broadcast (91) versus brooding spawners (51) (Table 1). The majority
292 of records were based on genetic markers from sequence data (82) followed by microsatellite
293 data (46) with 10 or less records for the other types (Table 1). Observations were recorded in
294 all but one ecoregion (ecoregion 21) reaching a maximum of 70 observations in ecoregion 5
295 (Figure 1) with a median of 17 observations per ecoregion (Table 1). The observations were
296 not distributed evenly among ecoregions with the highest observations in the central IAA in
297 ecoregion 3, 5, 10, 11, 14 and 15 (Figure 1). Generally, a lower number of genetic clusters were
298 identified relative to the number of ecoregions sampled (68% of observations) but in 17% of
299 the observations the number of genetic clusters was the same as the number of ecoregions
300 sampled and in 14% there were more ecoregions sampled than genetic clusters
301 (http://dx.doi.org/10.25958/4sj2-sw67).

302 **3.2. Predictors of genetic structure: F_{ST}**

303 Using F_{ST} as the response variable for the full dataset, 5-8 different models were supported by
304 the GLMs. This was also the case for the subset, where records with PLD of 0 were removed.
305 Each model had a slightly different set of predictor variables although *PLD* and *adult* were
306 always present (Table 2). These two predictors were consistently > 0.8 based on the model-

307 averaged importance of terms (Figure 2) indicating high significance. F_{ST} declined with
308 increasing PLD (p-value= 0.04; Appendix Figure S1).

309 Pairwise analysis on the effect of categorical variables found no evidence of any significant
310 differences due to marker type (Appendix Table S3), but in the *adult* categories, migratory
311 species were significantly different to the other adult life habit categories (Appendix Table S4).
312 The average F_{ST} for migratory species was 0.05, compared to 0.18 for motile species, 0.26 for
313 sedentary and 0.22 for sessile species. For Subset 2 (PLD>0) the predictor *distance* was also
314 identified as an important predictor passing the 0.8 threshold and present in 7 of the 8 supported
315 models. In this case, a greater over-water distance resulted in a higher F_{ST} (p-value=0.015;
316 Appendix Figure S2). In the full dataset, *ecoregion* was the next most supported variable, in 4
317 out of the 5 models and approaching the 0.8 threshold of relative importance (Figure 2); if more
318 ecoregions were sampled, the F_{ST} was higher.

319 When the drivers of F_{ST} were assessed on subsets based on the adult life habit, five models
320 were supported for motile species with PLD present in all models and identified as the most
321 important predictor of genetic structure (Table 2, Figure 2). For sedentary species, three models
322 were supported, all containing PLD and ecoregion, and both these predictors passed the 0.8
323 threshold for relative importance (Table 2, Figure 2). The relationship of PLD and ecoregion
324 to F_{ST} followed similar patterns to that described for the full model (p-value for PLD=0.02, p-
325 value ecoregion= 0.001; Appendix Figure S3). However, for sessile species four models were
326 supported and in this case the variables overwater *distance* and reproductive strategy were
327 present in all models (Table 2), but did not meet the 0.8 threshold for variable importance, but
328 both were close at ~0.7 (Figure 2).

329 **3.3. Predictors of genetic structure: number of clusters**

330 Genetic structure based on the number of genetic clusters had the same or very similar predictor
331 variables for global F_{ST} in the GLMs analyses (Table 3, Figure 2). Different types of genetic
332 markers also had no significant effect on the number of genetic clusters identified (Appendix
333 Table S5). For the full dataset, the number of supported models and the important predictor
334 variables were identical to those in the F_{ST} analysis; *PLD* and *adult* life habit (Table 3, Figure
335 2). However, the relationship between PLD and the average number of genetic clusters
336 identified was weak. The average number of genetic clusters was 1.4 for migratory species
337 (range 1-3), 1.9 for motile species (range 3-6), 2.5 for sedentary (range 1-6) and 3.2 for sessile
338 species (range 1-13). Sessile species were significantly different to migratory and motile
339 species (Appendix Table S6).

340 When only the records with PLD>0 were analysed, there was a slight difference compared to
341 the full dataset; only *adult* life habit and not *PLD* was present in all the models, almost reaching
342 the 0.8 threshold for variable importance. Genetic structure for sessile species was the same as
343 the result for F_{ST} with overwater *distance* in all the models and it passed the 0.8 threshold
344 indicating its importance as a predictor (Table 3, Figure 2, p-value=0.05; Appendix Figure S4).
345 For motile (free swimming) species, the results were also very similar to those for $F_{ST,}$ with
346 *PLD* in all three models and *ecoregion* identified as important (> 0.8 threshold). There were
347 more genetic clusters when more ecoregions were sampled (p-value= 0.001; Appendix Figure
348 S5). For this particular variable we returned to the original papers and identified which
349 ecoregions within a study were allocated to each genetic cluster, and hence between which
350 pair-wise ecoregion comparisons a barrier had been identified based on being allocated to
351 different genetic clusters. Six ecoregions (1, 5, 6, 10, 12 and 14) had genetic structure identified
352 within a region. Barriers that were repeatedly identified were between ecoregions 3 & 5, and 5
353 & 10 (Appendix Figure S6 and S7). For sedentary species only *PLD* was in all models and

passed the threshold of 0.8 for importance, unlike the GLM analysis of F_{ST} ecoregion was not supported as an important variable.

4. DISCUSSION

4.1. The influence of dispersal-related traits

Our synthesis of 99 marine species in the Indo-Australian Archipelago, representing a diverse group of organisms with a range of life-history traits, has identified that species dispersal biology linked to adult mobility and maximum larval/propagule dispersal has the greatest influence on population genetic structure (Figure 3). This was supported through two measures of genetic structure, F_{ST} and the number of genetic clusters. These findings are congruent with well established, theoretical predictions that longer dispersal via larvae or propagules and more mobile adult life stages promote greater connectivity among populations (Cowen and Sponaugle, 2009). This is however, not always observed from studies of one or a few species (Weersing and Toonen, 2009; Liggins et al., 2016) highlighting the value of multi-species synthesis for providing insights into drivers of genetic structure at a regional scale, the scale important for informing conservation and management (Kelly and Palumbi, 2010; Treml et al., 2015).

It is not just the absence of a pelagic larval or propagule phase that influences genetic structure (as measured by $F_{ST)}$ but the maximum time that the larvae or propagule can disperse. This was evident because analysis of observations including marine organisms with a pelagic dispersal phase as well as without a pelagic dispersal phase were significant. We used only one PLD class, the maximum PLD, which Treml et al. (2015) identified as one of the key drivers of connectivity of representative marine species in the region. Weersing and Toonen (2009) also argued that the tails on the variation of larval duration were more informative than the mean PLD as they account for rare or extreme events as genetic structure is influenced by multiple

successful dispersal events over successive generations. PLD was often estimated from a few individuals at one sampling site and generally under laboratory conditions (Wellington and Victor, 1992; Macpherson and Raventos, 2006; Weersing and Toonen, 2009), and this basic biological trait was not known for all species collated in this analysis (9/99). Despite this limitation, and considering the importance of this predictor for understanding genetic structure, particularly in the IAA, more work is warranted to quantify this trait across marine plants and animals, as well as other reproductive strategy traits such as density and timing of larval or propagule release (Treml et al., 2015).

Pelagic larval duration alone was not the best predictor of genetic structure but support for its importance improved in combination with the adult life history category related to mobility (Figure 3). When all observations were assessed, migratory species had lower levels of genetic structure than motile, sessile or sedentary species. This would be expected as adults can disperse freely, and the population connectivity is less likely constrained by dispersal barriers, larval dispersal and other dispersal-related traits but more influenced by the behavioural ecology of the adults. For example, the spawning/reproductive behaviour and feeding migration have been shown to account for strong population connectivity in some species of salmonids, sharks, and herrings (Gaggiotti et al., 2009; Frisk et al., 2014). In this analysis, migratory species were the least represented, with only 15 observations out of 150 but despite this they had a strong, consistent pattern of lower F_{ST} and less genetic clusters. This could occur if the spatial scale of sampling was quite different between these different groups but migratory species were sampled over a similar spatial scale to motile and sessile species, on average 3,500 km maximum overwater distance between sites compared to 3,820 km and 3,630 km respectively.

