Title: Emergent Patterns of Population Genetic Structure for a Coral Reef Community

Running title: Coral reef community genetics

Author list with address numbers:

Kimberly A. Selkoe [1,2]
Oscar E. Gaggiotti[3],
ToBo Laboratory[4],
Brian W. Bowen[1],
Rob J. Toonen[1]

Address List:

1. Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, Hawai'i 97644, USA
2. National Center for Ecological Analysis and Synthesis, 735 State St., Santa Barbara, California 93101, USA
3. School of Biology, Scottish Oceans Institute, University of St Andrews, St Andrews, Fife KY16 8LB, UK
4. Author names and affiliations listed in Appendix 1.

Abstract

What shapes variation in genetic structure within a community of co-distributed species is a central but difficult question for the field of population genetics. With a focus on the isolated coral reef ecosystem of the Hawaiian Archipelago, we assessed how life history traits influence population genetic structure for 35 reef animals. Despite the archipelago's stepping stone configuration, isolation by distance was the least common type of genetic structure, and regional structuring (i.e., division of sites into genetically and spatially distinct regions) was most common, detected in 20 of surveyed species, and nearly all endemics and habitat specialists. IBD only occurred in four shallow, non-endemic invertebrates. Seven species displayed chaotic (spatially unordered) structuring, and all were non-endemic generalist species. Chaotic structure also associated with relatively high global F_{ST}. Pelagic larval duration (PLD) was not a strong predictor of variation in population structure (R^2= 0.22), but accounting for higher F_{ST} values of chaotic and invertebrate species, compared to regional structuring and fish species, doubled the power of PLD to explain variation in global F_{ST} (adjusted R^2=0.50). Multivariate correlation of eight species traits to six genetic traits highlighted dispersal ability and habitat specialization as strongest influences on genetics, but otherwise left much variation in genetic traits unexplained. Considering that the study design controlled for many sampling and geographical factors, the extreme interspecific variation in spatial genetic patterns observed for Hawai'i marine species may be generated by demographic variability due to species-specific abundance and migration patterns and/or seascape and historical factors.

KEYWORDS: Community genetics, stepping stone dispersal, chaotic genetic heterogeneity, Hawai'i , pelagic larval duration, marine connectivity

INTRODUCTION

The structuring of species into genetically distinct populations has many impacts on a species' demography and evolution (Kokko & López-Sepulcre 2007). In turn, ecological and environmental factors influence population genetic structuring (Avise 2000; Storfer *et al.* 2007). Understanding linkages between ecological, genetic and environmental patterns is central to many current challenges in organismal biology and conservation (Taberlet et al. 2012). Uncovering generalities about these linkages requires comparison across multiple species, habitats and scales. Meta-analyses can test for meaningful relationships between genetic structuring and ecological traits across many species, but are hindered by the large number of possible confounding variables. In fact, an early finding of the rapidly expanding field of landscape genetics is that genetic structuring is highly species specific, influenced by the individual's interaction with landscape features according to life history and demographic factors, such that generalities may be few (Manel *et al.* 2003). Marine systems are known for harboring diverse and often surprising spatial population genetic patterns (Selkoe *et al.* 2008). Here we characterize the variation in population genetic structure across species within a single marine community, which share a basic habitat array, environmental gradients and key study sampling design elements. Further, we examine whether life history traits associate with genetic patterns, perhaps pointing to mechanisms maintaining the diversity in genetic patterns across species.

There is great interest in determining what drives spatial patterns of population genetics for marine species, and the extent to which life history traits associate with particular types of structuring. Theory suggests that the scale and pattern of genetic structure reflects long-term rates of gene flow driven primarily by migration, drift and selection. A longstanding focus of the field of population genetics is the relationship of dispersal potential to gene flow, because dispersal is a difficult trait to study directly but central to many basic and applied questions in ecology. Across studies of marine species, dispersal traits show significant correlation to genetic structuring, albeit often only weakly (Bradbury *et al.* 2008; Weersing & Toonen 2009; Riginos *et al.* 2011; Selkoe & Toonen 2011; Faurby & Barber 2012), leaving open the question of whether the remaining variation in genetic differentiation between populations may be explainable by factors such as taxonomy, life history, sampling design or historical effects. Despite hundreds of single-species marine population genetics studies across the globe, it is still unclear whether stronger or more coherent links between genetic and life history traits might emerge if variables such as history, habitat array or taxonomy could be constrained.

Two basic categories of population genetic structure are historically recognized, the island model (discrete structuring in which individuals exist in genetically homogeneous "islands" with limited gene flow between them) and the stepping stone model (continuously increasing differentiation along spatial gradients) (Wright 1943). A third possibility is extensive dispersal and low genetic drift, whereby genetic differentiation is statistically insignificant and the entire geographic domain is genetically homogeneous. This is particularly likely in the ocean, where populations can be very large and migration is favored by the high-dispersal medium (e.g. Theisen et al. 2008). A fourth model has emerged out of empirical marine genetics, "chaotic" population structuring (Johnson & Black 1982; Hedgecock & Pudovkin 2011) in which the level of genetic structuring has been found to be highly variable with no obvious spatial patterning, possibly indicating non-equilibrium conditions, sweepstakes recruitment (Hedgecock 1994, Arnoud-Haond et al. 2008), drift-dominated structuring (Johnson & Black 1982; Puritz & Toonen 2011; Broquet & Yearsley 2012; Yearsley *et al.* 2013), unaccounted for seascape drivers or selection (Baums *et al.* 2006; White *et al.* 2010; Galindo *et al.* 2010; Selkoe *et al.* 2010; Foster *et al*. 2012), or some combination of all these factors (Toonen & Grosberg 2011). The relative frequencies of these four types of genetic structuring and their main drivers are unknown for marine ecosystems.

The present study leverages recent genetic studies of Hawaiian coral reef species to examine the range of population genetic patterns across reef animals within a single community and investigate whether species traits co-vary with metrics and models of genetic structure. The geography of the Hawaiian Archipelago provides an especially tractable system in which to study marine population structure, because it is the remotest archipelago in the world, composed of a nearly linear 2,400 km long array of discrete habitat patches of islands, atolls and seamounts. Insofar as possible in a natural system, these factors constrain the patterns of population structure, connectivity across patches, and genetic history of populations (Fig. 1). We began by categorizing datasets for 37 diverse coral reef species sampled with standard genetic markers at >5 islands. Based on habitat array, we hypothesized that a stepping stone pattern of dispersal producing an isolation by distance pattern of spatial genetic structure would be prevalent. There are no obvious known physical barriers or strong oceanographic discontinuities that might lead to hierarchical genetic structuring. However, a precursor to this study found that locations of significant pairwise F_{ST} were highly variable across 27 species in Hawai'i, but occurred most commonly in the main Hawaiian Islands where islands are more closely spaced (Toonen *et al.* 2011).

Riginos *et al.* (2011) outlined two approaches to studying life history effects on population structure: planned multispecies comparisons using a common sampling regime and geography, and post-hoc compilation of published studies. This study represents a hybrid, in which compilation of raw genetic datasets for co-distributed species within a single discrete study region allowed a large degree of standardization. One would expect that by controlling for basic habitat array, environmental gradients and many shared historical influences, much of the noise in this relationship might be eliminated. In this way, more nuanced multivariate influences on gene flow could emerge, enhancing our understanding of the feedbacks between life history, ecology and genetics.

By comparing life history and genetic data across a broad taxonomic range of species, we hope to gain insights into the mechanisms that drive geographic population structuring in marine systems. We characterize variation in population structure across species in two ways. First, we constrain the question by theory, evaluating how the established models of genetic structuring are represented by the 39 species. Second, we use unconstrained ordination to determine natural divisions among the datasets based on a suite of genetic metrics of spatial structure. These two approaches are complimentary in that the first is using significance testing of how spatial distributions of genetic diversity fit with models, whereas the second is based on combinations of the metrics themselves, perhaps revealing additional divisions in the database which may not map well to the theory-based categories (i.e., if the values of the genetic metrics in each category show different ranges or variances).

Next we test whether taxonomy, life history or sampling effects contribute to the observed variation in genetic structure. Using canonical analysis we estimate how much variation in genetic metrics across species can be explained by available life history and taxonomic traits. This broad-brush assessment is followed by alternative model testing of relationships between particular genetic and life history traits, to get insight into mechanisms driving structuring in this system. Previous empirical studies comparing large numbers of marine genetic datasets have reported that pelagic larval duration (PLD) shows positive correlation with genetic differentiation (Weersing & Toonen 2009; Selkoe & Toonen 2011) and that body size and depth preference show negative correlations with genetic differentiation (Bradbury *et al.* 2008; Kelly & Palumbi 2010; Riginos *et al.* 2011). We test each of these relationships here, and separately examine species with strong vs. weak genetic structuring, and Hawaiian endemics vs. widespread species, as endemism occurs at a high rate in Hawaii and could be associated with distinct genetic characteristics.

Methods

1. Dataset preparation format

Data sets were assembled primarily from collections made on NOAA expeditions throughout the Hawaiian archipelago from 2005 to 2012, and subsequent publications of mtDNA and nuclear DNA (usually microsatellite) data sets. Genetic datasets were contributed in ARLEQUIN format (vers. 3.5.1.2, Excoffier & Lischer 2010), or an Excel format that allowed easy conversion to ARLEQUIN format. A modified version of PGDSpider (vers. 2.0.5.1, Lischer & Excoffier 2012) was used to convert between file formats for genetic analyses. For coral species only, GENETIX (vers. 4.05.2, Belkhir *et al.* 2002) was used to estimate and filter out clonal replicates within sites. ARLEQUIN files were modified to give all sites standardized four letter name codes and standardized ordering from SE to NW along the island chain. Because we have data at the scale of the island/atoll, we focus hypotheses at this spatial resolution. In most cases, allele/haplotype frequencies at adjacent islands are statically indistinguishable, indicating that island/atoll is an appropriate spatial scale for our study.

Any distinct sub-island or sub-atoll samples were kept separate, with distinct names, except when F_{ST} was statistically indistinguishable from zero, in which case sub-localities were lumped. Several species showed samples collected in the vicinity of Kona to be distinct from those near Hilo on Hawai'i Island, and *Acanthaster planci* showed two distinct populations at Pearl & Hermes Atoll.