When adult mobility categories were assessed independently, some subtle differences in the predictors of genetic structure were identified (Figure 3). Surprisingly, for sessile species,

403 where dispersal occurs during the early life stages (Cowen and Sponaugle, 2009), genetic
404 structure was not associated with pelagic larval duration but rather with maximum overwater
405 distance. In contrast, the genetic structure of motile and sedentary species was explained
406 strongly by PLD but also ecoregion, highlighting that in these groups of marine species,
407 different habitats and oceanographic-geological features among ecoregions act as barriers to
408 gene flow. Our study identified that when more ecoregions are sampled there is likely to be
409 more genetic structure, and when this did occur, it was not a function of the area sampled. A
410 greater distance did not necessarily mean more genetic structure, except for the case of sessile
411 species. While interactions between larval life history, habitat heterogeneity, and
412 oceanographic-geological barriers on genetic connectivity and diversity have been
413 demonstrated for single taxa, such as corals (Baums et al., 2006), fishes (Galarza et al., 2009;
414 Watson et al., 2010) and molluscs (Miller et al., 2013), our synthesis confirmed that they are
415 important across a wide range of species from a number of taxonomic groups including corals,
416 crustaceans, echinoderms, molluscs, reptiles and fish, but not migratory species. Larval life
417 history provides a means for dispersal, but the spatial scale and direction of dispersal is
418 influenced by oceanographic or geologic barriers that may be contemporary or historical, like
419 past changes in sea level and connectivity. For example, populations might be separated during
420 Pleistocene glaciations, then re-joined as sea levels rise, but genetic signatures of this historical
421 separation can appear in genetic structure analysis. Even in the absence of dispersal barriers,
422 individuals may reach new habitats, but local environmental selection may prevent them
423 settling, recruiting and reproducing thus preventing gene flow (Hunt and Scheibling, 1997;
424 Bierne et al., 2003).

425 **4.2. Influence of genetic markers on genetic structure**

426 Across all studies, we found no difference in the F_{ST} values nor number of genetic clusters
427 identified based on marker type. This contradicts previous studies that have shown differences
428 between mtDNA sequence data and other marker types in measuring genetic structure
429 (Weersing and Toonen, 2009; Riginos et al., 2011). As the set of genetic markers are from
430 different regions of the genome they coalesce over different timescales due to: (i) the
431 uniparental inheritance of mtDNA leading to fixation faster than biparental inheritance of
432 nuclear markers (thus higher F_{ST}); and (ii) differences in mutation rates, time to reach
433 migration-drift equilibrium, and degree of polymorphisms among the marker types (reviewed
434 in more details by Ballard and Whitlock, 2004; Zink and Barrowclough, 2008; Weersing and
435 Toonen, 2009). In this study, global F_{ST} was used as a descriptor of the genetic structure
436 summarising the variation across all sites in the study. This measure could vary due to a variety
437 of processes. For example, a higher F_{ST} could occur if sampling occurred at sites that spanned
438 a genetic break or if populations were spread over a large distance with limited gene flow and
439 genetic drift created divergence. Crandall et al. (2019) highlighted through simulations of
440 marine species across the Indian and Pacific Oceans that F-statistics were often an unreliable
441 indicator of divergence among populations. As the number of genetic clusters generated very
442 similar predictors of genetic structure to F_{ST}, this gives confidence for the general importance
443 of these predictors at the scale of this region and for the diversity of organisms sampled.

444 **4.3. Implications for marine conservation and future research**

445 The strong relationship between genetic structure and ecoregion, specifically for free
446 swimming and sedentary species warrants consideration for incorporating a genetic dimension
447 into the definition of Marine Ecoregions of the World (MEOW). Currently genetic diversity in
448 conservation planning is not explicitly included despite increasing awareness of its value (Sgrò

449 et al., 2011; Rivers et al., 2014). When exploring the barriers between ecoregions for the free-
450 swimming species they were congruent with a number of well documented barriers,
451 particularly the Sunda Shelf and Java Sea, south of Borneo between ecoregions 5 & 11, the
452 Halmahera Eddy at the boundary of the ecoregions 9, 11 and 13 and the southern barrier to the
453 Java Sea between ecoregions 14 and 15 (Appendix S6 and S7). Two commonly found barriers
454 were between ecoregions 3 and 5, and 5 and 10, either side of the Sunda Shelf. Treml et al.
455 (2015) identified that there is low larval connectivity between ecoregions 5 and 10, supporting
456 this observation and providing an additional potential mechanism for this barrier. However,
457 there were also a number of ecoregions where there was genetic structure within a single
458 ecoregion (Appendix Figure S6). If marine ecoregions were reassessed in order to incorporate
459 genetic cohesiveness more representative genetic data would be required for a range of taxa
460 across appropriate spatial scales.

461 This study has enabled identification of gaps in our knowledge to inform future sampling.
462 Although sequencing data was best represented, followed by microsatellite markers, studies
463 using single nucleotide polymorphisms (SNPs) are increasing in recent years. These genomic
464 approaches and the rapid development of more cost-effective whole genome sequencing will
465 enable interrogation of drivers of genetic structure, adaptation and evolution of biodiversity in
466 the region with insights into more recent timescales (Liggins et al., 2019; Nielsen et al., 2020).
467 Habitat formers such as corals, seagrass and algae which support high biodiversity through
468 their structure and food provision were least represented in this analysis with fish contributing
469 to over 50% of the observations, a similar finding identified from the synthesis of Keyse et al.
470 (2014) which only focused on animals. Future research should target these habitat forming
471 species considering they are a key focus of management and conservation measures, especially
472 in the face of rapidly changing environment (Underwood et al., 2013; Bulleri et al., 2018;
473 Babcock et al., 2019). Genetic structure data were available for most ecoregions, although it

474 was not evenly distributed (Keyse et al., 2014), and a number of areas either had no samples
475 or were poorly represented e.g. north western or eastern parts in the IAA and Papua New
476 Guinea. Furthermore, the spatial extent of data was not extensive, with most studies covering
477 only three ecoregions and a median overwater distance of 3,050 km. This reinforced the
478 recommendation of Crandall et al. (2019) that future studies would greatly benefit from co-
479 sampling from sites across the entire IAA.

480 Despite increases in the number of genetic studies since 2000 (Carpenter et al., 2011), there are
481 clearly still many gaps both spatially and within certain taxa (e.g. marine macrophytes) (Keyse
482 et al., 2014). Synthesis of existing and new studies may provide justification for the
483 incorporation of genetic cohesiveness into the classification of marine ecoregions. To assist
484 with this process, we recommend future genetic studies should sample sites that are
485 representatively nested in each ecoregion within a marine province or a marine realm and cover
486 the spatial extent of the IAA.

487 **4.4. Conclusions**

488 The genetic structure of marine species across the IAA, from a broad range of taxonomic
489 groups (corals, crustaceans, echinoderms, molluscs, reptiles, fish) was best predicted by traits
490 related to dispersal of larvae or propagules and the mobility of adults. The synthesis of these
491 101 studies from a biodiversity hotspot indicated that spatial management and conservation
492 should consider these key traits to help guide decision-making. The strong relationship between
493 genetic structure and ecoregion for free-swimming and sedentary species suggests that
494 historical geological and oceanographic processes, current oceanography and contemporary
495 environmental characteristics are also important drivers. Consideration for incorporating a
496 genetic dimension into the definition of marine ecoregions was supported in our study, as when
497 more ecoregions are sampled there is likely to be more genetic structure. There were still many

gaps in our understanding of genetic structure, both spatially and within certain taxa, and we recommended future genetic studies focus on habitat-forming taxa and sample sites that are representatively nested in each ecoregion within a marine province or a marine realm, over the spatial extent of the IAA.

ACKNOWLEDGMENTS

We thank the Botany writing group at the Ocean Institutes, The University of Western Australia, Prof Michelle Waycott at the State Herbarium of South Australia and the University of Adelaide and anonymous reviewers for comments improving the manuscript. This work was within the G100379 project of the Department of Education and Training Collaborative Research Network Program (Funding Agreement CRN2011:5, ECU and UWA) and supported by the Indonesia Endowment Fund for Education (LPDP-Indonesia).

AUTHOR CONTRIBUTIONS

Udhi E. Hernawan and Kathryn McMahon: Conceptualization, Funding acquisition, Project administration, Data curation, Formal analysis, Methodology, Validation, Visualization, Writing - original draft, review & editing.

Paul S. Lavery, Gary A. Kendrick, Kor-jent van Dijk, Yaya I Ulumuddin, and Teddy Triandiza: Formal analysis, Methodology, Writing- review & editing.

REFERENCES

Avise J.C. (1992) Molecular population structure and the biogeographic history of a regional fauna: a case history with lessons for conservation biology. *Oikos*, 63, 62–76.

Ayre D.J., Minchinton T.E., & Perrin C. (2009) Does life history predict past and current connectivity for rocky intertidal invertebrates across a marine biogeographic barrier? *Molecular Ecology*, 18, 1887–1903.

Babcock R.C., Bustamante R. H., Fulton E. A., Fulton D. J., Haywood M. D. E., Hobday A. J., Kenyon R., Matear R. J., Plagányi E. E., Richardson A. J., & Vanderklift M. A. (2019) Severe continental-scale impacts of climate change are happening now: Extreme climate events impact marine habitat forming communities along 45% of Australia's coast. *Frontiers in Marine Science*. 6, 411. doi: 10.3389/fmars.2019.00411

Ballard J.W.O. & Whitlock M.C. (2004) The incomplete natural history of mitochondria. *Molecular Ecology*, 13, 729–744.

Barber P.H. (2009) The challenge of understanding the Coral Triangle biodiversity hotspot. *Journal of Biogeography*, 36, 1845–1846.

Barber P.H., Cheng S.H., Erdmann M. V, Tenggardjaja K., & Ambariyanto (2011) Evolution and conservation of marine biodiversity in the Coral Triangle. *Crustacean Issues: Phylogeography and Population Genetics in Crustacea*, 19, 129–156.

Baums I.B., Paris C.B., & Chérubin L.M. (2006) A bio-oceanographic filter to larval dispersal in a reef-building coral. *Limnology and Oceanography*, 51, 1969–1981.

Bierne N., Bonhomme F., & David P. (2003) Habitat preference and the marine-speciation paradox. *Proceedings of the Royal Society B: Biological Sciences*, 270, 1399–1406.

Bradbury I.R., Laurel B., Snelgrove P.V.R., Bentzen P., & Campana S.E. (2008) Global patterns in marine dispersal estimates: the influence of geography, taxonomic category and life history. *Proceedings of the Royal Society B Biological Sciences*, 275, 1803–1809.

Bulleri F., Eriksson B. K., Queirós A., Airoldi L., Arenas F., (2018) Harnessing positive species interactions as a tool against climate-driven loss of coastal biodiversity. *PLOS Biology* 16(9):

e2006852. https://doi.org/10.1371/journal.pbio.2006852

Burnham K.P. & Anderson D.R. (2002) *Model Selection and Multimodel Inference: A practical Information-Theoretic Approach.* Springer-Verlag New York.

Calcagno V. & Mazancourt C. de (2010) glmulti: An R package for easy automated model selection with (Generalized) Linear Models. *Journal of Statistical Software*, 34, 1–29.

Carpenter K.E., Barber P.H., Crandall E.D., Ambariyanto, Ablan-Lagman C.A., Mahardika G.N., Manjaji-Matsumoto B.M., Juinio-Menez M.A., Santos M.D., Starger C.J., & Toha A.A. (2011) Comparative phylogeography of the Coral Triangle and implications for marine management. *Journal of Marine Biology*, 2011, 1–14.

Charlesworth B., Charlesworth D., & Barton N.H. (2003) The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology Evolution and Systematics*, 34, 99–125.

Costantini F., Ferrario F., & Abbiati, M. (2018) Chasing genetic structure in coralligenous reef invertebrates: patterns, criticalities and conservation issues. *Scientific Reports* 8, 5844.

Cowen R.K. & Sponaugle S. (2009) Larval dispersal and marine population connectivity. *Annual Review of Marine Science*, 1, 443–466.

Crandall, E. D., Riginos, C., Bird, C. E., Liggins, L., Treml, E., Beger, M., … Gaither, M. R. (2019) The molecular biogeography of the Indo-Pacific: Testing hypotheses with multispecies genetic patterns. *Global Ecology and Biogeography*, 28(7), 943–960. https://doi.org/10.1111/geb.12905

de Juan S., Demestre M., & Thrush S. (2009) Defining ecological indicators of trawling disturbance when everywhere that can be fished is fished: A Mediterranean case study. *Marine Policy*, 33, 472–478.

Doherty P., Planes S., & Mather P. (1995) Gene flow and larval duration in seven species of

fish from the Great Barrier Reef. *Ecology*, 76, 2373–2391.

Donati G. F. A., Parravicini V., Leprieur F., Hagen O., Gaboriau T., Heine C., Kulbicki M., Rolland J., Salamin N., Albouy N., & Pellissier L. (2019) A process-based model supports an association between dispersal and the prevalence of species traits in tropical reef fish assemblages. *Ecography,* 42, 2095–2106

Evans SM, McKenna C, Simpson SD, Tournois J, Genner MJ. (2016) Patterns of species range evolution in Indo-Pacific reef assemblages reveal the Coral Triangle as a net source of transoceanic diversity. *Biology Letters*, 12(6), 20160090.

Fisher R., Radford B. T., Knowlton N., Brainard R. E., Michaelis F.mB., & Caley M. J. (2011) Global mismatch between research effort and conservation needs of tropical coral reefs. *Conservation Letters*, 4, 64–72.

Foggo A., Bilton D.T., & Rundle S.D. (2007) Do developmental mode and dispersal shape abundance-occupancy relationships in marine macroinvertebrates? *Journal of Animal Ecology*, 76, 695–702.

Fox J., Weisberg S., Adler D., Bates D., Baud-Bovy G., Ellison S., Firth D., Friendly M., Gorjanc G., Graves S., Heiberger R., Laboissiere R., Monette G., Murdoch D., Nilsson H., Ogle D., Ripley B., Venables W., & Zeileis A. (2015) *Package "car" 2.0-25 : Companion to Applied Regression.* (http://cran.r-project.org/web/packages/car/index.html),

Frisk M.G., Jordaan A., & Miller T.J. (2014) Moving beyond the current paradigm in marine population connectivity: Are adults the missing link? *Fish and Fisheries*, 15, 242–254.

Gaggiotti O.E., Bekkevold D., Jørgensen H.B.H., Foll M., Carvalho G.R., Andre C., & Ruzzante D.E. (2009) Disentangling the effects of evolutionary, demographic, and environmental factors influencing genetic structure of natural populations: Atlantic herring as a case study. *Evolution*, 63, 2939–2951.

Galarza J., Carreras-Carbonell J., Macpherson E., Pascual M., Roques S., Turner G., & Rico C. (2009) The influence of oceanographic fronts and early-life-history traits on connectivity among littoral fish species. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 1473–1478.

Harrell Jr F.E. (contributions from C.D. and many others) (2015) *Hmisc: Harrell Miscellaneous, R package version 3.15-0.* (http://cran.r-project.org/web/packages/Hmisc/index.html)

Hoegh-Guldberg O. (2010) Coral reef ecosystems and anthropogenic climate change. *Regional Environmental Change*, 11, 215–227.

Hoeksema B. (2007) Delineation of the Indo-Malayan centre of maximum marine biodiversity: the Coral Triangle. In: Renema, W. (Ed.) *Biogeography, Time, and Place: Distributions, Barriers, and Islands*. Springer Netherlands, pp. 117–178.

Hothorn T., Bretz F., & Westfall P. (2008) Simultaneous inference in general parametric models. *Biometrical Journal*, 50, 346–363.

Hu Z.M., Zhang J., Lopez-Bautista J., & Duan D.L. (2013) Asymmetric genetic exchange in the brown seaweed *Sargassum fusiforme* (Phaeophyceae) driven by oceanic currents. *Marine Biology*, 160, 1407–1414.

Hunt H.L. & Scheibling R.E. (1997) Role of early post-settlement mortality in recruitment of benthic marine invertebrates. *Marine Ecology Progress Series*, 155, 269–301.

Jacobs D.K., Haney T.A., & Louie K.D. (2004) Genes, diversity, and geologic process on the Pacific Coast. *Annual Review of Earth and Planetary Sciences*, 32, 601–652.

Kelly R.P. & Palumbi S.R. (2010) Genetic structure among 50 species of the northeastern Pacific rocky intertidal community. *PloS ONE*, 5, e8594.

Keyse, J., Crandall, E.D., Toonen, R., Meyer, C., Treml, E.A., & Riginos, C. (2014). The scope

of published population genetic data for Indo-Pacific marine fauna and future research opportunities in the region. *Bulletin of Marine Science*, 90, 47-78.

Liggins, L., Treml E.A., Possingham, H.P. & Riginos C. (2016) Seascape features, rather than dispersal traits, predict spatial genetic patterns in co-distributed reef fishes. *Journal of Biogeography*, 43, 256–267.

Liggins, L., Treml, E. A., & Riginos, C. (2019). Seascape Genomics: Contextualizing Adaptive and Neutral Genomic Variation in the Ocean Environment. In M. Oleksiak & O. Rajora (Eds.), Population Genomics: Marine Organisms. Springer, Cham. Switzerland, pp. 171–218. https://doi.org/10.1007/13836_2019_68

Lohman D.J., de Bruyn M., Page T., von Rintelen K., Hall R., Ng P.K.L., Shih H.-T., Carvalho G.R., & von Rintelen T. (2011) Biogeography of the Indo-Australian Archipelago. *Annual Review of Ecology, Evolution, and Systematics*, 42, 205–226.