2. Sampling filters

For inclusion in the analyses, a dataset required at least 5 sites sampled with at least 10 individuals per site. For datasets that meet these criteria, sites with fewer than 10 individuals were also excluded. We also analyzed results for a sample size minimum of because F_{ST} can be inflated at small sample size, and allele frequency estimates are less reliable for low frequency alleles at highly polymorphic loci. Using a minimum sample size of 20 individuals per site excluded 90 of 533 samples in the dataset using 10 or more samples per site (17% of samples). We comment below on how the two sampling filters affect results.

3. Summary statistics

Nuclear loci with significant deviation from Hardy-Weinberg equilibrium were excluded before analysis. GENODIVE (vers. 2.0b23, Meirmans & van Tienderen 2004) was used to calculate estimates of global and pairwise F_{ST} based on Weir and Cockerham’s (1984) θ, with AMOVA using 9999 permutations. ARLEQUIN was used to calculate AMOVA based on φ_{ST}, using AIC criteria from jModelTest (vers. 2.1.4, Darriba *et al.*

2012) to choose the most appropriate mutational model in ARLEQUIN. SMOGD online calculator (Crawford 2010) was used to calculate D_{EST} and effective alleles. GENODIVE's K-means clustering was run for number of clusters (K) from 1 to N-2 using AMOVA based simulated annealing with 50,000 steps and 20 repeats. Cluster membership was examined to determine whether adjacent sampling sites clustered together, highlighting where genetic boundaries (i.e., genetic discontinuities) between regions might exist. Genetic boundaries were considered where AMOVA estimation of F_{CT} across the boundary was statistically significant. The largest number of clusters of spatially discrete samples that returned significant F_{CT} results with AMOVA was recorded. In most cases, this was K=2 or 3. In some cases, K-means clustering showed slightly spatially-mixed clustering. For example if Midway, a Northwestern Hawaiian Islands (NWHI) site, grouped with the Main Hawaiian Islands (MHI) but otherwise MHI and NWHI sites were in two distinct clusters, a "spatially strict" versions of the clusters (e.g., Midway was placed in the NWHI cluster) were tested with AMOVA to confirm that F_{CT} values were significant after the regrouping. This procedure was only used when 1-2 samples were geographically incongruent in the clustering results. Clusters made up of spatially mixed samples were considered evidence that genetic structuring was not regionally organized. Pairwise geographic distance between sites based on coordinates were generated using GENODIVE. Isolation by distance analyses were generated using linearized F_{ST} [$F_{ST}/(1-F_{ST})$] vs. Euclidean distance. Significance testing was based on Mantel tests with 999 replicates performed in GENODIVE.

Nine species were represented by 2 datasets, one using a mtDNA marker and a second using one or more nuclear markers (e.g., microsatellite panels or nuclear intron sequence). Genetic summary statistics were calculated for each marker class independently and then compared to gauge congruence. The mtDNA dataset was preferentially chosen to represent the species in subsequent ordination analyses (which required one dataset per species to avoid double counting), except where sampling power of the nuclear dataset was superior, see results for details.

4. Categories of spatial genetic structure

Based on the above summary statistics, datasets were placed into the following categories of spatial genetic structuring, summarized in Table 1:

(1) Panmixia -- defined as a lack of spatial genetic structuring, indicated here when global F_{ST}, φ_{ST} and D_{EST} p>0.05, spatial groupings based on K-means clustering show F_{CT} p>0.05, and IBD testing shows Mantel r p>0.05.

(2) Chaotic genetic heterogeneity -- defined as genetic differentiation of samples with no apparent spatial organization, indicated here when global F_{ST}, φ_{ST} and/or D_{EST} p<0.05, but neither IBD nor any spatial clustering are statistically significant.

(3) IBD -- a significant IBD Mantel correlation (p<0.05) without significant spatial clustering, or within clusters, indicates auto-correlated spatial variation, regardless of the global tests of differentiation.

(4) Regional genetic structure -- when K-means clustering identified groupings of adjacent populations with F_{CT} p<0.05, regardless of IBD, and global differentiation.

It is possible that a species could conform to more than one category in different regions of the archipelago. Most datasets could not be properly evaluated for this possibility due to limited sampling in remote portions of the archipelago. However, for every case of regional structure, we tested for the joint presence of IBD and regional groups. As illustrated by Meirmans (2012), IBD and hierarchical structure can be confounded. Hierarchical structure can mimic IBD when differentiation with regions is low and distant pairs are cross-regional comparisons, whereas IBD spatial autocorrelation can mimic hierarchical population structure if sampling is sparse and uneven. These scenarios were distinguished (albeit with low power in our case) using stratified Mantel tests in GENODIVE to permute the locations of populations within the clusters.

5. Clustering datasets by genetic summary statistics

The above categorizations are based on labeling datasets according to their fit with existing models of genetic structure derived from genetic theory. The designations are based on the statistical significance at alpha=0.05 of a small number of genetic metrics. This approach ignores possibly useful information contained in the continuous range of values of the metrics themselves. It is also sensitive to sample size, which influences statistical significance. As an unconstrained alternative, we conducted a principle components analysis (PCA) with JMP ver. 10 (SAS). These included all genetic summary statistics (F_{ST}, φ_{ST}, D_{EST}, F_{CT}, IBD r, and the number of genetic regions; Table S1) to find natural divisions in the datasets which are unconstrained by any pre-existing labels or theory. Genetic metrics were linearized and log transformed to homogenize scales prior to all analyses. Negative values of F_{ST}, φ_{ST} and D_{EST} were set to zero to avoid a confounding influence on ordinations. PCA allowed us to visualize the main trends in summary statistics, ascertain redundancy in summary statistics and visualize natural breaks or clusters of datasets by genetic traits.

6. Life History data

Published literature and FishBase were searched for each species to gather basic life history data (Table 2). Any life history traits available for a great majority of species were included, producing nine variables in the initial analyses. Estimates of mean PLD were available in the literature for 32 of the 37 species. To fill in missing values, a mean based on congenerics (n=16) was used for the two *Chaetodon* spp. lacking PLD data and a mean based on confamilials (n=7) was used for the two groupers (family Serranidae). There is little information on tropical subtidal hermit crab PLD. As they typically go through 4-6 larval stages lasting a few days at least (Lang and Young 1977), we estimated the mean PLD to be 50. The log transformation minimizes the effects of imprecise large values, and this one point has little leverage on the linear fit. Depth range (in m), maximum total length (body or colony size in cm) and estimates of generation time (in years) were available for all species and used on both a continuous scale and $\log_{10}$ transformed. Species were divided into habitat specialists and generalists. Generalists utilize sand, rubble or reef whereas specialists are restricted to, or limited by, specific habitat features which may have small total area and distinct spatial arrays of habitat that differ greatly from the array of shallow habitat across the archipelago (e.g., damselfish requiring nesting sites, hermit crabs sheltering in certain corals, limpets limited to basalt which is patchy or absent at islands , corallivores requiring live coral). Five basic trophic categories were designated: corallivore, detritivore/sediment, invertivore, piscivore, algivore and planktivore, but analyses collapsed these into a binary categorization (algivore and planktivore vs. others) given the sample size of the dataset. Other binary categorizations were examined: predator (invertivore and piscivore vs. others), and benthic feeders (corallivores, detritivore/sediments, invertivores and algivores vs. piscivores and planktivores) but provided no further insights to the analyses. Remaining life history categorizations were endemic to Hawai'i vs. non-endemic, and free-floating eggs vs. attached to body or substrate. Higher taxonomic affiliation was also used as a categorical variable (fish vs. invertebrate and dolphin) representing fundamental but unspecified generalities that may tend to be shared across these highly diverse species, such as adult mobility, mutation rates or mating systems. A PCA using the 4 continuous variables and the 5 binary variables allowed us to assess the variation in the life history traits across species and visualize colinearities between life history traits, which were then confirmed with univariate linear regression or t-tests.

7. Redundancy analysis

Canonical analysis was used to assess how much the suite of life history traits explains the variation in genetic traits as a whole, and to visualize which traits most closely associate (Legendre & Legendre 2012). Redundancy analysis (RDA) is an ordination with regression; we used the package VEGAN in R for calculations. The genetic metrics (Y) are first transformed to Y' by fitting the values to a linear regression of each life history trait (X). A PCA is then carried out on the Y' values. Colinearity of life history traits was examined before proceeding, leading to the elimination of generation time, which was correlated to maximum length but generally measured with much less precision (OLS r=0.71). The genetic summary statistics used were the same as described above for PCA (Table S1). F_{ST} was used in place of φ_{ST} for the 4 nuclear marker datasets as missing data is not allowed in the analysis. F_{ST} and φ_{ST} were correlated (OLS r=0.71) but both were included in the analysis to reveal differences in their responses to species traits, as was our primary goal for the RDA instead of statistical hypothesis testing. All other genetic traits showed low colinearity. Two datasets with outlier F_{ST} values (F_{ST}>0.2) were removed (*Cellana exerata* and *Chaetodon lunulatus*) because outliers have disproportional influence on ordinations. We examined sampling factors as covariates in the analysis by performing a partial RDA with all factors (alleles, marker type, number of sites sampled, recent arrival species), but no effects were found. Adjusted R^2 was calculated following the Ezekiel method (Legendre & Legendre 2012). The RDA triplot provided guidance on where to concentrate tests of particular associations of life history and genetic traits (i.e., it showed which traits have the strongest associations) to avoid large ratio of alternative models to sample size (Burnham & Anderson 2002).

8. Linear models of genetic differentiation

Based on the RDA results, correlations of several genetic and species traits were examined. Univariate correlations of continuous variables were made using Ordinary Least Squares (OLS) regression, and t-tests were used for assessing significant association of genetic traits with categorical variables. Multivariate explanatory models combining categorical and continuous variables were made with generalized linear models (GLM) using normal distribution and identity link function in the software program JMP. Akaike's information criterion (AIC_c) was used to select the most parsimonious models. The same two datasets were removed (*C. exerata* and *C. lunulatus*) because they were extreme outliers (i.e., their F_{ST} values were more than twice the value of the next highest values).

RESULTS

1. Genetic Categorizations

Datasets were divided into all four possible categories of genetic structuring: regional, IBD, chaotic and panmictic (Fig. 2, Table S1). Among species sampled with two marker types, five of the nine showed congruent categorization of both datasets. The remaining four species all had one dataset with a low number of alleles that showed panmixia, and the other dataset had a high number of alleles that showed structuring. The association of low polymorphism with panmixia was the most prominent sampling factor associated with placement of a dataset into one of the four categories (Table S2). To avoid double counting in analyses requiring one dataset per species, we selected the dataset with the more polymorphic marker(s) because of its greater statistical power. For the congruent pairs, the mtDNA datasets were preferentially selected to increase consistency in marker type across datasets.

Regional grouping was the most common type of spatial genetic structure, observed in 20 of the 37 species. Eight species showed support for two spatial regions, eight for three regions, three for four regions and one for five regions (Table S1). In some cases, "regions" comprised only one sample separated from others by a significant genetic break. In all but three of these species, hierarchical AMOVA showed no evidence of significant finer scale structuring within regions (i.e., significant F_{SC} values; exceptions were *Stenella longirostris* and *Montipora capitata*). Although ten of the 20 regionally structured species showed significant overall IBD results (uncorrected p<0.05), none showed a significant stratified Mantel test, which would indicate IBD within regions. Thus, these IBD signals are likely an artifact of the regional structuring (the increased mean pairwise F_{ST} across regions compared to within regions), although for a minority, the stratified Mantel test may have lacked power to detect a true within-region IBD signal. Interestingly, twelve of the 20 regionally structured species showed global F_{ST} values not significantly different from zero. Add assessment of sampling gaps contributing to significance of IBD results.

Only four species were categorized as IBD, because they showed significant IBD without any regional structuring. One of these, *Acanthurus nigrofuscus*, had a very weak signal (p=0.05) and most pairwise F_{ST} <0. The other three species classified as IBD datasets were invertebrates: a sea star (*Acanthaster planci*), a coral (*Porites lobata*) and a brittlestar (*Ophiocoma pica*).

Seven species categorized as chaotic showed highly significant global differentiation among sample sites and many significant pairwise F_{ST} values, but with no obvious spatial organization. However, one of these species, the brittlestar *Ophiocoma erinaceus*, showed a nearly significant IBD test (r=0.42, p=0.07) and nearly significant test for 2 regions (F_{CT}=0.049, p=0.15) that might have gained significance with more specimens. Two of the chaotic datasets yielded F_{ST} with p>0.05, but D_{EST} and/or φ_{ST} were highly significant.

The remaining six species were panmictic, with nonsignificant and very low global F_{ST}, φ_{ST} and D_{EST} values. Three of these had low allele counts such that their results may be considered inconclusive (*Acanthurus olivaceus*, *Heterocentrotus mammilatus* and *Chaetodon multicinctus*).

Changing minimum sample size from 10 or more individuals per location to 20 or more affected the categorization of only 1 dataset (*Chaetodon miliaris* lost four sites and switched from panmictic to regionally structured). Several other species lost enough sites to be excluded from analysis.

2. Ordinations of genetic traits and life history traits

The six genetic summary statistics (F_{ST}, φ_{ST}, D_{EST}, F_{CT}, number of significant regions, IBD fit) showed only moderate to low colinearity. The most correlated values were φ_{ST} and F_{ST} (OLS r=0.71). D_{EST} was uncorrelated with φ_{ST} and F_{ST}. A PCA using these six genetic summary statistics showed two datasets (limpet *C. exerata* and butterflyfish *C. lunulatus*) to be outliers to the rest because their values of φ_{ST}, F_{ST} D_{EST} and F_{CT} were much larger than the others (e.g., F_{ST}>0.2 vs. <0.09, Table S1). These two datasets were removed from all further analyses and the PCA was repeated to lessen the influence of skew on the analysis. The first 4 PCs showed eigenvalues>1 (Fig. 3a). The first PC, which showed high loadings for both φ_{ST} and F_{ST}, separated out six datasets for which differentiation among sites is largest (e.g., F_{ST}>0.02; red markers in Fig. 3a). PC2 separated most of the datasets with 1 region and high F_{CT} values (purple in Fig. 3a) from those with multiple regions and low F_{CT} values (blue in Fig. 3a). PC3 was correlated with IBD r and PC4 with D_{EST}. To show how the four categories of genetic structuring map to the PCA results, the biplot is recolored in Fig. 3b; it indicates that chaotic and IBD spatial organization are not clustered into a small range of values of F_{ST} or φ_{ST}.

A PCA of the life-history traits show that species have diverse combinations of traits instead of a few clusters of associated types (see Fig. S1 for biplot and more detail). Pairwise linear regressions testing revealed two notable apparent correlations among life history traits. As previously known, Generation Time and Maximum Length positively associate (OLS R^2=0.52, p<0.0001). We excluded Generation Time from the RDA analysis due to colinearity. Also, fishes showed significantly broader depth ranges than invertebrates (R^2=0.30, p<0.0001) but both were retained in the RDA as correlation was weak. The biplot shows that the genetic types contain a mixture of life history traits, but some tend to be absent from certain quadrants: for PC1 v PC2, chaotic datasets tend to be in the upper left (all were non-endemic and habitat generalists), panmictic datasets in the upper half (broad and deep depth ranges and mostly fish), and IBD datasets in the lower half (shallow and invertebrate), whereas regional species are widely distributed over the plot.

4. Redundancy analysis of life history and genetic traits

The multivariate linear relationship between eight life history and six genetic traits was significant but not strong (R^2=0.35, adj. R^2=0.11, p=0.037). When φ_{ST} is excluded to reduce redundancy with F_{ST}, R^2 is unchanged but p=0.055. In both cases, all eigenvectors were <1 indicating lack of principle component interpretability. The triplot is nonetheless useful for visualizing which life history traits associate with genetic traits (Fig. 4). Three traits (PLD, Fish, Habitat) are located near the edges indicating strongest explanatory power. A reduced model with only these 3 traits raises the adjusted R^2 (R^2=0.22, adj. R^2=0.15, p=0.001). Three genetic traits (Regions, D_{EST} and F_{CT}) sit close to the center of the ordination indicating that they are poorly explained by the life history traits. F_{ST} aligns strongly with the PLD vector, D_{EST} weakly with the PLD vector, and φ_{ST} is more influenced by Fish and Endemic than are F_{ST} and D_{EST}.

The vector for IBD is opposite the trajectory of Fish, Endemic and Depth range, indicating negative relationships of these traits to stepping stone dispersal. As in Fig. 3a, the plot separates the species with strong differentiation on the right side, away from the majority of other datasets. These species are a shark, a dolphin, a sea cucumber, a coral, a sea star and a brittlestar, a group encompassing all 3 types of spatial genetic structuring.

5. Linear modeling of genetic traits

The RDA indicates that Fish, Endemic and Depth have strong negative impacts on IBD r. A comparison of multivariate GLM models for these three traits and their interactions showed the most parsimonious model of IBD r includes Endemic and Depth only (adj. R^2=0.22, p=0.006); shallow, non-endemic invertebrates show stronger IBD patterns. Maximum depth, not minimum depth drives the correlation with depth range.

A linear fit of PLD vs. F_{ST} was highly significant (OLS R^2=0.45, p<0.0001, Fig. 5), but when the two species (shark and dolphin) that lack larval development are excluded, the fit drops (R^2=0.19, p=0.012). The fit strengthens slightly (R^2=0.22) without the 6 panmictic datasets, for which F_{ST} is less informative because it is measured with larger error and is likely an artefact of low marker polymorphism for many of the datasets (Table S2)..

For this subset of 27 non-panmictic species with PLD>0, a comparison of multivariate GLM models based on AIC_c showed that a model with Fish and PLD is more parsimonious than a model of PLD alone (adj. R^2=0.31, p=0.044, ΔAIC_c=2.5, Table 4). Adding Fish improved the fit because invertebrates have a higher intercept than fishes due to their generally higher F_{ST} values. Adding in the other 8 species traits as additional factors does not improve the model (only individual additions were tested to minimize number of models compared). However, adding an indicator of whether the dataset shows spatially organized structure (regional or IBD) or disorganized structure (chaotic) as a covariate improves the model significantly (adj. R^2=0.50, p<0.0001, ΔAIC_c=6.8, Table 4). The model improvement occurs because chaotic datasets have a higher F_{ST} values on average and thus a higher intercept for PLD vs. F_{ST}. Categorization of datasets as regional or chaotic was made based on the p value of F_{CT} in a hierarchical AMOVA, which shows no correlation with F_{ST} (R^2=0, p=1.0). With the two direct developers (PLD=0) included, the fit of this model is boosted (adj. R^2=0.65). Interactions were not significant and thus excluded from the best fit model.

The 10 endemics show no relationship between PLD and F_{ST}, although all endemic PLD values are fairly large (PLD>23 days). No species traits significantly explain F_{ST} for the endemics. Excluding endemics from the non-panmictic, PLD>0 group, results in the same best fit model, but with higher explanatory power (adj. R^2=0.58, Table 4).

We examined how GLM models explain variation in φ_{ST} for the mtDNA datasets compared to those for F_{ST}. Fish and structure type (i.e., spatially organized vs. disorganized) without PLD best explained φ_{ST} for the PLD>0 non-panmictic set of species (adj. R^2=0.55, p<0.002). φ_{ST} also shows no relationship to PLD for endemics, but shows a highly significant positive relationship to both maximum length and herbivory for endemics (adj. R^2=0.66, p=0.009). D_{EST}, F_{CT} and regions showed no significant linear relationships to the species traits collected for all species combined, as indicated by the RDA.

Discussion

It is well known that marine species exhibit extensive variation in their genetic patterns that is poorly predicted by ecological or species traits. By focusing on a single isolated region with a simplistic habitat array and calculating genetic metrics in a standardized way from raw data of many species, we investigated the extent to which variation in genetic patterns is constrained across species using two complimentary approaches. The dual approaches help illuminate the extent to which our assessment is sensitive to the chosen metrics, categories, and statistical framework. First, the PCA analysis focused on the strength of genetic differentiation and the spatial scale of structuring (i.e., number of regions), and revealed 3 clusters within the species set: single region/low differentiation, multi-regional/moderate differentiation and a small number of high differentiation species with a mix of single and multiple regions. Second, the categorization of the datasets into four *a priori* types focused primarily on the spatial organization of the structuring and not the strength of differentiation (i.e., regional, IBD, chaotic and panmictic). Comparison of these two approaches revealed that species with strongest structuring show diverse spatial organization of structuring and likely a diversity of causes for that high structuring. Life history analyses revealed that chaotic species were all non-endemic and habitat generalists, IBD occurred for four shallow invertebrate habitat generalists, regionally structured species showed a variety of life history associations, and panmixia was mostly limited to fishes with broad and deep depth ranges or associated with low allelic diversity indicating low statistical power. Polymorphism creates precision much the way increasing the number of samples would (Kalinowski 2002).