Lowe A., Harris S., & Ashton P. (2004) *Ecological Genetics: Design, Analysis and Application.* Blackwell Publishing, Malden, MA USA.

Macpherson E. & Raventos N. (2006) Relationship between pelagic larval duration and geographic distribution of Mediterranean littoral fishes. *Marine Ecology Progress Series*, 327, 257–265.

Maguire J.-J., Sissenwine M., Csirke J., & Grainger R. (2006) The state of the world highly migratory, straddling and other high seas fish stocks, and associated species. *FAO Fisheries Technical Paper*, 495, 67.

McLeod E., Moffitt R., Timmermann A., Salm R., Menviel L., Palmer M.J., Selig E.R., Casey K.S., & Bruno J.F. (2010) Warming seas in the Coral Triangle: Coral reef vulnerability and management implications. *Coastal Management*, 38, 518–539.

Meirmans P.G. & Hedrick P.W. (2011) Assessing population structure: F(ST) and related

measures. *Molecular Ecology Resources*, 11, 5–18.

Miller A.D., Versace V.L., Matthews T.G., Montgomery S., & Bowie K.C. (2013) Ocean currents influence the genetic structure of an intertidal mollusc in southeastern Australia - implications for predicting the movement of passive dispersers across a marine biogeographic barrier. *Ecology and Evolution*, 3, 1248–1261.

Nanninga, G. B., & Manica, A. (2018). Larval swimming capacities affect genetic differentiation and range size in demersal marine fishes. *Marine Ecology Progress Series*, 589, 1–12.

Nielsen E. S., Henriques R., Beger M., Toonen, R & von der Heyden S. (2020). Multi-model seascape genomics identifies distinct environmental drivers of selection among sympatric marine species. *BMC Evolutionary Biology* 20, 121. doi:10.1186/s12862-020-01679-4.

Palumbi S. R. (1992) Marine speciation on a small planet. *Trends in Ecology & Evolution*, 7, 114–118.

Palumbi S. R. (1994) Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics*, 25, 547–572.

Palumbi, S. R. (2004) Marine reserves and ocean neighbourhoods: the spatial scale of marine populations and their management. *Annual Review of Environment and Resources*, 29, 31–68.

Parker P.G., Snow A. a, Schug M.D., Booton G.C., & Fuerst P.A. (1998) What molecules can tell us about populations: Choosing and using a molecular marker. *Ecology*, 79, 361–382.

Pelc R., Warner R., & Gaines S. (2009) Geographical patterns of genetic structure in marine species with contrasting life histories. *Journal of Biogeography*, 36, 1881–1890.

R Development Core Team (2015) *R: A Language and Environment for Statistical Computing*.

Reiss, H., Hoarau, G., Dickey-Collas, M., & Wolff, W. J. (2009). Genetic population structure

of marine fish: Mismatch between biological and fisheries management units. *Fish and Fisheries*, 10 (4), 361–395.

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

Riginos C., Douglas K.E., Jin Y., Shanahan D.F., & Treml E.A. (2011) Effects of geography and life history traits on genetic differentiation in benthic marine fishes. *Ecography*, 34, 566–575.

Rivers M.C., Brummitt N.A., Nic Lughadha E., & Meagher T.R. (2014) Do species conservation assessments capture genetic diversity? *Global Ecology and Conservation*, 2, 81–87.

Schlötterer C. (2004) The evolution of molecular markers--just a matter of fashion? *Nature Reviews: Genetics*, 5, 63–69.

Selig E.R., Turner W.R., Troëng S., Wallace B.P., Halpern B.S., Kaschner K., Lascelles B.G., Carpenter K.E., & Mittermeier R. (2014) Global priorities for marine biodiversity conservation. *PLoS ONE*, 9, e82898.

Selkoe K.A., Gaggiotti O.E., Bowen B.W., & Toonen R.J. (2014) Emergent patterns of population genetic structure for a coral reef community. *Molecular Ecology*, 23, 3064–79.

Sgrò C.M., Lowe A.J., & Hoffmann A.A. (2011) Building evolutionary resilience for conserving biodiversity under climate change. *Evolutionary Applications*, 4, 326–337.

Shanks A., Grantham B., & Carr M. (2003) Propagule dispersal distance and the size and spacing of marine reserves. *Ecological Applications*, 13, S159–S169.

Shanks A.L. (2009) Pelagic larval duration and dispersal distance revisited. *Biological Bulletin*, 216, 373–385.

Short F., Carruthers T., Dennison W., & Waycott M. (2007) Global seagrass distribution and diversity: A bioregional model. *Journal of Experimental Marine Biology and Ecology*, 350, 3–20.

Siegel D., Kinlan B., Gaylord B., & Gaines S. (2003) Lagrangian descriptions of marine larval dispersion. *Marine Ecology Progress Series*, 260, 83–96.

Simpson S.D., Harrison H.B., Claereboudt M.R., & 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.

Spalding M.D., Fox H.E., Allen G.R., Davidson N., Zach A., Finlayson M., Halpern B.S., Jorge M.A., Lombana A.L., Lourie S.A., Martin K.D., McManus E., Molnar J., Cheri A., & Robertson J. (2007) Marine Ecoregions of the World: A bioregionalization of coastal and shelf areas. *BioScience*, 57, 573–583.

Treml, E. A., Roberts J., Halpin P. N., Possingham H. P., & Riginos, C. (2015) The emergent geography of biophysical dispersal barriers across the Indo-West Pacific. *Diversity and Distributions*, 21, 465–47

Underwood, J. N., Wilson, S. K., Ludgerus, L., & Evans, R. D. (2013). Integrating connectivity science and spatial conservation management of coral reefs in north-west Australia. *Journal for Nature Conservation*, 21 (3), 163–172. https://doi.org/10.1016/j.jnc.2012.12.001

von der Heyden, S., Beger, M., Toonen, R. J., van Herwerden, L., Juinio-Meñez, M. A., Ravago-Gotanco, R., … Bernardi, G. (2014). The application of genetics to marine management and conservation: examples from the Indo-Pacific. *Bulletin of Marine Science*, 90(1), 123–158.

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

Watson J.R., Mitarai S., Siegel D.A., Caselle J.E., Dong C., & McWilliams J.C. (2010) Realized and potential larval connectivity in the southern California bight. *Marine Ecology Progress Series*, 401, 31–48.

Weersing K. & Toonen R. (2009) Population genetics, larval dispersal, and connectivity in marine systems. *Marine Ecology Progress Series*, 393, 1–12.

Wellington G.M. & Victor B.C. (1992) Regional differences in duration of the planktonic larval stage of reef fishes in the eastern Pacific Ocean. *Marine Biology*, 113, 491–498.

Wright S. (1931) Evolution in Mendelian populations. *Genetics*, 16, 97–159.

Zink R. M. & Barrowclough G. F. (2008). Mitochondrial DNA under siege in avian phylogeography. *Molecular Ecology*, 17, 2107-2121

FIGURE CAPTIONS

Figure 1. Marine ecoregions in the Indo-Australian Archipelago based on Spalding *et al*. (2007). Dashed lines correspond to boundaries of each ecoregion and circles indicate the number of observations collated in this review for each ecoregion. (1)-Andaman Sea Coral Coast, (2)-Western Sumatra, (3)-Malacca Strait, (4)-Gulf of Thailand, (5)-Sunda Shelf/Java Sea, (6)-Southern Java, (7)-Southern Vietnam, (8)-South China Sea Oceanic Islands, (9)-Eastern Philippines, (10)-Palawan/North Borneo, (11)-Sulawesi Sea/Makassar Strait, (12)-Northeast Sulawesi, (13)-Halmahera, (14)-Banda Sea, (15)-Lesser Sunda, (16)-Arafura Sea, (17)-Papua, (18)-Bismarck Sea, (19)-Solomon Sea, (20)-Gulf of Papua, (21)-Southeast Papua New Guinea, (22)-Torres Strait Northern GBR, (23)-Coral Sea, (24)-Central and Southern GBR, (25)-Arnhem Coast-Gulf of Carpentaria, (26)-Bonaparte Coast, (27)-Exmouth to Broome, (28)-Cocos-Keeling/Christmas Island. NEC=North Equatorial Current, KC=Kuroshio Current, MC=Mindanao Current, HE=Halmahera Eddy, SEC=South Equatorial Counter current.