Regional Boundaries across the Archipelago

The finding that regional structuring was most common, and IBD least common, was surprising given the stepping stone habitat array. Although every inter-island channel along the chain was a possible boundary for at least 1 dataset in the study, the most frequent site of a regional boundary, shared by 13 of the 20 regionally structured species, occurred at the center of the archipelago, in the vicinity of French Frigate Shoals. This trend could lend insight into the factors enabling regional structuring for a diversity of taxa in this system. First, it might be possible to produce such a boundary in a stepping stone dispersal system with finite ends, because this would concentrate genetic differences on either end, especially when gene flow is relatively high, elevating the importance of the increased drift at the edges (Rousset 1994). However, the stratified Mantel test results indicate that this scenario is unlikely, because IBD within regions is rare. Second, there may be an oceanographic divergence zone at the center of the chain. Larval dispersal might be biased away from the center due to the eastern flowing Subtropical Countercurrent splitting as it encounters the archipelago, combined with the westerly North Hawai'i Ridge Current which may drive larvae westward (Fig. 1; Qiu et al. 1997; Kobayashi 2006). However, complex eddying indicated by meso-scale circulation modeling and simulated larval dispersal results suggest this scenario may also be overly simplistic and unlikely (Kobayashi 2006; Rivera *et al.* 2011). Finally, heterogeneity in demographic processes might drive departure from an IBD pattern, perhaps due to differences in habitat area or effective population size (N_e) among locations. The sampling gaps in the datasets are a key consideration in this context. Despite a bias toward selecting study species that are abundant and easy to sample, many of the sampling gaps were caused by absence or very low density of organisms at sites. Thus, uneven abundance or density distributions across islands may lead to hierarchical structuring despite stepping stone dispersal for some species. Almost all endemics and all habitat specialists were regionally structured (except a few cases of panmixia). Both groups are more likely to have variable abundance across the chain due to spatially varying micro-habitats, supporting this cause for regional structuring. This phenomenon begs for marine population genetic studies to carefully consider sampling design and run simulations to test effects of sampling factors on results. Additional factors and analysis approaches might add insight to the current results. For example, if ocean currents are important drivers of gene flow, the seasonal timing of larval dispersal, and other larval traits for which we were unable to find data, may help generate variation in genetic structuring across these species. We will explore the relative roles of history, oceanography, sampling and habitat factors in generating the observed variation in genetic patterns across species in future studies.

Statistical considerations of characterizing spatial genetic structure

Our dataset proved to be a good example of the 'trouble with isolation by distance,' described recently by Meirmans (2012). Nine datasets showed significant test results for IBD that on closer examination were driven only by regional structuring, evident both by examining site membership of data points on the IBD plot and by a stratified Mantel test. Our algorithm for categorizing a dataset by its spatial genetic structuring was inspired by this study, and at least in the case of marine species, it appears that understanding whether the dataset is spatially auto-correlated, regionally structured or chaotically structured is an important first step to interpreting population genetic analyses. Meirmans (2012) found that 70% of a sample of studies testing for IBD found it. These were mostly terrestrial or aquatic studies. It is already known that marine species show much lower rates of IBD, and in this study, despite an uncommonly clear-cut stepping stone habitat array, only 10% of species showed IBD. It is unlikely sampling gaps biased the categorization of datasets as IBD vs. regionally structured, as the average number of samples, number of sites in the MHI, NWHI and whole chain, and the size of the largest sampling gap were nearly identical for the two groups (Table S2). However, IBD datasets tended to have slightly larger effective number of alleles than regional datasets (4.1 vs. 3.6; Table S2). Similarly, chaotic and panmictic datasets

Many species showed global estimates of F_{ST} and φ_{ST} near zero despite strong regional structuring. Despite the fact that island-scale differentiation (global F_{ST}) correlated with PLD, many species in the highest PLD category showed 2 or 3 genetically distinct regions, perhaps indicating that regional boundaries are not caused by dispersal related processes and instead may be a product of historical events (Marko 2004) and/or local adaptation. The K-means clustering approach to guide hierarchical AMOVA has not been widely used, but is more sensitive than *a priori* designation of groups. While it has the potential to uncover large-scale structure that is missed by other approaches (e.g., Kelly & Eernisse 2007; Díaz-Ferguson *et al.* 2010), it is also possible that the approach has inflated type 1 error.

Regional, taxonomic and life history variation in the correlation of PLD and FST

We found that accounting for whether a dataset is spatially organized improves insight into the relationship between genetic structure and species traits. It is interesting that the chaotic datasets showed a significant correlation of F_{ST} and PLD, but with higher mean F_{ST} values than species with spatially organized structure. The pattern suggests that these datasets are not chaotic simply because they are out of drift-migration equilibrium, but rather that they have an additional factor inflating differentiation. Consistent with this idea, recent simulation studies indicate that chaotic genetic patchiness can arise via small local effective population size and mildly-aggregated dispersal of kin

(Broquet & Yearsley 2012), which may occur even in species with extremely long pelagic developmental periods (Iacchei *et al.* 2013).

The correlation of PLD and F_{ST} (R^2=0.22 for species with pelagic larvae) was lower than the value for a global sample of studies (R^2~0.30) derived from a variety of spatial scales, habitat configurations, regions and environmental settings (Selkoe & Toonen 2011). Furthermore, our sample of φ_{ST} showed no significant relationship to PLD, instead correlating just with taxon, consistent with φ_{ST} having higher sensitivity to demographic history and mutation than F_{ST} (Bird *et al.* 2009, 2011; Meirmans & Hedrick 2011). For any isolated marine habitat, retention strategies are crucial to persistence, but PLD may be less indicative of realized dispersal distance in this extremely isolated ecosystem compared to other places (but see Schultz & Cowen 1994; Robertson 2001). This possibility is supported by two additional insights. First, endemics, which may be more dependent on larval retention for persistence than widespread species, showed no relationship of F_{ST} to PLD. However, all endemics in the study have PLD>20 days, after which the linear relationship saturates. Thus, retention of larvae may not be highly related to the mean PLD in this setting. Second, comparison of the correlation strength for Hawai'i to that found in other regions shows that the PLD v. F_{ST} correlation is relatively weak in Hawai'i (Fig. 7).

Aside from poor correlation of PLD and dispersal distance, there are many other factors that can decouple PLD and FST values. One recent focus has been the influence of coalescent time on FST, such that holding coalescent time constant should improve PLD vs. FST correlation (Dawson *et al.* 2014). An anlysis using hierarchical approximate Bayesian computation (Hickerson & Meyer 2008; Beaumont 2010) of our dataset indicate nearly uniform timing and rate of expansion following the last glacial maximum (Chan *et al.* in prep). Therefore, we conclude that different coalescent histories are likely not driving variation in genetic patterns of these species. Contrary to claims that comparisons among synchronously diverging co-distributed (SDC) species “consistently evince higher gene flow in species with higher dispersal potential” (Dawson 2014), results from SDC taxa in Hawai'i (and previous global analyses; Selkoe and Toonen 2011) mandate a more nuanced treatment of the many forces impacting the population genetics of marine species.

We found that fishes show less structure and less organized structure than invertebrates. This point has not been previously highlighted, but the pattern is evident in other marine datasets (Carpenter *et al.* 2011; Toonen *et al.* 2011; Selkoe & Toonen 2011). Compared to invertebrates, fishes generally are more capable of behaviors that promote dispersal, and adult and juvenile migration is possible (Eble *et al.* 2011; Poortvliet *et al.* 2013). In addition, we found that species with deeper depth ranges tend to show less structuring than shallow species (also see Etter *et al.* 2005; Kelly & Palumbi 2010; Gaither *et al.* 2012; Andrews *et al.* 2014), perhaps because shallow habitat is smaller in total area, harder for larvae to intercept, and subject to more frequent disturbance, contributing to higher rates of genetic drift. The reason that only shallow invertebrates showed IBD may be due to the double constraint of limited dispersal ability and smaller habitat patch sizes.

In our analyses, egg type, body size, and trophic group showed little influence on genetic traits. However, egg type and body size correlated with F_{ST} in other synthesis studies of marine species (Bradbury *et al.* 2008; Riginos *et al.* 2011). The great variation in findings for ecological correlates with F_{ST} and φ_{ST} indicates that such syntheses are sensitive to the species composition and/or genetic markers in the dataset, as well as the statistical approach. A shift of focus away from linear modeling of global F_{ST}, which is a relatively uninformative metric, toward a deeper understanding of what drives variation in spatial patterns of genetic differentiation will bring new insights to this line of inquiry (Lowe & Allendorf 2010; Marko & Hart 2011a; b).

Multivariate estimates of the covariation of species and genetic traits

Canonical analysis of how 8 species traits associated with genetic traits revealed that PLD, taxonomy (fish vs. invertebrate) and habitat specialization had the strongest influences on F_{ST}, φ_{ST} and IBD fit, but overall explanatory power was low. Our use of RDA to uncover associations of life history and genetic traits followed a similar study of 27 co-distributed high-alpine plants of the European Alps (Meirmans *et al.* 2011). That study used AFLP data to generate 8 genetic summary statistics describing spatial genetic diversity, paired with 6 species traits related to dispersal and habitat preference. The 6 species traits together explained a very similar fraction of variation in genetic traits relative to our finding (R^2=0.30, adj. R^2=0.17). Considering the diversity of ecological, organismal, and historical factors that can impact the distribution of genetic diversity, the authors interpreted this as a large fraction. Our dataset included a wider diversity of species in terms of life history and taxonomy. Interestingly, Meirmans *et al.* (2011) analysis showed the same qualitative main results we report here. First, F_{ST} was the most strongly predicted trait, and was driven by dispersal factors. Plants with multiple dispersal modes showed higher gene flow, similar to our finding that fishes, which can disperse both as adults and larvae, show higher gene flow than invertebrates. Second, Mantel r was the only other strongly predicted genetic metric aside from F_{ST} in both studies, and rather than associating with dispersal factors as would be expected, it was best predicted by habitat factors (soil type for plants, depth range for coral reef species). Historical factors (i.e., size and distribution of refugia) may drive both the depth and soil type effects, and retrospective analyses using coalescent models are needed to distinguish ancient connectedness from contemporary gene flow. A final similarity to Miermans *et al.* (2011) is that Jost's D_{EST} showed no correlation with life history or other genetic traits, thus providing little insight in either context.