Figure 2. Relative evidence weight of predictors generated for F_{ST} (left-hand panel) and genetic cluster (right-hand panel) using the full dataset (top), species with PLD>0 (2nd), free-swimming (3rd), sedentary (4th) and sessile (bottom). The x-axis indicates relative weight of evidence. A vertical dashed line at 0.8 is the threshold above which the predictor is significant. Abbreviations, eco= *ecoregion*, dist= *distance*, rs= *rep. strategy*, ad= *adult* (adult life habit), egg= *egg* type.

Figure 3. Summary of the key drivers of genetic structure in the IAA based on this analysis. Dark blue indicates the driver was significant in based on both genetic structure measures (F_{ST} and the number of genetic clusters) whereas light blue indicates it was significant for only one measure and black indicates that this driver was not assessed.

Figure 1

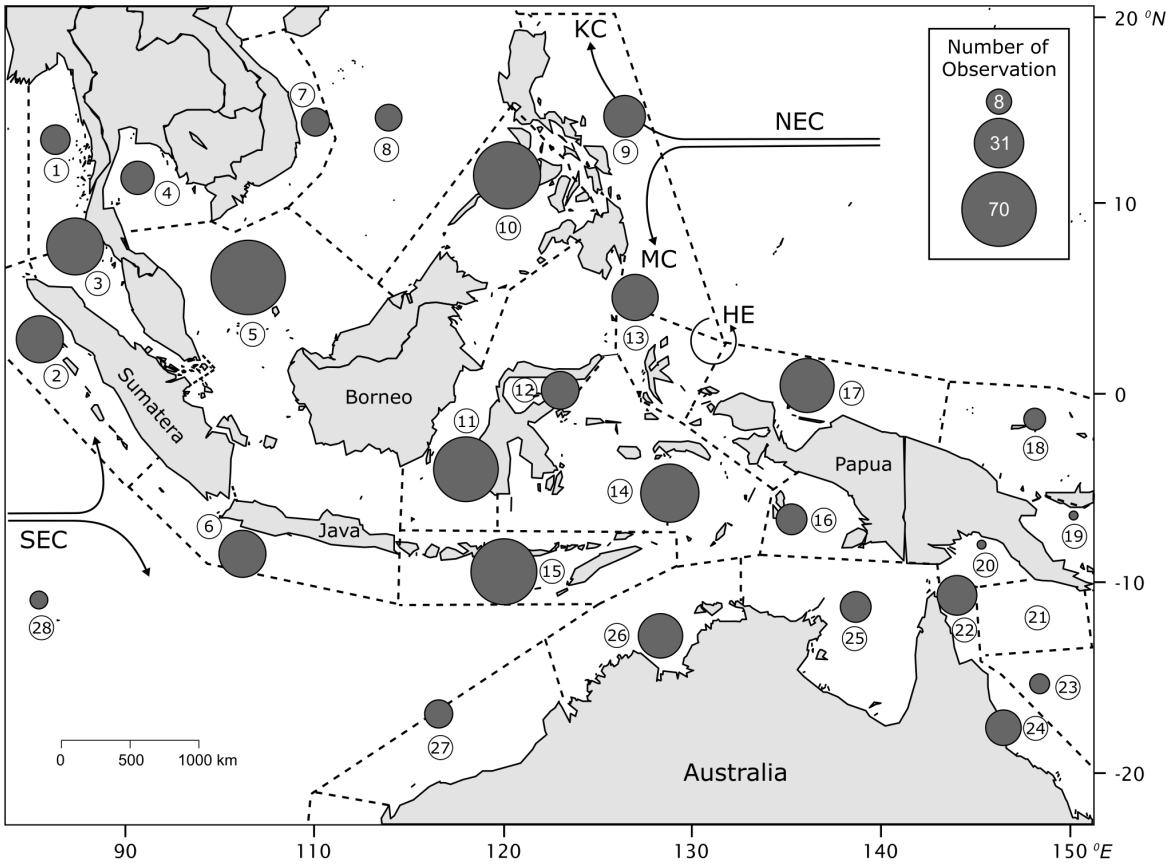

KC
NEC
MC
HE
SEC
Number of Observation
8
31
70
Sumatera
Borneo
Java
Papua
Australia
0 500 1000 km
20 °N
10
0
-10
-20
90
110
120
130
140
150 °E

Figure 2

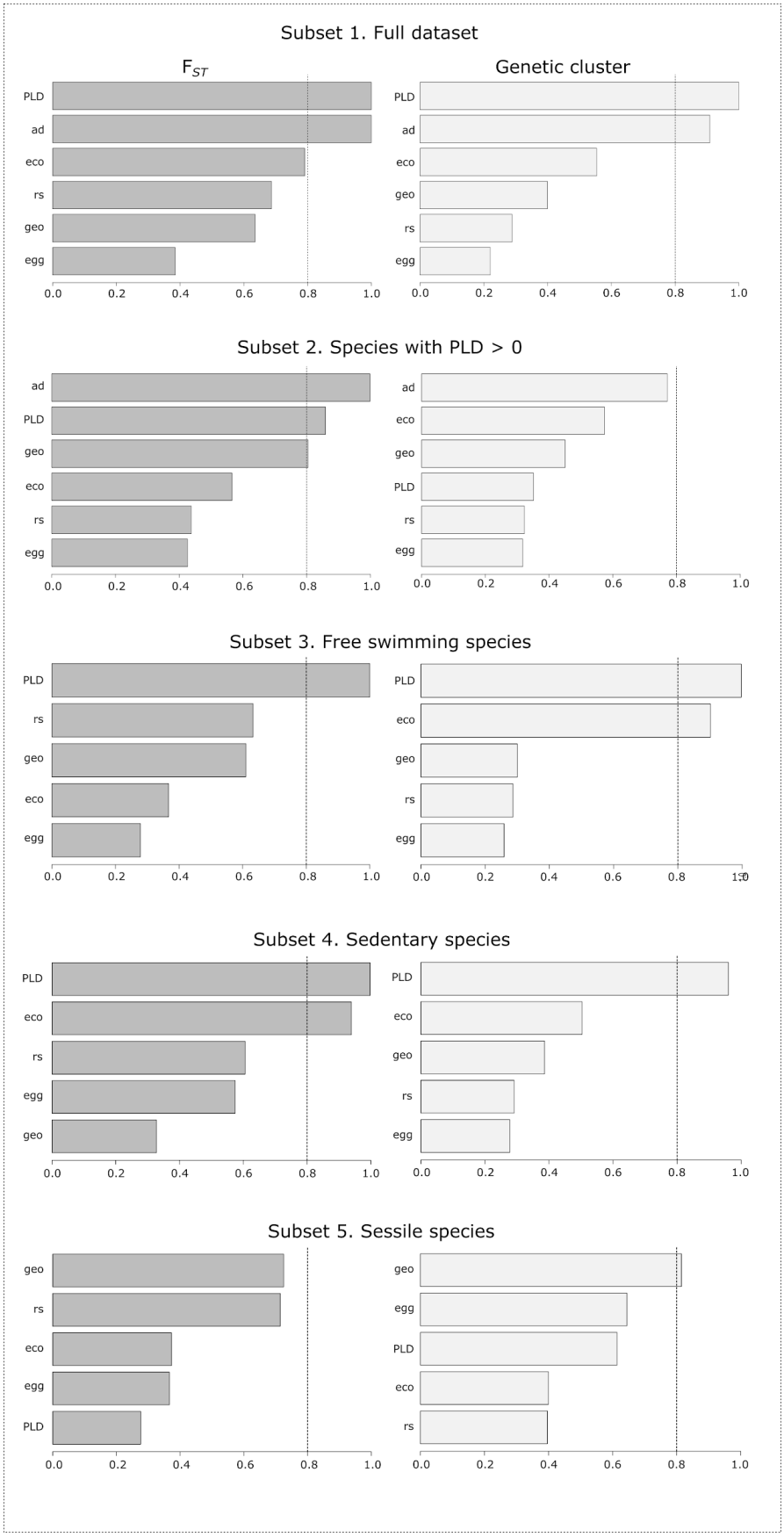

Subset 1. Full dataset
F_{ST}
Genetic cluster
PLD
ad
eco
rs
geo
egg
0.0
0.2
0.4
0.6
0.8
1.0
Subset 2. Species with PLD > 0
Subset 3. Free swimming species
Subset 4. Sedentary species
Subset 5. Sessile species

Figure 3

Drivers of genetic structure	Taxa groups				
	All	PLD>0	Motile 	Sedentary 	Sessile 
PLD					
Adult					
Distance					
Ecoregion					