Despite the study design to limit sources of natural variance, the species included in this study showed great variation in genetic structure, and species traits were not highly predictive of that variation. Two of the species showed extremely high spatial structuring relative to all others, one regionally structured and one chaotically structured. Their exclusion from the analysis serves only to weaken the link between genetic variation and species traits. In sum, the question of what maintains the extreme diversity in spatial genetic patterns across marine species remains largely unanswered by this study, but is narrowed by the finding that it persists despite controlling for sampling design, marker type, habitat array, major environmental and oceanographic gradients and recent history to a greater extent than possible in meta-analyses of published works. Even within a single reef community, life history of marine species is extremely diverse and likely drives high diversity of demographic and genetic patterns.

Genetic diversity is a crucial foundation for biodiversity, with demonstrated influence on fitness, persistence, species diversity, and ecosystem functioning (reviewed in Hughes *et al.* 2008; Taberlet *et al.* 2012). There is great interest in integrating population genetics into community ecology to understand the forces controlling community assembly and species interactions (Avise 2000; Wares 2002; Cavender-Bares *et al.* 2009). Continuing to characterize the forces controlling spatial genetic structure in more marine and terrestrial communities and regions is an important first step.

Acknowledgements

Funding was provided to K. Selkoe and R. Toonen by the National Science Foundation (BioOCE Award Number 1260169) and the National Oceanic and Atmospheric Administration (NMSP MOA#2005-008/66882). O. Gaggiotti was supported by the Marine Alliance for Science and technology for Scotland (MASTS). We are grateful to Heidi Lischer for modifying the software PGDSpider on our request, to the following for contributing ideas and feedback: Eric Crandall, Mary Donovan, and Alan Friedlander; and to the following for contributing raw data: Giacomo Bernardi, Mathew Craig, Marina Ramon, and Nicholas Whitney.

Author Contributions

KAS, RJT & OG designed research, KAS performed research, KAS analyzed data and KAS, RJT, OG and BB wrote the paper. Tobo Laboratory authors contributed data and analyzed data.

Data Accessibility

Dryad DOI, scripts, datasets

Literature Cited

Arnaud-Haond S, Vonau V, Rouxel C, Bonhomme F, Prou J, Goyard E, Boudry P (2008) Genetic structure at different spatial scales in the pearl oyster (*Pinctada margaritifera cumingii*) in French Polynesian lagoons: beware of sampling strategy and genetic patchiness. *Marine Biology*, **155**, 147–157.

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

Baums ILB, Miller M, Hellberg ME (2006) Geographic variation in clonal structure in a reef-building Caribbean coral, *Acropora palmata*. *Ecological Monographs*, **76**, 503–519.

Beaumont MA (2010) Approximate Bayesian Computation in Evolution and Ecology. *Annual Review of Ecology, Evolution, and Systematics*, **41**, 379–406.

Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F (2002) *GENETIX 4.04, logiciel sous WindowsTM pour la genetique des populations*. Laboratoire Genome, Populations, Interactions, CNRS UMR 500, Universite de Montpellier II, Montpellier, France.

Bird CE, Karl SA, Toonen RJ (2009) Detecting and Measuring Genetic Differentiation . *Phylogeography and Population Genetics in Crustacea*, **19**, 31–55.

Bradbury IR, Laurel B, Snelgrove PVR, Bentzen P, Campana SE (2008) Global patterns in marine dispersal estimates: the influence of geography, taxonomic category and life history. *Proceedings of the Royal Society Series B, Biological Sciences*, **275**, 1803–1809.

Broquet T, Yearsley JM (2012) Genetic drift and collective dispersal can result in chaotic genetic patchiness. *Evolution*, 1660–1675.

Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: a Practical Information-Theoretic Approach*. Springer-Verlag, New York.

Carpenter KE, Barber PH, Crandall ED, Ablan-Lagman MCA, Mahardika GN, Manjaji-Matsumoto BM, Juinio-Meñez MA, Santos MD, Starger CJ, Toha AHA (2011) Comparative Phylogeography of the Coral Triangle and Implications for Marine Management. *Journal of Marine Biology*, **2011**, 1–14.

Cavender-Bares J, Kozak KH, Fine PV a, Kembel SW (2009) The merging of community ecology and phylogenetic biology. *Ecology letters*, **12**, 693–715.

Craig M, Eble J, Bowen B, Robertson D (2007) High genetic connectivity across the Indian and Pacific Oceans in the reef fish *Myripristis berndti* (Holocentridae). *Marine Ecology Progress Series*, **334**, 245–254.

Craig MT, Eble J a., Bowen BW (2010) Origins, ages and population histories: comparative phylogeography of endemic Hawaiian butterflyfishes (genus *Chaetodon*). *Journal of Biogeography*, **37**, 2125–2136.

Crawford N (2010) SMOGD: software for the measurement of genetic diversity. *Molecular ecology resources*, **10**, 556–557.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772–772.

Dawson MN, Hays C, Grosberg RK, Raimondi PT (2014) Dispersal potential and population genetic structure in the marine intertidal of the eastern North Pacific. *Ecological Monographs*, **in press**.

Díaz-Ferguson E, Haney R a, Haney R, Wares JP, Wares J, Silliman BR, Silliman B (2010) Population genetics of a trochid gastropod broadens picture of Caribbean Sea connectivity. *PloS one*, **5**.

Etter RJ, Rex MA, Chase MR, Quattro JM (2005) Population differentiation decreases with depth in deep-sea bivalves. *Evolution; international journal of organic evolution*, **59**, 1479–91.

Excoffier L, Lischer HE (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources*, **10**, 564–7.

Faurby S, Barber PH (2012) Theoretical limits to the correlation between pelagic larval duration and population genetic structure. *Molecular ecology*, **21**, 3419–3432.

Foster NL, Paris CB, Kool JT, Baums IB, Stevens JR, Sanchez JA, Bastidas C, Agudelo C, Bush P, Day O, Ferrari R, Gonzalez P, Gore S, Guppy R, McCartney MA, McCoy C, Mendes J, Srinivasan A, Steiner S, Vermeij MJA, Weil E, Mumby PJ (2012) Connectivity of Caribbean coral populations: complementary insights from empirical and modelled gene flow. *Molecular Ecology*, **21**, 1143-1157.

Galindo HM, Pfeiffer-Herbert AS, McManus MA, Chao Y, Chai F, Palumbi SR (2010) Seascape genetics along a steep cline: using genetic patterns to test predictions of marine larval dispersal. *Molecular ecology*, **19**, 3692–707.

Hedgecock D, Pudovkin AI (2011) Sweepstakes Reproductive Success in Highly Fecund Marine Fish and Shellfish: A Review and Commentary. *Bulletin of Marine Science*, **87**, 971–1002.

Hedrick PW (2005) A standardized genetic differentiation measure. *Evolution*, **59**, 1633–8.

Hickerson MJ, Meyer CP (2008) Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. *BMC evolutionary biology*, **8**, 322.

Hughes AR, Inouye BD, Johnson MTJ, Underwood N, Vellend M (2008) Ecological consequences of genetic diversity. *Ecology letters*, **11**, 609–23.

Johnson MS, Black R (1982) Chaotic genetic patchiness in an intertidal limpet, *Siphonaria sp. Marine Biology*, **70**, 157–164.

Kalinowski ST (2002) How many alleles per locus should be used to estimate genetic distances? *Heredity,* **88**, 62-65.

Kelly RP, Eernisse DJ (2007) Southern hospitality: a latitudinal gradient in gene flow in the marine environment. *Evolution*, **61**, 700–7.

Kelly RP, Palumbi SR (2010) Genetic structure among 50 species of the northeastern Pacific rocky intertidal community. *PloS one*, **5**, e8594.

Kokko H, López-Sepulcre A (2007) The ecogenetic link between demography and evolution: can we bridge the gap between theory and data? *Ecology letters*, **10**, 773–82.

Legendre P, Legendre L (2012) *Numerical Ecology*. Elsevier B.V., Amsterdam.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298–9.

Long (I'll fill this in)

Lowe WH, Allendorf FW (2010) What can genetics tell us about population connectivity? *Molecular ecology*, **19**, 3038–51.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003) Landscape genetics: combining landscape ecology and population genetics. *Trends in Ecology & Evolution*, **18**, 189–197.

Marko PB, Hart MW (2011a) Retrospective coalescent methods and the reconstruction of metapopulation histories in the sea. *Evolutionary Ecology*, **26**, 291–315.

Marko PB, Hart MW (2011b) The complex analytical landscape of gene flow inference. *Trends in ecology & evolution*, **26**, 448–56.

Meirmans PG (2012) The trouble with isolation by distance. *Molecular ecology*, **21**, 2839–46.

Meirmans PG, Goudet J, Gaggiotti OE (2011) Ecology and life history affect different aspects of the population structure of 27 high-alpine plants. *Molecular ecology*, **20**, 3144–55.

Meirmans PG, Hedrick PW (2011) Assessing population structure: F(ST) and related measures. *Molecular ecology resources*, **11**, 5–18.

Poortvliet M, Longo GC, Selkoe KA, Barber PH, White C, Caselle JE, Perez-Matus A, Gaines SD, Bernardi G (2013) Phylogeography of the California sheephead, *Semicossyphus pulcher:* the role of deep reefs as stepping stones and pathways to antitropicality. *Ecology and Evolution*, **3**, 4558-4571.

Puritz JB, Toonen RJ (2011) Coastal pollution limits pelagic larval dispersal. *Nature communications*, **2**, 226.

Qiu B, Koh D, Lumpkin C, Flament P (1997) Existence and formation mechanism of the North Hawaiian Ridge Current. *Journal of Physical Oceanography*, **27**, 431–444.

Riginos C, Douglas KE, Jin Y, Shanahan DF, Treml EA (2011) Effects of geography and life history traits on genetic differentiation in benthic marine fishes. *Ecography*, **34**, 566–575.

Robertson DR (2001) Population maintenance among tropical reef fishes: Inference from small island endemics. *Proc. Natl. Acad. Sci.*, **98**, 5667-5670.

Rousset F (2004) *Genetic Structure and Selection in Subdivided Populations*. Princeton University Press, Princeton, NJ.

Schultz ET, Cowen RK (1994) Recruitment of coral-reef fishes to Bermuda: local retention or long-distance transport? *Marine Ecology Progress Series*, **109**,15-28.

Selkoe KA, Henzler CM, Gaines SD (2008) Seascape genetics and the spatial ecology of marine populations. *Fish and Fisheries*, **9**, 363–377.

Selkoe KA, Toonen RJ (2011) Marine connectivity: a new look at pelagic larval duration and genetic metrics of dispersal. *Marine Ecology Progress Series*, **436**, 291–305.