759 Table 1. A summary of the data extracted from the peer-reviewed studies based on genetic
760 structure (global F_{ST} and the number of genetic clusters) and the potential predictor variables
761 for genetic structure. Texts in brackets are the standard terms used throughout the document
762 to summarise these variables. The summary statistics include the median and range for each
763 variable based on all the observations in the dataset or the number of observations for each
764 category of predictor variables.
765
766

Criteria	**Variable**	**Median from dataset**	**Range from dataset**
Genetic structure	Global F_{ST} (F_{ST})	0.077	0 - 0.905
	Number of genetic clusters (*cluster*)	2	1 - 13
Habitat heterogeneity & oceanographic-geologic features	Number of marine ecoregions (*ecoregion*)	3	1 - 15
Geographic distance	Maximum overwater distance among sampling sites in km (*distance*)	3050	23 - 9728
Dispersal-related traits	Pelagic larval duration in hr (*PLD*)	360	0 - 5640
		Number of observations per category	
Dispersal-related traits	Adult life habit (*adult*)	sessile: 42 motile: 50	sedentary: 43 migratory: 15
	Reproductive strategy (*rep. strategy*)	broadcaster: 91	brooder: 59
	Egg type (*egg*)	pelagic: 102 direct: 16	benthic: 32
Genetic marker	Type of genetic markers used (*marker*)	sequence: 82 allozyme: 8	MSat: 46 other: 4 SNP: 10

767

768 Table 2. Best models generated for F_{ST} using full and restricted dataset. Only models within
769 the lowest two AIC units are shown in the table.
770

Model	ΔAIC	AIC Weight
Full dataset (n=117)		
Fst ~ (marker) + adult + PLD + rep. strategy + ecoregion + distance	0.000	0.242
Fst ~ (marker) + adult + PLD + rep. strategy + ecoregion	0.676	0.173
Fst ~ (marker) + adult + PLD + rep. strategy + distance	1.309	0.126
Fst ~ (marker) + adult + PLD + egg + ecoregion	1.653	0.106
Fst ~ (marker) + adult + PLD + egg + ecoregion + distance	1.925	0.092
Species with a pelagic larval stage / PLD > 0 (n=103)		
Fst ~ (marker) + adult + PLD + distance + rep. strategy	0.000	0.124
Fst ~ (marker) + adult + PLD + distance	0.348	0.104
Fst ~ (marker) + adult + PLD + distance + egg	0.382	0.102
Fst ~ (marker) + adult + PLD + distance + rep. strategy + ecoregion	0.427	0.100
Fst ~ (marker) + adult + PLD + distance + ecoregion	0.447	0.099
Fst ~ (marker) + adult + PLD + distance + ecoregion + egg	0.793	0.083
Fst ~ (marker) + adult + PLD + distance + egg+ rep. strategy	1.559	0.057
Fst ~ (marker) + adult + PLD + ecoregion	1.808	0.050
Free swimming species (n=35) excluding predictor adult life habit		
Fst ~ (marker) + rep. strategy + PLD + distance	0.000	0.208
Fst ~ (marker) + rep. strategy + PLD	0.864	0.135
Fst ~ (marker) + rep. strategy + PLD + distance + ecoregion	1.430	0.102
Fst ~ (marker) + rep. strategy + PLD + ecoregion	1.507	0.098
Fst ~ (marker) + PLD + distance	1.561	0.095
Sedentary species (n= 35) excluding predictor adult life habit		
Fst ~ (marker)+ PLD + ecoregion + rep. strategy + egg	0.000	0.274

Fst ~ (marker) + PLD + ecoregion	0.653	0.197
Fst ~ (marker) + PLD + ecoregion + rep. strategy + egg + distance	1.329	0.141
Sessile species (n= 34) excluding predictor adult life habit		
Fst ~ (marker) + rep. strategy + distance	0.000	0.180
Fst ~ (marker) + rep. strategy + distance + egg	1.400	0.089
Fst ~ (marker) + rep. strategy + distance + ecoregion	1.598	0.081
Fst ~ (marker) + rep. strategy + distance + PLD	2.000	0.066

771
772

773 Table 3. Best models generated for genetic clusters using full and restricted dataset. Only
774 models within the lowest two AIC units are shown in the table.
775

Model	ΔAIC	AIC Weight
Full dataset (n=142)		
cluster ~ (marker) + adult + PLD	0.000	0.182
cluster ~ (marker) + adult + PLD + ecoregion	0.400	0.149
cluster ~ (marker) + adult + PLD + ecoregion + distance	0.624	0.133
cluster ~ (marker) + adult + PLD + distance	1.998	0.067
cluster ~ (marker) + adult + PLD + rep. strategy	1.998	0.067
Species with a pelagic larval stage / PLD > 0 (n=119)		
cluster ~ (marker) + adult	0.000	0.088
cluster ~ (marker) + adult + ecoregion + distance	0.502	0.069
cluster ~ (marker) + adult + ecoregion	0.581	0.066
cluster ~ (marker) + adult + PLD	1.288	0.046
cluster ~ (marker) + ecoregion + distance	1.608	0.039
cluster ~ (marker) + adult + rep. strategy	1.730	0.037
cluster ~ (marker) + adult + PLD + ecoregio	1.838	0.035
cluster ~ (marker) + adult + egg	1.894	0.034
cluster ~ (marker) + adult + distance	1.998	0.033
Free swimming species (n=50) excluding predictor adult life habit		
cluster ~ (marker) + PLD + ecoregion	0.000	0.345
cluster ~ (marker) + PLD + ecoregion + rep. strategy	1.802	0.140
cluster ~ (marker) + PLD + ecoregion + distance	1.988	0.128
Sedentary species (n= 40) excluding predictor adult life habit		
cluster ~ (marker) + PLD + ecoregion	0.000	0.179
cluster ~ (marker) + PLD + distance	0.706	0.126

cluster ~ (marker) + PLD	0.774	0.121
cluster ~ (marker) + PLD + ecoregion + rep. strategy	1.752	0.074
cluster ~ (marker) + PLD + ecoregion + egg	1.923	0.068
cluster ~ (marker) + PLD + ecoregion + distance	1.931	0.068
Sessile species (n= 37) excluding predictor adult life habit		
cluster ~ (marker) + distance + PLD + egg	0.000	0.130
cluster ~ (marker) + distance + PLD + egg + rep. strategy	0.379	0.108
cluster ~ (marker) + distance + PLD	1.441	0.063
cluster ~ (marker) + distance + PLD + egg + ecoregion	1.453	0.063
cluster ~ (marker) + distance + egg	1.495	0.062
cluster ~ (marker) + distance	1.637	0.057

SUPPLEMENTARY TABLES AND FIGURES

Appendix Table S1. Pairwise comparison of Hoeffding's D values (lower triangle) and its p-value (upper triangle).

	ecoregion	*distance*	*PLD*
ecoregion	-	0.24	0
distance	0	-	0.0083
PLD	0.416	0.615	-

#Hoeffding's D ranges -.05 to 1 => 1 means absolute dependency, 0 means total independency

Appendix Table S2. Test of multicollinearity on *ecoregion, distance,* and *PLD* based on Variable Inflation Factor (VIF) with dependent variable *Fst* and *cluster*

	VIF		
	ecoregion	*distance*	*PLD*
Fst	1.837547	1.844075	1.007233
cluster	1.872152	1.786853	1.070240

Note: a *VIF* value of 1 means that the predictor is not correlated with other variables. The higher the value, the greater the correlation of the variable with other variables. Values of more than 4 or 5 are sometimes regarded as being moderate to high, with values of 10 or more being regarded as very high (meaning a strong correlation among variables)

Appendix Table S3. Pairwise comparisons of means on the influence of genetic marker type (lower triangle) and p-values (upper triangle) on the dataset using F_{ST}

Marker	**allozyme**	**EPIC**	**msat**	**Seq**	**SNP**
allozyme		ns	ns	ns	ns
EPIC	0.65847	-	ns	ns	ns
msat	-0.01136	-0.66983	-	ns	ns
mtDNA	0.7237	0.06524	0.73507	-	ns
SNP	-0.08408	-0.74254	-0.07272	-0.80778	-

Note: formula used for this comparison was the best model in Table 2 with full dataset: *logFst1 ~ 1 + marker + adult + rep.strategy + PLD + ecoregion + distance*

Appendix Table S4. Pairwise comparisons of means on the influence of adult life habit (lower triangle) and p-values (upper triangle) on the dataset using F_{ST}.

Adult life habit	free swimming	migratory	sedentary	sessile
free swimming	-	< 0.001	ns	ns
migratory	-2.6716	-	0.014	<0.001
sedentary	-0.5837	2.0879	-	ns
sessile	0.6515	3.3231	1.2353	-

Note: formula used for this comparison was the best model in Table 2 with full dataset: *logFst1 ~ 1 + marker + adult + rep.strategy + PLD + ecoregion + distance*

Appendix Table S5. Pairwise comparisons of means on the influence of genetic marker (lower triangle) and p-values (upper triangle) on the dataset using the number of genetic clusters.