Selkoe KA, Watson JR, White C, Horin T Ben, Iacchei M, Mitarai S, Siegel DA, Gaines SD, Toonen RJ (2010) Taking the chaos out of genetic patchiness: seascape genetics reveals ecological and oceanographic drivers of genetic patterns in three temperate reef species. *Molecular ecology*, **19**, 3708–26.

Storfer A, Murphy MA, Evans JS, Goldberg CS, Robinson S, Spear SF, Dezzani R, Delmelle E, Vierling L, Waits LP (2007) Putting the “landscape” in landscape genetics. *Heredity*, **98**, 128–42.

Taberlet P, Zimmermann NE, Englisch T, Tribsch A, Holderegger R, Alvarez N, Niklfeld H, Coldea G, Mirek Z, Moilanen A, Ahlmer W, Marsan PA, Bona E, Bovio M, Choler P, Cieślak E, Colli L, Cristea V, Dalmas J-P *et al.* (2012) Genetic diversity in widespread species is not congruent with species richness in alpine plant communities. *Ecology letters*, **15**, 1439–48.

Toonen RJ, Andrews KR, Baums IB, Bird CE, Concepcion GT, Daly-Engel TS, Eble JA, Faucci A, Gaither MR, Iacchei M, Puritz JB, Schultz JK, Skillings DJ, Timmers MA, Bowen BW (2011) Defining Boundaries for Ecosystem-Based Management: A Multispecies Case Study of Marine Connectivity across the Hawaiian Archipelago. *Journal of Marine Biology*, **2011**, 1–13.

Toonen RJ, Grosberg RK (2011) Causes of chaos : spatial and temporal genetic heterogeneity in the intertidal anomuran crab Petrolisthes cinctipes . In: *Phylogeography and Population Genetics in Crustacea* (eds Koenemann S, Held C, Schubart C), pp. 75–107. CRC Press Crustacean Issues Series.

Wares JP (2002) Community genetics in the Northwestern Atlantic intertidal. *Molecular ecology*, **11**, 1131–44.

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

Weir BS, Cockerham CC (1984) Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, **38**, 1358–1370.

White C, Selkoe KA, Watson J, Siegel DA, Zacherl DC, Toonen RJ (2010) Ocean currents help explain population genetic structure. *Proceedings of the Royal Society Series B, Biological Sciences*, **277**, 1685–94.

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

Yearsley JM, Viard F, Broquet T (2013) The effect of collective dispersal on the genetic structure of a subdivided population. *Evolution*, **67**, 1649–59.

Appendix 1: ToBo Laboratory authors and addresses:

Kimberly Andrews, School of Biological Sciences, Durham University, South Road, Durham, DH1 3LE, UK

Moises Bernal, California Academy of Sciences, 55 Music Concourse Drive, Golden Gate Park, San Francisco, CA 94118, USA

Christopher Bird, Marine Biology Program, Department of Life Sciences, Texas A & M University–Corpus Christi, 6300 Ocean Drive, Corpus Christi, Texas 78412, USA

Holly Bolick, Bishop Museum, 1525 Bernice St, Honolulu, HI, 96817, USA

Iliana Baums, Department of Biology, The Pennsylvania State University, 208 Mueller Laboratory University Park, PA, 16802, USA

Richard Coleman, Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, HI 97644, USA

Greg Concepcion, Pacific Biosciences, 1380 Willow Rd, Menlo Park, CA, 94025, USA

Joseph DiBattista, Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, HI 97644, USA

Jeffry Eble, Center for Environmental Bioremediation and Diagnostics, University of West Florida, Pensacola, FL, 32561, USA

Iria Fernandez-Silva, California Academy of Sciences, 55 Music Concourse Drive, Golden Gate Park, San Francisco, CA 94118, USA

Michelle Gaither, School of Biological and Biomedical Sciences, Durham University, South Road, Durham DH1 3LE, UK *and* California Academy of Sciences, Ichthyology, 55 Music Concourse Drive, San Francisco, CA 94118, USA

Mathew Iacchei, Department of Oceanography, University of Hawai'i at Manoa, 1000 Pope Rd., Honolulu, HI, 96822, USA

Nicholas R Polato, Department of Ecology & Evolutionary Biology, 215 Tower Rd., Cornell University, Ithaca, NY 14853, USA

Malia Rivera, Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, HI 97644, USA

Luiz Rocha, California Academy of Sciences, 55 Music Concourse Drive, Golden Gate Park, San Francisco, CA 94118, USA

Derek Skillings, Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, HI 97644, USA

Molly Timmers, Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, HI 97644, USA

Zoltan Sbazo, Hawai'i Institute of Marine Biology, University of Hawai'i, Kāne'ohe, HI 97644, USA

Table 1. Summary of the criteria used to categorize datasets by type of spatial genetic structuring. Significance tests used p<0.05 without correction for multiple tests.

	Global F_{ST}, φ_{ST} or $\mathrm{D}_{\mathrm{EST}}$ test significant?	Spatial clustering (F_{CT}) significant?	IBD test significant?
1. Panmixia	no	no	no
2. Chaotic	yes	no	no
4. IBD	yes or no	no	yes
5. Regional groups	yes or no	yes	yes or no

Table 2. Taxonomic and life history traits of 37 species used in the study. ^ indicates recent arrivals to Hawai'i (<60yrs); taxon indicates vernacular group; PLD = estimates of mean pelagic larval duration in days; End = endemic to Hawai'i , IP = Indo-Pacific wide, Pac = Pacific wide (including eastern Indian Ocean); depth given in m; maximum length refers to body or colony size in cm; Gen. Time = generation time or minimum doubling time in years; att = eggs attached to substrate or body, free = eggs spawned into the water column, int = direct development; Habitat = habitat association, S = specialist and G = generalist categories.

Genus species	Taxon	PLD	Range	Depth Range	Max. Length	Gen. Time	Eggs	Habitat	Trophic Group
1. *Abudefduf abdominalis*	Damselfish	24	End	49	30	2.5	att	S: rubble	planktivore
2. *Abudefduf vaigiensis^*	Damselfish	20	IP	49	20	1	att	G: reef	planktivore
3. *Acanthurus nigrofuscus*	Surgeonfish	31	IP	25	20	2.5	free	G: reef	herbivore
4. *Acanthurus nigroris*	Surgeonfish	58	End	89	25	1	free	G: reef	herbivore
5. *Acanthurus olivaceus*	Surgeonfish	60	Pac	43	35	2.5	free	G: reef	herbivore
6. *Acanthaster planci*	Sea Star	14	Pac	3	30	2	free	G: reef	corallivore
7. *Calcinus hazletti*	Crab	50	Pac	15	1	4	att	S: coral	detritivore
8. *Cellana exarata*	Limpet	6	End	2	7	3	free	S: intertidal	herbivore
9. Cephalopholis argus	Grouper	28	IP	39	60	2.5	free	S: high relief	lrg. predator
10. *Chaetodon fremblii*	Butterflyfish	40	End	61	13	1	free	S: coral	invertivore
11. *Chaetodon lunulatus*	Butterflyfish	40	Pac	17	14	2.5	free	S: coral	corallivore
12. *Chaetodon miliaris*	Butterflyfish	60	End	250	13	1	free	G: reef	omnivore
13. *Chaetodon multicinctus*	Butterflyfish	40	End	109	12	1	free	G: coral & rubble	corallivore
14. *Ctenochaetus strigosus*	Surgeonfish	58	End	112	14	1	free	G: coral & rubble	herbivore
15. *Hyporthodus quernus*	Grouper	40	End	360	122	15	free	S: high relief	lrg. predator
16. *Etelis coruscans*	Snapper	40	IP	157	120	8	free	G: deep reef	lrg. predator
17. *Etelis marshi*	Snapper	40	IP	128	127	8	free	G: deep reef	lrg. predator
18. *Halichoeres ornatissimus*	Wrasse	40	End	11	18	2.5	free	S: coral	invertivore
19. *Heterocentrotus mammillatus*	Urchin	8	IP	49	8	1	free	G: reef	herbivore
20. *Holothuria atra*	Cucumber	15	IP	30	60	1	free	G: sand	sediments
21. *Holothuria whitmaei*	Cucumber	15	IP	20	30	1	free	G: sand	sediments
*22. Lutjanus kasmira^*	Snapper	31	IP	262	40	2.5	free	G: reef	lrg. predator
23. *Monitpora capitata*	Coral	3	Pac	17	200	10	free	G: reef	planktivore
24. *Mulloidichthys flavolineatus*	Goatfish	60	IP	75	43	2.5	free	G: sand & reef	invertivore
25. *Mulloidichthys vanicolensis*	Goatfish	36	IP	112	38	2.5	free	G: sand & reef	invertivore
26. *Myripristis berndti*	Squirrelfish	55	IP	12	30	1	free	G: high relief	planktivore
27. *Ophiocoma erinaceus*	Brittle star	50	IP	27	20	1	free	G: sand & reef	sediments
28. *Ophiocoma pica*	Brittle star	50	IP	27	10	1	free	G: sand & reef	sediments

29. *Panulirus marginatus*	Lobster	365	End	3	40	4	att	S: high relief	invertivore
30. *Panulirus penicillatus*	Lobster	270	IP	3	40	4	att	S: reef & rock	invertivore
31. *Parupeneus multifasciatus*	Goatfish	44	Pac	158	35	2.5	free	G: sand	invertivore
32. *Porites lobata*	Coral	3	IP	23	200	8	free	G: reef	planktivore
33. *Pristipomoides filamentosus*	Snapper	45	IP	360	100	5	free	G: deep reef	lrg. predator
34. *Stegastes fasciolatus*	Damsel	30	End	29	16	1	att	S: reef & rock	herbivore
35. *Stenella longirostris*	Dolphin	0	IP	250	200	13	int	S: all	lrg. predator
36. *Triaenodon obesus*	Shark	0	IP	32	200	8	int	S: all	lrg. predator
37. *Zebrasoma flavescens*	Tang	54	Pac	43	20	1	free	S: reef	herbivore

Table 4. Upper: Alternative model comparison for linear modeling of F_{ST} using 35 species (left side; 2 outliers excluded) and 25 non-endemic species (right side). Lower: Detailed results of final GLM model. ΔAIC_c = delta AIC_c, the difference in AIC_c value between the model and the top model.