Marker	allozyme	EPIC	msat	Seq	SNP
allozyme	-	ns	ns	ns	ns
EPIC	0.4709	-	ns	ns	ns
msat	0.5763	0.1054	-	ns	ns
mtDNA	0.3555	-0.1154	-0.2208	-	ns
SNP	0.4786	0.0078	-0.0977	0.1231	-

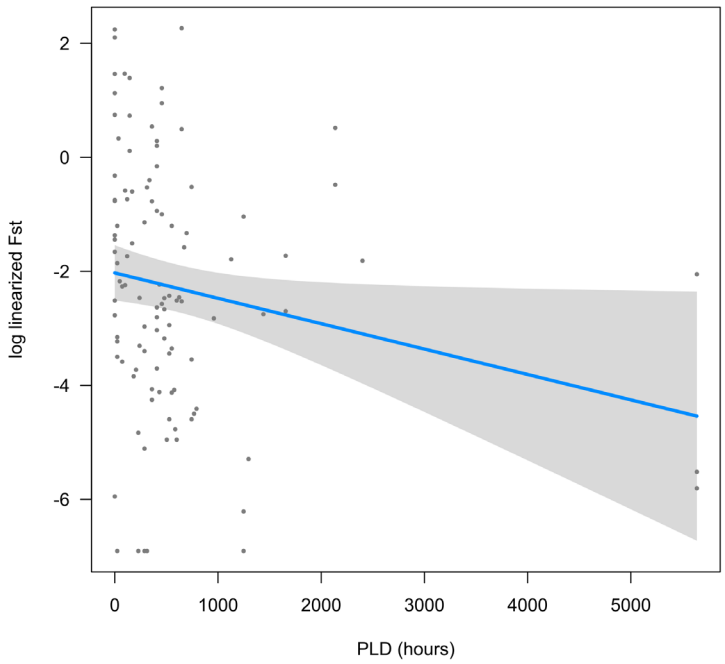
Note: formula used for this comparison was the best model in Table 3 with full dataset: *cluster ~ (marker) + adult + PLD*

Appendix Table S6. Pairwise comparisons of means on the influence of adult life habit (lower triangle) and p-values (upper triangle) on the dataset using the number of genetic clusters.

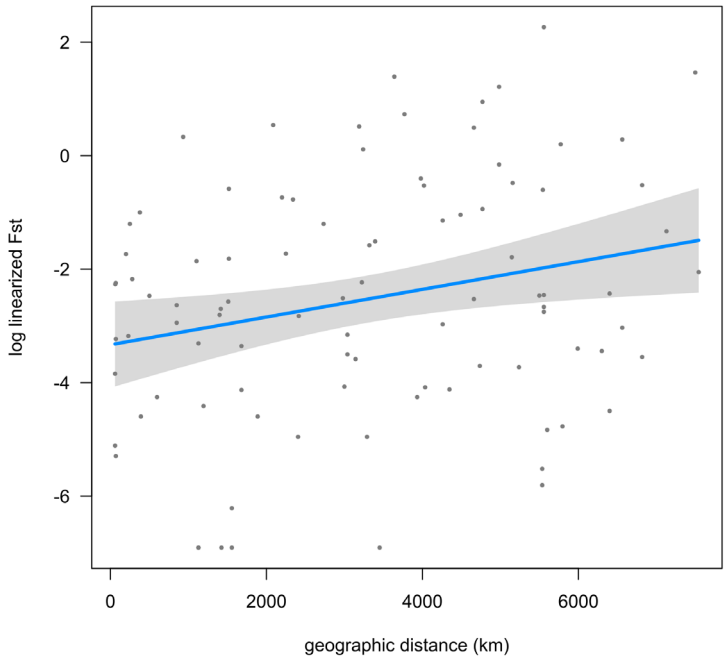
Adult life habit	free swimming	migratory	sedentary	sessile
free swimming	-	ns	ns	0.0183
migratory	-0.3351	-	ns	0.0152
sedentary	0.2258	0.5609	-	ns
sessile	0.4105	0.7456	0.1847	-

Note: formula used for this comparison was the best model in Table 3 with full dataset: *cluster ~ (marker) + adult + PLD*

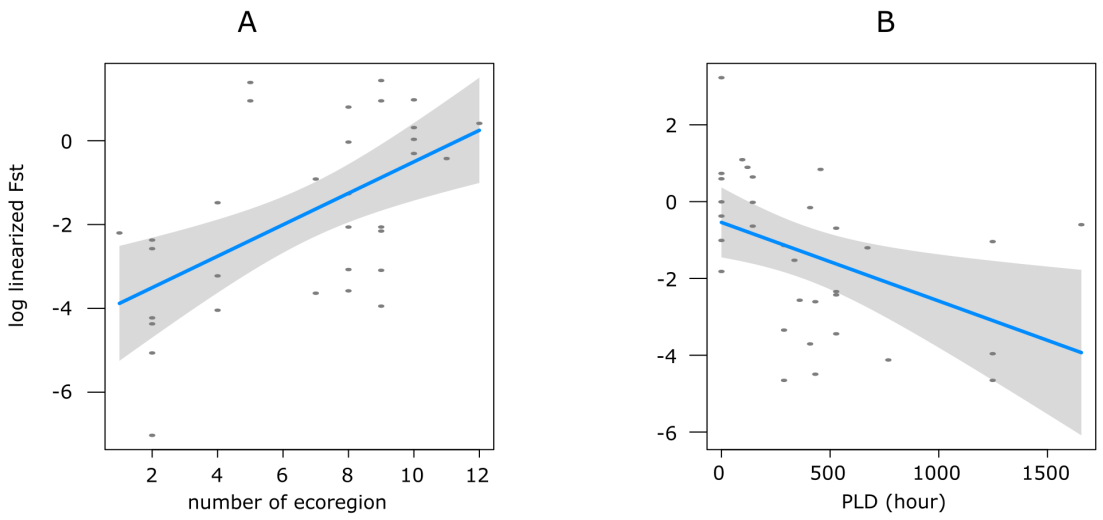
APPENDIX FIGURES



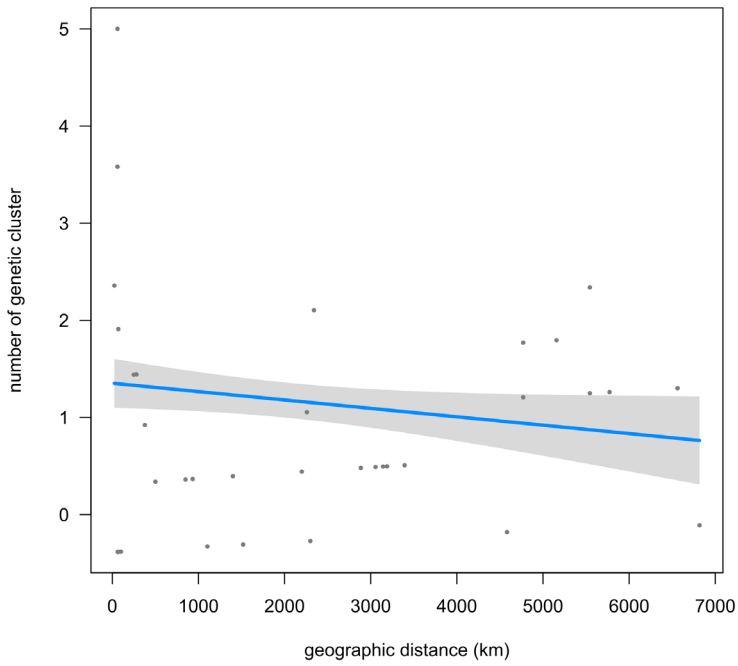
Appendix Figure S1. Relationship between F_{ST} and *PLD* using the full dataset with species with no PLD included (p-value=0.04). On the y-axes, F_{ST} was log linearized using formula $\log((F_{ST}+0.001)/(1-(F_{ST}+0.001))$. When the extremely long PLD points were removed there was still a significant relationship.