	All species				Endemics omitted			
Parameter	P value	R^2	adj. R^2	ΔAIC_c	P value	R^2	adj. R^2	ΔAIC_c
Fish+PLD+Structure	0.004	0.555	0.497	0.0	0.000	0.645	0.578	0.0
PLD+Structure	0.009	0.398	0.348	5.1	0.009	0.451	0.387	5.1
Fish+PLD	0.044	0.359	0.305	6.8	0.082	0.391	0.319	7.1
Fish	0.010	0.238	0.207	8.7	0.019	0.269	0.228	7.6
PLD	0.013	0.216	0.185	9.5	0.024	0.251	0.210	8.1
Structure	0.025	0.185	0.152	10.5	0.072	0.169	0.122	10.2

Term	Coefficient estimate	Std. Error	χ^2 P value
Intercept	0.030376	0.005377	<.0001
Fish	-0.00778	0.002516	0.0042
PLD	-0.00863	0.003273	0.0128
Structure	-0.00974	0.002817	0.0017

Figure 1. Map of the Hawaiian archipelago. The number of species sampled per island or atoll is indicated next to each. Major currents are represented by arrows. 1000m and 2000m isobaths delineated.

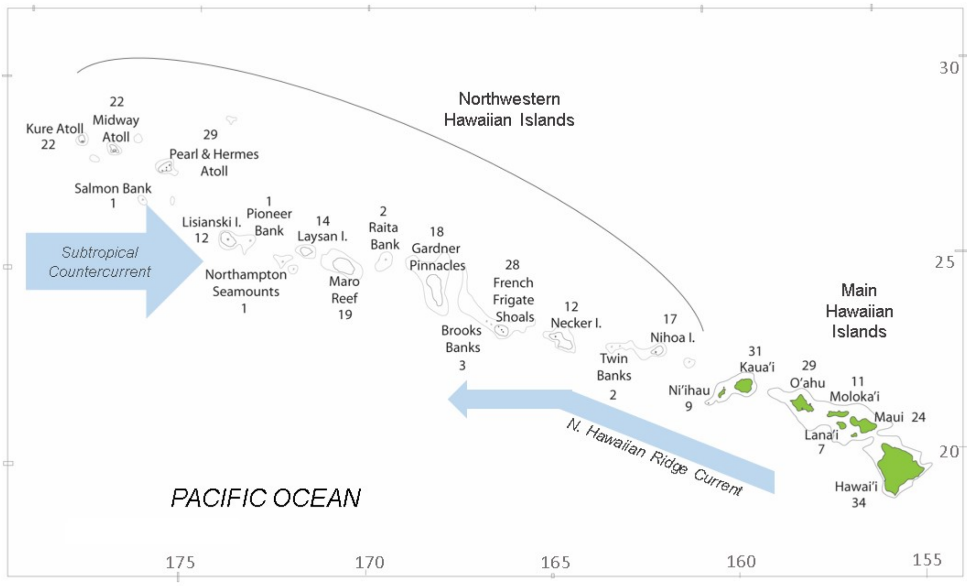
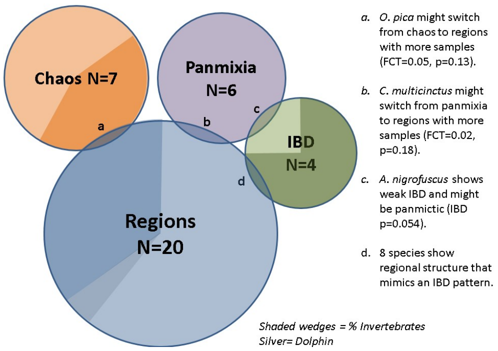


Figure 2. Venn diagram showing the categorization of 37 species into four models of population genetic structuring. Size of circles correspond to number of species, shaded wedges correspond to proportion of invertebrates, and shaded sliver in largest circle indicates the dolphin dataset. Overlapping edges of circles indicate grey areas where categorization of one or more datasets was borderline between the two models; each overlap is lettered and explained on right side.



a. *O. pica* might switch from chaos to regions with more samples (FCT=0.05, p=0.13).

b. *C. multicinctus* might switch from panmixia to regions with more samples (FCT=0.02, p=0.18).

c. *A. nigrofuscus* shows weak IBD and might be panmictic (IBD p=0.054).

d. 8 species show regional structure that mimics an IBD pattern.

Shaded wedges = % Invertebrates
Silver= Dolphin

34 Figure 3. PCA biplot for six genetic summary statistics (F_{ST}, φ_{ST}, D_{EST}, F_{CT}, number of regions, IBD fit). A:
35 Datasets are color coded to show inherent clustering of datasets by genetic trait values (red = large F_{ST}
36 values, purple = single region, blue = multiple regions. B. Datasets are color coded by genetic structure
37 categorizations as in Fig. 2 (purple=panmixia, green = IBD, orange = chaos, blue = regions).

38

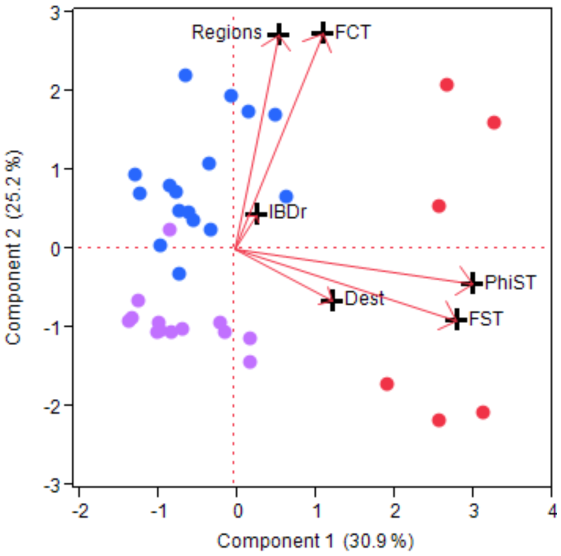


Fig. 3b

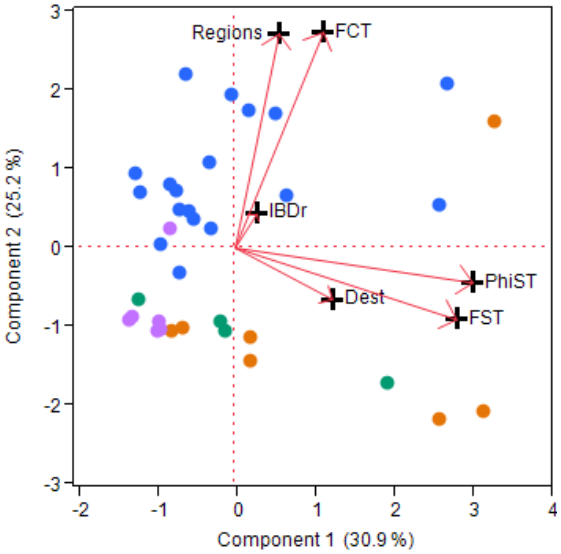

Regions
FCT
IBDr
PhiST
Dest
FST
Component 2 (25.2 %)
Component 1 (30.9 %)
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
4

Figure 4. RDA triplot showing the associations of life history traits (blue font) with genetic traits (red font). Labels for Regions and F_{CT} are obscured because they overlap at coordinates (0.0, 0.9). Label for Herbivore is also obscured because it falls within the cloud of species names in the upper left quadrant. Species names are coded as in Table 1.

Kimberly Selkoe 28/4/2014 09:38

Comment [1]: Increase fonts of vectors, use point and adjust positions of point labels to avoid overlap.

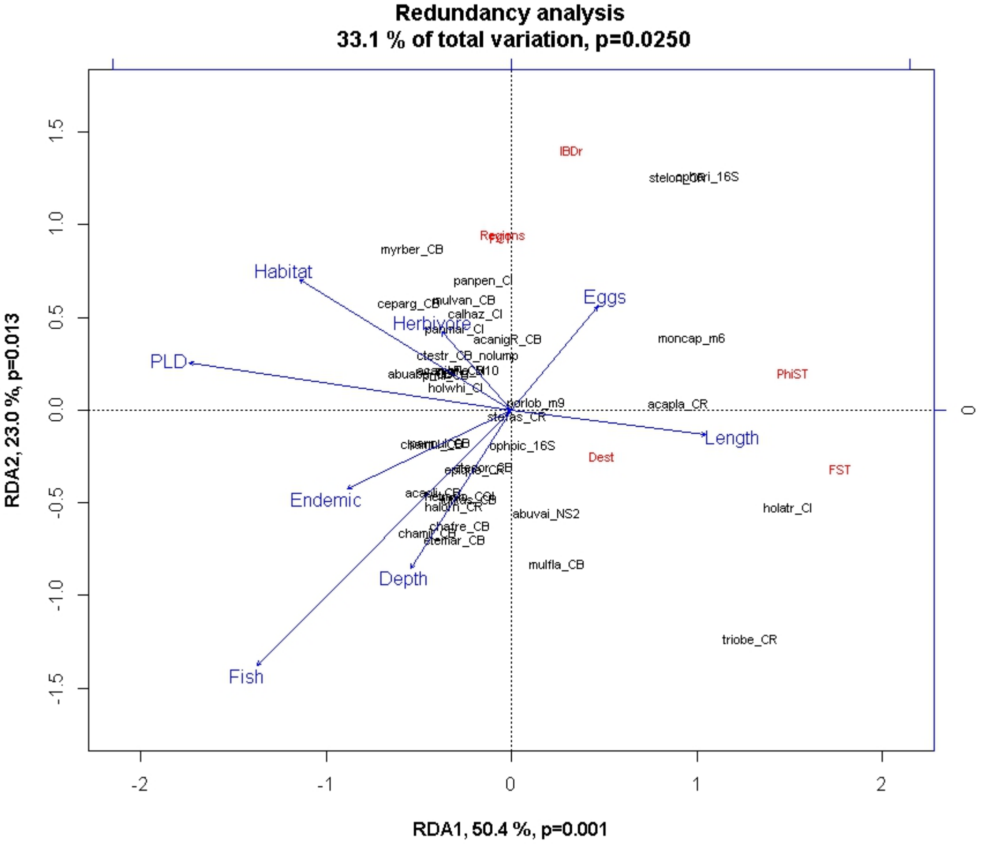


Figure 5. Plot of PLD vs. F_{ST} for all species. Line excludes two outliers at top edge (*Cellana exerata, Chaetodon lunulatus*). Species are coded by their taxon (fish = squares, invertebrate = circles, dolphin = triangle) and type of spatial structuring (purple=panmixia, green = IBD, orange = chaos, blue = regions).