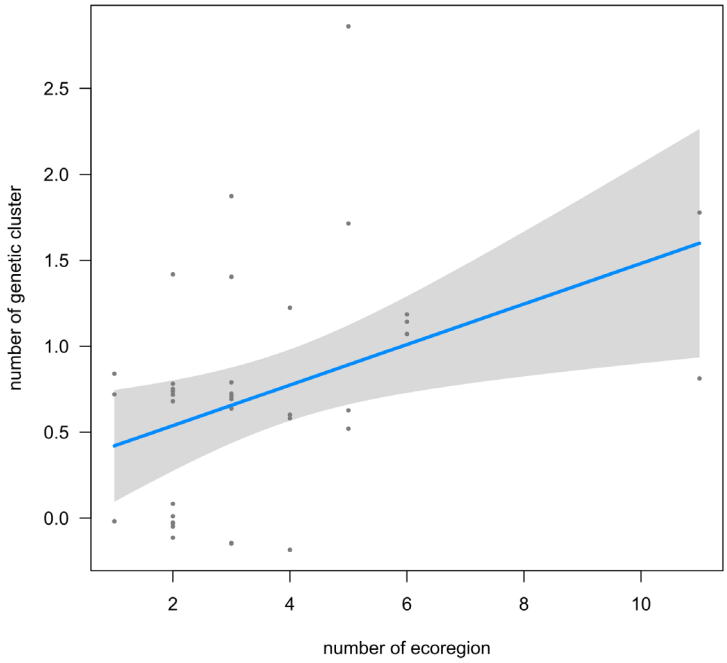
Appendix Figure S2. Relationship between F_{ST} and geographic (over-water) distance (km) using Subset 2 with PLD>0 observations only (p-value=0.015). On the y-axes, F_{ST} was log linearized using formula log((F_{ST}+0.001)/(1-(F_{ST}+0.001)).



Appendix Figure S3. Relationship of F_{ST} with ecoregion (panel A; p-value=0.001) and PLD (panel B; p-value= 0.02) using Subset 4, sedentary species (p-value=0.015). On the y-axes, F_{ST} was log linearized using formula $\log((F_{ST}+0.001)/(1-(F_{ST}+0.001))$.



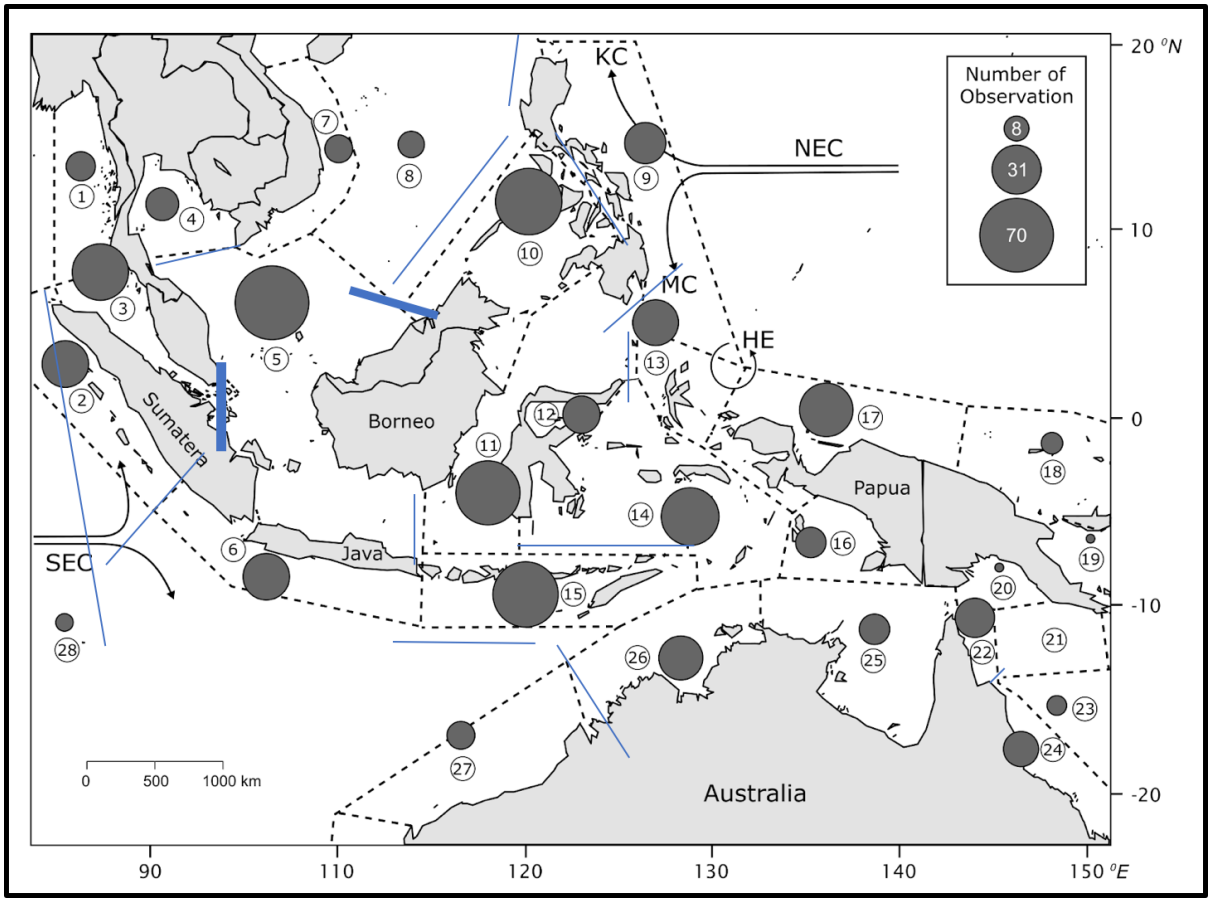
Appendix Figure S4. Relationship between genetic structure (cluster) and geographic (over-water) distance using Subset 5, sessile species (p-value<0.05), as the distance was shown to be the most important continuous predictor in sessile species (Figure 2).



Appendix Figure S5. Relationship between genetic structure (cluster) and ecoregion using Subset 3 (p-value=0.001), as ecoregion was shown to be the most important continuous predictor in free swimming species.

Ecoregion	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	1	4	2	6	8	5	5	3	1	7	7	3	0	5	5	1	3	0	0	0	0	0	1	1	2	0	1	1
2			9	4	22	6	3	3	4	17	22	7	12	16	23	3	20	2	0	0	0	3	2	4	2	1	2	2
3				5	33	15	3	5	4	17	21	4	10	13	16	1	14	3	0	0	0	0	0	2	1	1	1	2
4		1	1		9	6	5	3	1	6	5	1	2	4	5	3	1	0	1	1	0	3	0	1	1	4	0	0
5	1	2	7	1	2	24	5	4	10	37	40	16	21	35	44	8	31	5	0	0	0	5	1	6	5	5	1	1
6	1	1	2			1	5	4	3	11	21	5	13	18	18	4	13	4	1	0	0	4	1	3	3	3	1	1
7	1							3	1	6	4	0	2	2	4	2	0	0	1	1	0	3	1	1	2	2	1	1
8		1	1						4	8	6	1	1	2	2	0	1	0	0	0	0	0	0	0	0	0	0	0
9		1	1		1			1		16	11	6	5	6	10	0	6	2	0	0	0	2	2	3	1	0	1	1
10		1	4	1	4	1		1	1	2	33	11	12	25	29	8	18	2	0	0	0	4	3	3	1	2	2	1
11	1	1	1									17	23	35	41	7	33	5	0	0	0	3	1	3	2	2	2	1
12					1					1		1	5	17	17	2	13	5	0	0	0	2	1	2	0	0	0	0
13		1	1		1	1		1	1	1	1			18	24	8	21	5	1	1	0	3	1	2	0	2	0	0
14			1		2	1				1				1	39	8	28	5	0	0	0	3	1	4	2	3	2	1
15		2	3	1	1					2			1	1		9	32	5	0	1	0	8	2	6	3	6	3	3
16																	4	0	1	1	0	2	0	0	0	2	0	0
17		1	1		1	1		1	1	1	1				1			5	0	0	0	2	2	4	1	1	2	2
18																			0	0	0	2	1	3	0	0	0	0
19																				0	0	1	0	0	0	1	0	0
20																					0	1	0	0	0	1	0	0
21																						0	0	0	0	0	0	0
22					1	1				1					1								2	8	3	14	1	0
23										1												1		3	1	2	4	2
24			1		1	1									1										8	8	3	3
25	1				2	2	1				1															8	2	2
26										1												2	1	1			6	1
27																	1					3	2	1	1	2		3
28			1												1		1						1	2				

Appendix Figure S6. Pairwise matrix indicating pairwise comparisons between sites for the number of observations (top triangle) and the count of the number of barriers identified between ecoregions for free-swimming taxa (bottom triangle). Total number of barriers identified were 113. Yellow squares with counts indicate where more than one genetic cluster was identified within a single ecoregion.



Appendix Figure S7: The spatial location of the barriers identified for free-swimming taxa, thicker lines indicate more observations with these barriers.