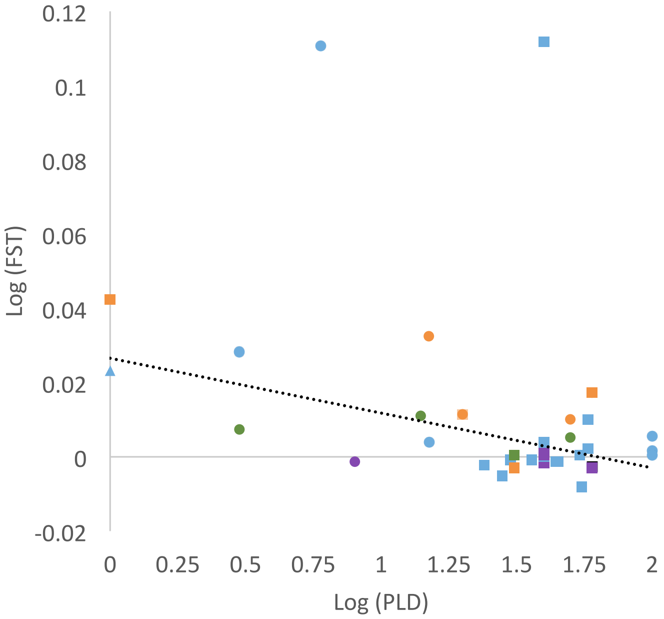


Figure 7. Log PLD vs. log F_{ST} correlation (OLS R^2) for regional subsets of genetic studies sampled from the literature, which calculated F_{ST} using at least 5 sites per study. Data for all regions except Hawai'i taken from Selkoe & Toonen (2011). R^2 value for Hawai'i used an analogous calculation and sample filtering of the present dataset (i.e., F_{ST} <0.001 excluded, PLD=0 included, and outliers *Cellana exerata and Chaetodon lunulatus* included). Numbers of studies used to calculate the R^2 values of each region are indicated above columns.

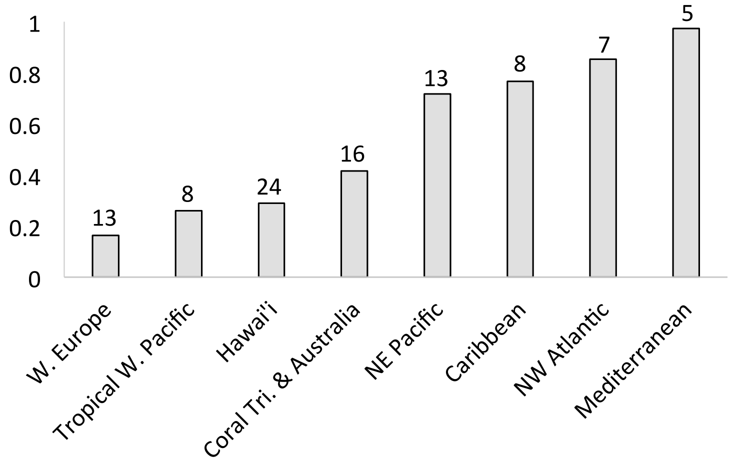


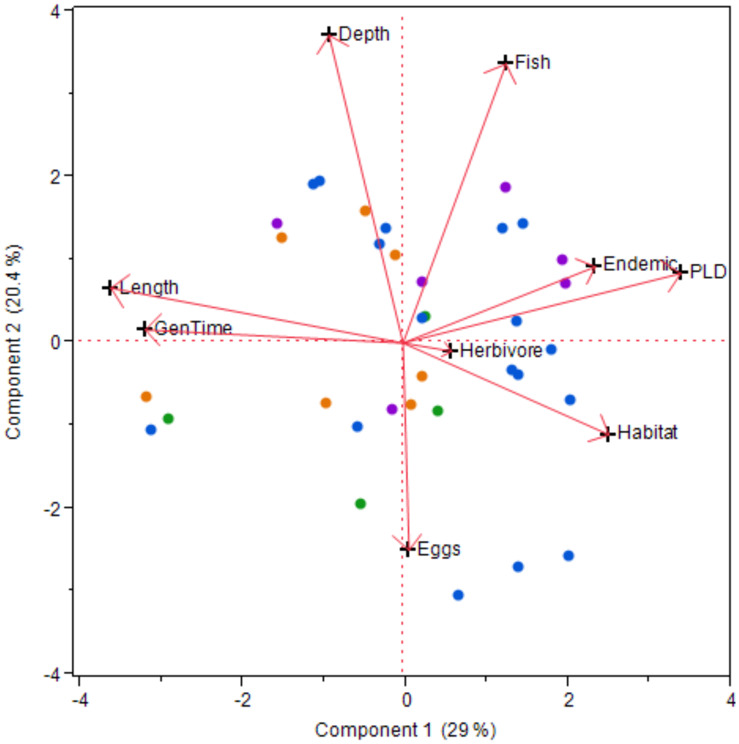
Table S1. Basic genetic results for the subset of datasets used in the analyses (excludes duplicate datasets; see Dataset S1). Marker codes: CB=cytochrome B, CI= cytochrome oxidase I, CR=control region, M#=microsatellites, NS#=nuclear sequence data, 16S=ribosomal unit 16S. Regions = number of spatially discrete K-means clusters; ^=result considered inconclusive due to low allele count leading to insufficient power.

Genus species	Marker	F_{ST}	φ_{ST}	D_{EST}	F_{CT}	Regions	IBD r	Genetic Structure
1. Abudefduf abdominalis	CB	-0.005	0.005	-0.154	0.014***	2	0.122	regional
2. Abudefduf vaigiensis	NS2	0.026***	n/a	0.015*	0.004	1	0.022	chaos
3. Acanthurus nigrofuscus	CB	-0.007	-0.005	-0.043	0	1	0.519*	IBD
4. Acanthurus nigroris	CB	0.023*	0.009	-0.02	0.035**	3	-0.013	regional
5. Acanthurus olivaceus	CB	-0.007	-0.007	-0.026	0	1	0.082	panmixia^
6. Acanthaster planci	CR	0.025***	0.087***	0.255*	0	1	0.543**	IBD
7. Calcinus hazletti	CI	0.013	-0.003	-0.026	0.045***	2	0.101	regional
8. Cellana exarata	CI	0.225***	0.138***	0.067*	0.282***	3	0.187	regional
9. Cephalopholis argus	CB	-0.012	-0.009	-0.034	0.037***	3	0.029	regional
10. Chaetodon fremblii	CB	0.003	0.000	-0.015	0	1	-0.023	panmixia
11. Chaetodon lunulatus	CB	0.227***	0.259***	0.032*	0.316*	5	0.889**	regional
12. Chaetodon miliaris	CB	-0.006	-0.004	-0.043	0	1	-0.084	panmixia
13. Chaetodon multicinctus	CB	-0.004	-0.007	-0.03	0.024	1	-0.023	panmixia^
14. Ctenochaetus strigosus	CB	0.005	0.004	-0.028	0.015**	3	0.103	regional
15. Hyporthodus quernus	CR	0.009**	0.008*	-0.017	0.016*	2	-0.146	regional
16. Etelis coruscans	CB	0.003	0.01**	0.009*	0	1	0.202	chaos
17. Etelis marshi	CB	0.001	0.002	-0.017	0	1	-0.083	panmixia^
18. Halichoeres ornatissimus	CR	0	-0.009	0.064*	0.004***	2	-0.122	regional
19. Heterocentrotus mammillatus	CI	-0.003	0.013	-0.041	0	1	0.053	panmixia
20. Holothuria atra	CI	0.072***	0.131***	0.052*	0	1	0.135	chaos
21. Holothuria whitmaei	CI	0.009	-0.003	-0.006	0.025**	4	-0.260	regional
22. Lutjanus kasmira	CR	0.001	0.003	0.018*	0.001	1	0.090	chaos
23. Monitpora capitata	M6	0.063***	n/a	0.092	0.023**	3	0.233	regional
24. Mulloidichthys flavolineatus	CB	0.039***	0.019**	-0.018	0	1	-0.100	chaos
25. Mulloidichthys vanicolensis	CB	-0.002	0.002	-0.046	0.011***	2	0.49*	regional
26. Myripristis berndti	CB	-0.019	-0.015	-0.053	0.016*	2	0.604*	regional
27. Ophiocoma erinaceus	16S	0.023***	0.171***	-0.046	0.049	2	0.424	chaos
28. Ophiocoma pica	402	0.012*	0.034*	-0.037	0.002	1	0.225*	IBD
29. Panulirus marginatus	CI	0.004**	-0.001	-0.10	0.005*	3	0.287*	regional
30. Panulirus penicillatus	CI	0.001	0.008	-0.03	0.01*	2	0.595**	regional
31. Parupeneus multifasciatus	CB	-0.003	-0.004	-0.027	0.008*	2	0.029	regional
32. Porites lobata	M9	0.017***	n/a	0.017	0	1	0.467***	IBD

33. Pristipomoides filamentosus	CB	-0.003	-0.004	-0.01	0.008**	3	0.134	regional
34. Stegastes fasciolatus	CR	-0.002	0.013**	0.315*	0.006**	4	-0.036	regional
35. Stenella longirostris	CR	0.052***	0.087***	-0.016	0.035*	4	0.425**	regional
36. Triaenodon obesus	CR	0.093**	0.091**	-0.041	0	1	-0.320	chaos
37. Zebrasoma flavescens	M10	0.001**	n/a	0.004	0.002***	3	0.212	regional

Table S2. Sampling statistics by category of structure for all datasets. Values are means for the number of datasets indicated (several species have two datasets included); standard errors in parentheses.	datasets	sites	total alleles	effective alleles	samples	MHI sites	NWHI sites	largest gap
Panmixia	10	11.6 (0.80)	1.5 (0.14)	2.9 (0.55)	377	4.0	5.5	2.9
Chaos	8	9.5 (0.89)	1.7 (0.16)	3.4 (0.62)	391	4.3	6.6	2.7
Regional	24	9.3 (0.53)	1.7 (0.08)	3.6 (0.36)	363	3.9	5.3	3.1
IBD	4	10.75 (1.28)	2.0 (0.21)	4.1 (0.89)	357	4.5	5.5	3.0

Figure S1. PCA biplot of life history traits for all species (n=37). Species are color coded by the genetic structure category (purple=panmixia, green = IBD, orange = chaos, blue = regions).


Depth
Fish
Endemic
PLD
Length
GenTime
Herbivore
Habitat
Eggs
Component 2 (20.4 %)
Component 1 (29 %)